LREC 2026

The 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026

Workshop Proceedings

Editors
Byung-Doh Oh, Tatsuki Kuribayashi, Giulia Rambelli, Ece Takmaz, Philipp Wicke, Jixing Li, Ryo Yoshida

16 May 2026

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026

This workshop has been supported by the National Institute for Japanese Language and Linguistics (NINJAL), Japan.

National Institute for Japanese Language and Linguistics
Advanced Language Science
(E3P) Research Center

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-55-5

Preface

Welcome to the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2026)!

Now in its 15th edition, CMCL continues to be one of the leading workshops at the intersection of computational (psycho-)linguistics, cognitive modeling, and natural language processing. Building on the success of last year's event, we are thrilled to gather once again this year to explore the latest advancements in this exciting, interdisciplinary field.

This year, we received 48 submissions in total, including 38 regular archival submissions and 10 non-archival submissions. We accepted 26 archival papers, resulting in an acceptance rate of 26/38=68.4%, which is higher than that of last year (62.9%). As you will see, we believe this is largely due to the higher quality of submissions in this edition. We also accepted 7 non-archival papers for presentation. The program features a wide array of topics, spanning topics such as real-time sentence processing, language acquisition, and neuroscience-inspired evaluation of language models. We also see an increased focus on bilingual processing and pragmatic reasoning compared to last year's edition. This diversity reflects the growing engagement in our field and the increasing importance of understanding human language through computational methods.

We thank the Program Committee members for their timely and constructive reviews, which were crucial for shaping the program. We also extend our gratitude to our invited speakers, Dr. Claire Stevenson and Dr. Cory Shain, for enriching this year's program with their inspiring talks and expertise. We would also like to thank the LREC Workshop Chairs for selecting CMCL as one of the co-located workshops, as well as for helping us organize the workshop.

Last but not least, we are immensely grateful to our sponsor, the National Institute for Japanese Language and Linguistics (NINJAL). Their generous support allowed us to host the invited speakers for this year's event.

Thank you all for being a part of CMCL 2026. We look forward to an engaging and stimulating exchange of ideas!

The CMCL 2026 Organizing Committee

Organizing Committee

Byung-Doh Oh, Nanyang Technological University
Tatsuki Kuribayashi, Mohammed bin Zayed University of Artificial Intelligence
Giulia Rambelli, Aix-Marseille University
Ece Takmaz, Utrecht University
Philipp Wicke, AURYAL
Jixing Li, City University of Hong Kong
Ryo Yoshida, The University of Tokyo

Invited Speakers

Claire Stevenson, University of Amsterdam
Cory Shain, Stanford University

Program Committee

Giorgia Albertin, University of Bologna
Suhas Arehalli, Macalester College
Masayuki Asahara, National Institute for Japanese Language and Linguistics
Anna Bavaresco, University of Amsterdam
Philippe Blache, LPL CNRS
Emmanuele Chersoni, Hong Kong Polytechnic University
Rendi Chevi, Mohammed bin Zayed University of Artificial Intelligence
Marianne de Heer Kloots, University of Amsterdam
Aniello De Santo, University of Utah
Cui Ding, University of Zurich
Luca Dini, Istituto di Linguistica Computazionale "Antonio Zampolli" - CNR Pisa
Jakub Dotlacil, Utrecht University
Nadine El-Naggar, Mohammed bin Zayed University of Artificial Intelligence
Micha Elsner, The Ohio State University
Diego Frassinelli, LMU Munich
Richard Futrell, University of California, Irvine
Hai Hu, City University of Hong Kong
Shinnosuke Isono, National Institute for Japanese Language and Linguistics
Cassandra L. Jacobs, University at Buffalo
Kohei Kajikawa, Georgetown University
Shunsuke Kando, The University of Tokyo
Jordan Kodner, Stony Brook University
Junlin Li, Mohammed bin Zayed University of Artificial Intelligence
Tal Linzen, New York University
Agnese Lombardi, University of Pisa
Veronica Mangiaterra, University School for Advanced Studies IUSS Pavia, Italy
Sathvik Nair, University of Maryland, College Park
Miyu Oba, Nara Institute of Science and Technology
Yohei Oseki, The University of Tokyo
Francesca Padovani, University of Groningen
Ludovica Pannitto, University of Bologna
Laurent Prévot, University of Aix-Marseille

William Schuler, The Ohio State University
Alice Suozzi, Ca' Foscari University of Venice
Weijie Xu, University of California, Irvine
He Zhou, Hong Kong Polytechnic University

Table of Contents

Transformer Attention as a Unified Model of Encoding and Retrieval in Human Sentence Processing
Dan Parker 1

ChineseDevBench: A Chinese Developmental Benchmark for Language Development
Shaonan Wang, Yiwen Wu, Na Li, Zesheng Chen, Gan Wang, Shuchen Zhang, Xin Sun, Luan Li and Yaran Chen 10

Predicting Sentence Acceptability Judgments in Multimodal Contexts
Hyewon Jang, Nikolai Ilinykh, Sharid Loaiciga, Jey Han Lau and Shalom Lappin 25

ToM in LLM is not ToM, but a Pragmatic Effect
Agnese Lombardi and Alessandro Lenci 35

What Kind of Language is Easy to Language-Model Under Curriculum Learning?
Nadine El-Naggar, Tatsuki Kuribayashi and Ted Briscoe 48

Modeling semantic association in self-paced reading with language model embeddings
Sara Møller Østergaard, Kenneth Enevoldsen, Afra Alishahi and Bruno Nicenboim 60

Character-aware Transformers Learn an Irregular Morphological Pattern Yet None Generalize Like Humans
Akhilesh Kakolu Ramarao, Kevin Tang and Dinah Baer-Henney 74

Comparing Transformer Model Interpretability with Human Cognition: A Dual Analysis of Attention and Attribution
Lingchen Kong, Jinnie Shin and Pavlo Antonenko 86

Correlating Language Model Surprisal With Cloze and Plausibility: Getting the Best of Both Measures
Kate Rebecca Belcher and Matthew Crocker 99

Is Cross-Lingual Transfer in Bilingual Models Human-Like? A Study with Overlapping Word Forms in Dutch and English
Iza Škrjanec, Irene Elisabeth Winther, Vera Demberg and Stefan L. Frank 110

Brain-to-Text Decoding with Brain Atlases and Brain Foundation Models
Haruka Akama, Ryo Yoshida, Max Müller-Eberstein and Yohei Oseki 124

Floating or Suggesting Ideas? A Large-Scale Contrastive Analysis of Metaphorical and Literal Verb–Object Constructions
Prisca Piccirilli, Alexander Fraser and Sabine Schulte im Walde 145

Social Meaning in Large Language Models: Structure, Magnitude, and Pragmatic Prompting
Roland Mühlenbernd 162

Headlines You Won't Forget: Can Pronoun Insertion Increase Memorability?
Selina Meyer, Magdalena Abel and Michael Roth 172

Aggregated Transformer Attention Measures Predict Reading Times Beyond Surprisal
Lukas Mielczarek and Laura Kallmeyer 188

'Layer su Layer': Identifying and Disambiguating the Italian NPN Construction in BERT's family
Greta Gorzoni, Ludovica Pannitto and Francesca Masini 203

Teaching LLMs to unveil tendentious implicit contents of Italian political communication
Walter Paci, Lorenzo Gregori and Alessandro Panunzi 221

Examining Algebraic Recombination for Compositional Generalisation
Joaquin Cardona Ruiz, Antske Fokkens and Lucia Donatelli 230

Can LLMs Simulate Human Behavioral Variability? A Case Study in the Phonemic Fluency Task
Mengyang Qiu, Zoe Brisebois and Siena Sun 250

Translation from the Information Bottleneck Perspective: an Efficiency Analysis of Spatial Prepositions in Bitexts
Antoine Taroni, Ludovic Moncla and Frederique Laforest 264

Causal Inferences Are Driven by Noun Concept Specificity Evidence from Self-Paced Reading and Large Language Model Surprisal
Fabian Schlotterbeck and Raphael Barth 276

Topic-context Dependency on Continuous Semantic Reconstruction of Language from fMRI Signals
Fermin Travi, Agustin Delmagro, Diego Fernandez Slezak, Bruno Bianchi and Juan E. Kamienkowski 286

Disentanglement and Compositionality of Letter Identity and Letter Position in Variational Auto-Encoder Vision Models
Bruno Bianchi, Aakash Agrawal, Stanislas Dehaene, Emmanuel Chemla and Yair Lakretz 293

Plausibility as Commonsense Reasoning: Humans Succeed, Large Language Models Do not
Sercan Karakas 309

Benchmarking Source-Sensitive Reasoning in Turkish: Humans and LLMs under Evidential Trust Manipulation
Sercan Karakas and Yusuf Şimşek 329

Toward Cognitive Alignment in Large Language Models: Integrating Linguistic Theory and Human Data
Wajdi Zaghouani 347

Non-archival papers

Quantifying the Pragmatics of Discourse Markers: A Surprisal-Based Analysis of Taiwanese "Ah"
Ri-Sheng Huang

Acoustic Representations Support Statistical Learning of Syllable Sequences: A Computational Study
Nika Jurov and Thomas Schatz

Simulating Cross-Linguistic Influence in Bilingual Reading - a Knowledge Distillation Approach
Irene E. Winther and Stefan L. Frank

Investigating Brain-LLM Alignment with Category-Deprived Text
Shantanu Nath, Marco Tettamanti and Roberto Zamparelli

Semantic Norms and Structural Relations in Chinese Semantic Composition: A Dataset and Analysis
He Zhou, Emmanuele Chersoni and Yu-Yin Hsu

Polar Questions in SPA–TTR: Linking Dialogue, Acquisition, and Neurosemantics
Jonathan Ginzburg, Shiyun Dong, Robin Cooper, Andy Luecking and Staffan Larsson

Generics as Useful Defeasible Rules of Inference
Justin Helms

Workshop Program

Saturday, 16 May 2026

09:10–09:20 **Opening Remarks**

09:20–10:20 **Invited Talk: Claire Stevenson**

Saturday, 16 May 2026 (continued)

10:30–11:40 **Session 1: Poster Presentations**

Transformer Attention as a Unified Model of Encoding and Retrieval in Human Sentence Processing
Dan Parker

ChineseDevBench: A Chinese Developmental Benchmark for Language Development
Shaonan Wang, Yiwen Wu, Na Li, Zesheng Chen, Gan Wang, Shuchen Zhang, Xin Sun, Luan Li and Yaran Chen

Predicting Sentence Acceptability Judgments in Multimodal Contexts
Hyewon Jang, Nikolai Ilinykh, Sharid Loaiciga, Jey Han Lau and Shalom Lappin

ToM in LLM is not ToM, but a Pragmatic Effect
Agnese Lombardi and Alessandro Lenci

What Kind of Language is Easy to Language-Model Under Curriculum Learning?
Nadine El-Naggar, Tatsuki Kuribayashi and Ted Briscoe

Modeling semantic association in self-paced reading with language model embeddings
Sara Møller Østergaard, Kenneth Enevoldsen, Afra Alishahi and Bruno Nicenboim

Quantifying the Pragmatics of Discourse Markers: A Surprisal-Based Analysis of Taiwanese "Ah"
Ri-Sheng Huang

Acoustic Representations Support Statistical Learning of Syllable Sequences: A Computational Study
Nika Jurov and Thomas Schatz

Simulating Cross-Linguistic Influence in Bilingual Reading - a Knowledge Distillation Approach
Irene E. Winther and Stefan L. Frank

Investigating Brain-LLM Alignment with Category-Deprived Text
Shantanu Nath, Marco Tettamanti and Roberto Zamparelli

Saturday, 16 May 2026 (continued)

11:40–13:00 **Session 2: Oral Presentations**

Character-aware Transformers Learn an Irregular Morphological Pattern Yet None Generalize Like Humans
Akhilesh Kakolu Ramarao, Kevin Tang and Dinah Baer-Henney

Comparing Transformer Model Interpretability with Human Cognition: A Dual Analysis of Attention and Attribution
Lingchen Kong, Jinnie Shin and Pavlo Antonenko

Correlating Language Model Surprisal With Cloze and Plausibility: Getting the Best of Both Measures
Kate Rebecca Belcher and Matthew Crocker

Is Cross-Lingual Transfer in Bilingual Models Human-Like? A Study with Overlapping Word Forms in Dutch and English
Iza Škrjanec, Irene Elisabeth Winther, Vera Demberg and Stefan L. Frank

13:00–14:00 **Lunch Break**

14:00–15:00 **Invited Talk: Cory Shain**

15:00–16:00 **Session 3: Oral Presentations**

Brain-to-Text Decoding with Brain Atlases and Brain Foundation Models
Haruka Akama, Ryo Yoshida, Max Müller-Eberstein and Yohei Oseki

Floating or Suggesting Ideas? A Large-Scale Contrastive Analysis of Metaphorical and Literal Verb–Object Constructions
Prisca Piccirilli, Alexander Fraser and Sabine Schulte im Walde

Social Meaning in Large Language Models: Structure, Magnitude, and Pragmatic Prompting
Roland Mühlenbernd

Semantic Norms and Structural Relations in Chinese Semantic Composition: A Dataset and Analysis
He Zhou, Emmanuele Chersoni and Yu-Yin Hsu

Saturday, 16 May 2026 (continued)

16:00–17:00 **Session 4: Poster Presentations**

Headlines You Won't Forget: Can Pronoun Insertion Increase Memorability?
Selina Meyer, Magdalena Abel and Michael Roth

Aggregated Transformer Attention Measures Predict Reading Times Beyond Surprisal
Lukas Mielczarek and Laura Kallmeyer

'Layer su Layer': Identifying and Disambiguating the Italian NPN Construction in BERT's family
Greta Gorzoni, Ludovica Pannitto and Francesca Masini

Teaching LLMs to unveil tendentious implicit contents of Italian political communication
Walter Paci, Lorenzo Gregori and Alessandro Panunzi

Examining Algebraic Recombination for Compositional Generalisation
Joaquin Cardona Ruiz, Antske Fokkens and Lucia Donatelli

Can LLMs Simulate Human Behavioral Variability? A Case Study in the Phonemic Fluency Task
Mengyang Qiu, Zoe Brisebois and Siena Sun

Translation from the Information Bottleneck Perspective: an Efficiency Analysis of Spatial Prepositions in Bitexts
Antoine Taroni, Ludovic Moncla and Frederique Laforest

Causal Inferences Are Driven by Noun Concept Specificity Evidence from Self-Paced Reading and Large Language Model Surprisal
Fabian Schlotterbeck and Raphael Barth

Polar Questions in SPA–TTR: Linking Dialogue, Acquisition, and Neurosemantics
Jonathan Ginzburg, Shiyun Dong, Robin Cooper, Andy Luecking and Staffan Larsson

Generics as Useful Defeasible Rules of Inference
Justin Helms

Saturday, 16 May 2026 (continued)

17:00–17:50 **Session 5: Oral Presentations**

Topic-context Dependency on Continuous Semantic Reconstruction of Language from fMRI Signals
Fermin Travi, Agustin Delmagro, Diego Fernandez Slezak, Bruno Bianchi and Juan E. Kamienkowski

Disentanglement and Compositionality of Letter Identity and Letter Position in Variational Auto-Encoder Vision Models
Bruno Bianchi, Aakash Agrawal, Stanislas Dehaene, Emmanuel Chemla and Yair Lakretz

Plausibility as Commonsense Reasoning: Humans Succeed, Large Language Models Do not
Sercan Karakas

Benchmarking Source-Sensitive Reasoning in Turkish: Humans and LLMs under Evidential Trust Manipulation
Sercan Karakas and Yusuf Şimşek

Toward Cognitive Alignment in Large Language Models: Integrating Linguistic Theory and Human Data
Wajdi Zaghouani

17:50–18:00 **Closing Remarks**

Transformer Attention as a Unified Model of Encoding and Retrieval in Human Sentence Processing

Dan Parker
Department of Linguistics
The Ohio State University
parker.1758@osu.edu

Abstract

Recent work suggests that the attention mechanism in Transformer-based language models operates as a cue-based memory architecture, exhibiting behavior consistent with leading memory-based accounts of human sentence processing. On this view, attention implements a cue-based retrieval procedure that is subject to similarity-based interference, which, in principle, should encompass both syntactic and semantic interference effects. Prior research has shown that attention captures classic syntactic interference effects during retrieval, such as those observed during relative clause processing. However, it remains unclear whether attention also captures encoding-based interference driven by semantic properties. The present study addresses this issue by testing whether attention is sensitive to encoding interference arising from referential class distinctions that modulate relative clause difficulty. Three attention-based metrics derived from GPT-2 (Earth Mover's Distance, attention-to-target, attention entropy) were evaluated. All three recapitulate human behavioral patterns across referential manipulations. These findings indicate that attention can capture both retrieval- and encoding-based interference along both syntactic and semantic dimensions, providing broader empirical coverage than existing memory-based models (e.g., ACT-R), and supporting a unified computational account of how linguistic representations are encoded and accessed in memory during sentence processing.

Keywords: retrieval, encoding, syntactic interference, semantic interference, transformer attention

1. Introduction

A central question in sentence processing research concerns how linguistic representations are encoded and retrieved in real time. Leading memory-based models of human sentence processing suggest that items in a sentence representation are accessed using a cue-based retrieval procedure (Lewis et al., 2006; Van Dyke and Johns, 2012). These models account for core empirical patterns, such as locality effects (e.g., Gibson, 2000) and similarity-based interference (e.g., Lewis et al., 2006). However, they leave underspecified how sentence representations are encoded in memory and how constraints on the encoding shape processing dynamics.

Recently, Ryu and Lewis (2021, 2025) proposed that the attention mechanism of Transformer-based language models functions as a cue-based memory architecture that is subject to similarity-based interference effects like those observed in human sentence processing. On this view, attention implements a retrieval process that is guided by learned, distributed cues and exhibits behavior consistent with core predictions of memory-based theories of human sentence processing. Empirically, attention has been shown to capture a range of classic psycholinguistic phenomena, including agreement attraction, difficulty with center embedding, ambiguity, and the asymmetry between object and subject relative clauses (Jacobs and MacDonald, 2024; Ryu and Lewis, 2021, 2025).

A limitation of these studies is their narrow focus on retrieval-based syntactic effects. Memory-based theories of human sentence processing aim to provide a unified account of both retrieval and encoding effects driven by syntactic and semantic properties of the sentence. A mechanism that captures retrieval-based effects alone therefore provides only partial support for its status as a cognitively plausible memory system. A critical test of the attention-based account (Ryu and Lewis, 2021, 2025) is whether it also captures encoding-based interference and semantically driven effects.

Evidence for semantic interference during encoding is well documented. A classic finding in psycholinguistics is that the typical difficulty observed for sentences with an object relative clause (e.g., *The banker [that the barber praised] climbed the mountain.*) is reduced when the head noun and embedded noun differ in referential class (Gordon et al., 2001, 2004). When the nouns overlap in referential class (e.g., both are definite descriptions), they compete, increasing difficulty; when they differ, this competition is reduced, easing processing. Crucially, this pattern reflects interference at encoding: overlapping semantic properties degrade the quality of the corresponding memory representations, increasing subsequent processing difficulty (Van Dyke and McElree, 2006).

Encoding-based interference effects of this sort are not captured by current memory-based models of human sentence processing. The leading model, ACT-R (Lewis and Vasishth, 2005), provides a de-

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 1–9
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

tailed specification of retrieval processes, but lacks a comparable specification of encoding processes and does not offer a mechanistic account of how semantic overlap affects memory representations.

The present study tests whether attention captures encoding-based interference driven by semantic properties. If it does, this would extend prior work beyond retrieval-based effects and provide the additional evidence required to evaluate attention as a comprehensive model of memory in human sentence processing, encompassing both encoding and retrieval processes.

2. Memory-based sentence processing

Memory-based accounts of sentence processing attribute difficulty to limitations in how linguistic information is encoded and retrieved during comprehension. Long-distance dependencies illustrate these demands. In (1), the anaphor *herself* requires access to its antecedent *the girl* encoded in memory for interpretation:

(1) The girl will buy herself some ice cream.

Memory-based accounts formalize this process as content-addressable retrieval, in which retrieval cues are matched in parallel against the discrete feature units of each item in working memory. The best-matching item is retrieved (modulo fluctuations in activation). When multiple items overlap with the cues, they compete, giving rise to similarity-based interference (Lewis et al., 2006; Van Dyke and Johns, 2012).

Crucially, the degree of interference depends on how representations are encoded in memory: their specific features determine both cue overlap and the extent of competition at retrieval.

The interdependence of encoding and retrieval is particularly evident in relative clause (RC) processing. A classic finding is that object relative clauses (ORCs; (2a)) are more difficult than subject-relative clauses (SRCs; (2b)) (e.g., King and Just, 1991; Gordon et al., 2001; Grodner and Gibson, 2005). This asymmetry is typically attributed to greater memory demands in ORCs, where multiple nouns compete for retrieval to fill the object gap.

(2) a. **ORC:** The banker [that the barber praised] climbed the mountain.
b. **SRC:** The banker [that praised the barber] climbed the mountain.

Equally important, the ORC/SRC asymmetry is modulated by the referential class of the nouns. Gordon et al. (2001, 2004) found that the disruption associated with ORCs depends on whether the nouns belong to the same referential class. When the RC noun and head noun overlap in referential class (e.g., both common nouns), the typical ORC/SRC asymmetry is observed. When they differ, as in (3), where the RC noun is a name, pronoun, or quantified pronoun, the asymmetry is reduced. This pattern reflects a categorical distinction in the behavioral data based on referential class.

(3) The banker [that Joe/you/everyone praised] climbed the mountain.

This effect has been attributed to modulation of semantic interference at encoding. When the nouns are from the same referential class, their overlapping properties induce interference within the memory representations themselves, degrading the quality of the representations and increasing processing difficulty. When they differ, overlap is reduced, minimizing interference and facilitating processing (Van Dyke and McElree, 2006). Because referential class is unlikely to serve as a retrieval cue for these dependencies (e.g., the RC verb *praised* does not select for referential class), these effects are not readily explained by interference at retrieval alone.

This type of encoding interference is standardly characterized as a process of "feature overwriting" (Nairne, 1990; Oberauer and Kliegl, 2006): when two items overlap along a dimension (e.g., referential class), they compete for representational resources associated with that dimension, leading to degradation or loss of the corresponding feature units in their memory traces and making subsequent retrieval more difficult.

These findings expose a limitation of current memory-based models of human sentence processing. For example, ACT-R (Lewis and Vasishth, 2005) provides a detailed formal account of retrieval but lacks a comparable specification of encoding. Memory representations are defined in terms of discrete hand-specified features, without a principled account of how those features interact or how overlap along dimensions irrelevant for retrieval (e.g., referential class) affects accessibility. This abstraction affords theoretical flexibility, but it does not provide a mechanistic explanation of how interference arises during encoding. As a result, encoding-based interference effects, such as those observed in RC processing, fall outside its explanatory scope.

2.1. Attention as cue-based memory

Recently, Ryu and Lewis (2025) proposed that the attention mechanism of Transformer-based language models provides a more flexible implementation of cue-based memory processing that ad-

dresses many key challenges facing existing implementations (e.g., ACT-R).

Transformers are distinguished from other language models by their use of self-attention to compute context-sensitive embeddings, whereby each token attends directly to prior tokens (Vaswani et al., 2017). This process is implemented through attention matrices that encode pairwise weighted connections, quantifying incoming attention (how much a token attends to another), outgoing attention (how much it is attended to), and self-attention.

Formally, attention is computed according to Equation 1, where Q is the matrix of queries, K is the matrix of keys, V is the matrix of values, and d_k is the dimensionality of the key/query vectors.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

Crucially, attention allows the model to dynamically weight connections between tokens as information accrues layer-by-layer. It therefore provides a structured view into the model's internal operations, revealing *what* contextual information is accessed and *how* it is weighted to guide next-word prediction.

As a model of memory, attention parallels ACT-R: both access the prior context based on content similarity and exhibit interference when multiple items match the retrieval probe (Ryu and Lewis, 2021, 2025). Methodologically, however, attention differs from ACT-R in that it does not rely on discrete, hand-coded representations/cues. Instead, it operates over learned, distributed representations, allowing both cues and content to emerge from the training data. This framework thus provides a more flexible memory architecture in which interference may arise not only at retrieval but also during encoding.

To support the proposal that attention implements a cue-based memory system, Ryu and Lewis (2021, 2025) demonstrated that attention patterns track retrieval interference in agreement processing, center-embedding difficulty, and the ORC/SRC asymmetry. Additional evidence comes from Jacobs and MacDonald (2024), who found that attention patterns align closely with reading times in classic syntactic ambiguity paradigms.

2.2. Beyond syntactic retrieval: Encoding-based semantic interference

A limitation of prior work on attention as a model of human sentence processing is its narrow focus on syntactic effects that arise at retrieval. That work provides important initial support for the claim that attention implements a cue-based memory system, but leaves open a central question, namely whether attention also captures interference that arises during encoding, particularly those driven by semantic properties.

This gap is theoretically significant. Memory-based theories of human sentence processing are intended to provide a unified account of both retrieval and encoding processes. A mechanism that captures retrieval-based interference alone therefore provides only partial support for its status as a cognitively plausible memory system.

Encoding-based interference constitutes a critical test case. If attention provides a general-purpose implementation of the memory representations and operations underlying sentence processing, as proposed by Ryu and Lewis (2021, 2025), then its behavior should reflect not only the dynamics of retrieval but also the structure of the representations encoded in memory. In particular, a memory mechanism that is sensitive to similarity among representations should therefore exhibit corresponding sensitivity to encoding-based interference.

Predictions: If attention implements a similarity-sensitive memory architecture, then it should capture not only retrieval-based syntactic interference, as previously demonstrated (e.g., Ryu and Lewis, 2021, 2025), but also effects of encoding-based semantic interference, such as those observed in relative clause processing (e.g., Gordon et al., 2001, 2004). Demonstrating this sensitivity would extend prior work by showing that attention captures both major loci of interference targeted by memory-based theories, thereby providing stronger evidence for its role as a comprehensive model of human sentence processing.

The present study tests these predictions by examining whether attention is sensitive to encoding interference driven by the referential class distinctions that modulate RC difficulty. Evidence of such sensitivity would indicate that attention captures not only retrieval-based effects but also core encoding effects, supporting a unified framework for modeling memory in human sentence processing.

3. Methods

3.1. Model

To test whether attention captures referential class effects in RC processing, three attention-based metrics were analyzed: Earth Mover's Distance, attention-to-target, and attention entropy. These metrics were derived from GPT-2 small (Radford et al., 2019), a 12-layer Transformer trained on ~8 billion tokens from WebText. GPT-2 small was selected based on recent evidence that its outputs strongly predict human reading times (Oh et al., 2022) and because it was the model used in the prior studies on attention and human sentence pro-

cessing (Jacobs and MacDonald, 2024; Oh and Schuler, 2022; Ryu and Lewis, 2021, 2025), enabling direct comparison across studies.

3.2. Materials

The stimulus set was comprised of the full experimental materials from Gordon et al. (2001, 2004), consisting of eight experiments designed to isolate structural and semantic effects in RC processing. The experiments compared SRC and ORC configurations across eight referential class conditions: definite descriptions (baseline), indefinite descriptions, generics, superordinate RC nouns, superordinate matrix nouns, pronouns, names, and quantified pronouns (Table 1). Each experiment included 24 item sets, yielding a total of 384 sentences. Gordon et al. reported that ORC difficulty was reduced when the embedded noun was a name, pronoun, or quantified pronoun, whereas the remaining conditions exhibited the standard ORC/SRC asymmetry in a categorical manner. All 384 sentences were embedded to obtain contextualized token representations, over which the attention-based metrics were computed.

3.3. Metrics

Earth Mover's Distance (EMD). Earth Mover's Distance (EMD; Rubner et al., 2000) is a symmetric distance metric that quantifies the amount of work required to transform one probability distribution into another. When applied to attention, EMD provides a measure of (dis)similarity between attention distributions over prior context at a given point in processing.

Prior work has used EMD to characterize changes in attention across time (Oh et al., 2022) or across interpretive states within a sentence (Jacobs and MacDonald, 2024). The present study extends the use of EMD to quantify the difference between attention distributions associated with two structural conditions (SRC vs. ORC). This use does not assume a transition between states, but rather treats EMD as a general measure of (dis)similarity between attention-based memory configurations.

Under this interpretation, attention distributions can be understood as encoding the model's memory state over prior tokens. EMD therefore provides a way of quantifying how different these memory configurations are across structures. The linking hypothesis adopted here is that greater dissimilarity in attention distributions corresponds to greater processing asymmetry, insofar as larger differences reflect more substantial reweighting of the underlying memory representations required to support one structure relative to another. In this sense, EMD indexes the degree of representational reconfiguration associated with the SRC/ORC contrast, extending prior work relating EMD to processing difficulty (Jacobs and MacDonald, 2024; Oh and Schuler, 2022) while not assuming that the compared states arise within a single incremental trajectory.

Formally, EMD is defined according to Equation 2, where A_i^{ORC} and A_i^{SRC} denote the attention weight assigned to prior position i in the ORC and SRC conditions. Following prior work (Jacobs and MacDonald, 2024; Oh and Schuler, 2022), EMD was computed from attention weights aggregated across heads in the top layers (9–12), as these layers most directly reflect the contribution of each previous token.

$$\text{EMD}(A^{\text{ORC}}, A^{\text{SRC}}) = \sum_i \left| \sum_{j \le i} \left(A_j^{\text{ORC}} - A_j^{\text{SRC}} \right) \right| \tag{2}$$

Attention-to-target. Attention-to-target is a metric developed by Ryu and Lewis (2021) that quantifies the amount of attention allocated to the retrieval target in the prior context at the critical words, denoted as the attention weight $Attn$ allocated from the critical word w_{cue} to the target word w_{target} by layer l: $Attn_l(w_{\text{cue}}, w_{\text{target}})$. This metric captures retrieval accuracy under interference. For instance, in ORCs, retrieval at the RC verb is more difficult in behavioral measures when the RC subject and head noun overlap in referential type, increasing competition. Lower attention-to-target values thus indicate greater interference, whereas higher values reflect more focused, accurate retrieval.

Attention entropy. Attention entropy quantifies uncertainty in the model's allocation of attention over the prior context. Introduced by Ryu and Lewis (2021, 2025), this measure captures how sharply focused (low entropy) or diffuse (high entropy) the attention pattern is at the critical words. It can be interpreted as the model's uncertainty about which prior word serves as the intended retrieval target. High entropy indicates a broad, uncertain distribution of attention, suggesting difficulty in identifying a unique target. Low entropy reflects a more focused and confident allocation of attention.

Formally, attention entropy is defined by Equation 3, where w_i is the current word, w_j ranges over the prior context ($j = 1, \ldots, i-1$), and $\text{Attn}(w_i, w_j)$ is the attention weight from w_i to w_j:

$$\text{AttnEntropy}(w_i) = -\sum_{j=1}^{i-1} \text{Attn}(w_i, w_j) \times \log_2 \text{Attn}(w_i, w_j) \tag{3}$$

Following Ryu and Lewis (2021, 2025), attention-to-target and attention entropy were aggregated across all 12 layers. Prior work suggests that intermediate layers are most sensitive to linguistic

Referential class	Structure	Sentence
Definite description	ORC	The banker [that the barber **praised**] climbed the mountain …
	SRC	The banker [that praised the **barber**] climbed the mountain …
Indefinite: *a barber*	**Generic:** *barbers*	**Superordinate noun:** *the person*
Pronoun: *you*	**Name:** *Sue*	**Quantified pronoun:** *everyone*

Table 1: Sample experimental stimuli from Gordon et al. (2001, 2004). Critical regions in **bold**.

manipulations (Hoover et al., 2019; Tenney et al., 2018; Zini and Awad, 2022). This approach provides a comprehensive assessment of how interference arises and propagates through the architecture during incremental processing. For present purposes, these metrics were evaluated only over the name, pronoun, and quantified pronoun conditions compared to the baseline definition description, as these conditions constitute the core empirical contrasts of interest.

3.4. Analysis

All three metrics were analyzed at the critical regions defined in the behavioral literature (Gordon et al., 2001, 2004): the RC verb in ORCs and the RC object noun in SRCs (Table 1). Effects were evaluated in a Bayesian framework by comparing each manipulation to the baseline definite description condition, using weakly informative priors (Gelman et al., 2017). Inferences were based on whether the 95% posterior credible interval (CrI) excluded zero and on the posterior probability that an effect was greater/less than zero (e.g., $P(\hat{\beta} > 0)$).

4. Results

All three metrics aligned with the behavioral patterns reported in Gordon et al. (2001, 2004). Specifically, the metrics captured both (i) the core structural ORC/SRC asymmetry and (ii) the modulation of the asymmetry by referential class.

Earth Mover's Distance (EMD). EMD was higher in conditions where the RC subject and head noun belonged to the same referential class (i.e., definite descriptions, indefinites, generics, superordinate NP1 and NP2), indicating a larger ORC/SRC asymmetry, than when the nouns differed in referential class (i.e., name, pronoun, quantified pronoun) (Figure 1). This pattern fully recapitulates the categorical behavioral contrasts reported in Gordon et al. (2001, 2004). Posterior estimates confirmed that conditions empirically associated with the asymmetry (i.e., indefinites, generics, and superordinate NPs) did not differ from the baseline definite description condition (95% CrI included zero in all cases). By contrast, conditions that empirically show a reduced asymmetry (i.e., name, pronoun,

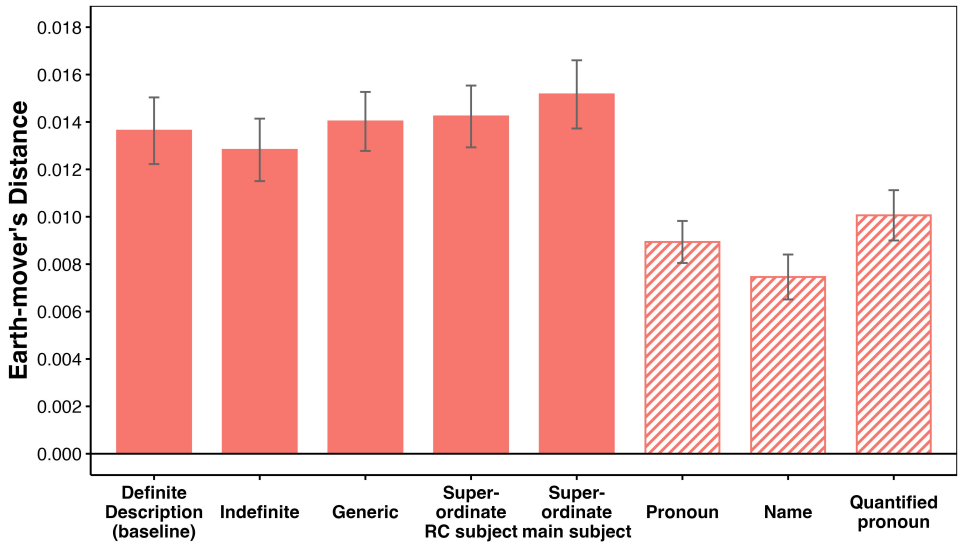


Figure 1: Earth Mover's Distance for each experimental manipulation. Positive values reflect greater redistribution of attention. Solid bars are conditions predicted to show an ORC/SRC asymmetry (higher EMD) due to encoding interference from referential similarity. Striped bars are conditions predicted to show an attenuated effect (lower EMD) due to reduced interference from referential dissimilarity. Error bars represent standard error of the mean.

and quantified pronoun) had reliably lower EMD values relative to baseline (95% CrI excluded zero in all cases; $P(\hat{\beta} < 0) > 0.98$), indicating a reduced asymmetry.

Attention-to-target. Conditions associated with a reduced ORC/SRC asymmetry in human processing (i.e., name, pronoun, quantified pronoun) exhibited higher attention to the target compared to the baseline condition (95% CrI excluded zero; $P(\hat{\beta} > 0) > 0.98$) (Figure 2). This indicates higher retrieval accuracy (i.e., interference is reduced) when the head noun and RC noun differ in referential class compared to when the nouns overlapped in referential class (e.g., both definite descriptions), consistent with the behavioral data.

Attention entropy. Attention entropy was lower in conditions associated with a reduced ORC/SRC asymmetry in human processing (i.e., name, pronoun, quantified pronoun) than in the baseline condition (95% CrI excluded zero; $P(\hat{\beta} < 0) > 0.98$) (Figure 3). These results indicate more focused attention over the prior context when the nouns differ in referential class, and more diffuse attention when they overlap, consistent with the behavioral data.

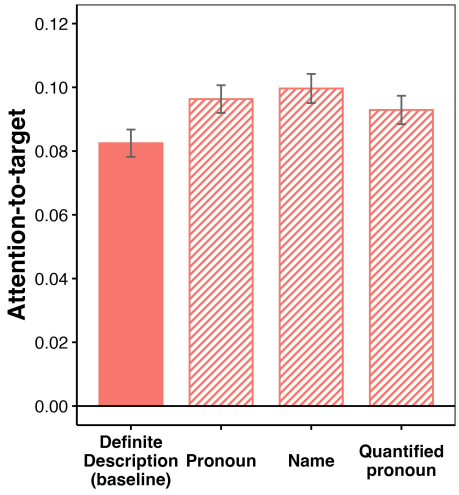


Figure 2: Attention to the target at the critical region for the baseline, pronoun, name, and quantified pronoun conditions. Higher values indicate greater attention to the target, reflecting more accurate retrieval. The solid bar (definite description) is predicted to show decreased attention to the target due to encoding interference from overlap in referential class. Striped bars are conditions predicted to show increased attention to the target due to reduced interference from referential dissimilarity. Error bars represent standard error of the mean.

5. Discussion

5.1. Summary of findings

Taken together, the results show that attention-based metrics derived from GPT-2 capture core patterns of human sentence processing, including both the structural ORC/SRC asymmetry and its modulation by referential class. These findings indicate that attention captures both retrieval- and encoding-based interference across syntactic and semantic dimensions, without being explicitly engineered to do so. The theoretical and empirical implications are discussed in the following sections.

5.2. Why attention should capture encoding interference

A central motivation of this study is theoretical. If Transformer attention constitutes a cognitively plausible general-purpose memory mechanism for sentence processing, as previously claimed (Ryu and Lewis, 2021, 2025), then its empirical adequacy cannot be evaluated solely on its ability to capture retrieval-based interference. Memory-based theories of human sentence processing aim to explain both retrieval and encoding effects. Encoding interference is therefore not an optional phenomenon, but a necessary test case for any proposed memory

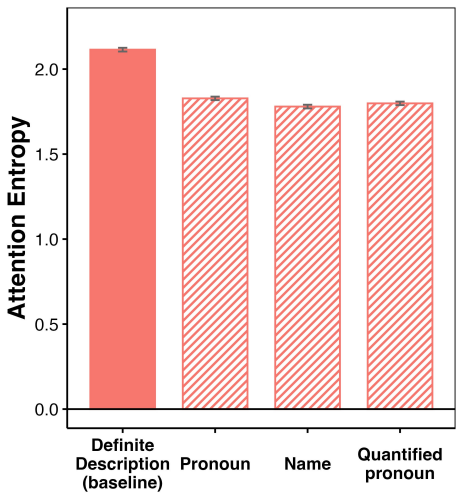


Figure 3: Attention entropy at the critical region for the baseline, pronoun, name, and quantified pronoun conditions. Higher values reflect more diffuse attention over the prior context, indicating greater uncertainty. The solid bar (definite description) is predicted to exhibit higher entropy due to encoding interference from overlap in referential class. Striped bars are conditions predicted to exhibit lower entropy due to reduced interference from referential dissimilarity. Error bars represent standard error of the mean.

architecture.

Under standard cue-based accounts, encoding interference arises when multiple items share overlapping features at the time of storage. This overlap reduces the distinctiveness of the resulting memory representations, which in turn disrupts subsequent access. Crucially, this form of interference is not driven by retrieval cue competition per se, but by the properties of the representations themselves as established during encoding.

If attention is a general-purpose memory system, then it should exhibit analogous sensitivity to the representational conditions that give rise to encoding interference. The present findings show that it does. Attention-based metrics reproduce the attenuation of the ORC/SRC asymmetry under conditions of referential dissimilarity, a hallmark signature of encoding-based semantic interference in human behavior. This result is therefore theoretically diagnostic. It demonstrates that attention captures not only competition among candidates at retrieval, but also the effects of representational similarity established at encoding.

5.3. Reinterpreting encoding interference as representational geometry

While attention captures encoding interference, it does so in a way that departs fundamentally from how such effects are typically characterized in human memory models. In standard accounts, encoding interference is understood as a consequence of feature-based overlap: when multiple items share features, those features are assumed to be degraded or overwritten, resulting in less distinct memory traces (Nairne, 1990; Oberauer and Kliegl, 2006).

This characterization does not transfer straightforwardly to Transformer models. At the point of retrieval, token representations are fixed: the key vectors associated with prior tokens are not modified or degraded. Encoding interference in Transformers therefore cannot arise from literal feature loss or representational corruption.

This creates a puzzle: If encoding interference in human models is attributed to feature loss, how can analogous effects arise in a system where no such loss occurs?

The answer lies in the geometry of the representational space. Encoding interference in Transformers emerges from two interacting components:

First, similar items are encoded into nearby regions of representational space. Nouns that share semantic properties, such as referential class, tend to have similar contextualized embeddings and therefore occupy proximate locations in key space.

Second, retrieval is implemented as a soft competition over this space. A query vector generated at a retrieval site (e.g., a verb) distributes attention weights across prior tokens as a function of their similarity to the query. When multiple candidates are located near one another in key space, they produce comparable query–key alignment scores. As a result, attention is distributed across them rather than concentrated on a single target.

Under these conditions, no item dominates retrieval. Attention is split, leading to higher entropy and reduced precision in identifying the intended referent. Importantly, this outcome arises without any loss or degradation of representational content. All features remain intact. Interference is instead a consequence of insufficient separability among representations in a continuous space.

On this view, we can establish the following functional equivalence between attention-based retrieval in Transformers and memory-based models of human sentence processing. In both systems: (i) the retrieval cue corresponds to a query vector, (ii) candidate items are accessed based on similarity in feature space (via key–query alignment), (iii) referential overlap leads to multiple items matching on cue-relevant dimensions, reducing diagnosticity.

This perspective reframes encoding interference. Rather than reflecting the loss of information, interference reflects the structure of representational geometry: when items are too similar in the dimensions relevant for retrieval, they compete for selection.

5.4. Implications for memory representations in sentence processing

Reinterpreting encoding interference in this way has important implications for memory-based theories of human sentence processing. In traditional models, representations are constructed from discrete, hand-specified features, and similarity is defined categorically in terms of feature overlap. Abstract semantic properties, such as referential class, must be specified as symbolic features that are either shared or not.

The attention-based results suggest a different view. Representations are distributed rather than bundles of discrete features, and similarity is gradient rather than categorical. Under this view, abstract semantic properties are not encoded as discrete feature units, but as patterns within a high-dimensional space. Similarity between items is a matter of distance within that space, not shared feature labels.

This shift has a direct consequence for how interference is understood in human sentence processing. Interference need not be implemented through explicit cognitive processes of feature loss, overwriting, or decay. Instead, it emerges naturally from

the geometry of the representational space: when items are insufficiently separated along relevant dimensions, retrieval becomes less precise because multiple candidates occupy nearby regions.

This perspective eliminates the need to posit dedicated cognitive mechanisms for feature degradation at encoding (e.g., feature overwriting). Consequently, it contrasts with recent proposals such as lossy-context surprisal (Futrell et al., 2021), which model noisy memory as the loss of information via erasure. The attention-based account suggests that interference effects can be explained without assuming that information is lost. What matters is how information is structured.

5.5. Encoding and retrieval as a unified mechanism

A related implication concerns the organization of memory operations. In cognitive models of sentence processing, encoding and retrieval are treated as distinct stages. Encoding determines what information is stored, and retrieval determines how that information is accessed using a set of cues. Interference can arise at either stage, but the mechanisms are conceptually separate.

In attention-based models, this distinction is collapsed. When each token is processed, it is projected into a representational space through learned transformations that produce key and value vectors. Simultaneously, subsequent tokens generate query vectors that probe that same space. Retrieval is implemented as a similarity computation between queries and keys, and the resulting attention weights determine how prior information is integrated.

Crucially, the same learned projections govern both processes. Encoding determines where items are placed in representational space, and retrieval consists of probing that space using similarity. There is no separate encoding module and retrieval module. Both operations are governed by a shared representational geometry.

This has direct consequences for interference. In the present study, interference arises when two nouns with similar referential properties are encoded into nearby regions of key space. When the verb generates a query, both candidates match the query, and attention is distributed across them. The interference originates in the similarity structure established at encoding, but it manifests during retrieval. Because both processes rely on the same underlying mechanism, they are necessarily intertwined.

This unified view offers a different way of understanding memory in sentence processing. Rather than a sequence of distinct stages, memory operations are better characterized as continuous interactions within a shared representational space.

5.6. Broader implications and future directions

The results support three primary conclusions: attention captures retrieval- and encoding-based interference within a single framework; attention is sensitive to syntactic and semantic distinctions; and interference emerges from representational geometry rather than feature loss.

Future work should prioritize a high-powered replication of the original behavioral findings reported in Gordon et al. (2001, 2004) to enable a direct comparison between model-derived and human effect sizes. Additionally, extending this analysis to a broader range of architectures (e.g., recurrent networks or bounded Transformers) will clarify whether these interference patterns are inherent to the attention mechanism itself or are a general property of learned sequence models under resource constraints.

Ultimately, these findings motivate a shift from discrete, feature-based representations used in traditional models like ACT-R toward distributed, similarity-based representations as the basis for encoding and retrieval in human sentence processing. This perspective opens new avenues for integrating computational models and cognitive theory in the study of language processing.

6. Bibliographical References

Richard Futrell, Edward Gibson, and Roger P. Levy. 2021. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44:1–54.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. 2017. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(555):1–13.

Edward Gibson. 2000. Dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126. MIT Press.

Peter C. Gordon, Randall Hendrick, and Marcus Johnson. 2001. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(6):1411–1423.

Peter C. Gordon, Randall Hendrick, and Marcus Johnson. 2004. Effects of noun phrase type on

sentence complexity. *Journal of Memory and Language*, 51:97–104.

Daniel Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, 29:261–290.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2019. Exbert: A visual analysis tool to explore learned representations in transformers models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Cassandra L. Jacobs and Maryellen C. MacDonald. 2024. Constraint satisfaction in large language models. *Language, Cognition and Neuroscience*, 39(10):1231–1248.

Jonathan King and Marcel A. Just. 1991. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30:580–602.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.

James S. Nairne. 1990. A feature model of immediate memory. *Memory & Cognition*, 18:251–269.

Klaus Oberauer and Reinhold Kliegl. 2006. A formal model of capacity limits in working memory. *Journal of Memory and Language*, 55:601–626.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5(777963).

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from gpt-2 attention patterns predict reading times over and above gpt-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121.

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71. Association for Computational Linguistics.

Soo Hyun Ryu and Richard L. Lewis. 2025. Memory for prediction: A transformer-based theory of sentence processing. *Journal of Memory and Language*, 145:1–19.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2018. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Julie A. Van Dyke and Clinton L. Johns. 2012. Memory interference as a determinant of language comprehension. *Language and Linguistics Compass*, 6(4):193–211.

Julie A. Van Dyke and Brian McElree. 2006. Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55:157–166.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31.

ChineseDevBench: A Chinese Developmental Benchmark for Language Development

Shaonan Wang[1,*†], **Yiwen Wu**[2,*], **Na Li**[3,*], **Zesheng Chen**[3], **Gan Wang**[3], **Shuchen Zhang**[3], **Xin Sun**[1], **Luan Li**[4], **Yaran Chen**[3,†]

[1]Department of Language Science and Technology, Hong Kong Polytechnic University

[2]Faculty of Arts and Education, University of Auckland

[3]Xi'an Jiaotong-Liverpool University

[4]School of Foreign Languages, Shanghai Jiao Tong University

shaonan.wang@polyu.edu.hk, yaran.chen@xjtlu.edu

Abstract

How similar are the learning trajectories of language models and children? Recent work has narrowed the data-efficiency gap by training language models on child-scale input—roughly 10^8 tokens by early adolescence. However, evaluation remains largely based on adult-oriented English benchmarks and rarely involves direct comparison with human developmental data. We introduce ChineseDevBench, a Mandarin developmental benchmark comprising eight tasks that probe word meaning comprehension, association structure, acquisition dynamics, and language production. Crucially, the benchmark includes behavioral data from both children and adults, enabling direct model–human comparison. We train Chinese GPT-2 models on child-scale Mandarin input using an age-based curriculum and evaluate alignment between model and human response patterns across training checkpoints. Across most tasks, alignment improves with training but developmental age does not predict performance. ChineseDevBench provides a framework for systematically characterizing where model learning converges with—and diverges from—human language development.

Keywords: language development, learning trajectory, language models

1. Introduction

Humans acquire language rapidly with limited explicit supervision, progressing from isolated words to structured, meaning-rich utterances. Explaining this trajectory—and identifying which aspects are universal versus language-specific—remains a central goal of developmental research (Bowerman and Levinson, 2001).

Mandarin Chinese provides a particularly important test case. As the world's most widely spoken native language, Mandarin differs typologically from English in morphology, argument structure, and form–meaning mappings. Prior work shows that developmental trajectories in Mandarin syntax and grammatical–semantic marking diverge from English patterns (Huang et al., 2022; Ma et al., 2009). Corpus studies further document systematic age-related shifts in semantic expression beyond vocabulary growth (Tang et al., 2023). These findings motivate benchmarks that track Mandarin language development across multiple dimensions of meaning.

In parallel, modern language models achieve strong performance through web-scale training far exceeding human input. A child encounters on the order of 10^8 tokens by early adolescence, whereas many models are trained on orders of magnitude more. Initiatives such as BabyLM attempt to match child-scale exposure, but evaluation remains largely adult-centric and English-based, with limited comparison to child behavioral data (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025; Tan et al., 2024). Moreover, recent work suggests a gap between linguistic competence and human-like processing. Steuer et al.

* Equal contribution.

† Corresponding author.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 10–24
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(2023) show that models achieving strong performance on linguistic benchmarks often exhibit a poorer fit to human reading times, pointing to a divergence between linguistic competence and psycholinguistic measures.

These findings point to two related concerns. First, it remains unclear whether results obtained from English-based models generalize cross-linguistically in the context of human language development. Second, it remains unclear whether this divergence between linguistic competence and human-like processing extends to other dimensions of language development, such as analogy and production.

To address these gaps, we introduce ChineseDevBench, a Mandarin developmental benchmark spanning eight tasks, including child and adult behavioral data probing lexical, semantic, and relational knowledge. Mandarin, with its distinct typological properties, offers an important test case for whether current findings generalize cross-linguistically. We train Chinese GPT-2 models on child-scale Mandarin input using a chronological curriculum and evaluate model–human alignment across training checkpoints, allowing us to track developmental trajectories and identify systematic convergences and divergences with human learning.

2. ChineseDevBench

ChineseDevBench consists of eight tasks organized into two groups: five child-oriented tasks (three unreleased and two publicly available) and three publicly available adult-oriented tasks. Together they probe compositional reasoning, associative structure, category knowledge, age of acquisition, production patterns, and perceptual feature structure.

2.1. Child-Oriented Tasks

CHILD-ANALOGY: Compositional Analogy

This task assesses morpho-semantic pattern completion through compositional analogy (The task originates from (Sun et al., 2022), and uses an unpublished dataset collected by co-author Sun). Children receive a short Mandarin prompt providing an example mapping from a description to a name, followed by a minimally changed description (15 in total). For example: "有一个又大又红的花叫大红花，那么有一个又大又黄的花叫什么? (There is a big, red flower called a big red flower; then what would you call a big, yellow flower?)"→ "大黄花 *(big yellow flower)*".

Population: 117 Mandarin-speaking children aged 3–6 years, divided into three age bins: 3–4 years (n=6), 4–5 years (n=63), and 5–6 years (n=48).

Evaluation: For each item, we compute the conditional log probability of the target word given the stimulus prompt (ending with 我们叫它 "we call it"). Performance is measured by Spearman correlation between children's accuracy and model scores.

CHILD-ASSOC: Word Association

This task evaluates cue–response association structure (574 in total) using child production data[1].

Population. 733 Mandarin-speaking children aged 2.5–7.5 years, divided into five age bins: 2.5–3.5 years (n=11), 3.5–4.5 years (n=102), 4.5–5.5 years (n=238), 5.5–6.5 years (n=306), and 6.5–7.5 years (n=76).

Evaluation. We use Representational Similarity Analysis (RSA). A human distance matrix is constructed from cue–response probability distributions; a corresponding matrix is derived from model embeddings. Alignment is measured via Spearman correlation between flattened matrices.

CHILD-FLUENCY: Verbal Fluency (Animals and Fruits)

This task assesses category (semantic) fluency. Children receive a category prompt (e.g., 请说出尽可能多的动物 "Please name as many animals as you can" / 请说出尽可能多的水果 "Please name as many fruits as you can") and produce an ordered list of category members. Data were collected from Li and Hills (2026), yielding 223 animal

[1] Unpublished data collected by co-author Li's group.

names and 90 fruit names.

Population: 733 Mandarin-speaking children aged 2.5–7.5 years, divided into five age bins consistent with CHILD-ASSOC.

Evaluation: For each category, we compute the conditional log probability of the target word given the stimulus prompt (e.g., 动物有 "Animals include" / 水果有 "Fruits include". Performance is measured by the Spearman correlation between the model's scores and children's production frequencies (occurrences divided by the number of children) at each age bin.

CHILD-WORD-ACQ: Word Acquisition

This task compares model-estimated age of acquisition (AoA) with human norms.

Source: Wordbank item data[2] and AOA data from Liu et al. (2011).

Population: Mandarin-speaking children from existing AoA norm datasets: 1,056 children aged 16–30 months from Wordbank (n≈70 per monthly age bin), and 262 children from Liu et al. (2011) across three kindergarten levels—K1 (n=99), K2 (n=64), and K3 (n=99).

Evaluation: Following (Chang and Bergen, 2022), we estimate model AoA by tracking token-level surprisal across training checkpoints. For each target word, we extract all sentences containing that word from adult speech in the CHILDES corpus and compute surprisal at each checkpoint. A sigmoid function is fitted to the resulting learning curve, and model AoA is operationalized as the training step at which surprisal crosses a threshold midway between baseline and minimum. Human–model alignment is assessed using Spearman rank correlation between model-predicted and child AoA. We selected 309 of 797 Wordbank words and 66 of 435 words from Liu et al. (2011), excluding items with insufficient occurrences for reliable estimation.

CHILD-PRODUCTION: Language Production

This task evaluates the model's ability to generate child-like continuations conditioned on short prefixes, and compares distributional properties of the generated output with developmental norms from child speech.

Source: CHILDES Mandarin corpus and CPCSLD (Chinese Preschool Children's Spoken Language Database)(Feng et al., 2026). Both corpora consist of transcribed child speech or lexical terms, and we use the existing transcriptions provided by these resources rather than raw audio.

Population: Mandarin-speaking children across developmental stages: <24, 24–30, 30–36, 36–42, 42–48, 48–60, 60–72, and 72+ months in CHILDES; 648 children from the CPCSLD dataset across three kindergarten levels—K1 (n=184, M=49.72 months, SD=4.08), K2 (n=181, M=60.60 months, SD=4.58), and K3 (n=283, M=71.36 months, SD=4.74).

Evaluation: We use short high-frequency prefixes selected from child utterances in the CHILDES Mandarin corpus (e.g., 我 喜 欢..."I like...") as prompts, and generate continuations using stochastic decoding (top-k = 50, top-p = 0.95, temperature = 0.9), with a maximum length of 50 tokens. For each checkpoint, 150 utterances are generated. To ensure linguistic validity and comparability with early child productions, we extract the first sentence segment (delimited by punctuation) from each generated continuation as the unit of analysis. The resulting utterances are segmented and part-of-speech tagged using jieba[3]. For each checkpoint, we compute lexical properties as well as shallow and sentence-level syntactic properties. These include the proportion of major part-of-speech categories (nouns, verbs, adjectives, adverbs), along with syntactic measures such as the proportion of clausal subjects. We then compare these distributions against benchmarks from two sources: lexical norms from CPCSLD (K1, K2, K3, ages 3–6), and distributional estimates computed from CHILDES child speech across eight age groups.

Note that the short prefixes used as generation prompts may also occur in the model's training data, as they correspond to high-frequency utterance-initial fragments commonly produced by

[2] https://wordbank.stanford.edu/data/?name=item_data

[3] https://github.com/fxsjy/jieba

young children. However, this does not constitute data contamination in the conventional sense. The prefixes are used solely as conditioning context, and the evaluation focuses on the distributional properties of the generated continuations rather than on the prefixes themselves. Moreover, such fragments are ubiquitous in child-produced speech, and their presence in training data is both unavoidable and ecologically appropriate.

2.2. Adult-Oriented Tasks

ADULT-STS: Semantic Textual Similarity

This task evaluates semantic textual similarity between sentence pairs.

Source: ChineseSTS[4] comprising 14,743 sentence pairs with human similarity ratings.

Evaluation: We compute Spearman's rank correlation between the model's cosine similarity scores and human gold-standard similarity scores for sentence pairs.

ADULT-ASSOC: SWOW-ZH Mandarin Word Association Norms

This task uses multiple-response free association (three responses per cue) with aggregate associative strengths.

Source: SWOW-ZH repository[5], providing word association data for 10,192 cues and over 2 million responses collected between 2016–2023 (Li et al., 2024).

Evaluation: RSA evaluation method identical to CHILD-ASSOC.

ADULT-SEMFEAT-6: Semantic Feature Norms

This task predicts semantic feature ratings across six dimensions: Vision, Motor, Socialness, Emotion, Time, and Space.

Source: Six Semantic Dimension Database[6], containing subjective ratings for 17,940 Chinese words (Wang et al., 2023).

Evaluation: For each target word, we extract its static representation by applying mean pooling over the last hidden layer. A linear regression probe maps these embeddings to human perceptual ratings using 5-fold cross-validation. Alignment is evaluated using Spearman's rank correlation between cross-validated model predictions and human ratings.

3. Language Model

3.1. Training Dataset

The training data comprise three categories: transcripts from children's cartoons, digitized children's picture books and audiobooks, and recordings of child-directed speech.

Children's Cartoons: We collected approximately 2,000 cartoon videos from Bilibili and YouTube, including roughly 1,000 videos targeting children aged 0–3 and 1,000 targeting children aged 3–6. Video selection was guided by age-specific keywords and titles, which were cross-referenced with children's media platforms[7] to ensure age appropriateness. Transcripts were generated using OpenAI's Whisper automatic speech recognition system.

Children's Picture Books and Audiobooks: We collected 1,158 digitized storybooks for children aged 0–6. Age labels and content were obtained from an open-access repository of Chinese children's literature that provides corresponding audio recordings for each story[8]. Audio content was transcribed using OpenAI's Whisper automatic speech recognition system.

CHILDES Mandarin Corpus: A core component of our conversational data was sourced from the TalkBank CHILDES Mandarin database[9]. We applied filtering criteria to select files with clear age markings and appropriate speaker roles. The selected data encompasses multiple

[4] https://github.com/IAdmireu/ChineseSTS

[5] https://github.com/lib314a/SWOWZH

[6] https://www.nature.com/articles/s41597-023-01995-6

[7] https://v.qq.com/channel/child

[8] https://www.limaogushi.com/

[9] https://talkbank.org/childes/access/Chinese/Mandarin/

sub-corpora including Beijing, Chang (Chang1, Chang2, ChangPlay), Erbaugh, LiReading, LiZhou, NSCtoys, Tong, and the comprehensive Zhou collection (Zhou1, Zhou2, Zhou3, ZhouAssessment, ZhouDinner, ZhouNarratives).

3.2. Training Settings

We implemented a decoder-only Transformer model based on the GPT-2 Small architecture, consisting of 12 Transformer blocks, 12 attention heads, and a hidden size of 768. We adopted the `bert-base-chinese` tokenizer (vocabulary size ≈21,128). Although originally derived from large-scale adult corpora, this tokenizer operates largely at the character level for Mandarin Chinese, where individual characters typically constitute stable and meaningful units and inflectional morphology is minimal. As a result, it approximates character-level modeling while maintaining compatibility with widely used Chinese NLP resources. The total parameter count is approximately 100M.

Training follows a chronological curriculum simulating child language acquisition, divided into two sequential stages:

Stage 1 (0–3 years): Focuses on simpler lexical items and sentence structures, comprising approximately 1.56 million tokens (across 5,126 extracted sentences). Training duration: 10 epochs (≈770 total steps). Checkpoints saved at each epoch end (11 checkpoints).

Stage 2 (3–6 years): Expands to approximately 2.97 million tokens (across 13,713 extracted sentences). Training duration: 10 epochs (≈5,700 total steps). Checkpoints saved at each epoch end (11 checkpoints).

Cumulative exposure totals approximately 4.53 million tokens. This checkpointing strategy enables reconstruction of the continuous learning trajectory across developmental phases.

3.3. Baseline Model

For comparative benchmarking of adult-level linguistic competence, we use UER GPT-2 Chinese[10]. This model shares identical architecture with our cognitive agent but was pre-trained on CLUECorpusSmall (approximately 14GB of general-domain Chinese text). This baseline quantifies the gap between models trained on limited, developmentally plausible child-directed speech and models trained on extensive open-domain data.

4. Results

We evaluated the developmental trajectory of our curriculum-trained model across ChineseDevBench. All model embeddings were extracted from the last layer. Results are organized into four sections: child-oriented comprehension and association tasks, word acquisition dynamics, language production, and adult-oriented semantic tasks.

4.1. Child-Oriented Tasks

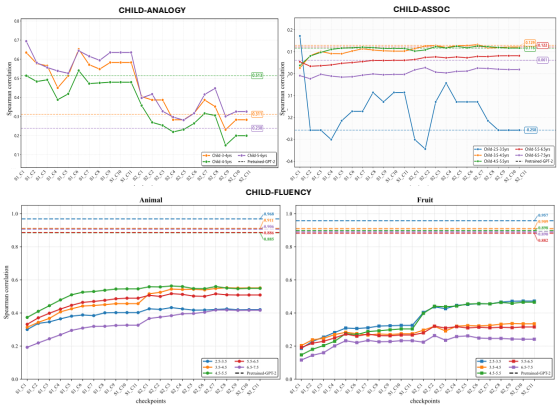


Fig. 1: GPT-2 developmental trends on CHILD-ANALOGY, CHILD-ASSOC, and CHILD-FLUENCY tasks.

Figure 1 presents results for three child-oriented tasks. For CHILD-ANALOGY, the first training stage achieves comparable or higher correlations with human responses than the pre-trained model, but performance drops to pre-trained levels during the second stage, suggesting that models trained on smaller datasets may

[10] https://huggingface.co/uer/gpt2-chinese-cluecorpussmall

better approximate child-like behavior for certain properties. For CHILD-ASSOC and CHILD-FLUENCY, model–human alignment generally improves with training, indicating that curriculum learning on child-directed input induces increasingly human-like semantic representations. For CHILD-ASSOC, RSA correlations remain low for both child-GPT and pretrained GPT-2, with greater variability in the 2.5–3.5 age bin due to smaller sample sizes; children's developmental age does not correspond to better model alignment, suggesting that alternative evaluation metrics may be needed. For CHILD-FLUENCY, a substantial performance gap exists between pretrained and child-GPT2 models, and developmental age bins show no consistent relationship with model performance.

4.2. Word Acquisition Dynamics (CHILD-WORD-ACQ)

We compared model and child word acquisition by tracking surprisal across training checkpoints and fitting sigmoid learning curves (Figure 2). Model AoA was defined as the training step at which surprisal crossed a threshold midway between baseline and minimum, analogous to the 50% production criterion in child studies.

Shared acquisition patterns. Both the model and children acquire certain word classes early: kinship terms (e.g., 奶奶 "grandmother," 哥哥 "older brother"), common animals (狗 "dog," 兔子 "rabbit"), basic household objects (电话 "phone"), and high-frequency everyday items (车 "car," 饭 "rice," 肉 "meat"). Conversely, both acquire abstract or low-imageability items late, such as the honorific classifier 位 and the plural pronoun 你们 "you-PL."

Divergent acquisition patterns. Two systematic dissociations emerged. First, grammatical function words (conjunctions, temporal markers) were acquired rapidly by the model but late by children (e.g., 已经 "already": model 142 tokens vs. child 29 months; 因为 "because": model 147 tokens vs. child 30 months). Second, embodied vocabulary—action verbs and sensory adjectives—showed the opposite pattern: children acquired these early (19–21 months), whereas the model required thousands of tokens (e.g., 带 "carry": model 4583 tokens vs. child 20 months; 香 "fragrant": model 2648 tokens vs. child 19 months).

These patterns suggest that the model efficiently exploits distributional regularities to learn function words, while children rely on conceptual development. Conversely, children leverage sensorimotor experience for action and sensory vocabulary, a grounding mechanism unavailable to text-only models. Detailed word lists are provided in Appendix Tables A1–A4.

Comparison with English. Consistent with Chang and Bergen (2022), we observed a negative effect of word length on model AoA: shorter words were acquired later in training. This contrasts with child language acquisition, where shorter words are typically learned earlier. Children are more likely to store short chunks of speech rather than chunks that are merely frequent or predictable (Grimm et al., 2019), a tendency shaped by phonological simplicity, memory constraints, and social interaction. Language models, in contrast, rely solely on distributional patterns in text, where shorter and more polysemous words are harder to represent consistently. Thus, the same surface property can have opposite effects depending on the underlying learning mechanism.

Unlike in English unidirectional models, where nouns tend to be acquired later than adjectives and verbs, our Chinese GPT-2 model acquired nouns relatively early. Conversely, verbs, acquired early by English models, were acquired later in the Chinese model. This pattern likely reflects several distributional challenges. Chinese verbs are often polysemous and context-dependent, making them more difficult to learn from distributional input alone. In contrast, Chinese nouns are more semantically stable, providing cleaner distributional signals. Additionally, the relatively flexible word order of Chinese allows verbs to appear in variable syntactic positions, potentially reducing the consis-

tency of their distributional contexts compared to nouns.

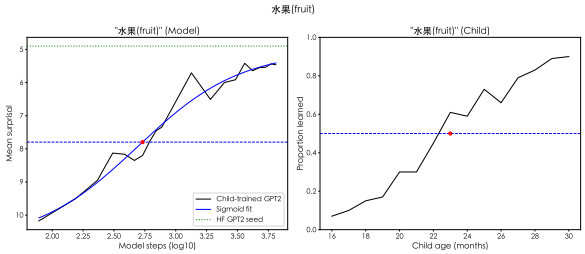


Fig. 2: Learning curves for the word "水果"(fruit) in a GPT-2 language model and human children. The blue horizontal lines indicate age of acquisition cutoffs. The green horizontal line reflects the baseline computed from the HF GPT-2 model prior to any child-directed training. The blue curve represents the fitted sigmoid function based on the language model surprisals during training (black). Child AoA values obtained from Hao et al. (2008).

4.3. Language Production (CHILD-PRODUCTION)

We examined whether GPT-2 exhibits child-like developmental trajectories in lexical complexity and syntactic category usage, comparing model outputs across checkpoints with CPCSLD benchmarks (kindergarten ages 3–6).

Lexical development. Table 1 and Figure 3 show that GPT-2 reproduces most qualitative developmental patterns: one-syllable words decrease while multi-syllable words increase, and noun proportions rise while adjective proportions fall. Five of six key indicators align with CPCSLD trajectories, providing partial support for human-like lexical development under curriculum learning.

Syntactic development. Figure 4 reveals a divergence between GPT-2 and children. Children in CHILDES exhibit low initial syntactic complexity followed by steady increases, reflecting staged acquisition of embedding structures. In contrast, GPT-2 shows substantial variability but no systematic growth, suggesting that the model accesses complex constructions early without undergoing human-like incremental syntactic development.

Table 1: Production trends in word length in syllables and POS distribution: GPT-2 checkpoints (Model) vs. CPCSLD benchmarks (Ref). *Align* indicates directional consistency with reference trajectories.

Feature	Model	Δ%	Ref	Align
Word Length				
1-syllable	↓	-7.63	↓	Yes
2-syllable	↑	+4.34	↑	Yes
3+ syllable	↑	+2.83	↑	Yes
POS Types				
Noun	↑	+3.97	↑	Yes
Verb	↑	+3.40	↓	No
Adjective	↓	-3.79	↓	Yes

4.4. Adult-Oriented Tasks

Figure 5 presents results for three adult-oriented tasks. Consistent with the child task results, a performance gap exists between pre-trained GPT-2 and child-GPT2 models, and model–human alignment generally improves with training for ADULT-STS and ADULT-SEMFEAT-6. For ADULT-STS and ADULT-ASSOC, performance drops sharply from the first to second training stage before recovering, suggesting that these tasks are sensitive to shifts in embedding distributions. Similar to CHILD-ASSOC, ADULT-ASSOC yields low correlation scores overall. For ADULT-SEMFEAT-6, the six semantic dimensions show differential alignment with human ratings: Vision, Socialness, and Space achieve higher correlations, while Time and Emotion prove most difficult to align—possibly because affective and temporal semantics require experiential grounding beyond what distributional statistics can provide.

5. Discussion

Human language acquisition has been explained by competing theoretical frameworks. Nativist accounts propose that children rely on innate linguistic constraints (Chomsky, 2002), while usage-based approaches argue that language emerges from actual use (Tomasello, 2003; Ib-

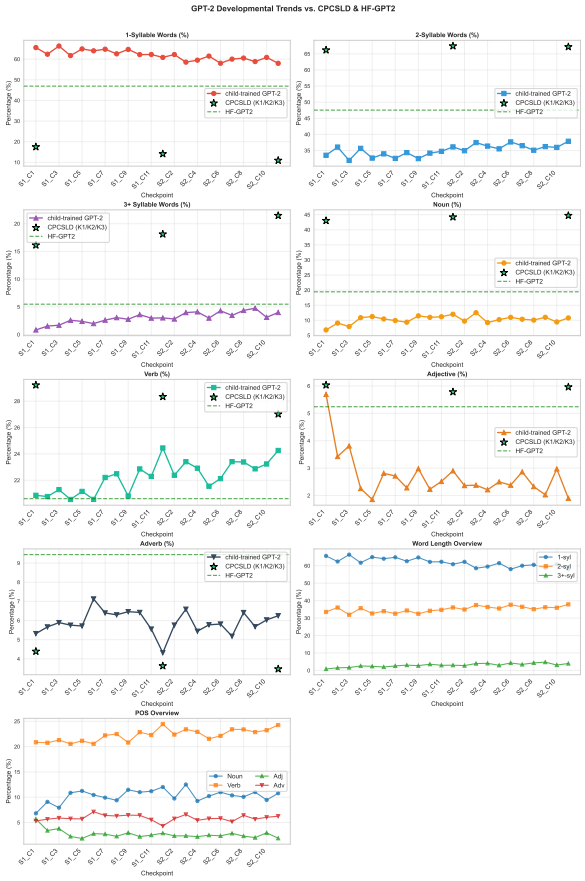


Fig. 3: Production: GPT-2 developmental trends compared with CPCSLD, focusing on lexical measures.

botson, 2013). Usage-based theories emphasize sensitivity to recurring patterns in input. Language models are likewise sensitive to such regularities in linguistic data. However, models lack the social-cognitive dimension that usage-based theorists consider equally essential to language development.

Following Portelance and Jasbi (2024), we interpret the model's behavior as jointly informative about the scope and limits of distributional learning. The model's strengths indicate aspects of language that can be captured from distributional information, while its limitations point to domains in which additional mechanisms—sensorimotor, social, or cognitive—likely play a crucial role. These mechanisms do not uniformly accelerate learning, but may shape different aspects of acquisition in distinct ways. The model's delayed acquisition of action verbs and sensory adjectives,

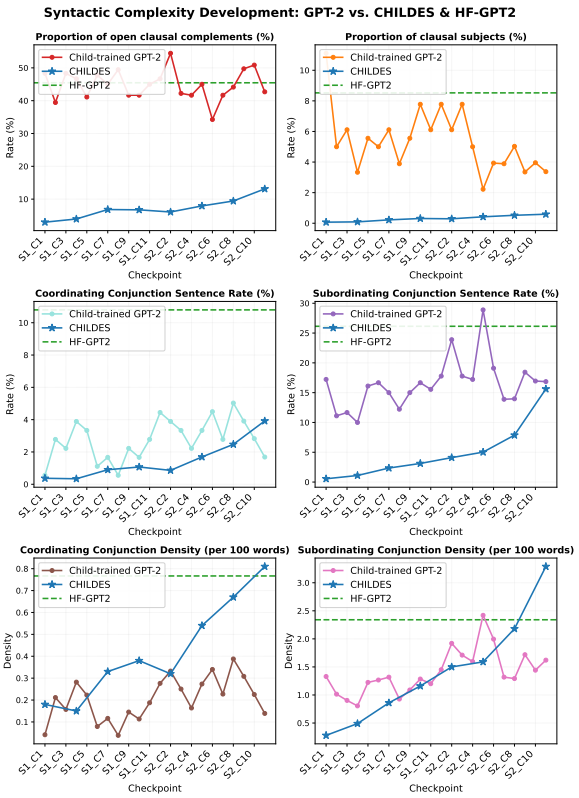


Fig. 4: Production: GPT-2 developmental trends relative to CHILDES-derived syntactic measures. Blue stars denote syntactic measures computed from CHILDES child speech for eight age groups (<24, 24–30, 30–36, 36–42, 42–48, 48–60, 60–72, and 72+ months), with one point per age group.

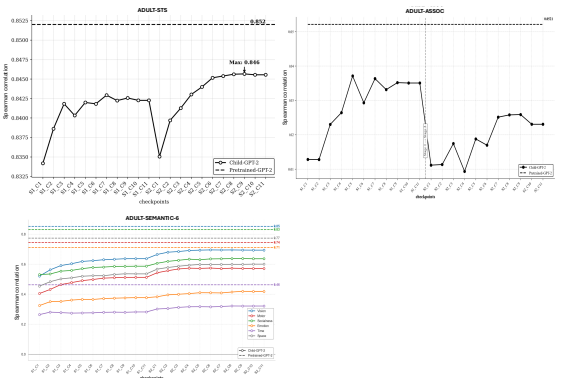


Fig. 5: GPT-2 developmental trends on ADULT-STS, ADULT-ASSOC and SEMFEAT-6 tasks.

which are typically acquired early in child development, implicates that sensorimotor grounding constitutes an important inductive bias in children's language acquisition. Conversely, children's comparatively late acquisition of function words, de-

spite their high distributional regularity in child-directed speech, suggests that their acquisition may depend on conceptual and pragmatic capacities that develop gradually and are not solely determined by input frequency. Taken together, these divergences suggest that human learners and language models may rely on different underlying mechanisms, even when models achieve similar outcomes.

Future work may explore how such mechanisms can be incorporated into computational models, for example through multimodal input or interactive learning settings, bringing models of language acquisition closer to the cognitive processes they aim to approximate.

6. Conclusion

ChineseDevBench provides a Mandarin developmental benchmark integrating child and adult behavioral data across eight tasks. By training GPT-2 on child-scale input with a chronological curriculum, we track model learning trajectories and compare them directly with human development. Results reveal both convergence and systematic divergence. For most child-oriented tasks, alignment improves with training, yet developmental age does not predict performance and syntactic complexity lacks human-like growth; the model rapidly acquires function words but lags on embodied vocabulary. For adult-oriented tasks, alignment also improves with training, though association tasks yield low correlations, and dimensions like Time and Emotion remain difficult to align, suggesting they require experiential grounding. Unlike English models, which acquire nouns later than adjectives and verbs, the Chinese GPT-2 model acquired nouns early and verbs later, highlighting language-specific effects on acquisition order. ChineseDevBench enables more precise evaluation of developmentally plausible language models and provides a foundation for future work on grounding, curriculum design, and cross-linguistic developmental comparison. *The data and evaluation code are publicly available at https://github.com/Evaisgreat/Chinese-DevBench.*

7. Ethical statements and limitations

All child data are drawn from publicly available datasets, datasets that will be made publicly available, or previously published datasets collected under appropriate ethical oversight. No new child data were collected for this study, and no personally identifiable information is included. The sampled populations may not fully represent the diversity of Mandarin-speaking children.

Part of the training data were automatically transcribed using OpenAI's Whisper speech recognition system, which may introduce transcription errors. Some evaluation metrics rely on specific prompting methods, and results may vary under different prompts. In addition, the available child data are relatively limited in size; more publicly released datasets will be necessary to obtain more robust and generalizable findings. Model–child comparisons are behavioral rather than mechanistic, and observed alignment does not imply shared underlying cognitive processes. The benchmark focuses on lexical and semantic development and does not comprehensively cover pragmatic or discourse-level abilities. Finally, the benchmark is intended for research purposes only and should not be used for clinical assessment.

8. Acknowledgements

This work was supported by the start-up fund project (1-BDE3) sponsored by the Faculty of Humanities of the Hong Kong Polytechnic University, the Research Development Fund (RDF) of Xi' an Jiaotong-Liverpool University (XJTLU) under Grant No. RDF-24-02-028 and Suzhou Science and Technology Development Planning Programme (Grant No. ZXL2025310）.

9. References

Melissa Bowerman and Stephen C Levinson. 2001. *Language acquisition and conceptual development*. 3. Cambridge University Press.

Tyler A Chang and Benjamin K Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Lucas Georges Gabriel Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candance Ross, et al. 2025. Findings of the third babylm challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420.

Noam Chomsky. 2002. *Syntactic structures*. Walter de Gruyter.

Chen Feng, Shuo Wang, and Shuang Li. 2026. Cpcsld: A lexical database of chinese preschool children's spoken words. *Behavior Research Methods*, 58(2):54.

Robert Grimm, Giovanni Cassani, Steven Gillis, and Walter Daelemans. 2019. Children probably store short rather than frequent or predictable chunks: Quantitative evidence from a corpus study. *Frontiers in psychology*, 10:80.

Min Hao, Hua Shu, Aiqing Xing, and Ping Li. 2008. Early vocabulary inventory for mandarin chinese. *Behavior Research Methods*, 40(3):728–733.

Michael Y Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2412.05149*.

Ruoyu Lexie Huang, Paul Fletcher, Zhixiang Zhang, Weilan Liang, Virginia Marchman, and Twila Tardif. 2022. Early grammatical marking development in mandarin-speaking toddlers. *Developmental psychology*, 58(4):631.

Paul Ibbotson. 2013. The scope of usage-based theory. *Frontiers in psychology*, 4:255.

Bing Li, Ziyi Ding, Simon De Deyne, and Qing Cai. 2024. A large-scale database of mandarin chinese word associations from the small world of words project. *Behavior Research Methods*, 57(1):34.

Luan Li and Thomas T. Hills. 2026. Exploration-versus-exploitation of semantic networks promotes early lexical development: Evidence from preschoolers' semantic fluency and computational simulations (unpublished pre-print). *Retrieved from https://osf.io/7t3hn/*.

Youyi Liu, Meiling Hao, Ping Li, and Hua Shu. 2011. Timed picture naming norms for mandarin chinese. *PloS one*, 6(1):e16505.

Weiyi Ma, Roberta Michnick Golinkoff, Kathy Hirsh-Pasek, Colleen McDonough, and Twila Tardif. 2009. Imageability predicts the age of acquisition of verbs in chinese children. *Journal of child language*, 36(2):405–423.

Eva Portelance and Masoud Jasbi. 2024. The roles of neural networks in language acquisition. *Language and Linguistics Compass*, 18(6):e70001.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. Large gpt-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 142–157.

Xin Sun, Kehui Zhang, Rebecca A Marks, Nia Nickerson, Rachel L Eggleston, Chi-Lin Yu, Tai-Li Chou, Twila Tardif, and Ioulia Kovelman. 2022. What's in a word? cross-linguistic influences on spanish–english and chinese–english bilingual children's word reading development. *Child development*, 93(1):84–100.

Alvin Tan, Chunhua Yu, Bria Long, Wanjing Ma, Tonya Murray, Rebecca Silverman, Jason Yeatman, and Michael C Frank. 2024. Devbench: A multimodal developmental benchmark for language learning. *Advances in Neural Information Processing Systems*, 37:77445–77467.

Tempo Po-Yi Tang, Dustin Kai-Yan Lau, and Man-Tak Leung. 2023. Corpus of mandarin child language: a preliminary study on the acquisition of semantic content categories in mandarin-speaking preschoolers. *Frontiers in psychology*, 14:1234525.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Shaonan Wang, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2023. A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1):106.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM challenge at the 27th conference on computational natural language learning*, pages 1–34.

Appendix: Word Acquisition Detailed Results

Table 2: Twenty words with the earliest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (months), based on Wordbank data. Words in **bold** indicate overlap between the model and children.

2*Rank	Model Earliest				Child Earliest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	**车 (car)**	**81.15**	**16**	**vehicles**	**车 (car)**	**16**	**81.15**	**vehicles**
2	火车 (train)	95.75	21	vehicles	**阿姨 (auntie)**	**16**	**103.45**	**people**
3	**阿姨 (auntie)**	**103.45**	**16**	**people**	**奶奶 (grandma)**	**16**	**122.23**	**people**
4	今天 (today)	115.67	23	time_words	哥哥 (brother)	16	146.94	people
5	告诉 (tell)	119.24	23	action_words	球 (ball)	16	353.36	toys
6	**奶奶 (grandma)**	**122.23**	**16**	**people**	抱 (hug)	16	379.43	action_words
7	喜欢 (like)	124.81	21	action/descriptive	妈妈 (mom)	16	693.57	people
8	小朋友 (kids)	125.87	20	people	爸爸 (dad)	16	962.13	people
9	桌子 (table)	126.12	19	furniture_rooms	拿 (take)	16	994.86	action_words
10	现在 (now)	126.27	23	time_words	一 (one)	16	1275.79	quantifiers
11	朋友 (friend)	126.32	22	people	灯 (light)	16	1348.66	household
12	张 (classifier)	127.61	23	classifiers	**叔叔 (uncle)**	**17**	**141.80**	**people**
13	干嘛 (what for)	129.94	21	question_words	蛋 (egg)	17	381.34	food_drink
14	长 (grow/long)	129.98	21	action_words	脚 (foot)	17	161.11	body_parts
15	漂亮 (pretty)	131.85	22	descriptive_words	笔 (pen)	17	170.87	supplies
16	就 (then)	135.40	23	connecting_words	饭 (rice)	17	174.70	food_drink
17	已经 (already)	141.73	29	time_words	肉 (meat)	17	327.37	food_drink
18	**叔叔 (uncle)**	**141.80**	**17**	**people**	飞 (fly)	17	333.03	action_words
19	飞机 (airplane)	144.21	20	vehicles	书 (book)	17	344.84	supplies
20	嘛 (particle)	144.99	21	final_particles	水 (water)	17	360.68	food_drink

Table 3: Twenty words with the latest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (months), based on Wordbank data. Words in **bold** indicate overlap between the model and children.

12*Rank	Model Latest				Child Latest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	带 (bring/carry)	4582.65	20	action_words	**位 (classifier)**	**30**	**1572.94**	**classifiers**
2	藏 (hide)	3687.71	20	action_words	得 (must/particle)	30	278.38	helping_verbs
3	路 (road)	3068.54	21	outside	因为 (because)	30	146.92	connecting_words
4	写 (write)	2987.26	20	action_words	已经 (already)	29	141.73	time_words
5	香 (fragrant)	2648.09	19	descriptive_words	以后 (after/later)	29	388.75	time_words
6	啦 (particle)	2413.04	22	final_particles	时间 (time)	29	184.85	time_words
7	小 (small)	2189.56	19	descriptive_words	这些 (these)	29	372.13	pronouns
8	**梯子 (ladder)**	**2093.66**	**24**	**outside**	他们 (they)	28	341.20	pronouns
9	舔 (lick)	2052.29	21	action_words	不错 (not bad)	27	492.72	descriptive_words
10	阳台 (balcony)	1764.43	23	furniture_rooms	全部 (all)	27	681.46	quantifiers
11	拉 (pull)	1711.31	20	action_words	些 (some)	27	422.81	classifiers
12	腿 (leg)	1588.03	19	body_parts	别的 (other)	27	688.07	pronouns
13	**位 (classifier)**	**1572.94**	**30**	**classifiers**	向 (toward)	26	563.22	directions
14	哭 (cry)	1462.79	20	action_words	晚 (late/night)	25	723.19	time_words
15	等 (wait)	1442.76	20	action_words	我们 (we)	25	514.47	pronouns
16	**你们 (you-plural)**	**1423.67**	**25**	**pronouns**	**你们 (you-plural)**	**25**	**1423.67**	**pronouns**
17	踢 (kick)	1406.86	19	action_words	人家 (others)	25	269.66	pronouns
18	开心 (happy)	1383.12	22	descriptive_words	次 (times/classifier)	25	315.10	classifiers
19	滚 (roll)	1346.09	22	action_words	层 (floor/layer)	25	963.55	classifiers
20	龙 (dragon)	1312.99	19	action_words	**梯子 (ladder)**	**24**	**2093.66**	**outside**

Table 4: Twenty words with the earliest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (years), based on Liu et al. (2011). Words in **bold** indicate overlap between the model and children.

2*Rank	Model Earliest				Child Earliest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	**狗 (dog)**	**-0.53**	**2.48**	**animals**	**兔子 (rabbit)**	**1.94**	**65.83**	**animals**
2	火 (fire)	17.73	5.19	natural	**眼睛 (eyes)**	**1.94**	**98.46**	**body**
3	牛 (cow)	44.10	7.71	animals	蝴蝶 (butterfly)	1.94	346.22	animals
4	杯子 (cup)	46.86	4.44	supplies	**飞机 (airplane)**	**2.25**	**67.60**	**vehicles**
5	鼻子 (nose)	49.39	3.74	body	苹果 (apple)	2.25	123.63	food
6	桌子 (table)	62.55	2.85	furniture	门 (door)	2.25	179.40	furniture
7	熊 (bear)	65.26	10.45	animals	**房子 (house)**	**2.30**	**69.67**	**buildings**
8	**兔子 (rabbit)**	**65.83**	**1.94**	**animals**	**火车 (train)**	**2.30**	**75.23**	**vehicles**
9	**飞机 (airplane)**	**67.60**	**2.25**	**vehicles**	**狗 (dog)**	**2.48**	**-0.53**	**animals**
10	**房子 (house)**	**69.67**	**2.30**	**buildings**	碗 (bowl)	2.48	159.83	supplies
11	盒子 (box)	71.96	5.68	supplies	**电话 (phone)**	**2.54**	**80.20**	**electronics**
12	**火车 (train)**	**75.23**	**2.30**	**vehicles**	**太阳 (sun)**	**2.54**	**88.28**	**natural**
13	**电话 (phone)**	**80.20**	**2.54**	**electronics**	乌龟 (turtle)	2.54	287.21	animals
14	头发 (hair)	80.91	10.40	body	长颈鹿 (giraffe)	2.63	321.48	animals
15	狮子 (lion)	85.37	3.78	animals	糖 (sugar)	2.63	1049.10	food
16	**太阳 (sun)**	**88.28**	**2.54**	**natural**	书 (book)	2.84	155.34	supplies
17	**眼睛 (eyes)**	**98.46**	**1.94**	**body**	床 (bed)	2.84	403.65	furniture
18	电视 (TV)	100.37	4.18	electronics	桌子 (table)	2.85	62.55	furniture
19	**老鼠 (mouse)**	**117.79**	**2.85**	**animals**	**老鼠 (mouse)**	**2.85**	**117.79**	**animals**
20	沙发 (sofa)	122.94	3.42	furniture	葡萄 (grape)	2.86	208.09	food

Table 5: Twenty words with the latest Age of Acquisition. Comparison between GPT-2 model learning steps and children's AoA (years), based on Liu et al. (2011). Words in **bold** indicate overlap between the model and children.

12*Rank	Model Latest				Child Latest			
	Word	Model AoA	Child AoA	Type	Word	Child AoA	Model AoA	Type
1	手臂 (arm)	1094.99	3.56	body	信 (letter)	10.68	357.11	abstract
2	糖 (sugar)	1049.10	2.63	food	**鹿 (deer)**	**10.64**	**275.26**	**animals**
3	心 (heart)	571.16	4.58	body	**苍蝇 (fly)**	**10.45**	**410.90**	**animals**
4	**针 (needle)**	**554.09**	**7.44**	**tools**	熊 (bear)	10.45	65.26	animals
5	腿 (leg)	462.44	3.41	body	头发 (hair)	10.40	80.91	body
6	斑马 (zebra)	429.76	4.12	animals	屋顶 (roof)	7.84	343.40	buildings
7	**苍蝇 (fly)**	**410.90**	**10.45**	**animals**	牛 (cow)	7.71	44.10	animals
8	床 (bed)	403.65	2.84	furniture	狼 (wolf)	7.45	247.23	animals
9	蚂蚁 (ant)	372.23	4.58	animals	**针 (needle)**	**7.44**	**554.09**	**tools**
10	猪 (pig)	365.25	3.68	animals	盒子 (box)	5.68	71.96	supplies
11	信 (letter)	357.11	10.68	abstract	尾巴 (tail)	5.19	155.75	body
12	**船 (boat)**	**355.66**	**4.20**	**vehicles**	火 (fire)	5.19	17.73	natural
13	蝴蝶 (butterfly)	346.22	1.94	animals	窗户 (window)	4.73	176.12	furniture
14	屋顶 (roof)	343.40	7.84	buildings	心 (heart)	4.58	571.16	body
15	龙 (dragon)	333.07	3.56	animals	蚂蚁 (ant)	4.58	372.23	animals
16	长颈鹿 (giraffe)	321.48	2.63	animals	礼物 (gift)	4.56	202.00	supplies
17	桥 (bridge)	313.87	3.73	buildings	杯子 (cup)	4.44	46.86	supplies
18	乌龟 (turtle)	287.21	2.54	animals	医生 (doctor)	4.42	253.32	people
19	**鹿 (deer)**	**275.26**	**10.64**	**animals**	**船 (boat)**	**4.20**	**355.66**	**vehicles**
20	**钟 (clock)**	**270.72**	**4.20**	**supplies**	**钟 (clock)**	**4.20**	**270.72**	**supplies**

Predicting Sentence Acceptability Judgments in Multimodal Contexts

Hyewon Jang[1], Nikolai Ilinykh[1]
Sharid Loáiciga[1], Jey Han Lau[2], Shalom Lappin[1,3,4]
[1]Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg,
[2]The University of Melbourne, [3]Queen Mary University of London, [4]King's College London
{hyewon.jang, shalom.lappin}@gu.se

Abstract

Previous work has examined the capacity of deep neural networks (DNNs), particularly transformers, to predict human sentence acceptability judgments, both independently of context, and in document contexts. We consider the effect of prior exposure to visual images (i.e., visual context) on these judgments for humans and large language models (LLMs). Our results suggest that, in contrast to textual context, visual images appear to have little if any impact on human acceptability ratings. However, LLMs display the compression effect seen in previous work on human judgments in document contexts. Different sorts of LLMs are able to predict human acceptability judgments to a high degree of accuracy, but in general, their performance is slightly better when visual contexts are removed. Moreover, the distribution of LLM judgments varies among models, with Qwen resembling human patterns, and others diverging from them. LLM-generated predictions on sentence acceptability are highly correlated with their normalised log probabilities in general. However, the correlations decrease when visual contexts are present, suggesting that a higher gap exists between the internal representations of LLMs and their generated predictions in the presence of visual contexts. Our experimental work suggests interesting points of similarity and of difference between human and LLM processing of sentences in multimodal contexts.

Keywords: sentence acceptability prediction, LLMs, visual context, multimodal effects on sentence acceptability

1. Introduction

There has been a considerable amount of work on using DNNs to predict human sentence acceptability judgments (Lau et al., 2017; Bernardy et al., 2018; Warstadt et al., 2019; Lau et al., 2020; Qiu et al., 2024). Some previous studies (Bernardy et al., 2018; Lau et al., 2020) consider the effect of document (textual) context on human acceptability judgments. We examined the impact of visual contexts on both human and LLM sentence ratings.[1] The primary question we ask in this work is *whether the presence and relevance of visual contexts affect sentence acceptability judgments in humans and LLMs*. This question is motivated from previous work that reports a compression effect for human ratings in document contexts, relative to null contexts (Bernardy et al., 2018; Lau et al., 2020). This involves raising acceptability at the lower end of the acceptability scale, and lowering them at the higher end. Two explanations have been suggested for this effect, one being due to cognitive load, and the other being due to discourse coherence effect. In the cognitive load account, acceptability ratings gather around the middle because humans experience cognitive load when faced with more input to process, making them more conservative in their judgment. In the discourse coherence account, the ratings for bad sentences are raised because they look less bad when placed in the relevant discourse.

In our experiments using visual contexts instead of document contexts, we observe no such effect for human acceptability judgments, although there is some indication of a slight raising effect at the lower end, in relevant visual image contexts. We discuss the implications of this finding in §3.

We selected recent text from the Internet in a variety of genres to prevent data contamination. We follow Lau et al. (2020) in performing round trip machine translation on sentences, through several languages, using an older statistical MT system, *Moses*, to introduce a variety of syntactic, semantic, and lexical infelicities into the English output.

We employed Prolific[2] crowd sourcing to obtain native speaker human ratings on these sentences, together with a subset of the original English sources, in the context of preceding visual images. GPT-5 generated images for the original English versions of the sentences in our test data. We looked at contexts where the images were relevant to the content of the sentences, contexts where they were irrelevant, and null contexts where no visual images appear. We then tested closed LLMs

[1]Code and data are available at: https://github.com/GU-CLASP/multimodal-sentence-acceptability.

[2]https://www.prolific.com/

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 25–34
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

and open-source LLMs to predict the human ratings in these contexts. Four of the seven LLMs that we tested score well on predicting human sentence ratings (0.8-0.9 for Spearman correlations). However, in contrast to humans, two out of four of these models exhibit a clear compression effect for the visual contexts, which closely resembles the one reported in Lau et al. (2020) for human judgments in document contexts (see the three regression graphs in Figure 3). Overall, our models' predictions are much higher in the absence of visual contexts. We also find varying patterns in the distribution of LLM judgments, with only one of them, Qwen2.5-7B, resembling the human rating clusters to a degree.

Our experiments suggest that while current LLMs have greatly improved in their ability to identify levels of sentence acceptability relative to earlier DNNs, including first generation transformers, they process sentences in multimodal contexts differently than humans do.

2. Related Work

Prior studies employ different settings for sentence acceptability judgment. Lau et al. (2017); Warstadt et al. (2019); Qiu et al. (2024) use DNNs to predict human sentence acceptability ratings independently of context. Lau et al. (2020) and Bernardy et al. (2018) address the effect of document contexts on sentence acceptability judgment by both humans and DNNs. They employ various probabilistic methods to approximate acceptability judged by DNNs, such as applying normalising score functions to raw probabilities to filter out the effects of lexical frequency and sentence length.

Different approaches have been employed in previous work for collecting sentence data. Lau et al. (2020) and Bernardy et al. (2018) crawl natural sentences from Wikipedia and introduce infelicities by round-trip machine translation. Warstadt et al. (2019) construct a test corpus (CoLA) consisting largely of linguists' example sentences in minimal pairs, with ratings given by linguists.

We follow Lau et al. (2020) and Bernardy et al. (2018) in using naturally occurring text (modulated through round trip MT), crowd sourced non-expert human rating, and assessment relative to context. We focus on the impact of visual image contexts on both human and LLM acceptability ratings. We also consider the similarities and differences in rating distributions between humans and models.

More recent studies have been reported that are relevant to our approach in this work. Qiu et al. (2024) apply GPT-3 to Lau et al. (2017)'s human rated sentences. They show that it performs well in predicting these ratings. But they do not address the possibility of contamination in which GPT-3 may have been trained on some of this data, which has been publicly available since 2017. To minimise the influence from data contamination we include sentences published online after the training dates of the models for acceptability prediction. Furthermore, Qiu et al. (2024) only report results based on prompting experiments. Our work provides more robust findings following Kauf et al. (2024), who compare prompting and logprobs for human judgments of semantic plausibility. They report that logprobs are a good predictor of these judgments.

3. Human Acceptability Judgments

We collected 75 English sentences between the length of 25 and 40 words, from news (the Guardian, CNN, Washington Post, Wallstreet Journal, BBC), books (Google Books) and Wikipedia. To minimize bias from data contamination, we selected sentences from 2025, with the exception of Wikipedia, for which it is hard to extract new sentences only. We subjected these original English sentences to a round-trip translation using publicly available *Moses* models (Koehn et al., 2007),[3] to introduce lexical, syntactic, and semantic infelicities to the sentences, following Lau et al. (2020). We used *Moses* because most decoder only LLM-based MT systems generate fluent well-formed text. We employed three non-English languages as the pivot language (i.e., en→cs→en, en→fr→en, en→de→en) to create 75 $\times$ 3 round-trip translated sentences (see Table 1 for examples). We generated images for 75 original sentences, using GPT-5 with the prompt "*Generate an image that describes or is relevant to this sentence*". We inspected the generated images and observed that the quality was high in terms of their relevance to the original sentence (see Table 2).[4] The total of 300 sentences (75 original, 225 modified) were split into multiple batches, so that each human participant would provide sentence acceptability ratings for 20 sentences (5 original, 15 modified) on a scale from 1 (very unnatural) to 4 (very natural).[5] The participants also indicated how concrete/abstract they found each sentence on a scale of 1 (very concrete) to 4 (very abstract). The sentences were presented in three

[3] https://www.statmt.org/moses-release/RELEASE-3.0/

[4] We acknowledge that the AI-generated images may contain hallucinations and imperfections. However, we opted for this approach — as opposed to using existing image captioning dataset with natural images — to mitigate the risk of data contamination. Additionally, generating the images allowed us to ensure tight congruence between the textual and visual modalities, so that the text is directly relevant to the image and vice versa.

[5] Following Lau et al. (2017), we asked participants to evaluate *naturalness* rather than *acceptability*, a non-expert-friendly term.

Book	Orig	One piece of felt, folded in the corner, will be spliced, unfurled and dangled from the ceiling like a canopy.
	Mod	A piece in a corner, spliced were put forward and believe that the upper ceiling as a canopy.
News	Orig	But the answer seems to be no: Reeves lets it be known she requests no costings on raising the three forbidden taxes.
	Mod	But the reply does not seem: reeves suggests that it would not cost to raise the three banned taxes.
Wiki	Orig	In response, the US deployed an additional 170,000 troops during the 2007 troop surge, which helped stabilize parts of the country.
	Mod	In response, we deployed with further 170 000 soldiers in 2007 increase troops, which help to stabilize parts of the country.

Table 1: Examples of sentences used in our experiment, sourced from books, news, and Wikipedia in original form (Orig) and modified (Mod) through round-trip machine translation with Moses 3.0.

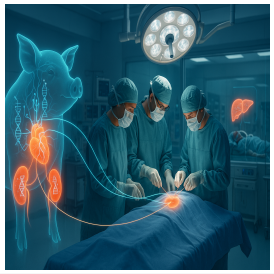
Original Sentence	**Relevant Image**	**Irrelevant Image**
One piece of felt, folded in the corner, will be spliced, unfurled and dangled from the ceiling like a canopy.		

Table 2: An example sentence with relevant and irrelevant image as context.

different conditions (*null*, *relevant*, *irrelevant*). In the *null* condition, participants rated each sentence in terms of naturalness without any preceding visual context. In the *relevant* and *irrelevant* condition, participants rated each sentence after having seen a relevant or irrelevant visual context, respectively. The irrelevant images were paired with the sentences by a simplified permutation design (e.g., s_1-im_2, s_2-im_3, s_3-im_1). To make sure that participants pay attention to the images, we asked them to select the most foregrounded object in the image from multiple choices, before rating the sentence. We batched the sentences such that no participant would see duplicate sentences or images. Each participant only rated sentences in one of three conditions. We added two attention check questions, where a prompt to choose a specific answer was embedded in a usual sentence rating task. Every sentence was rated in all three conditions (e.g., s_1-*null*, s_1-*relevant*, s_1-*irrelevant*). We recruited 20-25 native English-speaking participants for each condition and batch (gender-balanced). For quality control, we discarded data from participants who a) failed the attention check questions, b) rated original sentences as unnatural more than 40% of the time, or c) selected incorrect answers to the image question more than 25% of the time. As a result of such filtering criteria, 10% of the total participants were removed from the collected data. Participants were 45 years old on average ($\pm$ 13), with 71% British, 15% American, and 14% English speakers from other countries. The participants were paid GBP 9/hour.

3.1. Results

	Original			**Modified**		
	N	R	I	N	R	I
mean	3.54	3.54	3.53	1.96	2.02	1.94
sd	0.76	0.75	0.77	1.05	1.03	1.01

Table 3: Descriptive statistics for human ratings on sentence acceptability (1 - 4) in *null* (N), *relevant* (R), and *irrelevant* (I) conditions.

Table 3 shows the average acceptability ratings for original and modified sentences. The average ratings for original sentences are higher than for modified sentences, showing that modified sentences are infelicitous, as expected. We calculated

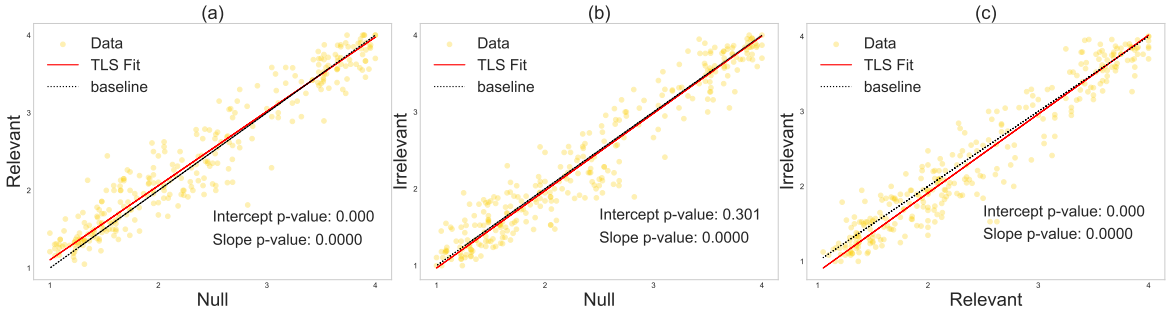


Figure 1: Human ratings of sentence acceptability in different conditions with regression lines generated from total least square regression. P-value alpha: 0.017 (Bonferroni correction).

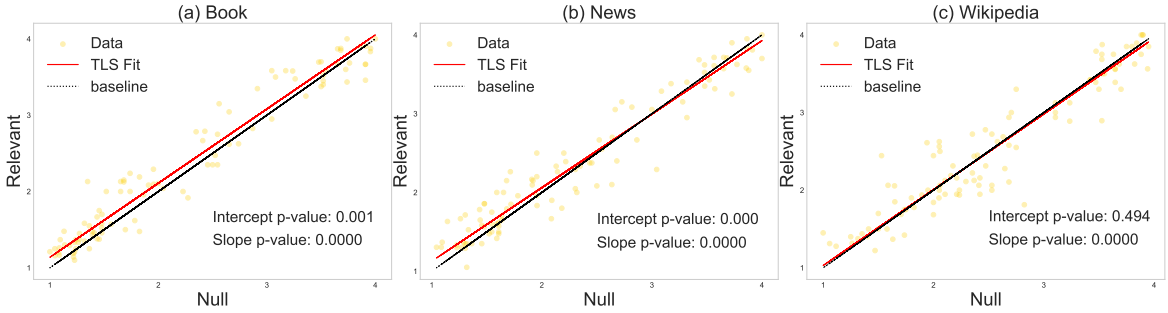


Figure 2: Human ratings of acceptability of sentences from different genres (books, news, Wikipedia) for *null-relevant* condition pair.

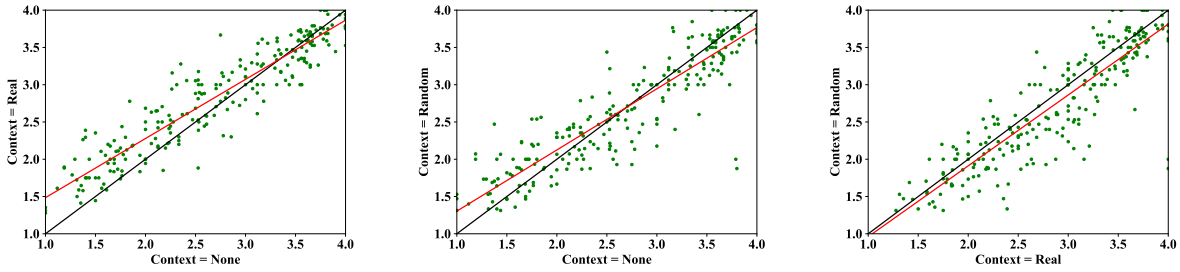


Figure 3: Human ratings of sentence acceptability in different **textual** context. "Context = None" means null context, "Context = Real" means relevant context and "Context = Random" means irrelevant context. Image reproduced from Lau et al. (2020).

mean sentence acceptability scores for each sentence across multiple human raters in each of the three conditions. We ran total least square regressions (TLR) for these average scores, with a pair of conditions at a time. Since there are no dependent variables serving as the ground truth between the condition pairs, TLR is a suitable regression analysis for this task. It accounts for errors in both X and Y variables, by minimising orthogonal distances rather than vertical ones when fitting a regression line. Figure 1 shows the average human ratings for each sentence in different conditions. Each data point (yellow dot) represents the average acceptability rating for a sentence. The red line in each pane represents the total least square regression fit for the data points in two out of three conditions at a time (*null-relevant*, *null-irrelevant*, *relevant-irrelevant*). In (a), relative to the *null* condition, we observe a slight raising effect in the *relevant* condition, where sentence acceptability scores in the lower end are higher. This indicates that human raters tend to judge a bad sentence slightly less bad when presented with relevant visual context prior to the judgment. We observe no such effect for *irrelevant* contexts (b), with the ratings for a sentence being almost identical in both conditions. The comparison between relevant and irrelevant in (c) confirms this, as the ratings for unnatural sentences are higher in relevant conditions, which mirrors (a).

Our results generally do not exhibit the compres-

sion effect for humans, reported in Lau et al. (2020) (Figure 3), where sentence acceptability ratings on both ends were compressed, causing bad sentences to be judged less bad, and good sentences to be judged less good, when a textual context precedes the target sentence.[6] In our results, we observe a raising effect only for the lower end (bad sentences) when relevant visual contexts are presented (see Figure 1(a)). This would seem to indicate that a discourse coherence effect is operative in visual contexts for human processing, as it works only to (slightly) raise lower end judgments when the preceding image renders the text more accessible. On the discourse coherence account, infelicitous sentences look more natural when placed within a context that makes them appear more coherent (see §1 for discussion regarding the cognitive load and discourse coherence accounts of compression).

Genre is also relevant. One possibility is that the characteristics of the sentences we tested, represented by the genre (news, books, Wikipedia) produces different effects for preceding contexts on human sentence processing. We looked at the effect of genre of sentences. Figure 2 shows the same information about human sentence acceptability judgment, on sentences from books, news, and Wikipedia, respectively. For sentences sourced from books, we observe a more uniform raising effect, while for sentences from news we observe a compression effect, though the magnitude is much smaller than the one reported in Lau et al. (2020). Visual contexts for sentences of distinct genres may affect acceptability judgmens differently. Sentences from books are rated as more abstract than sentences from news or wikipedia (see Table 4). Sentences from books are rated as more natural when relevant visual contexts are presented prior to the sentences, because images may render the content of ill-formed sentences more accessible. Sentences that are more concrete show a pattern comparable to the compression effect, which was manifest in textual contexts, but this effect is minimal. We do not observe it in irrelevant visual contexts, unlike Lau et al. (2020), where it is apparent in both non-null textual environments.

Another possible factor conditioning the absence of the compression effect involves cognitive load. According to this account, interpreting textual context puts additional informational processing demands on humans, forcing the acceptability ratings to be more conservative, clustering towards the middle range. This mechanism does not appear to operate in the case of visual images in conjunction with sentences. It is easier for humans to ignore them if they find no connection between an image and a sentence. Ignoring *textual* context, relevant or irrelevant, is presumably more difficult, as it is interpreted through the same processing mode. This may explain why humans exhibit a clear compression effect only with textual contexts. As a general note, the characteristics of visual contexts and textual contexts may differ in informational role. A visual context can be a depiction of a sentence's content, while a textual context is a preceding set of sentences that provides a prologue to the target sentence.

The observations made in this experiment need further reinforcement in order to reach a clear attribution for the presence and absence of compression effect. We leave this to future work, where we will directly compare the influence of a document context and visual context for a given sentence.

[6] Lappin (2021) describes a total least square regression test which shows that the compression effect that Lau et al. (2020) report is actual, and it cannot be reduced to regression to the norm.

	Books	News	Wikipedia
mean	1.95	1.63	1.70
sd	1.03	0.86	0.92

Table 4: Mean abstractness scores reported by human participants on original English sentences from books, news, and Wikipedia.

4. LLM Acceptability Judgments

We conducted two types of analyses with LLMs. First, we prompted LLMs (§4.1) to provide sentence acceptability judgments in the identical setup as the human experiment in §3. We then looked at the internal representations of the open-source LLMs by calculating normalised logprob values (§4.2), based on token probabilities, following prior work (Lau et al., 2020; Kauf et al., 2024). We used five open-source models that can take image and text as input (vision and language models) – InternVL3-1B, InternVL3-8B (Chen et al., 2024), Qwen2.5-3B, Qwen2.5-7B (Team, 2025), and llava-1.5-7b (Liu et al., 2024) – and two closed models – gpt-4o & gpt-4o-mini (OpenAI et al., 2024). InternVL3-1B, InternVL3-8B, have 1B and 8B parameters, respectively. Qwen2.5-3B, Qwen2.5-7B have 3B and 7B, respectively, and llava-1.5-7b has 7B parameters. We also made sure to select models released in early 2025 at the latest, to minimize the possibility of data contamination, although it should be noted that the chance of contamination is high for sentences from Wikipedia. We did our probability-based experiment (§4.2) on open-source models, because closed models do not allow for the extraction of logprob values for their output. Our experiments with open-source LLMs were conducted on 1 NVIDIA Tesla A40 GPU with 48GB RAM. The

experiments lasted about 30 hours in total.

4.1. Prompting-based analysis

We prompted the LLMs in a zero-shot setting to judge the naturalness of each of the 300 sentences used in §3 (75 original sentences + 75 $\times$ 3 modified sentences), under three different conditions. In the *null* condition, we presented only the sentences to the LLMs. In the *relevant* and *irrelevant* conditions, we first prompted the LLMs to see an image and solve the same task given to the human participants in §3. We then prompted them to judge the naturalness of the subsequent sentence after feeding their response to the next generation call (i.e., in the second call the input concatenates the image attention check task and response, and the sentence acceptability task instruction). We did not provide any explicit instructions to connect the image with the following sentence. We used 10 initialization seeds for each model, and we averaged the sentence acceptability ratings across the seeds. We used hyper-parameter settings of temperature=0.7, top-p=1.0, top-k=50 for all open-source models, and the default settings for closed models.[7]

		All	N	R	I
LLM ratings $\sim$ Human ratings	gpt-4o	0.87	**0.89**	0.87	0.88
	gpt-4o-mini	0.89	**0.89**	**0.89**	0.88
	InternVL3-1B	0.32	**0.66**	0.26	0.16
	InternVL3-8B	0.83	**0.88**	0.85	0.85
	Qwen2.5-3B	0.65	**0.70**	0.61	0.66
	Qwen2.5-7B	0.78	**0.84**	0.75	0.76
	llava-1.5-7b	0.25	**0.39**	0.21	0.15

Table 5: Spearman (ρ) correlations between average human sentence acceptability ratings (1-4) and LLM-prompted ratings (1-4) in null (N), relevant (R) and irrelevant (I) conditions. All correlations significant ($p<0.001$). Highest correlations per row marked in bold.

4.2. Probability-based analysis

For this analysis we used the sentence probability as estimated by the LLMs directly (Ide et al., 2025; Hu et al., 2024; Kauf et al., 2024). We passed the images and sentences through the open-source models (forward pass without generation) and extracted logits for only the input sentence in all three conditions (*null*, *relevant*, *irrelevant*). We calculated the average log probabilities of a sentence, normalised by the number of tokens (MeanLP), by summing the log-probabilities of each token in a sentence, conditioned on the preceding tokens (Lau et al., 2020; Kauf et al., 2024). We considered the MeanLP as a proxy for the LLM's internal representation of its sentence acceptability judgment, following prior work. We experimented with normalising functions employed in Lau et al. (2020), and found them to be unnecessary based on the high correlations of most LLMs with the human ratings.

		All	N	R	I
MeanLP $\sim$ Human ratings	InternVL-1B	0.70	**0.72**	0.70	0.70
	InternVL-8B	0.70	**0.72**	0.71	0.70
	Qwen-3B	0.61	**0.70**	0.61	0.57
	Qwen-7B	0.73	**0.79**	0.76	0.71
	llava-7B	0.74	0.74	**0.76**	0.73

Table 6: Spearman (ρ) correlations between Mean Logprobs from open-source LLMs and average human sentence acceptability ratings. All correlations significant ($p<0.001$). Highest correlations per row marked in bold.

4.3. Results

Table 5 shows correlations (Spearman ρ) between average human sentence acceptability ratings and LLM-generated ratings (§4.1).[8] The two GPT models show overall correlations of over 0.87, with little to no variations across conditions. Of the open-source LLMs, models with higher parameter size (7B-8B) perform almost equally well, with the exception of llava-1.5-7b. The poor performance of the smaller models (InternVL3-1B, Qwen2.5-3B) and llava-1.5-7b might be attributed to their reduced capability to follow instructions (llava-1.5-7b in particular is not trained to follow instructions). The open-source LLMs also seem to perform better without any visual context (N).

Table 6 presents correlations between average human sentence acceptability ratings and probability measures from the LLMs described in §4.2. Compared to the prompting results (Table 5), MeanLP generally performs on par or better for the open-sourced LLMs. Most interestingly, the performance of these models *across conditions* (N, R, I) also appears to be much more consistent — a contrast compared to the prompting results in the case of llava-1.5-7b.

To better understand the similarity between MeanLP versus prompted model ratings, we present their correlation in Table 7. The larger models (InternVL3-8B and Qwen2.5-7B) have better correlations (both 0.70 for *All*), while the others have much lower correlations. Either way, these results show that these two approaches produce

[7] We ran the experiments with varying temperatures posthoc (0.5 to 1.0 in the increments of 0.05), and the results maintained the same patterns.

[8] We only report Spearman ρ, as Pearson r yielded very similar results. The same applies to Tables 6 and 7.

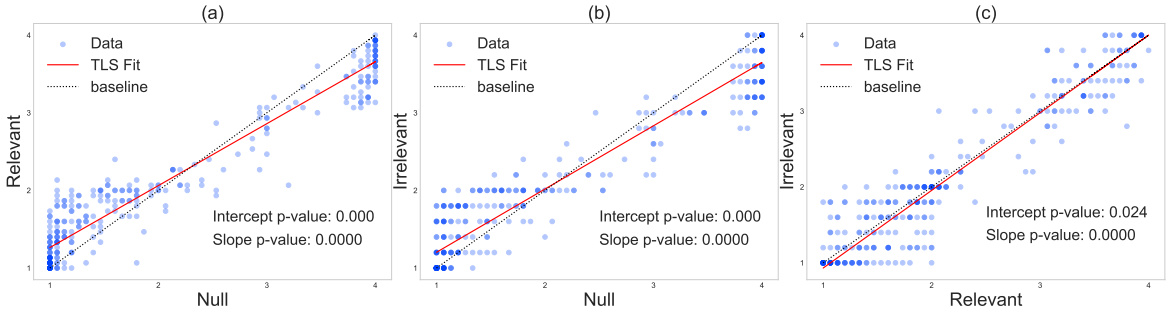


Figure 4: Sentence acceptability ratings generated by gpt-4o.

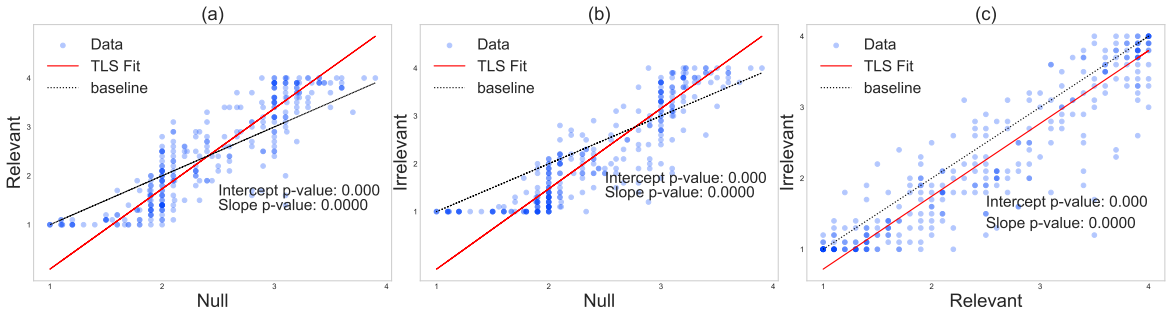


Figure 5: Sentence acceptability ratings generated by InternVL3-8B.

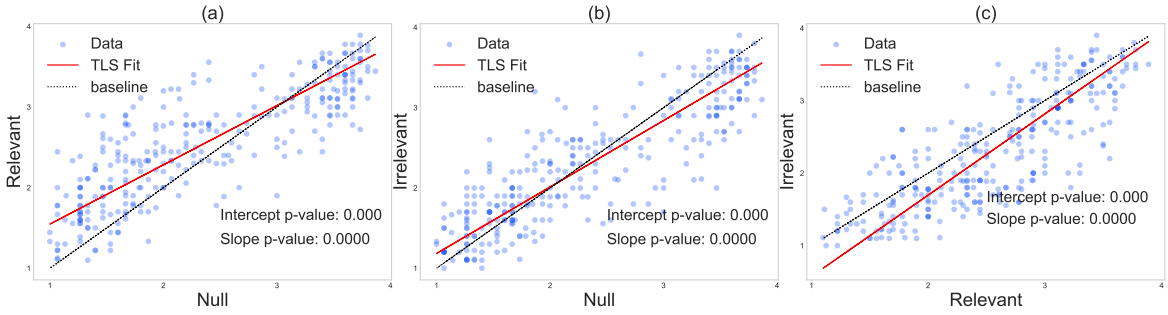


Figure 6: Sentence acceptability ratings generated by Qwen2.5-7B.

somewhat distinct predictions of acceptability ratings. But, Tables 5, 6, and 7 all suggest that correlations are higher in the *null* condition compared to the non-*null* conditions (with the exception of one case involving llava-1.5-7b).

		All	N	R	I
MeanLP ∼ Model ratings	InternVL-1B	0.30	**0.58**	0.30	0.17
	InternVL-8B	0.70	**0.75**	0.74	0.71
	Qwen-3B	0.52	**0.56**	0.48	0.45
	Qwen-7B	0.70	**0.76**	0.71	0.68
	llava-7B	0.28	**0.35**	0.24	0.22

Table 7: Spearman (ρ) correlations between Mean Logprobs from open-source LLMs and model prompted ratings. All correlations significant ($p<0.001$). Highest correlations per row marked in bold.

Based on the correlation analyses, we consider Qwen2.5-7B, InternVL3-8B, and gpt-4o to have yielded the most reliable results. In Figures 4, 5, and 6 we present the scatterplots of the sentence acceptability ratings produced by these models in each condition pair (*null-relevant*, *null-irrelevant*, *relevant-irrelevant*) for more detail, parallel to the human results in §3. Our intention is to examine these scatterplots to understand the qualitative nature of the model rating distribution, and also to make sure the correlation numbers are not skewed by a few outliers. In general, we see that the distributions of sentence acceptability ratings produced by the LLMs appear quite different from those of humans. gpt-4o -generated ratings cluster around top and bottom ends, and InternVL3-8B -generated ratings cluster around 2 and 3. Qwen2.5-7B shows

distributions that are most similar to humans,[9] but with a much higher variance than humans. Interestingly, we observe the compression effect for both gpt-4o and Qwen2.5-7B. Lau et al. (2020) show a very similar compression effect for human sentence ratings in textual contexts, as we see in Figure 3. But we observe the opposite of compression for InternVL3-8B. We do not have an explanation for this contrast. We will be exploring it in future work. The general divergence in the distributions of model-generated ratings from human-generated ratings may be due to the different mechanisms by which LLMs process sentences in comparison to humans. LLMs have less memory interference, and they retain perfect access to previous words (Oh and Linzen, 2025).

5. Discussion

The work that we present here indicates that LLMs achieve a high degree of accuracy in predicting human sentence acceptability judgments. It also shows that the normalised logprob values that these models assign to sentences are a reliable predictor of human ratings for sentence acceptability. Despite the strong convergence of LLM logprob scoring and human naturalness rating, most of the LLMs that we consider in this work display different data clustering patterns than humans in the distributions of these judgments (cf. Figures 1 and 5 for two illustrative examples). In particular, gpt-4o exhibits a more polarised pattern, while Qwen2.5-7B approaches human distributions (Figure 4 vs 6). However, although Qwen2.5-7B most closely approximates the human distributions, it still differs from them, both in our experiments and in those reported by Lau et al. (2020) (cf. Figures 1, 3, and 6). Due to their greater memory capacity compared to humans, Oh and Linzen (2025) argue that LLMs' superhuman next-token prediction ability makes them unsuitable as cognitive models of human linguistic prediction. However, the tasks differ. Oh and Linzen (2025) focus on linguistic continuation tasks which effectively involve next-token prediction. In contrast, our experiments examine acceptability judgments, where LLMs predict evaluation scores and not upcoming tokens. It therefore remains an open question whether the differing patterns we observe reflect the LLMs' superhuman predictive capacity or instead stem from fundamentally different processing strategies.

Our experiments, viewed from the perspective of the work reported in Bernardy et al. (2018); Lau et al. (2020), suggest an interesting difference in the way that humans and LLMs process information in different modalities. Textual context, whether relevant or not to the following sentence, influences the way in which humans assess the naturalness of sentences, both well-formed ones, and those containing infelicities. Interpreting these contexts requires processing resources that influence judgments concerning the naturalness of following sentences.

By contrast, for humans, interpreting images seems to proceed through alternative processing mechanisms, which permit the image to be discarded, or suppressed, when rating the naturalness of a following sentence. This effect is particularly clear when the image has no clear relation to the sentence. However, LLMs, at least the vision and language models we used in this work, incorporate images into their sentence processing environment. They include the images when assigning probability values to sentences, and in rating them for naturalness. They do not discard them, even when they are irrelevant. Cognitive load appears to affect their judgments across visual and textual modalities. This is not surprising since their architectures are generally designed to maximise contexts (Oh and Linzen, 2025). The vectors that encode these contexts include elements of all modalities that the model attends to.[10]

6. Future Work

In future work we will examine the role of different types of text genre more closely in determining the impact of both textual and visual contexts on human sentence acceptability rating. We will also explore the mechanisms through which humans suppress images, but not preceding text, when processing sentences. Our objective here is to obtain a clear sense of the way in which cognitive load is conditioned by different sorts of context, relative to distinct text genres. For this, we will also consider other linguistic tasks than sentence acceptability judgment and experiments with larger data for better generalizations.

Another issue that we will explore is how LLMs can be modified to converge on human processing with respect to the suppression of images in the assignment of logprob values to sentences, without losing the content of these images. This would cause the modified model to more closely approximate observed human sentence processing in multimodal environments.

[9] This might be an artifact of Qwen2.5-7B producing more varied ratings with different initialisation seeds than the other LLMs.

[10] Our experimental results are broadly consistent with the view of the relation between LLMs and human processing presented in Oh and Linzen (2025). Simlarly, our suggestions for future work align well with their approach.

Acknowledgements

The work reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. The computation and data storage for our experiments were supported by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Limitations

The work reported in this paper only addresses English. We tried to minimise data contamination issues by experimenting with sentences from 2025, but there is a high chance of data contamination with sentences from Wikipedia, which are not as new. We reported that genre effects influence sentence acceptability judgment by humans, but the explored genres in this work are limited to three (news, books, wikipedia). We did not directly compare visual contexts and document contexts for the same target sentence, which may render the results inconclusive. Lastly, we did not fully address the characteristic differences of visual versus textual contexts in this work. We will follow up on these limitations in our future work.

Ethics statement

All our human participants took part in the experiments voluntarily, and received due compensations. We recognized that we might expose crowdworkers to disturbing/offensive images and texts. To minimize this risk, we manually examined the 75 sentences before administering the human experiments. GPT-5 also has its own guardrails for image generation to keep such risks low (https://cdn.openai.com/gpt-5-system-card.pdf). Some of our modeling experiments were done with closed models, which may render the results not entirely transparent. We did not rely on any AI-assistant tools for the generation of our article text.

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. 2018. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 456–461, Melbourne, Australia.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. How to make the most of LLMs' grammatical knowledge for acceptability judgments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7416–7432, Albuquerque, New Mexico. Association for Computational Linguistics.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Shalom Lappin. 2021. *Deep Learning and Linguistic Representation*. CRC Press, Taylor & Francis, Boca Raton, London, New York.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. 2020. How furiously can colorless green ideas sleep? sentence acceptability in context. *Transactions of the Association for Computational Linguistics*, 8:296–310.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41:1202–1241.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Byung-Doh Oh and Tal Linzen. 2025. To model human linguistic prediction, make LLMs less superhuman. *arXiv preprint arXiv:2510.05141*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2024. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 189–198, Bangkok, Thailand. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5-vl.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

ToM in LLM is not ToM, but a Pragmatic Effect

Agnese Lombardi, Alessandro Lenci
University of Pisa
Via Santa Maria 36, Pisa
agnese.lombardi@phd.unipi.it, alessandro.lenci@unipi.it

Abstract

Instruction tuning has been shown to improve large language models' performance on pragmatic tasks, and recent work suggests that additional training can also enhance Theory of Mind (ToM)–like abilities. However, existing studies rarely examine how different alignment techniques and training data contribute to ToM-related behavior in language models. In this work, we investigate the respective roles of instruction tuning and preference learning in shaping pragmatic and ToM abilities. Using the LLaMA 3 8B architecture, we fine-tune models on either pragmatic or ToM-specific data and subsequently align them via Direct Preference Optimization. We evaluate the resulting models on benchmarks targeting both pragmatics and ToM. Our results show that pragmatic training can substantially improve ToM performance even without explicit belief-related supervision, and that instruction tuning plays a central role in model alignment. These findings clarify the relationship between pragmatics and ToM in large language models.

Keywords: theory of mind, pragmatics, instruction-tuning

1. Introduction

Instruction tuning, understood as the further refinement of language models to perform specific tasks, has been shown to be an effective technique for improving performance on pragmatic tasks (Wu et al., 2024; Ruis et al., 2023). From a theoretical perspective, this is particularly relevant given the nature of pragmatics itself. The domain of **pragmatics** encompasses a wide range of phenomena, including speech acts, implicature recognition, presupposition processing, the interpretation of figurative language, and the maintenance of discourse coherence and cohesion. These aspects are inherently context-sensitive, where context extends beyond the linguistic input to include extralinguistic factors such as communicative intentions, situational cues, and world knowledge (Bambini and Lecce, 2025).

Similarly, several studies report that additional training can lead to improvements in language models' abilities that extend beyond pragmatics (Cai et al., 2024; Ranaldi and Freitas, 2024; Puerto et al., 2025; Aw et al., 2023; Lu et al., 2025), including **Theory of Mind** (ToM).

ToM refers to the capacity to represent others' beliefs, intentions, and desires. It relies on mentalizing processes, that is, the ability to reflect on one's own mental states and to attribute distinct or shared mental states to others within a given context. More broadly, ToM involves recognizing that others may hold beliefs, perspectives, or knowledge that differ from one's own, and reasoning about those mental states in order to predict and interpret behavior.

Although ToM is often associated with pragmatics, the nature of their relationship remains controversial and actively debated (Bambini and Lecce, 2025). In this work, we support the view that the two abilities only partially overlap, as suggested by evidence from psycholinguistics (Bosco et al., 2018; Kotila et al., 2025) and neurolinguistics (Schieche et al., 2025; Schieche and Uddén, 2025).

Given the nature of this partial overlap, it is important to clarify why further training of language models facilitates the emergence of ToM-like behaviors. In particular, studies reporting improvements in LLM performance following additional training (Lu et al., 2025) rarely examine the nature of the training data with respect to its pragmatic properties, nor do they explicitly consider how such properties relate to ToM.

Moreover, existing work does not directly address the distinction between different alignment techniques, specifically instruction tuning and preference optimization, and their respective contributions to ToM-related abilities.

In this work, we aim to clarify the roles of instruction tuning and preference learning in enhancing ToM capabilities, in comparison to their effects on pragmatic performance. We further investigate the quality and nature of the training data underlying these improvements by explicitly assessing whether pragmatic materials that do not contain belief-related information can nonetheless support improved performance on ToM tasks, without relying on explicit ToM supervision.

To this end, we adopt the Llama 3 8B base architecture and perform two supervised fine-tuning procedures: one using pragmatic data and one using ToM-specific data. Each model is subsequently aligned through preference learning using Direct Preference Optimization (DPO), while maintaining a clear distinction between pragmatic and ToM training materials across model variants.

We evaluate the resulting four models, together with the base model, on two benchmarks targeting

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 35–47
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

pragmatic and ToM abilities. Our results reveal a surprising contribution of pragmatic training to ToM performance and highlight the effectiveness of instruction tuning as an alignment strategy. These results suggest that the ToM-like behaviors observed in LLMs (Street et al., 2024; Kosinski, 2024a) may, at least in part, be driven by exposure to pragmatic material rather than reflecting advanced mentalizing capacities in the human sense. Rather than implementing a fully-fledged capacity for representing others' mental states, LLMs may be leveraging advanced pragmatic reasoning skills, generalizing from patterns learned during training and benefiting from an enhanced ability to follow task-specific instructions. In light of the partial functional segregation between pragmatics and ToM, our findings are consistent with psycholinguistic evidence indicating that pragmatic competence may act as a driving force in the development of ToM abilities, rather than the reverse (Del Sette et al., 2025).[1].

2. Related Works

Previous work has suggested that alignment with human beliefs, particularly through preference learning training, is a key mechanism underlying the emergence of ToM-like abilities in large language models (Street, 2024). Other studies have shown that larger, fine-tuned models encode mental state information more accurately than smaller ones, pointing either to a scaling effect at the level of internal representations (Bortoletto et al., 2024; Nguyen, 2025; Cui et al., 2024), or to the influence of instruction tuning (Aw et al., 2024), in-context learning or prompting techniques (Moghaddam and Honey, 2023).

Despite these findings, there is no consensus regarding which training or alignment technique is primarily responsible for increased ToM alignment. In this work, we aim to assess the contribution of instruction tuning by directly comparing it with preference-learning approaches.

Most prior studies in this area focus on reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). Here, we instead adopt **Direct Preference Optimization** (DPO) (Rafailov et al., 2023), as it is simpler and computationally more efficient, and avoids the instability associated with reward-model training. DPO is an alignment algorithm that provides an alternative to RLHF by optimizing the policy directly via gradient ascent on the implicit RLHF objective, without requiring a separate reward model or explicit reinforcement learning.

To our knowledge, this is the first work to explicitly investigate whether pragmatic alignment and ToM-like behavior in language models arise primarily from instruction tuning or from human feedback through preference-optimization techniques. Additionally, although the partial overlap between pragmatics and ToM is acknowledged in LLM evaluation (Hu et al., 2025), and several studies assess LLM performance on pragmatic abilities (Hu et al., 2023; Sravanthi et al., 2025; Ruis et al., 2023) and ToM tasks (Kosinski, 2024b; Ullman, 2023; Strachan et al., 2024), no previous work has systematically investigated the interaction between these two abilities during training.

3. Training

First, we performed two separate supervised fine-tuning (SFT) procedures on Llama 3 8B: one using pragmatic data and the other using ToM data. We then further trained each of these models using direct preference optimization (DPO) with pragmatic and ToM data. In total, this resulted in the evaluation of four model variants: SFT-Pragmatics, SFT-ToM, DPO-Pragmatics, and DPO-ToM.

3.1. SFT

We performed SFT using parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA). All experiments were conducted using the Hugging Face `Transformers` and `TRL` libraries.

Two instruction-tuning conditions were considered:

- **ToM**. We used examples from the ExploreToM dataset, generated within the same framework (Sclar et al., 2024). ExploreToM is specifically designed to assess ToM abilities, as it comprises narratives in which different characters hold potentially divergent beliefs about the current state of the world and about one another's beliefs. Each story is paired with questions intended to probe the model's capacity to track and reason about these mental states. For instance, a story might include questions testing nested beliefs, such as whether "Anne does not know that Charles knows that the apple has been salted", requiring the model to reason across multiple perspectives. From the training split, we used 6,000 instances for training and 5,000 for validation. The inputs consisted of structured or infilled story prompts, while the targets corresponded to the expected answers.

- **Pragmatics**. We combined two synthetic reasoning datasets: `argilla/sft`[2] and

[1] Code for training and analysis: https://github.com/agneselombardi/ToM_is_not_ToM

[2] https://huggingface.co/datasets/argilla/synthetic-concise-reasoning-sft

lighteval/natural[3]. Both datasets include examples requiring concise reasoning that do not involve ToM, as illustrated by extended examples in Appendix A.1. The first dataset contributed 5,000 training examples, while from the second we selected 1,000 examples for training and 5,000 for validation. In all cases, questions were used as prompts and targets as completions.

All datasets were normalized to a unified format consisting of a *prompt* and a *completion*. During training, prompts and completions were concatenated into a single input sequence (prompt examples are reported in Appendix A.1).

We initialized the pretrained Llama3-8B model in half precision (FP16). Tokenization was performed using the corresponding Llama tokenizer, with the end-of-sequence token used as the padding token and right padding applied. Input sequences were truncated or padded to a maximum length of 2048 tokens.

LoRA was applied to both attention and feed-forward projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj) with rank 16, scaling factor 16, and no dropout. During training, only LoRA parameters were updated.

Instruction tuning was carried out using the SFTTrainer. Training was conducted for three epochs with a learning rate of 2×10^{-4}, a per-device batch size of 2, and gradient accumulation over four steps. Evaluation was performed every 200 steps, and model checkpoints were saved every 600 steps, retaining the three most recent checkpoints. Early stopping with a patience of two evaluations was applied based on validation loss, and the best-performing checkpoint was selected at the end of training.

To stabilize training, we applied NEFTune noise with $\alpha = 5$. All experiments were logged using Weights & Biases.

3.2. DPO

Following instruction tuning, we further aligned the models using DPO. For each condition, DPO was applied to the corresponding instruction-tuned model, using preference datasets aligned with the same target domain (ToM or pragmatics).

In all experiments, the policy model consisted of the instruction-tuned Llama3-8B model with LoRA adapters, which remained trainable during alignment. Critically, we used the frozen SFT checkpoint from the previous stage as the reference model. Specifically, for ToM alignment, the ToM-SFT model served as the reference; for pragmatic alignment, the Pragmatics-SFT model served as the reference. This design choice ensures that the KL penalty regularizes the policy toward the task-specific distribution learned during SFT, rather than toward a generic instruction-following distribution. All reference model parameters were kept fixed throughout training. Tokenization was performed using the Llama3 8B tokenizer, with the end-of-sequence token used as the padding token and right padding applied.

For ToM preference learning, we constructed a DPO training set by combining three publicly available preference datasets targeting social and mental-state reasoning[4], as illustrated by examples in Appendix A.2. All datasets were normalized to a common format consisting of a *prompt*, a preferred (*chosen*) response, and a dispreferred (*rejected*) response. For dialogue-based datasets, user turns were used as prompts and assistant turns as candidate responses.

The combined dataset contained 14,369 examples total. We shuffled the data (seed 42) and split it into 90% training (12,932 examples) and 10% validation (1,437 examples). Training was conducted for up to three epochs using a learning rate of 5×10^{-6}, a per-device batch size of 1, and gradient accumulation over 32 steps (effective batch size of 32). We set the DPO temperature parameter to $\beta = 0.3$. The maximum prompt length was set to 1,024 tokens and the maximum sequence length to 2,048 tokens (matching SFT training). Training was conducted in bfloat16 precision with gradient checkpointing enabled using the AdamW optimizer with default hyperparameters.

To prevent overfitting, we evaluated on the validation set every 100 training steps and saved checkpoints at the same frequency, retaining the three most recent checkpoints. At the end of training, the checkpoint with the lowest validation loss was selected as the final model.

For pragmatic preference learning, we followed an analogous procedure using preference datasets focused on conversational and pragmatic quality. Specifically, we combined three different datasets[5].

[3] https://huggingface.co/datasets/lighteval/synthetic_reasoning_natural

[4] https://huggingface.co/datasets/Doctor-Shotgun/theory-of-mind-dpo (539 examples), https://huggingface.co/datasets/onyrotssih/social-i-qa-orpo-dpo-10k (10200 examples) and https://huggingface.co/datasets/shayanfirouzian/SocialReasoning_DPO (3630 examples)

[5] https://huggingface.co/datasets/flammenai/casual-conversation-DPO (3730 examples); https://huggingface.co/datasets/minchyeom/Thinker-DPO (2670 examples); and https://huggingface.co/datasets/clembench-playpen/DPO_dialogue_1neg_old (6700 examples).

Each dataset was normalized to the standard prompt-chosen-rejected format. For multi-turn dialogue data (clembench dataset), conversations were flattened into single sequences by concatenating all turns in chronological order, with each turn's content separated by newline characters.

The combined dataset contained 13,100 examples total, which we shuffled (seed 42) and split into 90% training (11,790 examples) and 10% validation (1,310 examples).

For both DPO conditions, we verified several implementation details to ensure correct training dynamics. First, we confirmed that gradients were applied exclusively to the policy model's LoRA parameters while the reference model remained frozen. Second, we verified that the policy and reference models were properly initialized from the same SFT checkpoint by checking that their initial outputs were identical (mean absolute logit difference $< 10^{-6}$). Finally, we monitored that the models diverged during training, with the mean absolute logit difference increasing substantially by the end of training.

All experiments were conducted using a single NVIDIA A100 GPU with 40GB memory. The choice of $\beta = 0.3$ follows standard practices in the DPO literature, balancing the preference learning objective with KL regularization to prevent excessive deviation from the reference distribution. The sequence lengths (1,024/2,048 tokens) ensure that longer reasoning chains and multi-turn dialogues are not truncated during training.

4. Evaluation

We evaluated all model variants on two benchmarks targeting distinct but related abilities: **PUB** (Sravanthi et al., 2024), which explicitly assesses pragmatic reasoning, and **FANTOM** (Kim et al., 2023), which is specifically designed to evaluate ToM reasoning.

The PUB benchmark comprises 14 tasks, primarily targeting implicatures. Ten tasks provide a detailed assessment of models' implicature-related abilities, while presupposition and reference are each evaluated through two tasks. Task 1 tests models' ability to distinguish between direct and indirect responses using multiple-choice questions. Tasks 2 and 3 involve response classification: Task 2 categorizes indirect answers into five labels, whereas Task 3 extends this by incorporating implied meanings as additional cues. Task 4 focuses on implicature recovery in non-polar question-and-answer contexts. Task 5 ("Agreement Detection") and Task 6 ("Understanding Sarcasm") evaluate the model's capacity to comprehend figurative language within dialogues. Tasks 7, 8, and 9 are based on the FLUTE dataset, presenting sentences or premises expressed figuratively alongside corresponding hypotheses in simpler language. Task 10 assesses whether models can draw correct inferences when implicatures are involved. Finally, Tasks 11 and 12 evaluate presupposition understanding, either framed as natural language inference or by asking the model to judge whether a given presupposition is valid or invalid within the conversational context.

In contrast, the FANTOM benchmark is based on asymmetric conversational settings.[6] In this setting, the full conversation is provided to the model, which thus has access to all exchanged information and effectively acts as an omniscient observer. However, a designated character (PersonX) is absent during part of the interaction, leading to an information asymmetry between the model and PersonX. Responses that attribute to PersonX information that was not available during their absence are taken as evidence of a failure to correctly model that character's mental state.

For PUB evaluation we report F1 metric for each task, while for FANTOM evaluation we report F1 scores on the following metrics:

- **Belief (MCQA):** The model must choose between an omniscient belief and a PersonX-centric belief. The correct answer corresponds to the PersonX-centric belief.
- **Answerability List** and **Accessibility Info:** A response is considered correct only if it includes all characters who have access to the relevant answer or information, while excluding all characters who do not. Partial credit is not assigned.
- **Answerability Binary** and **Accessibility Binary:** The model must respond with *yes*/*true* for characters who have access to the answer or information, and with *no*/*false* for those who do not.
- **First Order:** Evaluation of first-order ToM reasoning.
- **Second Order (Overall):** Evaluation of second-order ToM reasoning. To further

[6] Small-talk conversations involving multiple characters, each centered around a topic (e.g., pets, risk-taking, personal growth). Each topic comprises several subtopics (e.g., the topic "pets" may include subtopics such as "breed" and "special moves"). Conversations begin with two or three characters, and as they progress, participants may join or leave while the subtopic evolves. During the absence of a character, information is exchanged among the remaining participants, creating a natural information asymmetry. When the absent character later rejoins the conversation, they lack access to the information shared in their absence.

analyze performance on second-order belief tasks, FANTOM distinguishes between **Second-Order Cyclic** and **Second-Order Acyclic** patterns. Cyclic second-order questions probe a character's belief about another character's belief concerning the first character (e.g., *What does Linda think about Kailey's belief regarding the breed of Linda's dog?*), whereas acyclic second-order questions focus on a character's belief about another character's belief concerning a third character (e.g., *What does David think about Kailey's belief regarding the breed of Linda's dog?*) (Kim et al., 2023).

5. Results

Figures 1 and 2 report a comparison of the three model variants on the PUB and FANTOM benchmarks, respectively. The PUB evaluation reveals substantial variability across tasks.

In Task 1, which focuses on **direct and indirect response classification**, both DPO and SFT pragmatic-based training achieve the best performance. A similar trend is observed in Task 2 and 3 which address response classification without and with implied meaning.

Task 4, targeting **implicature recovery**, shows a clear advantage for pragmatics-trained models over their ToM-trained counterparts, despite the fact that implicature resolution may involve mentalizing abilities. Task 5 (**agreement detection**) does not exhibit meaningful differences across model variants, including relative to the base model. In Task 6, which evaluates **sarcasm understanding**, pragmatics-trained models again display substantial improvements, even though sarcasm interpretation is often assumed to rely on ToM.

A similar pattern emerges in Tasks 7–9, which assess **figurative language understanding**. In all three cases, pragmatics-trained models outperform the ToM-trained versions, despite the fact that these tasks plausibly require ToM reasoning.

Tasks 10 and 11, addressing **implicature NLI** and **presupposition NLI** respectively, do not explicitly require mentalizing abilities. Accordingly, the best performance is achieved by the pragmatics-trained SFT model. Interestingly, Task 12, which focuses on **presupposition in QA**, does not show substantial improvements from either SFT or DPO. In Task 13 (**deictic QA**), the pragmatics-trained SFT model again performs best. Finally, Task 14, concerning **reference via metonymy**, shows a decrease in accuracy for both SFT and DPO models relative to the base model. This task appears to rely more on extralinguistic knowledge than on inferential reasoning.

Notably, the **SFT ToM-trained model** exhibits performance degradation on several tasks, particularly those involving implied meaning (Task 3), implicature recovery (Task 4), sarcasm (Task 6), and figurative language understanding (Tasks 7 and 8, with Task 9 showing performance comparable to the base model). Similar degradation is also observed for presupposition and metonymy tasks [7].

Model	Data	Wins	F1 (M / Md)
SFT	ToM	4	0.363 / 0.334
SFT	Pragmatics	11	0.493 / 0.455
DPO	ToM	4	0.398 / 0.357
DPO	Pragmatics	12	0.491 / 0.435

Table 1: Performance of SFT and DPO aligned models on PUB compared to the base pretrained model. The *Wins* column indicates the number of tasks in which the model outperforms the reference model (in this case, the base model). The *Score* metric reports both the mean F1 and the median F1 achieved by the SFT or DPO model.

Table 1 further summarizes these findings by reporting the number of tasks in which each model outperforms the base model, together with mean and median F1 scores. Overall, **SFT and DPO with pragmatics-oriented training data yields the strongest improvements over the base model**.

Table 2 confirms that pragmatics-trained models consistently outperform ToM-trained models across tasks, with SFT (pragmatic) achieving the largest gains, while ToM-specific training shows limited or inconsistent effects.

Figure 2 reports the performance of all model variants on the FANTOM benchmark. The first task, which evaluates **belief interpretation**, shows the SFT model trained on ToM data substantially outperforming all other variants, whereas the DPO-trained models, particularly the DPO–ToM version, exhibit the weakest performance.

For the **Answerability List** and **Accessibility List** metrics, DPO pragmatic-based yelds the best results, followed by SFT training on pragmatic material. Notably, for the Answerability List task, only the pragmatic-based variants yield a slight improvement over the base model, while the corresponding **Answerability Binary** metric benefits more clearly from ToM-based training, likely due to the binary response format. In contrast, for **Accessibility**, ToM-based models do not show meaningful gains in either formulation. In the binary version, pragmatic training proves to be the most effective, a result that contrasts with the pattern observed for

[7] In some cases, such as figurative language (Task 7), implicature NLI, presupposition, and metonymy, the DPO-trained model also underperforms relative to the base model.

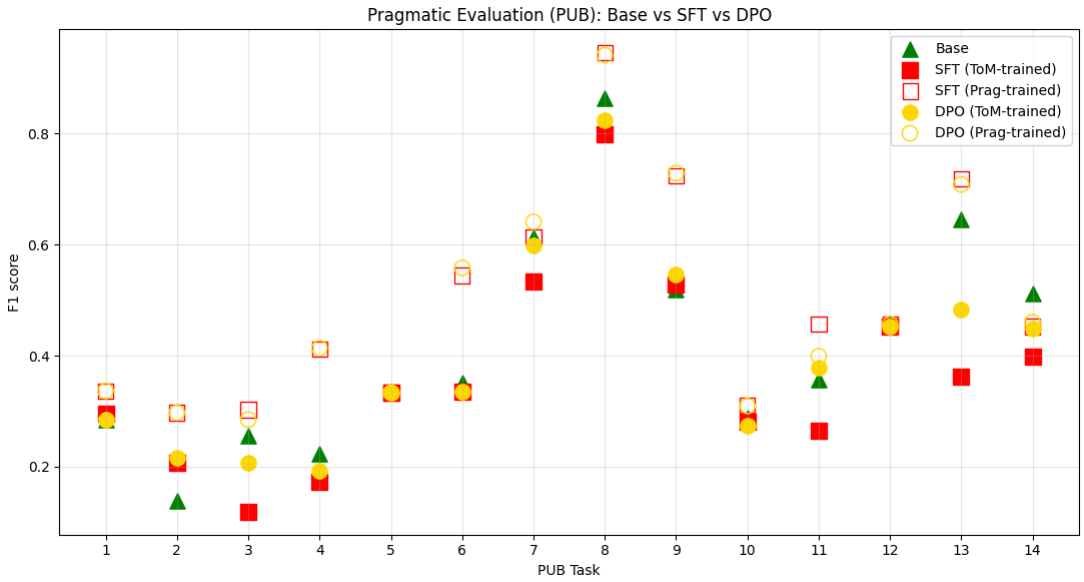


Figure 1: Results on the PUB benchmark for the base model, SFT, and DPO models.

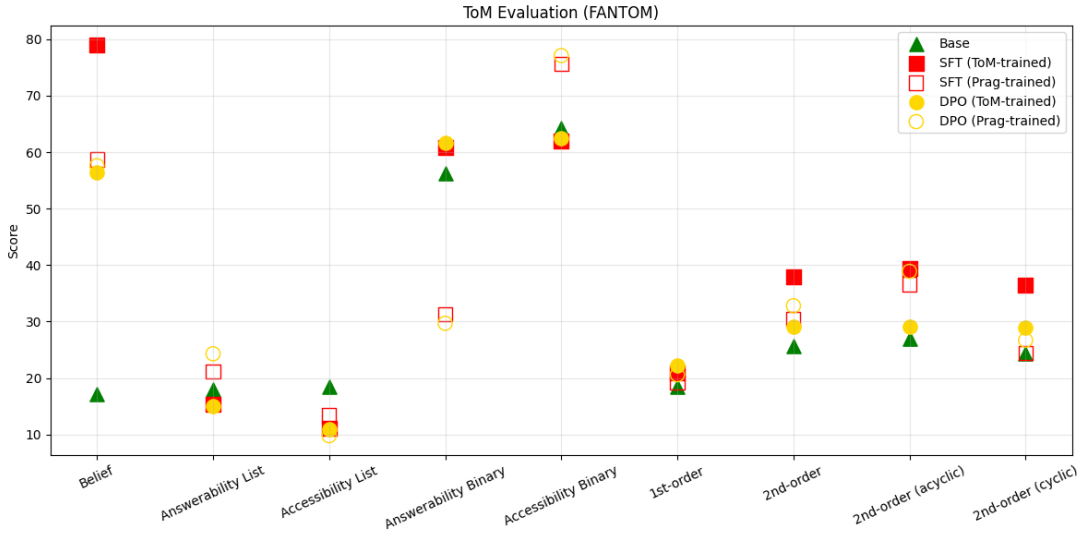


Figure 2: Results on the FANTOM benchmark for the base model, SFT, and DPO models.

Model / Task	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
SFT (ToM) vs Base	3.5	48.9	-53.5	-22.9	0.0	-4.5	-13.0	-7.5	1.5	-2.5	-26.0	-0.9	-43.9	-22.2
SFT (Prag) vs Base	18.1	113.7	18.2	84.1	0.0	55.3	0.0	9.5	39.1	8.0	28.0	0.0	11.2	-11.7
DPO (ToM) vs Base	30.2	92.4	27.6	2.7	0.0	-1.3	-7.4	1.1	7.3	-12.2	-4.7	-0.2	-1.7	-11.0
DPO (Prag) vs Base	18.0	67.3	12.0	74.1	0.0	64.3	7.4	9.3	42.5	1.5	1.7	0.0	-2.3	-7.9

Table 2: Per-task percentage change in F1 score relative to the base model on PUB.

Answerability under the same task format.

For both **first-order** and **second-order** ToM reasoning, the ToM-trained SFT model achieves the strongest performance, followed by both DPO models. Surprisingly, in all ToM-related metrics, the DPO–ToM model underperforms the SFT-ToM.

Table 3 corroborates these observations by summarizing the number of metrics in which each model improves over the base version, together with their mean and median scores. Overall, **DPO with pragmatics-oriented training data emerges as the most effective alignment strategy on FANTOM**. Both SFT variants also lead to improvements, with a slight advantage for ToM-based SFT training.

Table 4 confirms these trends: SFT (ToM) produces the largest improvements on first-order be-

Model	Data	Wins	Score (M / Md)
SFT	ToM	6	40.31 / 37.90
SFT	Pragmatics	6	34.54 / 30.50
DPO	ToM	6	35.10 / 29.10
DPO	Pragmatics	7	35.28 / 29.70

Table 3: Performance of SFT- and DPO-aligned models on the FANTOM benchmark compared to the base pretrained model. The *Wins* column indicates the number of tasks in which the model outperforms the reference model (in this case, the base model). The *Score* metric reports both the mean F1 and the median F1 achieved by the SFT or DPO model.

lief reasoning (M1), DPO (pragmatics) yields stable gains on later metrics (M7–M9) involving second-order cyclic and acyclic reasoning, and SFT (pragmatics) shows mixed results. Overall, SFT excels on core ToM tasks, whereas DPO provides more reliable performance on higher-order, multi-agent reasoning.

5.1. SFT vs. DPO

Tables 5 and 6 compare the effect of DPO alignment if compared to instruction tuning. In this case we want to see for ToM tasks wich training is the most effective and if model's correct answers on ToM data were determined from the ability of the model to follow instruction and to better understand the task or if they derive from preference learning alignment.

Results on the PUB dataset indicate that DPO is more effective than SFT only when training is performed with ToM data. In contrast, for pragmatic training, DPO does not produce a significant advantage over SFT.

Results on the FANTOM dataset instead show that DPO trained on ToM data is never effective for ToM tasks, whereas DPO applied to pragmatic training yields better results. Moreover, the comparison within the ToM setting shows that the SFT–ToM model consistently outperforms the DPO–ToM variant, indicating that preference learning is effective only when applied to pragmatic content.

5.2. Pragmatics vs. ToM

Tables 7 and 8 compare pragmatics-based and ToM-based training across the two evaluation benchmarks.

On the PUB benchmark, results clearly indicate that pragmatic training is more effective than ToM training across both alignment conditions. On FANTOM, a different pattern emerges. ToM-oriented data prove more effective only in the SFT condition, whereas pragmatic training yields superior performance for DPO-aligned models. It is clear that **ToM-specific training does not appear to provide a substantial advantage over pragmatic training** in enhancing ToM abilities.

6. Discussion

We can finally provide answers to the two research questions: i.) Is preference learning more effective when compared to instruction tuning? ii.) Is pragmatic training content sufficient to explain patterns observed in previous experiments?

Our results support the idea that what we observe in models might be the effect of instruction tuning rather than human alignment. Indeed, our first finding is that DPO yields better results when compared to SFT only in PUB evaluation (both DPO models) and in FANTOM if trained with pragmatic data. This suggests that **preference learning helps the model to solve pragmatic tasks, but not ToM tasks**.

Consequently, it is possible that what we observe in these models is not a genuine general capacity of reasoning or reproducing mentalizing processes, but instead an ability to learn how to perform specific tasks and to generalize this competence to closely related tasks. This phenomenon is also observed on the FANTOM benchmark, which is explicitly designed to isolate ToM abilities. Therefore, a more general methodological issue emerges, raising the question of how closely what we observe and evaluate in language models truly corresponds to the cognitive abilities we intend to test (Hu et al., 2025).

Our second finding suggest that training on pragmatic data is more effective at improving model performance than training on ToM data. This observation supports two points widely discussed in the psycholinguistic literature: i.) the necessity and difficulty of isolating ToM reasoning from pragmatic reasoning, and ii.) the dissociation between pragmatics and ToM in language models. We observed that models tend to dissociate these abilities at the output level, exhibiting stronger performance on pragmatic tasks and weaker performance on ToM tasks. We hypothesize that this is because models do not possess genuine ToM abilities; rather, what we observe in ToM evaluations is the manifestation of certain pragmatic inferences that emerge in some ToM tests.

This interpretation is further supported by the finding that pragmatic alignment benefits both ToM and pragmatic tasks. Therefore, what appears to be ToM reasoning in the models may instead reflect the generalization of specific pragmatic competencies. In this sense, our results align with Grice's view that ToM reasoning can be considered an enrichment that follows pragmatic interpretation (Grice,

Model / M	M1	M2	M3	M4	M5	M6	M7	M8	M9
SFT (ToM) vs Base	61.8	-2.5	-7.3	4.6	-2.3	2.5	12.3	12.5	12.0
SFT (Prag) vs Base	41.6	3.3	-5.0	-24.9	11.3	0.8	4.9	9.7	0.0
DPO (ToM) vs Base	33.5	-3.1	-9.4	3.8	-3.9	-5.2	-4.2	-1.2	-7.4
DPO (Prag) vs Base	37.6	3.6	-5.1	-34.3	13.5	10.9	22.8	29.7	15.9

Table 4: Per-metric score differences on FANTOM across training conditions.

Comp.	W	F1 M	F1 Md
DPO(ToM) vs SFT(ToM)	10	0.398	0.357
DPO(Prag.) vs SFT(Prag.)	7	0.491	0.435

Table 5: DPO vs SFT on PUB. The *W* column indicates the number of tasks in which the DPO model outperforms the SFT model. The mean F1 (*F1 M*) and the median F1 (*F1 Md*) refers to the DPO model.

Comp.	W	M	Md
DPO(ToM) vs SFT(ToM)	3	35.10	29.10
DPO(Prag.) vs SFT(Prag.)	6	35.28	29.70

Table 6: DPO vs SFT on FANTOM. The *W* column indicates the number of tasks in which the DPO model outperforms the SFT model. The mean F1 (*F1 M*) and the median F1 (*F1 Md*) refers to the DPO model.

1967). The models may then only perform the initial pragmatic step, and it is precisely this pragmatic information that produces apparent success on some ToM tasks.

7. Conclusion

In this work, we demonstrated that pragmatics can play a crucial role in eliciting ToM–like behaviour in LLMs. We also observed that preference learning appears to contribute differently to performance on ToM tasks compared to pragmatic tasks, and that SFT may be sufficient to improve performance on certain ToM evaluations. Although we did not construct the training and evaluation datasets ourselves, and although one potential limitation of this study is that pragmatic datasets may already embed ToM components, our findings nonetheless open the way to more fine-grained investigations into the role of specific linguistic features in training data. In addition, we did not systematically explore the effect of different dataset sizes for SFT and direct preference optimization DPO, which represents another avenue for future work. Future research could extend this analysis to pretraining data, enabling a more comprehensive understanding of how different stages of training contribute to ToM-related behaviour. More broadly, our results underscore the importance of disentangling ToM abilities from other concurrent capacities, such as pragmatic reasoning, when designing and interpreting ToM evaluations. This distinction is essential if we aim to make robust and theoretically grounded claims about the presence or extent of ToM abilities in large language models.

Comp.	W	F1 M	F1 Md
SFT (Prag.) vs SFT (ToM)	13	0.493	0.455
DPO (Prag.) vs DPO (ToM)	13	0.491	0.435

Table 7: Cross-training comparisons between pragmatic and ToM supervision on PUB. The *W* column indicates the number of tasks in which models trained on pragmatic material outperforms models trained on ToM material. The mean F1 (*F1 M*) and the median F1 (*F1 Md*) refers to SFT and DPO models with pragmatic training.

Comp.	W	M	Md
SFT (Prag.) vs SFT (ToM)	3	34.54	30.50
DPO (Prag.) vs DPO (ToM)	5	35.28	29.70

Table 8: Cross-training comparisons between pragmatic and ToM supervision on FANTOM. The *W* column indicates the number of tasks in which models trained on pragmatic material outperforms models trained on ToM material. The mean F1 (*F1 M*) and the median F1 (*F1 Md*) refers to SFT and DPO models with pragmatic training.

8. Limitation

While the EXPLORE-ToM dataset used for SFT is theory-driven and explicitly designed to target ToM, the composition of the DPO materials presents a more complex landscape.

Regarding the DPO datasets used for ToM optimization, we selected three datasets and manually inspected them to verify that their responses were grounded either in the recognition of others' beliefs or in the recognition of feelings and emotions. On this basis, we classified them as explicitly requiring ToM reasoning, ensuring that the ToM DPO condition was indeed targeting mentalizing processes.

Moreover, since the DPO datasets were not

custom-built for this study, we cannot definitively rule out the presence of latent ToM requirements in every entry. However, we do not view this as a primary pitfall. The central contribution of this work is demonstrating that non-ToM material enhances ToM elaboration more effectively than ToM-specific material itself. Consequently, any minor ToM overlap within the DPO set is acknowledged as a limitation rather than a structural failure.

Conversely, the inclusion of explicit ToM reasoning in our pragmatic datasets is highly improbable due to the following factors:

SFT Constraints: These datasets focus on concise, functional reasoning that lacks the recursive mentalizing characteristic of ToM.

DPO Domain Specificity: The datasets were selected based on three criteria that prioritize pragmatic coherence over mental state attribution: casual conversation (targets textual coherence rather than perspective-taking); mathematical reasoning (focuses on contextual interpretation and adherence to communicative norms) and game dialogue (while interactive, these entries require adherence to historical dialogue flow rather than an analysis of the player's internal perspective).

Isolating ToM is inherently difficult, as evidenced by the specialized framework required for EXPLORE-ToM. We define our pragmatic datasets as "pragmatic-only" because they lack explicit mentalizing prompts or reasoning elicitation. While some entries may involve phenomena that sit at the boundary of mentalizing (such as figurative language), this ambiguity is an intrinsic characteristic of the partial overlapping nature of pragmatics and ToM.

9. Bibliographical References

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. *ArXiv*, abs/2312.00575.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2024. Instruction-tuning aligns llms to the human brain.

Valentina Bambini and Serena Lecce. 2025. At the heart of human communication: new views on the complex relationship between pragmatics and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230486.

Matteo Bortoletto, Constantin Ruhdorfer, Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024. Limits of theory of mind modelling in dialogue-based collaborative plan acquisition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4856–4871, Bangkok, Thailand. Association for Computational Linguistics.

Francesca M. Bosco, Maurizio Tirassa, and Ilaria Gabbatore. 2018. Why pragmatics and theory of mind do not (completely) overlap. *Frontiers in Psychology*, 9.

Huanqia Cai, Yijun Yang, and Zhifeng Li. 2024. System-2 mathematical reasoning via enriched instruction tuning.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4302–4310, Red Hook, NY, USA. Curran Associates Inc.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Paola Del Sette, Valentina Bambini, Elena Tonini, and Serena Lecce. 2025. Are theory of mind and metaphor comprehension causally related? a training study in middle childhood. *Language Acquisition*, 32(4):432–452.

Paul H. Grice. 1967. The logic of conversation. *William James Lectures*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Phil. Trans. R. Soc.*

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages

14397–14413, Singapore. Association for Computational Linguistics.

Michal Kosinski. 2024a. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 121(45):e2405460121.

Michal Kosinski. 2024b. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 121(45):e2405460121.

A. Kotila, L. Mäkinen, E. Leinonen, and S. Loukusa. 2025. Interrelatedness of false-belief reasoning, structural language, and pragmatic communication in typically developing 4- to 6-year-old children. *First Language*, 45(1):47–65. Online first published 2024.

Yi-Long Lu, Chunhui Zhang, Jiajun Song, Lifeng Fan, and Wei Wang. 2025. Do theory of mind benchmarks need explicit human-like reasoning in language models?

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting.

Hieu Minh "Jord" Nguyen. 2025. A survey of theory of mind in large language models: Evaluations, representations, and safety risks.

Haritz Puerto, Tilek Chubakov, Xiaodan Zhu, Harish Tayyar Madabushi, and Iryna Gurevych. 2025. Fine-tuning on diverse reasoning chains drives within-inference CoT refinement in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3789–3808, Vienna, Austria. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Leonardo Ranaldi and Andre Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 2325–2347. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Christoffer Forbes Schieche, Manu Mahal, William Hedley Thompson, and Julia Uddén. 2025. Pragmatics partially segregated from theory of mind: evidence from resting-state functional connectivity. *Philos Trans R Soc Lond B Biol Sci*.

Christoffer Forbes Schieche and Julia Uddén. 2025. Neural adolescent pragmatic development mirrors pragmatic differences in adulthood: an fmri study. *bioRxiv*. Preprint.

Melanie Sclar, Jane Yu, Maryam Fazel-Zarandi, Yulia Tsvetkov, Yonatan Bisk, Yejin Choi, and Asli Celikyilmaz. 2024. Explore theory of mind: program-guided adversarial data generation for theory of mind reasoning. In *The Thirteenth International Conference on Learning Representations*.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities.

Settaluri Lakshmi Sravanthi, Kishan Maharaj, Sravani Gunnu, Abhijit Mishra, and Pushpak Bhattacharyya. 2025. Understand the implication: Learning to think for pragmatic understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23778–23790, Vienna, Austria. Association for Computational Linguistics.

James W. A. Strachan, Davide Albergo, Gabriele Borghini, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8:1285–1295.

Winnie Street. 2024. Llm theory of mind and alignment: Opportunities and risks.

Winnie Street, John Oliver Siy, Geoff Keeling, Adrien Baranes, Benjamin Barnett, Michael McKibben, Tatenda Kanyere, Alison Lentz, Blaise Aguera y Arcas, and Robin I. M. Dunbar. 2024. Llms achieve adult human performance on higher-order theory of mind tasks.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

A. Appendix

A.1. SFT: Dataset and Prompts

Example of a prompt from the dataset used in SFT training that involves ToM reasoning:

EXPLORE Prompt: The sun shone through the large glass doors of the hotel lobby, illuminating the marble floor and casting a warm glow over the comfortable seating areas. Soft music filled the air, mingling with the gentle hum of conversation and the occasional chime of the elevators in the bustling hotel. As Kaylee entered the hotel lobby her eyes quickly scanned the space, taking in every detail to ensure everything was in order for the upcoming event. Kaylee's task of tidying the lobby extended to the small, silver item, which she carefully stowed away in the desk drawer, and Liam, observing from across the room, felt his interest in the object grow, his mind racing with questions about its significance and purpose. With her tasks in the lobby complete, Kaylee stepped out into the fresh air, the sounds of the bustling hotel lobby fading into the background as the glass doors slid shut behind her. Liam moved across the lobby floor, his footsteps silent on the marble as he walked towards the empty reception desk where the mystery item was now hidden. Moments later, Kaylee stepped back through the glass doors of the hotel, joining Liam in the lobby once again. Liam's interest in the silver item led him to reposition it, now resting snugly within the hotel lobby's nearby leather briefcase. Meanwhile, Kaylee began to run a final check on the lobby, seemingly unconcerned by the briefcase's new contents. In which container was the silver letter opener at the beginning? **Completion**: wooden desk drawer

Examples of pragmatic reasoning prompts from datasets used in SFT training:

Argilla/SFT Prompt: What is the primary function of mitochondria within a cell? The primary function of mitochondria is to produce energy for the cell through a process called cellular respiration. **Completion**: That's correct. The primary function of mitochondria is indeed to generate energy for the cell in the form of adenosine triphosphate (ATP) through cellular respiration, which involves the breakdown of nutrients and oxygen to produce ATP. This process is crucial for powering various cellular activities, including muscle contraction, nerve impulses, and biosynthesis.

LightEval/Natural Prompt: If a poppy is boring, then the poppy is red. If a poppy is big or fast, then the poppy is weak. If a plant is smart, then the plant is hot. If a plant is good, then the plant is purple. If a plant is bad or beautiful, then the plant is strong. Fact: The poppy is fragile and beautiful. The following can be determined about the poppy: **Completion**: The poppy is strong.

These examples illustrate reasoning tasks that require general logical or factual inference rather than ToM reasoning.

A.2. DPO: Dataset and Prompts

Examples of ToM prompts from datasets used in DPO alignment:

Theory-of-mind-dpo Prompt: In a classroom, there were three students: Fiona, Marcellus, and Zhu. The teacher placed a blue notebook on the desk at the front and told the class that it contained the answers to the next day's test. The teacher then left the room, instructing the students not to peek inside. Fiona, who was really curious, couldn't resist and took a quick look while Marcellus and Zhu were focused on their own work. After Fiona sat back down, Marcellus went to the front to sharpen his pencil, briefly glanced at the blue notebook, but did not touch it. Meanwhile, Zhu was absorbed in writing an essay and didn't notice the interactions involving the notebook at all. Later, the teacher returned and found the notebook slightly open. Faced with the possibility of accusing a student of cheating, whom should the teacher question first based on the available information and why? **Chosen**: The teacher should question Fiona first about the slightly open notebook. Since Fiona actually went ahead and looked inside the notebook while her classmates were distracted, she demonstrated direct interest in the contents, clearly ignoring the teacher's instructions. Marcellus, on the other hand, although he approached the desk, merely glanced at the notebook and did not touch it, displaying no signs of having tried to open it. Zhu was preoccupied with writing an essay and had no interaction with the notebook or awareness of Fiona's actions, making Zhu the least likely suspect. Given Fiona's action of looking inside the notebook when she thought no one was watching, it follows that she is the

most plausible individual responsible for leaving it slightly open. **Rejected**: Based on the available information, the teacher should question Fiona first. Fiona was the only student who admitted to peeking inside the notebook while the teacher was away, making her the most likely suspect for opening the notebook and potentially altering its contents. Marcellus only briefly glanced at the notebook, and Zhu was not aware of any interactions involving the notebook at all. Therefore, Fiona is the most suspicious based on the information provided.

Social-i-qa Prompt: Riley was the best of friends with the boy with cancer. What will Riley want to do next? **Chosen**: visit the hospital **Rejected**: become friends with the boy with cancer too

SocialReasoning Prompt: How do you maintain relationships with the opposite party after a tough negotiation? **Chosen**: Forgiveness is key in a situation where a tough negotiation has taken place. Putting aside hard feelings after a difficult discussion is essential. One must separate their feelings about the person from their feelings about the issue. I try to remember that the problem is the problem, not the person. I approach them with love and compassion regardless of what we have just been through. **Rejected**: I will tell them that we agree to disagree, and that I appreciate a conflicting opinion that will not affect our friendship.

These datasets, respectively, require recognizing character's beliefs, intention and emotion.

Examples of pragmatic reasoning prompts from datasets used in DPO alignment:

Casual-conversation Prompt: exactly, it wouldn't be nice if it started raining. it's too hot. **Chosen**: i know, you're absolutely right. **Rejected**: I totally agree! A sudden downpour would be quite uncomfortable, especially when it's already hot outside. It might even cool things off a bit though, which could be refreshing in its own way.

Thinker Prompt: If two cars leave the same city heading in opposite directions at the same time, one traveling at 60 km/h and the other at 80 km/h, how far apart will they be after 2 hours? **Chosen**: *Step 1*: I start by visualizing the situation: two cars moving away from each other. *Step 2*: Car A is traveling at 60 km/h and Car B at 80 km/h. *Step 3*: I need to find the distance each car travels after 2 hours. *Step 4*: Distance is calculated by the formula: Distance = Speed $\times$ Time. *Step 5*: For Car A: $Distance_A = 60km/h \times 2h = 120km$. *Step 6*: For Car B: $Distance_B = 80km/h \times 2h = 160km$. *Step 7*: Now I have the distances for both cars. *Step 8*: Since they are moving in opposite directions, I add the two distances together to find the total distance apart. *Step 9*: $Totaldistance = Distance_A + Distance_B$. *Step 10*: $Totaldistance = 120km + 160km$. *Step 11*: This gives me $Totaldistance = 280km$. *Step 12*: I check my calculations to ensure they make sense. *Step 13*: I consider how both cars started from the same point and moved away from each other. *Step 14*: I note that the different speeds will result in them being farther apart than if they traveled at the same speed. *Step 15*: I can visualize the scenario with a diagram to confirm my reasoning. *Step 16*: After considering the time frame of 2 hours, my figures seem reasonable and accurate. *Step 17*: I reflect on whether any external factors could affect the distances, such as road conditions, but note these are not part of the problem. *Step 18*: I realize the problem is straightforward and purely mathematical. *Step 19*: I think back to similar problems I have encountered, which reassures me of my approach. *Step 20*: I wonder if there is a more complex way to approach this, but the simplicity here works perfectly. *Step 21*: I conclude the total distance apart is indeed correct and consistent with the rules of motion. *Output*: After 2 hours, the two cars will be 280 kilometers apart. Car A, going at 60 km/h, will have traveled 120 km, while Car B, moving at 80 km/h, will have covered 160 km. Adding those distances together gives the total distance apart. **Rejected**: After 2 hours, the two cars will be 280 kilometers apart. Car A, going at 60 km/h, will have traveled 120 km, while Car B, moving at 80 km/h, will have covered 160 km. Adding those distances together gives the total distance apart.

DPO-dialogue Prompt: You are a language wizard who likes to guess words by using the given rules. Welcome to Wordle! You have six attempts to guess the target word, a valid English word of five lowercase letters (a–z). Please use the tags `guess:` and `explanation:` to provide a concise explanation for each guess. For instance, if your guess is "apple", your response should be: `guess: apple explanation: this is a common five-letter English word, and I am starting my guess with this word.` After each guess, your answer will be validated, and you will receive feedback indicating which letters are correct (green), which letters are correct but in the wrong position

(yellow), and which letters are incorrect (red). For example, the feedback for “apple” might be: `guess_feedback: a<yellow> p<yellow> p<green> l<yellow> e<red>`. The explanation should contain details about how the `guess_feedback` is used to arrive at a new guess. Let us begin with your first guess. **Chosen**:

> **Assistant**: `guess: apple` — *This is a common five-letter English word, and I am starting my guess with this word.*
>
> **User**: `guess_feedback: a<red> p<red> p<red> l<yellow> e<red>`
>
> **Assistant**: `guess: table` — *This is a common five-letter English word; I am using the feedback to exclude the letters ‘a’, ‘p’, and ‘e’.*
>
> **User**: `guess_feedback: t<red> a<red> b<yellow> l<yellow> e<red>`
>
> **Assistant**: `guess: blend` — *This is a common five-letter English word; I am using the feedback to include the letters ‘b’, ‘l’, and ‘e’.*
>
> **User**: `guess_feedback: b<green> l<green> e<red> n<green> d<red>`
>
> **Assistant**: `guess: blink` — *This is a common five-letter English word; I am using the feedback to include the letter ‘k’ and exclude the letter ‘d’.*

Rejected:

> **Assistant**: `guess: adieu`
>
> **User**: `guess_feedback: a<red> d<red> i<green> e<red> u<red>`
>
> **Assistant**: `guess: ideas`
>
> **User**: `guess_feedback: i<yellow> d<red> e<red> a<red> s<red>`
>
> **Assistant**: `guess: idiot`
>
> **User**: `guess_feedback: i<red> d<red> i<green> o<red> t<red>`
>
> **Assistant**: `guess: idiot` [repeated]
>
> **User**: `guess_feedback: i<red> d<red> i<green> o<red> t<red>`
>
> **Assistant**: `guess: idiot` [repeated]

These, respectively, encompass discourse coherence and adherence to conversational maxims (specifically be clear and informative), general reasoning (which we include into pragmatic as general ability to produce inferences), and metalinguistic reasoning.

What Kind of Language is Easy to Language-Model Under Curriculum Learning?

Nadine El-Naggar, Tatsuki Kuribayashi, Ted Briscoe
Mohamed bin Zayed University of Artificial Intelligence
{nadine.naggar, tatsuki.kuribayashi, ted.briscoe}@mbzuai.ac.ae

Abstract

Many of the thousands of attested languages share common configurations of features, creating a spectrum from typologically very rare (e.g., object-verb-subject word order) or impossible languages to very common combinations of features (e.g., subject-object-verb word order). One central question is under what conditions such typological tendencies can be predicted, and specifically whether the learning bias of language models (LMs) is sufficient to reproduce such patterns. In this study, we add one dimensionality to such analysis — the learning scenario for LMs — to explore its interaction with the inductive bias of LMs. Specifically, as a first study, we examine the effect of curriculum learning (CL), as a developmentally motivated learning scenario, i.e., starting with simpler sentences rather than randomly-ordered input. We expand existing LM-based exploration (El-Naggar et al., 2025a,b) with a simple CL variant and find that CL substantially impacts the apparent inductive bias of LMs.

Keywords: Artificial languages, curriculum learning, generalization

1. Introduction

Natural languages (NLs) exhibit a range of properties, including different word order configurations, raising the question: which, if any, types of languages are easier for language models (LMs) to learn (Cotterell et al., 2018; Mielke et al., 2019; White and Cotterell, 2021; Borenstein et al., 2024; Arnett and Bergen, 2025)? Related questions include why some grammatical feature combinations are typologically common while others are rare (Dryer and Haspelmath, 2013), and to what extent LMs' domain-general learning biases replicate these linguistic tendencies (Chomsky et al., 2023; Xu et al., 2025). Existing studies have reported that, for example, the inductive bias of a specific class of LMs aligns with typological commonality, and the properties of such typologically-aligned models (e.g., working memory limitations) could partially explain such typological tendencies (Kuribayashi et al., 2024; El-Naggar et al., 2025a,b). Although such computational simulations of language learnability can serve as proof-of-concept for the idea that learning biases shaping language (Culbertson et al., 2012), LMs' and humans' language acquisition settings are generally different (e.g., the amount of data), and to make such simulations more relevant to human language science, it is necessary to align LM training scenarios with humans' ones (Warstadt and Bowman, 2022).

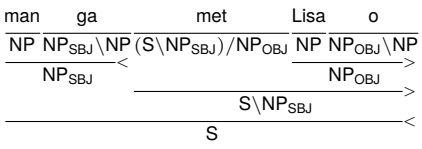
One factor that is relevant to human language acquisition and overlooked in existing LM-based simulations of typological patterns is the order of data presentation. For example, children with severe working memory limitations may take in relatively simple sentences first through language acquisition (Hudson Kam and Newport, 2005; Kam and Newport, 2009). Such effects have typically been simulated with curriculum learning (CL) in computational simulations, which present training data from simpler to more complex sentences.

In this study, we explore one basic CL setting in the LM-based simulation of typological patterns. That is, our research question is *what kind of language is easier for LMs to learn under CL*. As an initial foray, we adopt a simple length-based CL (Spitkovsky et al., 2009) and replicate existing studies (El-Naggar et al., 2025a,b) to analyze interaction effect between CL and LMs' inductive biases over diverse word orders. Our experimental results demonstrate that the word-order preferences indeed change with the CL setting, and, somewhat surprisingly, the CL-based results deviate more from typological commonality, which raises several implications (§ 5).

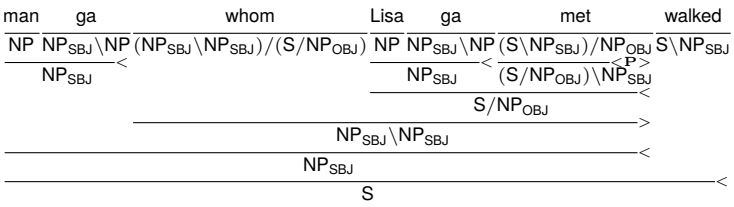
2. Background

Artificial Language Learning Artificial languages (ALs) are often used in the evaluation of LMs as they allow for more targeted evaluation of specific linguistic features. Typical investigations include analyzing what kind of structure LMs can model, using ALs of varying complexities (Suzgun et al., 2019; Weiss et al., 2018; El-Naggar et al., 2022; Kallini et al., 2024; Someya et al., 2024; Delétang et al., 2023). In addition to formal languages (e.g., Dyck languages), more linguistically grounded ALs, such as PCFG-based ones, have been developed to evaluate LM word-order inductive biases (White and Cotterell, 2021). El-Naggar et al. (2025a,b) extended PCFG-based ALs to generalized categorial grammar (GCG)-based

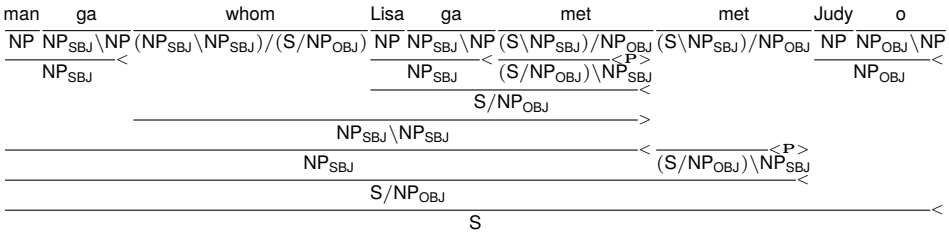
Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 48–59
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026



(a) Example with 5 words (train)



(b) Example with 7 words (train)



(c) Example with 9 words (MEDIUM test)

Figure 1: Examples of sentences and their GCG derivation (somewhat simplified for space limitations).

ALs, which support a broader range of grammatical phenomena. ALs are also often used to test LM generalization under controlled conditions (Weiss et al., 2018; Suzgun et al., 2019; El-Naggar et al., 2022, 2023, 2025b).

Curriculum Learning Research has long questioned whether the order in which training data is exposed to neural network models affects their learning (Elman, 1991, 1993; Rohde and Plaut, 1999; Krueger and Dayan, 2009; Bengio et al., 2009). Elman (1991, 1993) introduced the idea of "starting small" and conducted experiments where RNNs were trained in phases to learn English sentences, where sentence complexity was increased in each phase. The starting small hypothesis has been revisited in several NLP applications, like learning and generalization of grammatical patterns (Rohde and Plaut, 1999), unsupervised dependency parsing (Spitkovsky et al., 2009) and in a reinforcement learning framework (Krueger and Dayan, 2009), and even in visual applications like shape recognition (Bengio et al., 2009).

3. Experimental Settings

3.1. Original Setting

We first briefly introduce the base datasets adapted from prior works (El-Naggar et al., 2025a,b). A set of GCG grammars is first defined via multiple independent word-order parameters generating 96 languages for the GCG-based AL corpus. For example, one AL follows Japanese-like (head-final), and another follows English-like (mostly head-initial) word order. Specifically, there are 7 binary parameters to determine (i) subject–verb order, (ii) object–verb order, (iii) subject–object order, (iv) complementizer–clause, (v) noun-adposition, (vi) noun-adjective, and (vii) position of relativizer. These configurations are denoted as a sequence of digits (e.g., `0001010`). Configurations with more `0`s tend to be head-final, while the ones with more `1`s are head-initial. The binary parameter controls the directional slash of specific rules in the GCG grammar. Examples of the sentence "Kim said that John touched Lisa" in grammars `0000000` (Japanese-like) and `0101101` (English-like) are:[1]

`0000000`: *Ken ga John ga Lisa o touch that said*

`0101101`: *Ken ga said that John ga touch Lisa o*

Note that the ALs have case markers for subject (*ga*) and object (*o*). Figure 1 shows more examples of sentences derived by `0101101` (English-like) GCG grammar. Note that sentences are first generated fully randomly, and then grammatical (parsable) ones are selected based on the GCG parser, where GCG parser configurations differ in different word order configurations.

In each AL, the training data consists of 80K sentences with 3–8 words (uniform length distribution), and the LMs' inductive biases are evaluated on three types of evaluation sets: (i) SHORT test set with the same length distribution as the training data, (ii) MEDIUM test set with longer length distribution than the training set (9-10) to test generalization, and (iii) LONG test set that has further longer sentences created by several heuristics (11-20). [2] We follow this data split, but change how

[1] Note that while the words in the ALs are pseudo-words, real English words are used in these examples for readability.

[2] These are available at https://github.com/nadineelnaggar/gcg-based-artificial-languages

the training data is presented to the model and observe its effect on word-order preferences. We define ORIGINAL setting as the model training using this base dataset with a randomized data order.

3.2. Curriculum Learning Setting

To examine the CL effect on LMs' inductive biases, we introduce the CL setting in which the training data is presented based on a specific strategy. As a case study, we explore the widely-used token length-based CL, and thus our questions will be: do LMs prefer specific word order configurations independently of the ordering of training data? Or more generally, how reproducible are existing findings under different LM training scenarios? We do not aim to find the best, optimal learning scenario to control LMs' learning preferences or to perform a comprehensive analysis, trying every possible learning strategy.

Specifically, we split the base training data into six parts (lengths of 3, 4, 5, 6, 7, and 8) with the same number of sentences in each part. LMs are trained sequentially in stages corresponding to each training data part. That is, the first training stage is with sentences of length 3, and the second stage will be with sentences of length 4, and so on. To avoid catastrophic forgetting of previous data, we sample 5% of the data from each past part and inject it into the current part. For example, the data in the third stage consists of length-3 (5%), length-4 (5%), and length-5 (90%) sentences. For a fair comparison with the ORIGINAL setting, the total number of training sentences is set to match the base dataset (80K). For each of 96 ALs in the base dataset, we created a CL trained model.

4. Experiments

We adopt three evaluation settings and replicate El-Naggar et al. (2025b). For non-CL, ORIGINAL settings, results are excerpted from that study for a comparison, and our new CL-based results are superimposed on them. We focus on RNN, LSTM, and Transformer LMs.

Model Training. For the CL settings, we adopt the same setting as the original ones, except for CL (see Appendix C for hyperparameter details). We train LMs with three different random seeds for each model, using the Fairseq toolkit (Ott et al., 2019). Under CL, LMs are trained for 2 epochs in each stage, except the last, and the optimizer (e.g., learning rate scheduler) is inherited across all the stages. At the last stage, an early-stopping criterion of 5-epoch patience is adopted (i.e., if the validation PPL is not improved for 5 epochs, the training stops), following the non-CL settings.

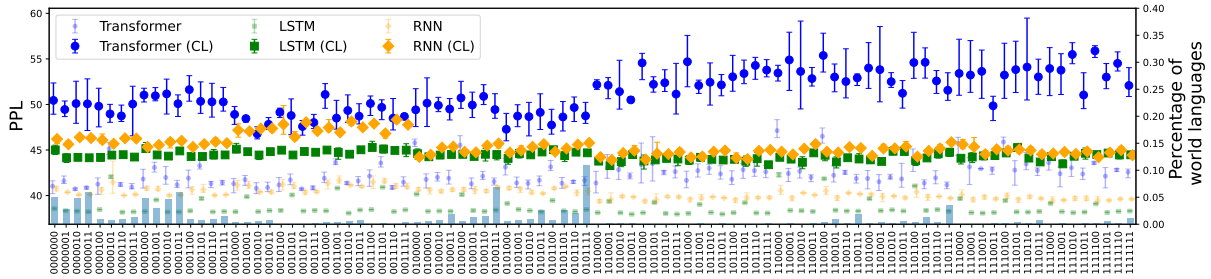
Evaluation. The SHORT test set evaluates the LMs' in-distribution modeling ability, and MEDIUM and LONG ones evaluate the out-of-distribution generalization ability of LMs. Note that the sentences in the SHORT test sets are unseen in the training data. We report average perplexity (PPL) among three runs for each word order configuration. We also report typological alignment (TA) scores — Pearson's correlation coefficient between model's average PPL and the frequency of respective word order configurations in the world (Kuribayashi et al., 2024; El-Naggar et al., 2025a,b). A negative TA score reflects a better alignment of a LM's learning preference with typological tendencies, i.e., typologically common word orders are easier to learn for the LMs. Our focus is on how the language-learning curriculum (ORIGINAL vs. CL) affects their learning preferences (PPLs and TAs).

4.1. Experiment 1: PPLs

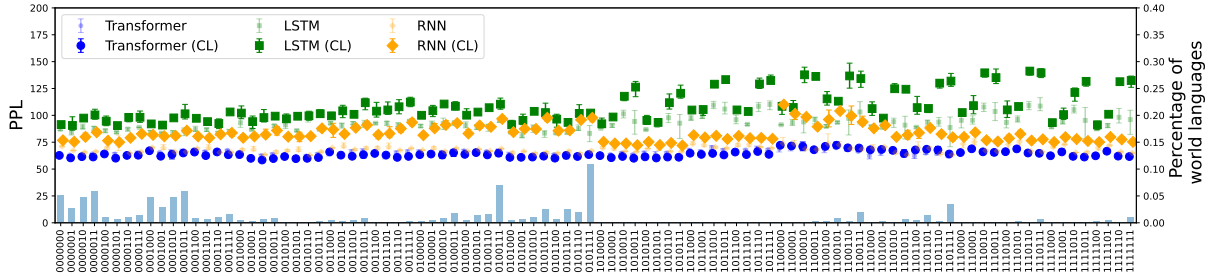
PPL variations and CL. We first evaluate LMs based on perplexity (PPL) on the SHORT, MEDIUM, and LONG test sets. Figure 2 shows the general PPL tendencies (y-axis) over different word order configurations (x-axis) in each test set. Here, PPLs from LMs without CL are also shown with smaller, semi-transparent markers.[3] As a quick check, we first confirm that word order preference still varies even when data are presented in a comprehensive way with CL. Compared to the original results (smaller, semi-transparent markers in Figure 2), we observe that LMs with CL tend to exhibit worse PPL on SHORT tests, but comparable or better PPL on MEDIUM and LONG tests, suggesting less overfitting to short sentences and a positive effect of CL for length generalization.

Typological variations and CL. Table 1 aggregates the results of our CL-based LMs, and the TA score is also computed. To summarize the results, first, the TA scores, especially in RNN/LSTM results on MEDIUM/SHORT sets, substantially change with and without CL, and this basically leads to a worse TA. Second, the TA scores in LONG set are still significantly negative; that is, typologically common word order configurations facilitate generalization to longer sentences, regardless of CL. Therefore the learning setting does affect the results, and in our case, this impact was not equally strong nor random across models/conditions, but rather altered specific results on individual ALs and test sets. Further experimentation with different approaches to CL, such as ordering by construction

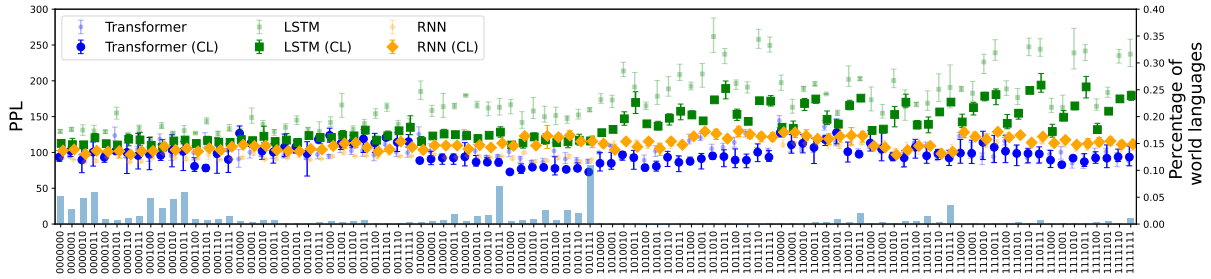
[3] These scores without CL are exactly the same as those reported in El-Naggar et al. (2025b). Our CL results are comparable with their study as the adaptation of CL is the only difference.



(a) SHORT test (length 3–8).



(b) MEDIUM test (length 9–10).



(c) LONG test (length 11–20).

Figure 2: Distributions of perplexities and typological plausibility across languages. The error bars indicate max and min PPLs within three runs. The smaller, semi-transparent markers correspond to the results without CL, which is excerpted from El-Naggar et al. (2025b).

complexity as opposed to token length, may shed further light on the interactions between models, learning scenarios and inductive bias.

Inter-model correlations. Figure 3 shows inter-correlation of PPL-vectors over 96 word orders from different LMs and CL settings. Note that the orthogonal elements in the matrices show the average correlation between PPL-vectors from the same LM architecture, but with different seeds. First, PPL correlation between the same model with and without CL is generally smaller than the seed variance (respective orthogonal elements), showing that CL does affect the word order preference of LMs. Second, CL has different effects depending on LM architectures; for example, in SHORT test, LSTM vs. LSTM (CL) shows nearly zero correlation (0.036), in MEDIUM test, RNN vs. RNN (CL) shows a small correlation (0.145), while in LONG test, Transformer vs. Transformer (CL) shows a relatively small correlation (0.276). This demonstrates subtle and non-obvious interactions between word order preferences, model architectures, and learning scenarios.

4.2. Experiment 2: Targeted Evaluations

We also replicate the targeted evaluations conducted in El-Naggar et al. (2025b) (see details in their Sections 6 and 7).

4.2.1. PPLs in Targeted Generalization Sets.

We measure PPLs on the test data with specific complex constructions. Here, the generalization is evaluated on unbounded dependency structures, specifically recursive relative clauses (RECURSIVE), where relative clauses are nested, and embedded relative clauses (EMBEDDED), where a relative clause is embedded in a subordinate clause. English examples are as follows:

Recursive Relative Clauses: *fruits ga which pasta ga which John ga promised nibbles received wall o*

Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
		Short						
Trans. (El-Naggar et al., 2025b)		41.8	41.6	42.3	42.6	42.7	43.3	−**27.7**†
Trans.	✓	50.2	48.8	49.3	52.6	53.3	53.4	−**22.3**†
LSTM (El-Naggar et al., 2025b)		38.7	38.8	38.7	38.4	39.1	38.5	−**14.2**
LSTM	✓	44.4	44.9	44.5	43.9	44.2	44.4	16.1
RNN (El-Naggar et al., 2025b)		40.4	41.0	40.6	39.7	40.1	39.7	13.0
RNN	✓	45.9	47.4	45.1	44.5	45.0	44.8	14.2
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-

(a) Short test set.

Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
		Medium						
Trans. (El-Naggar et al., 2025b)		65.2	63.5	64.2	65.9	66.1	65.0	−**10.4**
Trans.	✓	63.2	61.8	62.9	62.8	68.9	64.7	−**5.9**
LSTM (El-Naggar et al., 2025b)		85.9	91.7	88.0	97.5	92.9	97.9	−**31.0**†
LSTM	✓	95.6	102.3	101.9	112.4	119.8	117.7	−**20.1**†
RNN (El-Naggar et al., 2025b)		67.8	67.9	66.7	69.6	69.0	69.4	−**17.4**
RNN	✓	80.3	84.5	89.7	76.6	91.7	78.1	21.2
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-

(b) Medium test set.

Model	CL	SOV	OSV	SVO	OVS	VSO	VOS	TA ↓
		Long						
Trans. (El-Naggar et al., 2025b)		102.3	99.4	97.9	104.0	107.6	97.9	−**19.2**
Trans.	✓	95.1	112.9	83.1	89.9	106.2	96.2	−**18.0**†
LSTM (El-Naggar et al., 2025b)		131.9	141.5	160.7	205.5	180.9	207.5	−**33.4**†
LSTM	✓	113.8	122.0	119.3	153.6	151.9	163.7	−**32.3**†
RNN (El-Naggar et al., 2025b)		91.8	94.6	93.2	118.0	109.0	114.2	−**43.1**†
RNN	✓	102.9	107.2	113.0	117.6	113.7	117.8	−**20.2**†
NL (Prob. ↑)		0.54	0.04	0.23	0.01	0.12	0.05	-

(c) Long test set.

Table 1: Average PPLs within each base word order group and Pearson's correlation coefficient (TA) between PPL and typological frequency. Negative TA scores are highlighted in bold. Statistical significance ($p<0.05$) is marked with †. NL is the percentage of natural languages possessing each word order.

Embedded Relative Clause: *fruits ga which John ga said that pasta ga nibbles received wall o*

In the Recursive test set, two relative clauses are nested. In the Embedded test set, a relative clause is embedded in a subordinate clause, e.g., "he said". Through randomly sampling lexical items for the above two fixed construction templates, 500 examples for each construction are generated in each of 96 word order variations. Note that the resulting sentences are longer than 8 tokens (training data). PPLs are measured in each targeted test set, and these are aggregated as the TA score, correlation between average PPL over three model with different random seeds and word order frequency across 96 languages.

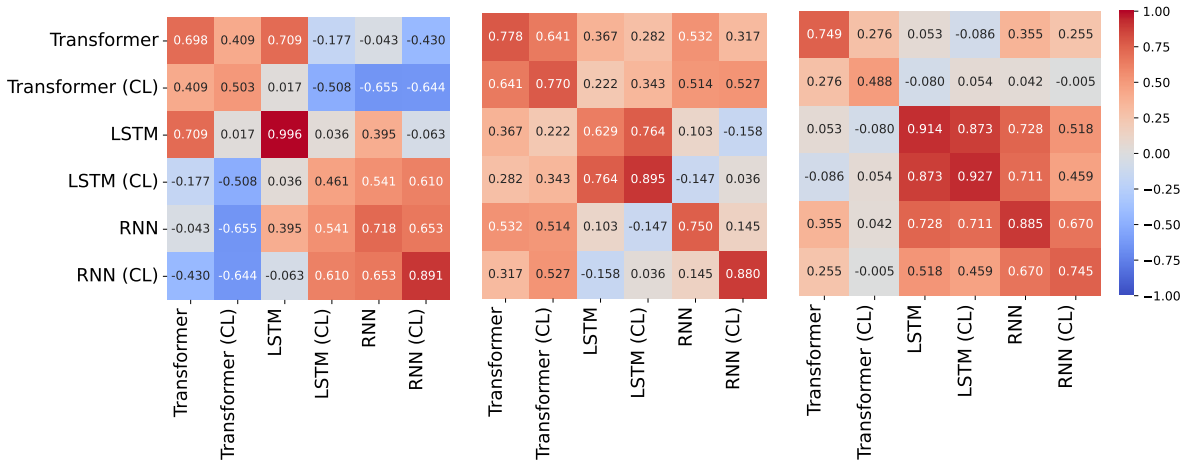
Table 2 (left-side) shows results with and without CL. CL typically yields comparable or slightly worse TA scores, while the relative order of TA scores among conditions is basically preserved, e.g., Transformer in Embedded yields the lowest TA, while RNN in Recursive yields the highest TA with or without CL.

4.2.2. Grammaticality Judgments.

We also perform grammaticality judgment evaluation with two test suites: (i) case-type accuracy, and (ii) verb-type accuracy. English-like (`0101101`) examples are as follows:

Case-type judgment: [4]

[4] *ga* and *o* in the examples are subject and object case



(a) SHORT (b) MEDIUM (c) LONG

Figure 3: Correlation of word order preference between different models×CL. Transformer (CL), for example, denotes Transformer-based LM trained with CL. Orthogonal elements denote seed variance, i.e., average correlations between the same model but different seeds.

		RECURSIVE	EMBEDDED	Case Judgment		Verb Judgment	
Model	CL	TA (↓)	TA (↓)	Acc. (↑)	Corr. (↑)	Acc. (↑)	Corr. (↑)
Trans. (El-Naggar et al., 2025b)		−5.1	−23.5†	97.7±1.5	0.14	81.0±14.7	0.27†
Trans.	✓	5.5	−19.2†	96.3±3.1	0.01	74.4±22.2	0.26†
LSTM (El-Naggar et al., 2025b)		9.2	−3.7	97.2±1.4	0.03	85.1±9.6	0.28†
LSTM	✓	16.2	0.4	95.6±1.9	-0.31†	78.1±11.7	0.32†
RNN (El-Naggar et al., 2025b)		12.9	−18.1†	97.4±1.4	0.21†	77.4±15.5	0.23†
RNN	✓	17.0	−18.9†	92.6±5.8	0.10†	63.0±25.1	0.31†

Table 2: Correlation between PPLs and word-order frequencies in targeted generalization sets (Recursive/Embedded) and correlation between accuracy and word-order frequencies in grammaticality judgment tests. Statistical significance ($p<0.05$) is marked with †.

- *fluffy soft and intelligent mango ga controls owl o*
- **fluffy soft and intelligent mango o controls owl o*

Verb-type judgment: [5]

- *scooter ga which green machine ga escorts walk*
- **scooter ga which green machine ga evolves walk*

For the case-type judgment, a randomly selected grammatical case marker is replaced with an ungrammatical one, e.g. “ga” is replaced with “o” or vice versa. For verb-type judgment, a transitive verb in a grammatical sentence is replaced with an intransitive verb, resulting in the sentence becoming ungrammatical (marked with * in the examples above). For each word order, 500 items are sampled from the MEDIUM test set, and their ungrammatical variants are created by applying specific modifications. Should there be more than one potential replacement candidate in a sentence, one is chosen at random. We report accuracy (Acc.) of grammatical judgments and correlation (Corr.) with typological commonality — the higher the better.

Table 2 (right side) shows the results. CL consistently worsened the accuracies, and the typological alignments are also weakened (Case-type judgment) or comparable (Verb-type judgment). The non-positive effect of CL is similar to the other experiments.

5. Discussion

Overall, CL led to a slightly negative effect for typological alignment between learnability and typological patterns, which leads to some interpretations. First, the results empirically demonstrated the considerable interaction effects between LMs’ word-

markers, respectively.

[5] *escorts* and *evolves* in the examples are transitive and intransitive verbs, respectively.

order preferences, or more generally, inductive biases, and how data are presented. It is worth revisiting LMs' learning biases in developmentally motivated language-learning scenarios. Second, if the length-based CL we adopted is truly a good proxy for human language acquisition, the alignment between LMs' inductive biases and typological patterns might have been overestimated in existing non-CL studies. This view, more or less, challenges the idea that learners' biases shape language design and, indirectly, evokes the dominance of alternative hypotheses associating language design with cognitive-external factors, e.g., historical accidents or geographical influence (Moravcsik, 1978; Atkinson, 2011); see Culbertson et al. (2012) i.a. for a more comprehensive overview. Third, otherwise, if we stand on the premise of the learner's bias shaping language design, the weaker typological alignment under the adopted length-based CL raises the concerns that the length-based control is too simplified CL strategy, given broader possibility of CL implementations (Wang et al., 2022), or CL alone is insufficient to simulate the plausible biases in language acquisition (Perfors, 2012). These views at least poses the next question: are there any other CL strategies or factors in addition to CL leading to better simulating the emergence of typological patterns?

6. Conclusion

In this paper, we revisit the prior work (El-Naggar et al., 2025b) by introducing one basic type of curriculum learning (CL). In our experiments, we ablate the CL effect on the inductive bias of RNN, LSTM, and Transformer LMs towards different word-order configurations. We confirm that the LMs' word-order preferences indeed change under CL, suggesting that the order of the training data affects apparent LM learning biases. Our results show subtle and somewhat inconsistent interactions between individual ALs, model architectures, learning scenarios, and inductive bias. In the future, we can explore several directions to extend this line of work and gain further insights. One of them is to investigate different CL strategies, such as the construction-complexity approach or model-based control of working memory growth (Wang et al., 2022; Mita et al., 2025).

Limitations

In addition to the future directions mentioned in § 6, the definition of typological alignment should be carefully addressed. For example, we have used Pearson's correlation to examine the relationship between typological commonality and perplexity in this study, following Kuribayashi et al. (2024) and El-Naggar et al. (2025b), but its validity should be re-investigated. Finally, this study explanatorily analyzed the effect of CL on LMs' inductive bias evaluation and does not derive any strong conclusion like CL clearly promotes/hinders typological alignment between LMs' learning biases and typological commonalities. We rather hope that our paper introduces and encourages this interdisciplinary exploration space combining LMs' training scenarios and typological alignment, particularly given the progressive interdisciplinary integration between language modeling and language science.

Ethics Statement

The data used in this paper is all artificial data based mostly on English words. We have no ethical concerns with the contents of this paper.

Acknowledgment

We would like to thank the anonymous reviewers for their insightful feedback.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Quentin D Atkinson. 2011. Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science*, 332(6027):346–349.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Nadav Borenstein, Anej Svete, Robin Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. 2024. What languages are easy to language-model? a perspective from learning probabilistic regular languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15115–15134, Bangkok, Thailand. Association for Computational Linguistics.

Ted Briscoe. 1997. Co-evolution of language and of the language acquisition device. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 418–427, Madrid, Spain. Association for Computational Linguistics.

Ted Briscoe. 2000. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, 76(2):245–296.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Culbertson, Paul Smolensky, and Géraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition*, 122(3):306–329.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. 2023. Neural networks and the chomsky hierarchy. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025a. Gcg-based artificial languages for evaluating inductive biases of neural language models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 540–556.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025b. Which word orders facilitate length generalization in LMs? an investigation with GCG-based artificial languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35599–35613, Suzhou, China. Association for Computational Linguistics.

Nadine El-Naggar, Pranava Madhyastha, and Tillman Weyde. 2022. Exploring the long-term generalization of counting behavior in RNNs. In *I Can't Believe It's Not Better Workshop: Understanding Deep Learning Through Empirical Falsification*.

Nadine El-Naggar, Pranava Madhyastha, and Tillman Weyde. 2023. Theoretical conditions and empirical failure of bracket counting on long sequences with linear recurrent networks. *EACL 2023*, page 143.

Jeffrey L Elman. 1991. Incremental learning, or the importance of starting small. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 13.

Jeffrey L Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Carla L Hudson Kam and Elissa L Newport. 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Learn. Dev.*, 1(2):151–195.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14691–14714. Association for Computational Linguistics.

Carla L Hudson Kam and Elissa L Newport. 2009. Getting it right by getting it wrong: when learners change languages. *Cogn. Psychol.*, 59(1):30–66.

Kai A Krueger and Peter Dayan. 2009. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14522–14543. Association for Computational Linguistics.

S. J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume*

1: Long Papers, pages 4975–4989. Association for Computational Linguistics.

Masato Mita, Ryo Yoshida, and Yohei Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9386–9399, Vienna, Austria. Association for Computational Linguistics.

Edith Moravcsik. 1978. Language contact. *Universals of human language*, 1:93–122.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Amy Perfors. 2012. When do memory limitations lead to regularization? an experimental and computational investigation. *J. Mem. Lang.*, 67(4):486–506.

Douglas LT Rohde and David C Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted syntactic evaluation on the chomsky hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15595–15605. ELRA and ICCL.

Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2009. Baby steps: How "less is more" in unsupervised dependency parsing. *NIPS: Grammar Induction, Representation of Language and Language Learning*, pages 1–10.

Mark Steedman. 1996. *Surface structure and interpretation*, volume 30 of *Linguistic inquiry*. MIT Press.

Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M Shieber. 2019. Lstm networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pages 44–54.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4555–4576.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press, Boca Raton.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 740–745. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 454–463. Association for Computational Linguistics.

Mary McGee Wood. 2014. *Categorial grammars (RLE linguistics b: Grammar)*. Routledge.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can language models learn typologically implausible languages?

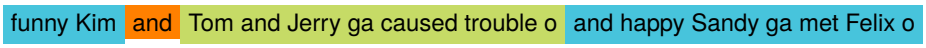
A. Categorial Grammar

Categorial Grammar (CG) is a formalism consisting of a lexicon, where each word is assigned a functor category, and a set of rules that define how the categories combine with each other. CG was introduced with the combinatory operation **application**, which has forward and backward variants. Extensions of CG have been introduced, such as combinatory categorial grammar (CCG) (Steedman, 1996) and generalized categorial grammar (GCG) (Wood, 2014). El-Naggar et al. (2025b) use a GCG with the following combinatory operations, which they detail in their paper:

- Application (forward and backward variants)
- Coordination
- Composition (forward and backward variants)
- Weak Generalized Permutation

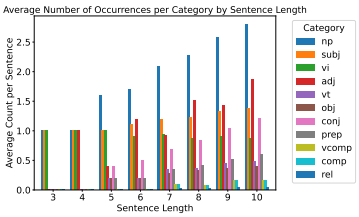
This GCG allows for the expression of mildly context-sensitive constructions, like cross-serial and unbounded dependencies (Briscoe, 1997, 2000). Furthermore, by using GCG-based ALs in experiments, we are able to test a wider attested range of word order configurations including VSO and OSV base word order ALs.

(a) Example of two valid sentences are concatenated with a conjunction to create a longer one.

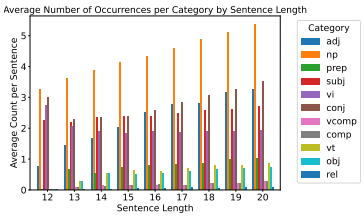


(b) Example of how a valid sentence is embedded into another valid sentence to create a longer one.

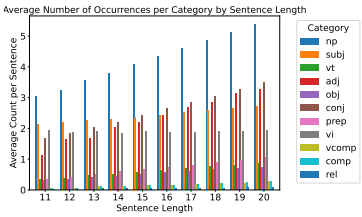
Figure 4: Examples of short templates (lengths 3-10) being combined to create longer templates (lengths 11-20). The first sentence is in blue, the second sentence is in green, and the conjunction is in orange.



(a) How many times each category appears in SHORT and MEDIUM per template on average.

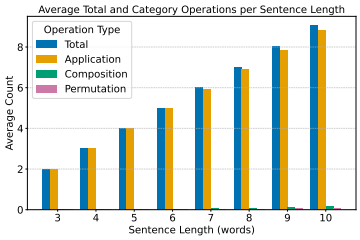


(b) How many times each category appears in templates used to create LONG set, per template on average.

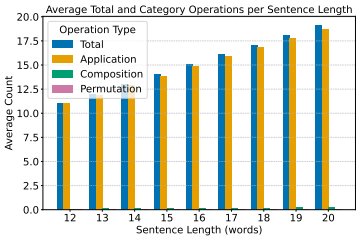


(c) How many times each category appears in LONG per template on average.

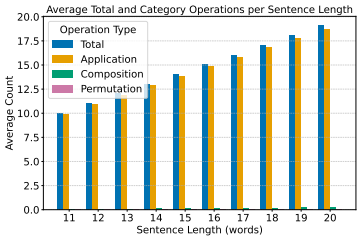
Figure 5: The number of occurrences of the different categories for all templates lengths 3-10 (a), the templates used to create the long dataset (b), and all the generated long templates (c).



(a) Template lengths 3-10 words (SHORT and MEDIUM).



(b) Sampled templates used to create the LONG dataset.



(c) All extended templates created of lengths 11-20.

Figure 6: The average number of combinatory operations in the GCG derivation for the templates of different lengths for the language configuration `0101101` by El-Naggar et al. (2025b).

B. Dataset Details

B.1. Dataset Generation

The ORIGINAL, SHORT, MEDIUM and LONG datasets are the same ones used by El-Naggar et al. (2025b).

- **SHORT Dataset:** The sentences in this dataset are created by generating all possible category combinations, which are referred to as templates, of length 3 to 8 words. The templates are then parsed and assigned to the ALs where they would produce a valid parse, i.e. result in an S node spanning the input. All lengths are represented equally in the dataset, and within each length, the different templates are also represented equally.

- **MEDIUM Dataset:** Like the SHORT dataset, this dataset is created by generating all possible templates of lengths 9-10 words. The templates are then parsed and assigned to ALs where they produce a valid parse. Similarly, all lengths are represented equally and within each length all templates are represented equally. The lexicon is then randomly sampled to create unique sentences.

- **LONG Dataset:** To generate the LONG dataset, the templates for the SHORT and MEDIUM datasets are combined in 3 ways, as shown in Figure 4:

  1. Appended end to end
  2. Concatenated with a conjunction (Fig. 4a),
  3. Embedded with a conjunction (Fig. 4b).

  The resulting longer templates are parsed to filter out ungrammatical ones. Because there are millions of valid templates of length 11-20, 20K templates are randomly sampled, and for each one, the lexicon is sampled. It is worth

Fairseq model	share-decoder-input-output-embed	True
	embed_dim	128
	ffn_embed_dim	512
	layers	2
	heads	2
	dropout	0.3
	attention_dropout	0.1
	#params.	462K
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	400
	warmup init learning rate	1e-7
Training	batch size	2,048 tokens
	tokens-per-sample	128 tokens
	sample-break-mode	none
	epochs	10

(a) Transformer.

Fairseq model	share-decoder-input-output-embed	True
	embed_dim	128
	hiden_size	512
	layers	2
	dropout	0.1
	#params.	3,547K
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	400
	warmup init learning rate	1e-7
Training	batch size	2,048 tokens
	tokens-per-sample	128 tokens
	sample-break-mode	none
	epochs	10

(b) LSTM.

Fairseq model	share-decoder-input-output-embed	True
	embed_dim	64
	hiden_size	64
	layers	2
	dropout	0.1
	#params.	49K
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	400
	warmup init learning rate	1e-7
Training	batch size	2,048 tokens
	tokens-per-sample	128 tokens
	sample-break-mode	none
	epochs	10

(c) RNN.

Table 3: Hyperparameters of LMs.

noting that for the English-like language configuration `0101101`, the vast majority of the valid extended templates will have been created by concatenation with a conjunction (b) or embedding with a conjunction (c). However, it may be possible for the end to end appending of templates to result in valid templates in other language configurations.

B.2. Dataset Statistics

We show the occurrences of the different categories in the templates in El-Naggar et al. (2025b) for lengths 3-10 in Figure 5a. The occurrences of the templates used to create the LONG dataset are shown in Figure 5b, and in Figure 5c for all generated templates of length 11-20. We can see in Figure 5 that the number of categories that appear in shorter sentences is smaller than in longer sentences, and that the number of category occurrences increases with sentence length. We show in Figure 6 the average number of different combinatory operations that are required to derive the templates of different lengths.

C. Model Hyperparameters

We use exactly the same model hyperparameters as El-Naggar et al. (2025b). These are summarised in Table 3.

Modeling semantic association in self-paced reading with language model embeddings

Sara Møller Østergaard*, **Kenneth Enevoldsen**†, **Afra Alishahi***, **and Bruno Nicenboim***
*Department of Computational Cognitive Science, Tilburg University
†Center for Humanities Computing, Aarhus University
s.m.ostergaard@tilburguniversity.edu

Abstract

Semantic association between a word and its context has been identified as an important component of reading comprehension, even when word predictability is accounted for. Recent research has highlighted the potential of language model (LM) embeddings to quantify semantic association. Yet, embedding-based semantic association have been operationalized in a myriad of ways. In this study, we use embeddings from LMs to estimate semantic association on a corpus of joint electroencephalography (EEG) and self-paced reading of natural, Dutch texts. Semantic association is calculated in ten different implementations that vary the embedding model and context lengths. The effects of semantic association across the different implementations on the N400 and self-paced reading times are examined using Bayesian hierarchical models and Bayes factor. The results show that the choice of embedding model can alter the estimated effect of semantic association on both the N400 and self-paced reading times. Furthermore, the results demonstrate a promising potential of sentence embeddings for capturing semantic association, as only implementations relying on sentence embeddings indicate reliable results of semantic association beyond word predictability on both neural and behavioral measures. Together, these findings highlight the importance of methodological choices in quantifying semantic association.

Keywords: semantic association, self-paced reading (SPR), electroencephalography (EEG), N400, sentence processing

1. Introduction

Humans process words in the context in which they are presented. How predictable a word is given its preceding context largely impacts the processing difficulty of the word (Kutas and Federmeier, 2011; Ehrlich and Rayner, 1981; Wong et al., 2024). For example, in the sentence pair, "By the end of the day, the hiker's feet were extremely cold and wet. It was the last time he would ever buy a cheap pair of *boot/jeans*.", the final word "boots" is highly predictable based on the preceding context and is therefore processed more easily than the alternative ending "jeans", which is comparatively unpredictable in this context (example from Federmeier and Kutas, 1999).

Predictability of a word, or its probability given a context, has been estimated using a range of probabilistic models, including probabilistic grammars (Hale, 2001) and, more recently, next token probabilities from language models (LMs) (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024). Additionally, word predictability has been estimated using the cloze task[1] (Luke and Christianson, 2018; Dambacher et al., 2006; Bulkes et al., 2020).

Word predictability has been able to explain important aspects of processing difficulty, however, it doesn't provide a full account. In addition to predictability, semantic association presents another factor that modulates reading comprehension (Kutas and Federmeier, 2011; Brouwer et al., 2012). Semantic association refers to the degree of semantic relatedness between a target word and the context in which it is presented. While this measure is related to the predictability of the word, it has distinct properties. Using the example context from above, "By the end of the day, the hiker's feet were extremely cold and wet. It was the last time he would ever buy a cheap pair of *sandals*.", the word "sandals" is unpredictable in the context, however, it is semantically associated with the context (which mentions feet). Federmeier and Kutas (1999) show that this distinction results in different processing of these target words.

Semantic illusion has been used to study the effects of semantic association beyond word predictability. Semantic illusions refer to a phenomenon where unpredictable (or incorrect) words are temporally unnoticed because the words are semantically associated with the context. The sentence "For breakfast the *eggs* would only eat toast and jam.", illustrates this effect, where the word "eat"

[1]The cloze task is a language comprehension task in which one or more words are removed from a text and must be filled in by the participant based on contextual cues.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 60–73
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

fails to elicit the expected neural response to an unpredictable word (Kuperberg et al., 2003). Studies on semantic illusion report that words semantically associated with their context are processed differently (as shown with electroencephalography; EEG) compared to words that lack such associations (Kuperberg et al., 2003; Nieuwland and Van Berkum, 2005; Stone and Rabovsky, 2025; Aurnhammer et al., 2023). Relatedly, Krieger et al. (2024) found that word predictability from LMs doesn't capture the complete role of contextual information in human sentence processing, particularly with respect to semantic association.

Processing difficulty is commonly indexed using behavioral measures such as reading times, as well as neural measures derived from EEG, including the N400 and the P600 event-related potential (ERP) components. Word predictability has been shown to have robust effects on reading times and the N400 (Kutas and Federmeier, 2011; Ehrlich and Rayner, 1981; Frank et al., 2015; Shain, 2024; Pimentel et al., 2023; Frank and Aumeistere, 2024; Federmeier and Kutas, 1999). In contrast, semantic association between target words and their context has been investigated primarily in ERP studies, with fewer studies examining its relationship to reading times.

ERP studies of semantic association have mostly focused on the N400 component, where semantic association decreases the negative amplitude of the component (Fischler et al., 1983; Kuperberg et al., 2003; Federmeier and Kutas, 1999; Xu et al., 2024; Broderick et al., 2018; Frank and Willems, 2017). However, studies have found that the effect of semantic association on the N400 disappears when there is a delay between the semantically related context and the critical word (Chow et al., 2018; Stone and Rabovsky, 2025). Furthermore, Salicchi and Hsu (2025) found that semantic association didn't explain variance in the N400 component when surprisal was accounted for, while it did in the P600 component, suggesting effects on later processing stages. Evidence of the effect on reading times is less explored. While some studies have found that stronger semantic association decreases reading times (Pynte et al., 2008; Mitchell et al., 2010), other studies indicate that semantic association has no effect on reading times when excluding the variance explained by word predictability (Traxler et al., 2000; Frank, 2017).

Studies of semantic association have mostly relied on stimuli consisting of handcrafted contexts and target words, where they are either semantically similar or not (Federmeier and Kutas, 1999; Fischler et al., 1983; Kuperberg et al., 2003; Stone and Rabovsky, 2025). However, recent studies have attempted to estimate the semantic association using embeddings from LMs (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024; Frank and Willems, 2017; Parviz et al., 2011). Thereby, enabling the quantification of semantic association as a continuous measure and facilitating analyses that can extend to naturalistic stimuli.

Embedding-based estimates of semantic association have been conceptualized in a myriad of ways. Firstly, studies deploy different embedding models for extracting the embeddings of the context and the critical word. Most studies use word embeddings, e.g., GloVe, word2vec or fastText (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024; Frank and Willems, 2017), however, these models vary in model architecture, embedding size, and training data. Secondly, the context embedding is defined in a variety of ways. Most commonly an average of the word embeddings are used, however, which words are included in the average varies: some studies use all the words in the context (Michaelov and Bergen, 2024; Michaelov et al., 2024; Xu et al., 2024; Broderick et al., 2018), others only content words (Mechtenberg et al., 2025; Frank and Willems, 2017) or a manually select subset of the words (Frank, 2017; Ettinger et al., 2016). Additionally, the length of the context varies. While most studies rely on sentence-level stimuli and use all the preceding words as the context, other studies relying on stimuli consisting of longer text have defined context windows. Frank (2017) defined the context in two separate ways: i) only the sentence preceding the critical word and ii) the four content words immediately preceding the critical word. Similarly, Mechtenberg et al. (2025) examined local and global effects of semantic association by defining context windows of one, two, five, and ten words preceding the critical word, excluding stop words. Finally, different functions for calculating the similarity between the embeddings of the critical word and the context have been employed: While the vast majority utilize the cosine similarity (Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024), Pearson's correlation has also been used (e.g., Broderick et al., 2018)

The present study investigated whether semantic association derived from LM embeddings captured aspects of language processing not accounted for by word predictability alone. To accommodate alternative formalizations of semantic association, we defined multiple implementations, varying the embedding model and the size of the contextual window used to compute semantic association. We evaluated these implementations using Bayesian model comparison (Bayes factor) and assess their effects on self-paced reading times and the N400

ERP component. The results of the study showed how the choice of embedding model and the conceptualization of the context can alter the conclusions across neural and behavioral signals.

2. Methods

2.1. Data

The study used data from the Tilburg corpus of Natural Dutch Texts (TiNT; Østergaard et al., 2025). The corpus consists of joined recordings of EEG and self-paced reading (SPR) from 71 participants (whereof 56 participants were included in the analysis of the current study). All participants read eight medium-length (approx. 600 words), natural, Dutch texts of different genres. Seven texts were read using a SPR paradigm, while a single text was read in a rapid serial visual presentation (RSVP) paradigm (the exact text changing from participant to participant). In this study, we only used data recorded during SPR.

Preprocessing of the EEG signal and extraction of ERPs were identical to that of Østergaard et al. (2025). Preprocessing included rereferencing of the electrodes, band-pass filtering, and artifact detection and exclusion. The N400 was defined as the mean amplitude of centroparietal electrodes in the time window 300-500ms after word onset.

2.2. Semantic association

Semantic association was defined as the similarity between the embedding of the context and the embedding of the critical word. Thus, three methodological decisions were required: (1) How to represent the embeddings of the context and the critical word, (2) what context length to use, and (3) which similarity function to apply. In this paper, we defined multiple implementations of semantic association by varying the first two factors, while we used the cosine similarity as the similarity metric across all implementations. Cosine similarity was used, as it is the standard similarity measure for distributional embedding models (Yamada et al., 2020; Reimers and Gurevych, 2019).

(1) Embeddings of the context and the word: Multiple approaches exist for deriving embeddings of text using LMs. Embeddings can be uncontextualized, such as, GloVe, word2vec, or fastText (Pennington et al., 2014; Mikolov et al., 2013; Bojanowski et al., 2017). Such models produce a single embedding for each word in isolation. Alternatively, embeddings can be contextualized. Contextualized embeddings can be derived from transformer models, including both encoders such as BERT (Devlin et al., 2019) and generative models such as GPT and LLama (Radford et al., 2018; Touvron et al., 2023) by retrieving embeddings from the last hidden state. However, embeddings derived directly from pre-trained models typically perform poorly, and thus it has become the norm to adapt contextualized transformer models for embedding tasks, such as semantic text similarity (Reimers and Gurevych, 2019; Gao et al., 2021; Li et al., 2025).

An initial exploration of implementations of semantic association using different embedding models was conducted with simple sentences where the differences in semantic association were handcrafted. The results of the exploration indicated that both contextualized and uncontextualized embedding models were able to differentiate words semantically associated with the context from unrelated words. The results from the models were similar within embedding type (i.e., contextualized or uncontextualized).[2] As such, we selected two candidate models: an uncontextualized word embedding model and a contextualized sentence embedding model.[3]

For the uncontextualized model, we used the word2vec model `wikipedia2vec_nlwiki_20180420_300d`[4] (Yamada et al., 2020). This model was chosen as the training procedure matched previous literature (Broderick et al., 2018; Ettinger et al., 2016; Frank, 2017; Frank and Willems, 2017) and its training data overlaps with the TiNT corpus (Østergaard et al., 2025). As the model only produces one embedding for each word, we used two methods for obtaining the embedding of the context: i) the average of the embeddings of all the words in the context (denoted *WE*), and ii) the average of all the content words in the context (denoted *CWE*). We used the *nl_core_news_sm* model from `spaCy` to extract the part of speech (POS) tags. Content words were identified as words with the POS tags noun, verb, adjective, or adverb.

For the sentence embedding we used `e5-large-trm-nl` (Banar et al., 2025) as it performed well on the Dutch embedding benchmark (MTEB(nld, v1); Banar et al., 2025; Enevoldsen et al., 2025). Sentence embeddings are trained to produce an aggregated embedding over multiple words, thus, it didn't require post-hoc averaging to obtain the embedding of the context. Implementations with sentence embeddings are denoted *SE*.

(2) Context length: Contexts of varying length were defined to examine local and global effects of semantic association. Four distinct context lengths

[2] Results of initial exploration can be found in appendices.

[3] For historical reasons, we call these sentence embeddings, as they initially were trained to embed sentences. However, they have since been expanded to embed entire documents.

[4] Model revisions can be found in appendices.

were used. First, a naive context consisting of all words preceding the critical word was used (*All*). Second, we defined a context consisting of all words in the preceding sentence, as well as in the sentence to which the critical word belonged (*Sentence(N=1)*). Finally, we defined a windowed context, where the context consisted of a fixed number of content words before the target. Here, we used windows of one and two (*Windowed(N=1)* and *Windowed(N=2)*). The windowed implementation was only defined with (content) word embeddings.

In addition to the contexts of different lengths, we defined a weighted average of the word embeddings. The weights followed an exponential forgetting curve, thus, assuming words appearing closer to the critical word were more important. This was implemented as in Equation 1:

$$\text{WE, Weighted} = \sum_{i=1}^{N} 2^{\frac{-i}{4}} \cdot \text{similarity}(w_c, w_i) \quad (1)$$

Here, w_c is the word embedding of the critical word and w_i the word embedding of the word i words away from the critical word. The equation sums over all words preceding the critical word. The denominator (4) determines the half-life of the decay and was chosen such that words at a distance of ten or more from the critical word receive minimal weights. As for the other implementations, the similarity was calculated with the cosine similarity. The weighted average was implemented with word embeddings of all words and word embeddings of content words only. All the implementations of semantic association are summarized in Table 1.

Correlations between semantic association for content words in the corpus extracted using the different implementations are shown in Figure 1. Implementations based on the same type of embedding (i.e., word or sentence embeddings) are strongly correlated, suggesting that they index similar sources of variance. In contrast, correlations across embedding types are substantially weaker.

2.3. Regression models and model comparison

Bayesian hierarchical models were fitted to examine the effect of the different implementations of semantic association on the two dependent variables: self-paced reading times and the N400. The models were fitted in Stan (version 2.32.2; Stan Development Team, 2023) using the `brms` package (version 2.22.0; Bürkner, 2017) in R (R Core Team, 2024). All predictors were z-score standardized. Words with reading times lower than 100 ms or greater than 3000 ms were excluded from analysis. Only content words (i.e., nouns, verbs, adjectives, and adverbs) were included in the analysis. As embeddings extracted from the word2vec model only exist for a finite number of words, the data loss slightly differed across different implementations of semantic association.[5] The models were fitted on complete cases across all implementations.

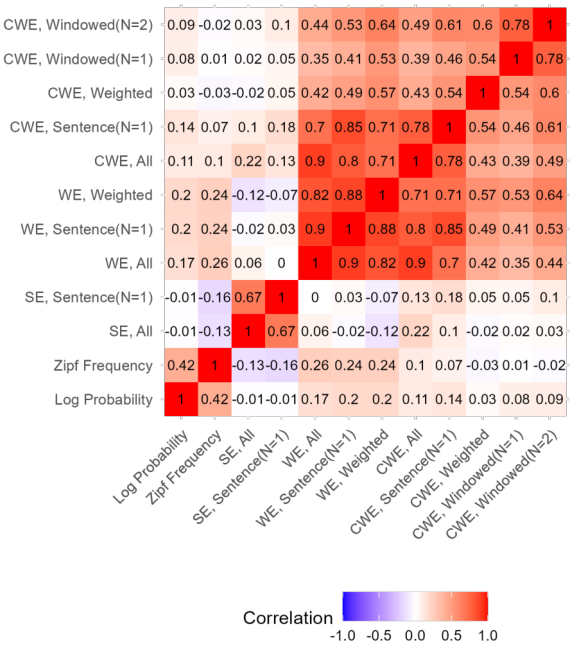


Figure 1: Pearson's correlation coefficients between implementations of semantic association, log-probability of words, and Zipf word frequency for all content words in the corpus.

The models were run with two predictors, log-probability lp, estimated by the average word probability from four GPT models (see Østergaard et al., 2025), and semantic association sem. One regression model for each of the two dependent variables and each of the 10 implementations of semantic association was fitted, resulting in 20 separate models. Uncorrelated group-level intercepts and slopes for both lp and sem were estimated for each participant, document, and word. For the reading time (RT) model, a log-normal likelihood was used, while for the N400 model, a Gaussian likelihood was used (See Equations 2, 3, and 4).

$$\text{RT} \sim LogNormal(\mu, \sigma) \quad (2)$$

$$\text{N400} \sim Normal(\mu, \sigma) \quad (3)$$

$$\begin{aligned} \mu =& \alpha + u_{participant,0} + u_{document,0} + \\ & u_{word,0} + (\beta_1 + u_{participant,1} + \\ & u_{document,1} + u_{word,1}) \cdot lp + (\beta_2 + \\ & u_{participant,2} + u_{document,2} + u_{word,2}) \cdot sem \end{aligned} \quad (4)$$

[5] Data loss across all implementations is reported in appendices.

Name	Embedding model	Context	Words
SE, All	Sentence Embeddings	All preceding words	All
SE, Sentence(N=1)	Sentence Embeddings	One sentence before the target sentence	All
WE, All	Word Embeddings	All preceding words	All
WE, Sentence(N=1)	Word Embeddings	One sentence before the target sentence	All
WE, Weighted	Word Embeddings	All preceding words (weighted)	All
CWE, All	Word Embeddings	All preceding words	Content words
CWE, Sentence(N=1)	Word Embeddings	One sentence before the target sentence	Content words
CWE, Weighted	Word Embeddings	All preceding words (weighted)	Content words
CWE, Windowed(N=2)	Word Embeddings	One content word preceding the target	Content words
CWE, Windowed(N=2)	Word Embeddings	Two content word preceding the target	Content words

Table 1: All implementations of semantic association used in the current study. The implementations will be referred to by the name in the name column.

Different priors were used for the reading times and the N400 model, as the scales of the dependent variables were different, i.e., reading times in ms and ERP components in μV. For all models, regularizing priors were used to ensure stable and plausible estimates (Nicenboim et al., 2025). The priors for the reading times model were as follows:

$$\begin{aligned}\alpha &\sim Normal(5.5, 1)\\ \beta &\sim Normal(0, .1)\\ u &\sim Normal(0, sd)\\ sd &\sim Normal_+(0, .5)\\ \sigma &\sim Normal_+(0, .5)\end{aligned}$$

The priors for the models of the ERP components were:

$$\begin{aligned}\alpha &\sim Normal(0, 20)\\ \beta &\sim Normal(0, 10)\\ u &\sim Normal(0, sd)\\ sd &\sim Normal_+(0, 10)\\ \sigma &\sim Normal_+(0, 10)\end{aligned}$$

To assess the influence of the various implementations of semantic association on reading comprehension (i.e., reading times and the N400), we used Bayes factors. Bayes factor provides a framework for Bayesian hypothesis testing by quantifying evidence in favor of a model (M_0) given another (M_1). This is calculated as the ratio between the marginal likelihoods of the two models, which in turn responds to two hypotheses. (see Equation 5).

$$BF_{01} = \frac{p(y|M_0)}{p(y|M_1)} \tag{5}$$

As such, a Bayes factor of one indicates no evidence for either model, a Bayes factor of 10 would be strong evidence for M_0, and a Bayes factor of $1/10$ indicates strong evidence for M_1. We used the Savage-Dickey density ratio method to calculate Bayes factor, as it provides a convenient method for computing Bayes factor for nested models with a point null hypothesis (Dickey and Lientz, 1970). We specifically tested the null hypothesis that there's no effect of semantic association in the models when log-probability is included. The Savage-Dickey ratio was calculated separately for each model as in Equation 6.

$$BF_{01} = \frac{p(\beta_2 = 0|y)}{p(\beta_2 = 0)} \tag{6}$$

Here, y denotes the observed data and β_2 is the coefficient for semantic association.

As Bayes factor is sensitive to the choice of prior, we conducted a sensitivity analysis by varying the width of the prior for β_2 while keeping the priors for all other parameters fixed. For the reading times models, we used additional priors of $\beta_2 \sim Normal(0, .05)$ and $\beta_2 \sim Normal(0, .5)$, and for the N400 models, $\beta_2 \sim Normal(0, 1)$ and $\beta_2 \sim Normal(0, 2)$. For a more elaborate explanation of Bayes factor and the Savage-Dickey ratio, see Nicenboim and Vasishth (2016).

Most models were fitted using four chains with 2,000 iterations, where half the iterations were warm-up samples. However, six models required 3,000 iterations to ensure stable posterior sampling. The models reported in this paper had no divergent transitions, $\hat{Rs} \leq 1.03$, and the number of bulk and tail effective samples was at least 119 and an average of 1,477.

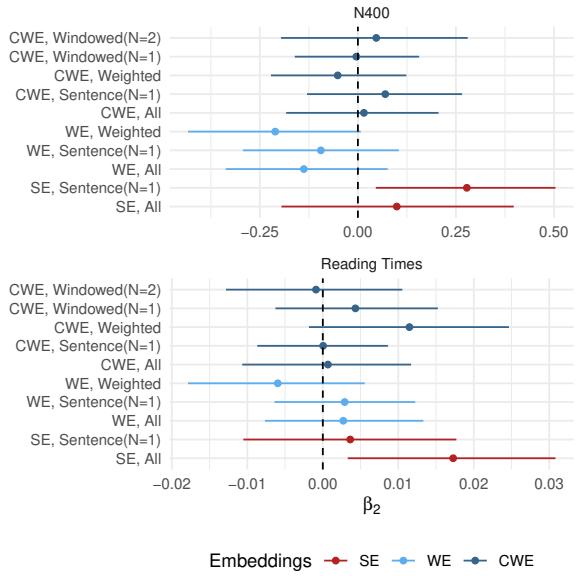


Figure 2: Regression coefficients and 95% credible intervals for semantic association β_2 as estimated by the different implementations. Every point represents β_2 from a separate regression model. The models with N400 as the dependent variable are measured in μV, while the reading times models are measured in ms (thus, the scales of the x-axis differ across the two).

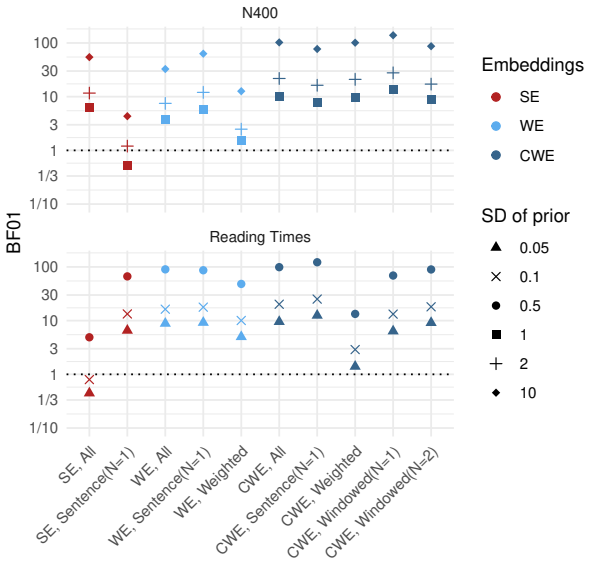


Figure 3: Bayes factor. $BF01 > 1$ indicates more evidence for the null hypothesis (i.e., no effect of semantic association) and $BF01 < 1$ indicates more evidence for the alternative hypothesis (i.e., an effect of semantic association). Each Bayes factor was calculated for separately fitted models with different standard deviations (SD) for the prior of β_2.

3. Results

Figure 2 displays the coefficients for semantic association (β_2 in Equation 4) estimated by the 20 regression models using the values from the different implementations of semantic association to predict the N400 and self-paced reading times. Bayes factor for β_2 across the regression models reported in Figure 3. The results from the Bayes factor showed anecdotal evidence ($BF01 \in \{1, 1/3\}$) for an effect of semantic association in only two models (*SE, Sentence(N=1)* for the N400 and *SE, All* for reading times). Across the rest of the models for both dependent variables, there was the most evidence for the null hypothesis, i.e., no effect of semantic association.

Embedding models: The results of the regression models indicate that the choice of embedding model when calculating semantic association impacts the estimated effects on neural and behavioral measures. This pattern was particularly pronounced for the models of the N400. The estimated effect of semantic association on the N400 when using sentence embeddings (*SE*) was positive, meaning that less semantically associated words elicited a more negative N400 amplitude, consistent with previous literature (Fischler et al., 1983; Kuperberg et al., 2003; Federmeier and Kutas, 1999; Xu et al., 2024; Broderick et al., 2018; Frank and Willems, 2017). In contrast, when semantic association was calculated using word embeddings (*WE*), the direction of the effect reversed, i.e., a negative estimate. When semantic association was calculated using the same word embeddings but only embeddings of the content words in the context (*CWE*), the estimated effect of semantic association was close to zero. For reading times, only the model of semantic association from the *SE, All* implementation indicated an effect. This model estimated a positive effect of semantic association on reading times; thus, reading times increased when words were more semantically associated to the context. The estimated coefficients for semantic association for the rest of the models were smaller and generally close to zero.

Context length: The results show that the length of the context matters only for the semantic association defined with sentence embeddings. The implementations of semantic association using word embeddings (both *WE* and *CWE*) showed similar effects on both the N400 and reading times across all contexts (*All*, *Sentence*, *Weighted*, and *Windowed*). For the implementations relying on sentence embeddings, the effect of context appeared to play a more substantial role. On the N400, the effect of semantic association was largest for the regression model with *SE, Sentence(N=1)*, while the largest effect of semantic association on reading

times was estimated by the model with *SE, All*.

4. Discussion

In this study, we employed both uncontextualized word embeddings and contextualized sentence embeddings to estimate semantic association. These two embedding types appear to capture distinct patterns in the text. This is both apparent from the correlations between the different implementations of semantic association (Figure 1) but, more importantly, for the estimated effects of semantic association on self-paced reading times and the N400 too (Figure 2 and Figure 3). The results of the regression models show that the type of embeddings used influences the estimated effect of semantic association. This finding is most pronounced for the N400, where the estimated effect reverses direction depending on whether semantic association is computed using sentence or word embeddings. A positive effect of semantic association on the N400 is found when using sentence embeddings, while the opposite (i.e., a negative effect) is estimated with word embedding-based semantic association. In contrast, previous literature using similar uncontextualized word embeddings to calculate semantic association reports a positive effect of semantic association on the N400 (Broderick et al., 2018; Frank and Willems, 2017; Xu et al., 2024). It is important to note that the Bayes factor indicated no evidence for the negative effects estimated for word embedding-based semantic association, however, anecdotal evidence for one of the models with a positive effect for semantic association from sentence embeddings.

What could be possible explanations for the observed difference in semantic association when computed with different types of embedding models? One important distinction between sentence embeddings and word embeddings when using them to create context representations lies in how information was aggregated. Although sentence embeddings output an aggregation of embeddings too, the model has been trained to produce a semantically coherent representation in which more informative words receive greater weight. In this study, the implementations of semantic association using word embeddings from word2vec relied on a naive approach, where an unweighted average was utilized, inspired by previous approaches (Michaelov and Bergen, 2024; Michaelov et al., 2024; Xu et al., 2024; Broderick et al., 2018; Mechtenberg et al., 2025; Frank and Willems, 2017; Frank, 2017; Ettinger et al., 2016). As such, important information could be lost in the context representations derived from the word embedding implementations — especially for the implementations of longer contexts (i.e., *WE, All* and *WE, Sentence(N=1)*). Studies finding positive effects of word embedding-based semantic association on the N400 have generally relied on shorter contexts (either because sentence-level stimuli were used or because they defined short context lengths), thus minimizing the information loss when averaging over embeddings. This speculation is supported by our initial exploration, where both sentence and word embeddings produced effects in the same direction using sentence pairs from Federmeier and Kutas (1999).[6] While the weighted implementations of semantic association (*WE, Weighted* and *CWE, Weighted*) were cognitively motivated implementations, discounting the influence of word embeddings on the overall average based an exponential forgetting curve, this approach didn't seem promising given the results of the current study. The weights were solely based on distances to the critical word, thus words were not weighted based on their semantic relevance.

To our knowledge, no previous works have used sentence embeddings for studying semantic association in sentence processing. The present findings suggest that this approach provides a promising method for estimating semantic relations between contexts and target words. Implementations based on sentence embeddings showed the most reliable effects on both the N400 and self-paced reading times, as reflected by the size of the regression coefficients and Bayes factors. Post-hoc qualitative analyses further indicated that sentence embeddings are more sensitive to the general themes of the texts compared to averaged word embeddings. Figure 4 illustrates a difference between sentence embeddings and word embeddings for calculating semantic association in two examples of sentence pairs from the corpus. In the first pair, the word "dragon" appears in both sentences. Only the *SE, All* implementation captures the association between "draak" (English: "dragon") and the story "Mijn Heer Zak met Rijst"[7] in the first sentence, while *SE, Sentence (N=1)* detects the association when the word reappears in the second sentence. In contrast, none of the word embedding implementations indicate a strong association for "dragon" in either instance. In sentence pair B from the text "Nomadisch pastoralisme", a similar pattern is observed with the two words "Nomadisme" (English: "Nomadism") and "nomadische" (English: "nomadic"). While these examples were selected for illustrative purposes, they suggest that sentence embeddings capture a thematically coherent representation of semantic association in natural texts.

Naturally, this interpretation depends on the operationalization of semantic association. In the

[6] Results of initial exploration can be found in appendices

[7] A fairy tale about a Dragon King.

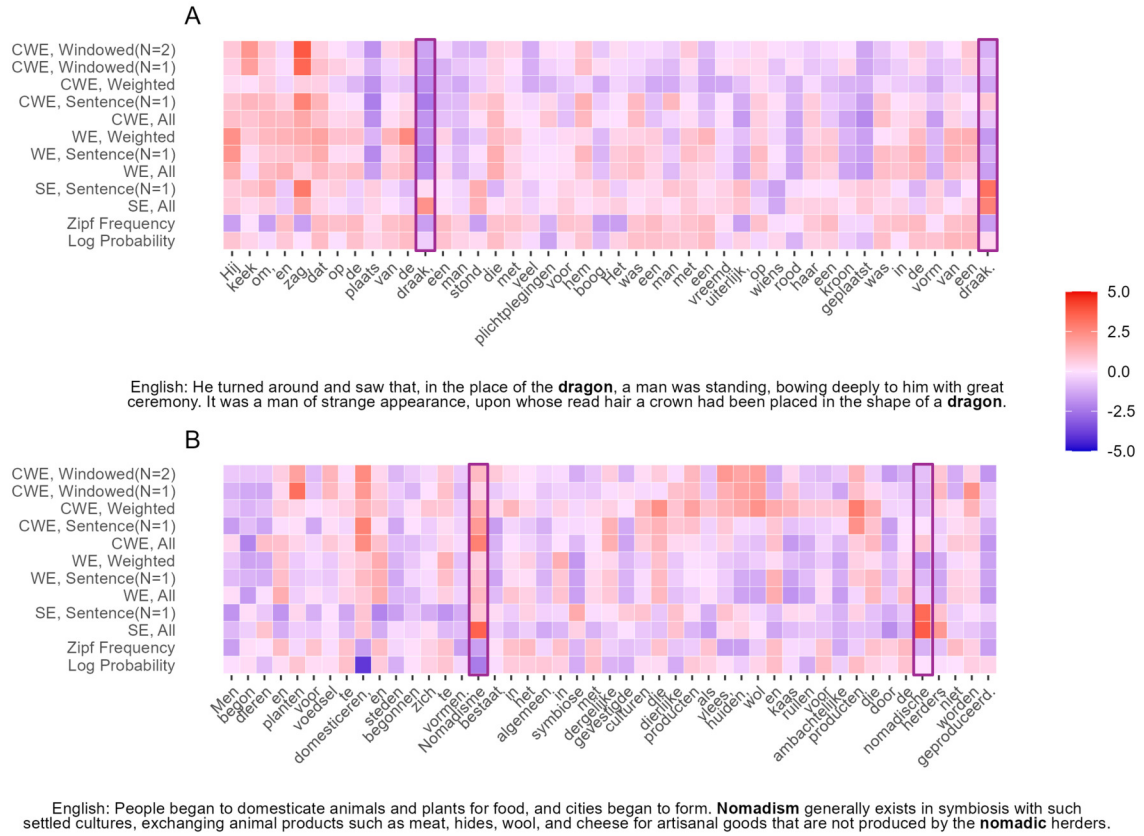


Figure 4: Semantic association (as estimated by different implementations), log-probability, and Zipf frequency of words in two sentence pairs from two different documents in the corpus. Highlighted are the words “draak” (English: “dragon”) in sentences A and “Nomadisme” (English: “Nomadism”) and “nomadische” (English: “nomadic”) in sentences B. All variables (i.e., log probability, Zipf frequency, and semantic associations) are z-score standardized.

present study, semantic association was defined as the similarity between embeddings of the context and the critical word, following prior work (Broderick et al., 2018; Ettinger et al., 2016; Xu et al., 2024; Michaelov et al., 2024; Frank, 2017; Michaelov and Bergen, 2024; Frank and Willems, 2017). This operationalization assumes that the embeddings encode multiple aspects of meaning, including both shared features (e.g., *nurse* and *mechanic* as occupations) and thematic relations (e.g., *nurse* and *hospital* as related). Consequently, embedding similarity captures similarities in the features of the word as well as their relatedness. Following this definition, repeated words will inflate semantic association, as the similarity between two identical embeddings is one (i.e., maximum semantic association). However, this property applies for all the implementations of semantic association considered in the current work, thus, the effect of repetition can’t account for the differences observed in example A in Figure 4.

The results of the current study are exploratory, and further work is required to identify under which conditions specific implementations of LM embedding-based estimates of semantic association differ. The analysis was based exclusively on texts from a single corpus (TiNT; Østergaard et al., 2025), which consists of medium-length, Dutch texts. Not only does this corpus stand in contrast to previously used stimuli by the length of the texts (as touched upon above), but also in the language. As Dutch has been less extensively studied than English, the quality of the embedding models may differ, potentially affecting their performance. As such, semantic association as estimated by the different implementations in this study should be validated on other corpora to determine whether it is possible to replicate previously reported effects of semantic association.

The most prominent finding of this study lies in the importance of the embedding model for estimating semantic association. Only one word embedding model and one sentence embedding model were included in the analysis, as initial explorations indicated minimal differences between models within each embedding type. However, in light of the results of the current study, further ex-

ploration of different embedding models would be interesting.

5. Conclusion

This study examined the effects of LM embedding-based semantic association on the self-paced reading of medium-length, Dutch texts. The findings demonstrate that the conclusions critically depend on how semantic association is implemented, particularly with respect to the embedding model. While uncontextualized word embeddings (e.g., word2vec) have previously been used to examine semantic association in sentence processing and showed effects on the N400 (Xu et al., 2024; Broderick et al., 2018; Frank and Willems, 2017), we observed no effects on either the N400 or self-paced reading times. In contrast, semantic association estimated with sentence embeddings was found to be predictive of processing difficulty. These results suggest sentence embeddings to be a promising approach for examining the effects of semantic association in natural reading.

6. Code availability

Code for reproducing the analysis is publicly available from GitHub at `https://github.com/saraoe/semantic_association`.

7. Ethics Statement

The study utilized human data from the Tilburg corpus of Natural Dutch Texts (TiNT) collected by Østergaard et al. (2025). The data has received an ethics approval and is licensed under a CC-BY-NC-SA license.

8. Bibliographical References

Christoph Aurnhammer, Francesca Delogu, Harm Brouwer, and Matthew W. Crocker. 2023. The P600 as a continuous index of integration effort. *Psychophysiology*, 60(9):e14302.

Nikolay Banar, Ehsan Lotfi, Jens Van Nooten, Cristina Arhiliuc, Marija Kliocaite, and Walter Daelemans. 2025. Mteb-nl and e5-nl: Embedding benchmark and models for dutch.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. 2018. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, 28(5):803–809.e3.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2012. Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, 1446:127–143.

Nyssa Z. Bulkes, Kiel Christianson, and Darren Tanner. 2020. Semantic constraint, reading control, and the granularity of form-based expectations during semantic processing: Evidence from ERPs. *Neuropsychologia*, 137:107294.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.

Wing-Yee Chow, Ellen Lau, Suiping Wang, and Colin Phillips. 2018. Wait a second! delayed impact of argument roles on on-line verb prediction. *Language, Cognition and Neuroscience*, 33(7):803–828.

Michael Dambacher, Reinhold Kliegl, Markus Hofmann, and Arthur M. Jacobs. 2006. Frequency and predictability effects on event-related potentials during reading. *Brain Research*, 1084(1):89–103.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

James M. Dickey and B. P. Lientz. 1970. The Weighted Likelihood Ratio, Sharp Hypotheses about Chances, the Order of a Markov Chain. *The Annals of Mathematical Statistics*, 41(1):214–226.

Susan F. Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua,

Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Allyson Ettinger, Naomi Feldman, Philip Resnik, and Colin Phillips. 2016. Modeling N400 amplitude using vector space models of word representation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 38(0).

Kara D. Federmeier and Marta Kutas. 1999. A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language*, 41(4):469–495.

I. Fischler, P. A. Bloom, D. G. Childers, S. E. Roucos, and N. W. Perry. 1983. Brain potentials related to stages of sentence verification. *Psychophysiology*, 20(4):400–409.

Stefan L. Frank. 2017. Word Embedding Distance Does not Predict Word Reading Time. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 39(0).

Stefan L. Frank and Anna Aumeistere. 2024. An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, 58(2):641–657.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Stefan L. Frank and Roel M. Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W. Crocker. 2024. On the limits of LLM surprisal as functional Explanation of ERPs. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Gina R. Kuperberg, Tatiana Sitnikova, David Caplan, and Phillip J. Holcomb. 2003. Electrophysiological distinctions in processing conceptual relationships within simple sentences. *Brain Research. Cognitive Brain Research*, 17(1):117–129.

Marta Kutas and Kara D. Federmeier. 2011. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, 62(1):621–647.

Litao Li, Leo Song, Steven Ding, Benjamin C. M. Fung, and Philippe Charland. 2025. Transforming Generic Coder LLMs to Effective Binary Code Embedding Models for Similarity Detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Hannah Mechtenberg, James Reilly, Emily B Myers, and Jonathan E Peelle. 2025. Measuring brain sensitivity to semantic distance in spoken narrative comprehension.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana

Coulson. 2024. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

James A. Michaelov and Benjamin K. Bergen. 2024. On the Mathematical Relationship Between Contextual Probability and N400 Amplitude. *Open Mind*, 8:859–897.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206, Uppsala, Sweden. Association for Computational Linguistics.

Bruno Nicenboim, Daniel J. Schad, and Shravan Vasishth. 2025. The influence of priors: sensitivity analysis. In *Introduction to Bayesian Data Analysis for Cognitive Science*, 1st edn edition, chapter 3.4, page 634. Chapman & Hall.

Bruno Nicenboim and Shravan Vasishth. 2016. Statistical methods for linguistic research: Foundational Ideas—Part II. *Language and Linguistics Compass*, 10(11):591–613.

Mante S. Nieuwland and Jos J. A. Van Berkum. 2005. Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. *Cognitive Brain Research*, 24(3):691–701.

OpenAI. 2025. GPT-5.2 via OpenAI API.

Mehdi Parviz, Mark Johnson, Blake Johnson, and Jon Brock. 2011. Using Language Models and Latent Semantic Analysis to Characterise the N400m Neural Response. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 38–46, Canberra, Australia.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger Levy, and Ryan Cotterell. 2023. On the Effect of Anticipation on Reading Times.

Joel Pynte, Boris New, and Alan Kennedy. 2008. On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research*, 48(21):2172–2183.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. *OpenAI Technical Report*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Lavinia Salicchi and Yu-Yin Hsu. 2025. Different Reading Processing Stages or Different Brain Areas? A Computational Cognitive Investigation on N400, P600, and PNP. In *Computational Psycholinguistics Meeting 2025*. Conference presentation abstract.

Cory Shain. 2024. Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, 8:177–201.

Stan Development Team. 2023. Stan Reference Manual.

Kate Stone and Milena Rabovsky. 2025. The Role of Syntactic and Semantic Cues in Preventing Temporary Illusions of Plausibility. *Journal of Cognitive Neuroscience*, 37(9):1535–1561.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Matthew J. Traxler, Donald J. Foss, Rachel E. Seely, Barbara Kaup, and Robin K. Morris. 2000. Priming in Sentence Processing: Intralexical Spreading Activation, Schemas, and Situation Models. *Journal of Psycholinguistic Research*, 29(6):581–595.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Roslyn Wong, Erik D. Reichle, and Aaron Veldre. 2024. Prediction in reading: A review of predictability effects, their theoretical implications, and beyond. *Psychonomic Bulletin & Review*.

Haoyin Xu, Masaki Nakanishi, and Seana Coulson. 2024. Revisiting Joke Comprehension with Surprisal and Contextual Similarity: Implication from N400 and P600 Components. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Sara Møller Østergaard, Lenneke Lichtenberg, Laura Boon, and Bruno Nicenboim. 2025. A Corpus of Joint EEG and Self-Paced Reading of Natural Dutch Texts.

A. Appendices

A.1. Hugging Face References and Revision

Hugging Face Reference	Revision
Word2vec/wikipedia2vec_nlwiki_20180420_300d (Yamada et al., 2020)	f7c83ecdf955a4f482a12517ca52a1f4b81e43cf
clips/e5-large-trm-nl (Banar et al., 2025)	683333f86ed9eb3699b5567f0fdabeb958d412b0
Word2vec/wikipedia2vec_enwiki_20180420_100d (Yamada et al., 2020)	7e4d6d224b95a5c351e2a47232701c4403ffbc16
fse/word2vec-google-news-300	528f381952a0b7d777bb4a611c4a43f588d48994
sentence-transformers/all-MiniLM-L6-v2	c9745ed1d9f207416be6d2e6f8de32d1f16199bf
intfloat/multilingual-e5-large (Wang et al., 2024)	0dc5580a448e4284468b8909bae50fa925907bc5
spacy/nl_core_news_sm (Honnibal et al., 2020)	b9d28fe480eeacf9809fbd5ead5ef1ff27d9394e

Table 2: Overview of Hugging Face models used for the analysis.

A.2. Data loss

The data loss for the different implementations of semantic association, i.e., the number of words in the corpus for which semantic association could not be calculated. The use of sentence embeddings resulted in a lower data loss compared to the word embedding implementations, as word embeddings extracted from a word2vec model only exist for a finite number of words. The data loss is largest for the *CWE, Windowed(N=1)*.

Implementation	Data loss	
	Overall	Content words
SE, All	0.17%	0.08%
SE, Sentence(N=1)	0.17%	0.08%
WE, All	1.78%	2.06%
WE, Sentence(N=1)	1.78%	2.06%
WE, Weighted	1.78%	2.06%
CWE, All	2.09%	2.33%
CWE, Sentence(N=1)	2.17%	2.38%
CWE, Weighted	2.09%	2.33%
CWE, Windowed(N=1)	3.91%	4.25%
CWE, Windowed(N=2)	2.21%	2.39%

Table 3: Data loss caused by the different implementations of semantic association. The table shows the overall data loss across all words in the corpus, and the data loss for the current analysis that only considers content words.

A.3. Validation of semantic association

To validate the implementations of semantic association, we used data from Federmeier and Kutas (1999). As seen in Ettinger et al. (2016), we wanted to validate that the model identified expected targets as more semantically similar to the context compared to the within-category and between-category target words. In addition to the original stimuli, we added an unrelated target word (identical to an expected target in another context). Moreover, using OpenAI's GPT-5.2 (OpenAI, 2025), we generated longer contexts of approx. 100 words to validate the models on contexts longer than the two-sentence context provided in the original data. The results of the validation are shown in Figure 5. All the implementations identify the correct ordering of semantic association between the target words and the contexts. The sentence embedding models, `all-MiniLM-L6-v2` and `intfloat/multilingual-e5-large`, generally exhibit less variance overlap between target words (and most notably between the expected and unexpected target) as compared to the word embedding models, `enwiki_20180420_100d` and `word2vec-google-news-300`.

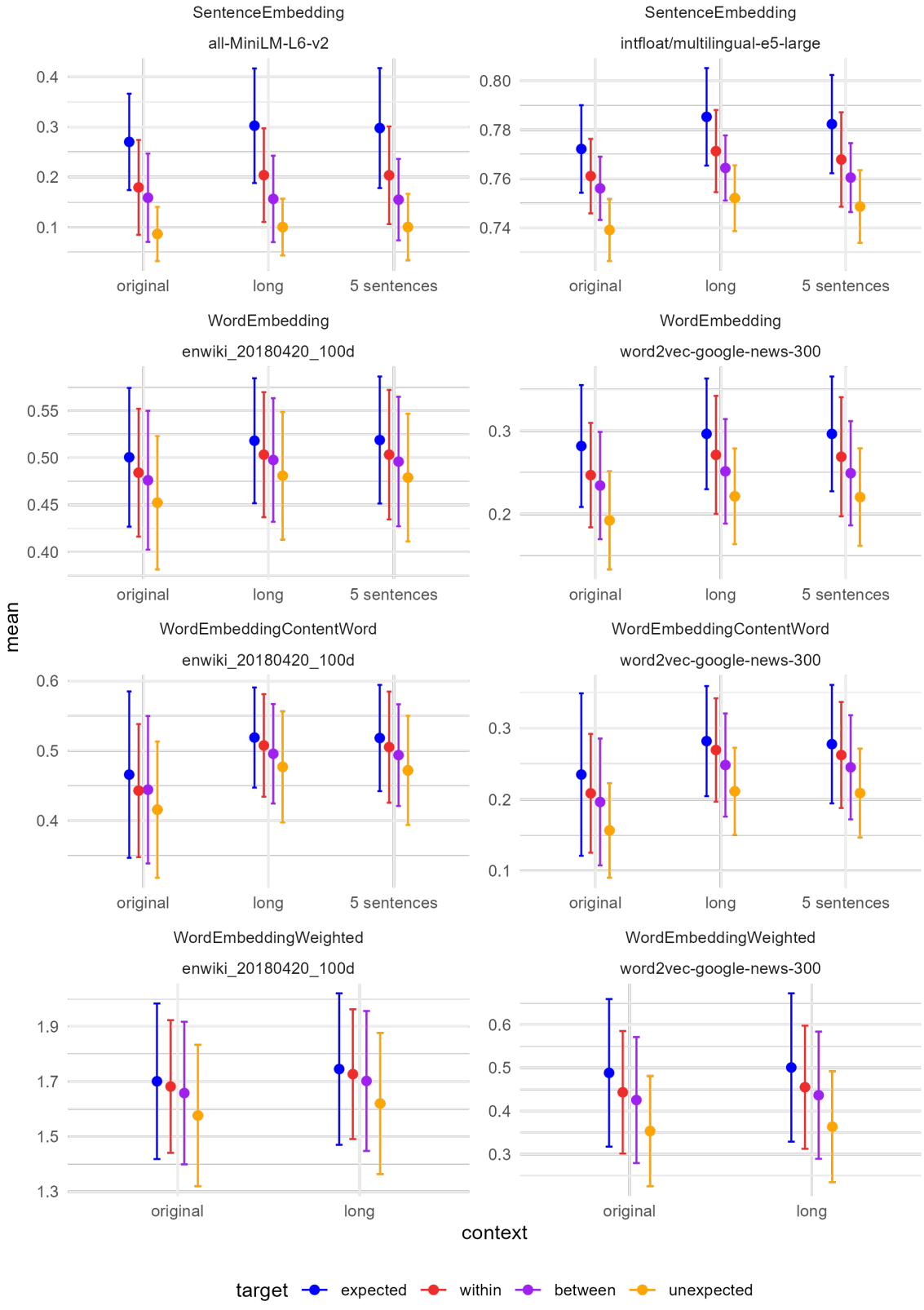


Figure 5: Average semantic association between context and target words (expected, within, between, and unexpected) from Federmeier and Kutas (1999) given different implementations of semantic association. The plot is divided by the embedding model and the implementation of semantic association. The x-axis indicates the context the target was associated with, where "long" means the original and the generated longer context, "original" means the original context, "5 sentences" means the two sentences in the original context and three more from the longer context. The error bars indicate the standard deviation. Note that the y-axes are different across plots.

Character-aware Transformers Learn an Irregular Morphological Pattern Yet None Generalize Like Humans

Akhilesh Kakolu Ramarao[1], Kevin Tang[1,3], Dinah Baer-Henney[2]
[1]Department of English Language and Linguistics, Heinrich Heine University Düsseldorf
[2]Institut für Germanistik, Philologische Fakultät, Ruhr-Universität Bochum
[3]Department of Linguistics, College of Liberal Arts and Sciences, University of Florida
{akhilesh.kakolu.ramarao, kevin.tang}@uni-duesseldorf.de, dinah.baer-henney@rub.de

Abstract

Whether neural networks can serve as cognitive models of morphological learning remains an open question. Recent work has shown that encoder-decoder models can acquire irregular patterns, but evidence that they generalize these patterns like humans is mixed. We investigate this using the Spanish *L-shaped morphome*, where only the first-person singular indicative (e.g., *pongo* 'I put') shares its stem with all subjunctive forms (e.g., *ponga, pongas*) despite lacking apparent phonological, semantic, or syntactic motivation. We compare five encoder-decoder transformers varying along two dimensions: sequential vs. position-invariant positional encoding, and atomic vs. decomposed tag representations. Positional encoding proves decisive: position-invariant models recover the correct L-shaped paradigm clustering even when L-shaped verbs are scarce in training, whereas sequential positional encoding models only partially capture the pattern. Yet none of the models productively generalize this pattern to novel forms. Position-invariant models generalize the L-shaped stem across subjunctive cells but fail to extend it to the first-person singular indicative, producing a mood-based generalization rather than the L-shaped morphomic pattern. Humans do the opposite, generalizing preferentially to the first-person singular indicative over subjunctive forms. None of the models reproduce the human pattern, highlighting the gap between statistical pattern reproduction and morphological abstraction.

Keywords: Cognitive Modeling, Morphological Inflection, Morphomes, Transformers, Positional Encoding

1. Introduction

Inflectional paradigms are not merely lists of forms but constitute a central organizing principle of morphological knowledge (Stump, 2001). This claim, that the paradigm rather than the morpheme is the primary object of morphological analysis (Matthews, 1972), has profound implications for how we understand the mental lexicon. If paradigms are cognitively primitive, then speakers do not assemble inflected words by concatenating stored morphemes; instead, they exploit implicational relations among whole forms, inferring unattested forms from partial paradigmatic knowledge (Ackerman et al., 2009). Yet how far does this organization extend to irregular verbs? Can paradigmatic patterns that lack phonological, semantic, or syntactic motivation be learned and productively generalized? In this paper, we use computational modeling to probe these questions, training neural network models that differ in architectural choices on inflectional paradigms, comparing how these choices affect pattern acquisition, and evaluating their generalization behavior against human speakers tested on the same items.[1]

We focus on *morphomic* patterns (Spencer and Aronoff, 1994), which provide the strongest test of paradigmatic learning because they lack any apparent phonological, semantic, or syntactic motivation. They challenge theories that require all morphological structure to be grounded in meaning or function, representing instead autonomous organization within the morphological system itself. Such patterns are well attested across Romance languages (Maiden, 2011, 2018, 2021), making them ideal candidates for investigating whether computational models can acquire purely morphological generalizations.

In Spanish, the *L-shaped morphome* is one such pattern: the first-person singular indicative (*pongo*) shares its stem with every subjunctive form (*ponga, pongas, . . .*), while all other indicative forms use the regular stem (*pones, pone, . . .*). Table 1 illustrates this pattern for the verb *poner* 'to put'. The defining property is the grouping of 1SG.IND patterns with all subjunctive forms rather than with its own mood category.

Whether morphomic patterns constitute a psychological reality for speakers remains contested. Nevins et al. (2015) argue against the cognitive status of the L-shaped morphome with evidence from wug-test experiments on Spanish, Italian, and Portuguese, but subsequent work challenges this view. Cappellaro et al. (2024) provide experimental evidence that Italian speakers exhibit sensitivity to L-shaped patterns in processing and production

[1]Code and Data: https://github.com/hhuslamlab/feature_informed_transformer_cmcl2026

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 74–85
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

	Indicative	Subjunctive
1SG	**pong**-o	**pong**-a
2SG	pon-es	**pong**-as
3SG	pon-e	**pong**-a
1PL	pon-emos	**pong**-amos
2PL	pon-éis	**pong**-áis
3PL	pon-en	**pong**-an

Table 1: Present tense paradigm of *poner* 'to put'. Hyphens mark stem-suffix boundaries. Bold marks the L-shaped stem.

tasks. Similarly, O'Neill (2025) offer corpus-based evidence that gaps in Portuguese defective verbs cluster specifically in L-pattern cells rather than being randomly distributed across the paradigm. Computational modeling offers a complementary approach to this debate. Kakolu Ramarao et al. (2025) train neural sequence-to-sequence models on Spanish verbal inflection, providing form-morphological feature pairs as input and generating inflected forms as output. They demonstrate that these models can reproduce L-shaped stem alternations, with frequency of L-shaped verbs playing a crucial role. However, their study employed a single architecture. We extend this line of work by comparing five model architectures to ask whether design choices affect the acquisition of morphomic patterns, and then evaluating their generalization behavior against human experimental data.

We focus on two architectural dimensions that could plausibly affect the acquisition of morphomic patterns: sequential vs. position-invariant positional encoding, and atomic vs. decomposed morphosyntactic tag representations. Character-level encoder-decoder transformers have become the dominant approach for morphological inflection (Wu et al., 2021), but the Unimorph schema (Sylak-Glassman, 2016), which provides feature specifications for inflectional morphology, treats morphosyntactic features as unordered bundles: `V;IND;PRS;1;SG` and `V;PRS;IND;SG;1` specify the same cell. This suggests that a tag should carry the same meaning regardless of its position in the sequence. We therefore compare architectures that assign tags sequential positions against ones that give all tags a fixed positional encoding, making their representation position-invariant. On the representation side, treating tags as atomic tokens requires the model to learn feature structure implicitly, whereas decomposed representations such as one-hot vectors or linguistically-informed feature geometries (Harley and Ritter, 2002) encode this structure directly. To test whether these choices affect morphomic pattern acquisition, we implement five encoder-decoder transformer variants that differ along these two dimensions. Because L-shaped verbs constitute a small minority of the Spanish lexicon, we additionally vary the proportion of L-shaped verbs in training across three frequency conditions. This allows us to assess whether architectural inductive biases compensate for the natural scarcity of L-shaped verbs or only matter when such data is abundant, and whether increased exposure to L-shaped patterns during training improves generalization to novel forms. We address the following research questions:

- **RQ1** How do positional encoding and morphosyntactic tag representation affect the acquisition of L-shaped morphomic patterns?
- **RQ2** Do models acquire the L-shaped paradigm as a structural pattern?
- **RQ3** Do models productively generalize the L-shaped pattern to nonce verbs, and how does their behavior compare to that of human speakers?

2. Background

The SIGMORPHON shared task series has driven rapid progress in neural morphological inflection since 2016 (Cotterell et al., 2016), with successive iterations expanding to typologically diverse languages (Cotterell et al., 2017, 2018; Vylomova et al., 2020). Character-level encoder-decoder transformers have emerged as the dominant approach (Wu et al., 2021), and design choices such as separate encoders for characters and morphosyntactic tags have produced better results (Anastasopoulos and Neubig, 2019; Peters and Martins, 2019). Yet despite high reported accuracies, productive generalization remains elusive: when models are evaluated on unseen lemmas rather than forms from familiar paradigms, performance drops sharply (Goldman et al., 2022; Kodner et al., 2023), and predicting inflections for unseen feature combinations proves particularly challenging (Kodner and Khalifa, 2022). This gap between memorization and productive generalization suggests that current transformer-based architectures may lack the inductive biases needed to acquire the abstract representations that morphological knowledge demands.

Whether neural networks can serve as cognitive models of morphological learning has been debated since connectionist models of the English past tense (Rumelhart et al., 1986; Kirov and Cotterell, 2018). But so far, encoder-decoder models correlate poorly with human wug-test data (Corkery et al., 2019; McCurdy et al., 2020), and transformers struggle to generalize in a wug-test setting (Liu and Hulden, 2022; Weissweiler et al., 2023). This gap between model and human behavior has motivated benchmarks, such as the BabyLM Challenge,

which trains models on developmentally plausible datasets to probe language acquisition (Warstadt et al., 2023; Hu et al., 2024).

The transformer architecture has no built-in sense of order, i.e., without positional encodings, it cannot, for example, distinguish *pongo* from *gopon* (Dufter et al., 2022). The original transformer introduced sinusoidal positional encodings (Vaswani et al., 2017). Since then, numerous alternatives have been proposed, including relative positional encoding (Shaw et al., 2018), learned absolute embeddings (Devlin et al., 2019), and rotary positional encoding (Su et al., 2024).

However, the necessity of positional encoding varies across tasks: Wang and Chen (2020) show that BERT without positional encoding functions essentially as a bag-of-words model in understanding tasks (classification, QA), while Haviv et al. (2022) find that causal language models remain competitive even without explicit positional encoding in generation tasks, suggesting some tasks are less sensitive to position.

How morphosyntactic tags are represented has received comparatively less attention. Since Kann and Schütze (2016), most systems concatenate morphological features as individual tokens to the character sequence and encode both through a single encoder. Other approaches separate the two inputs: Aharoni and Goldberg (2017) concatenate learned feature embeddings into a single vector that conditions the decoder at every step, Peters and Martins (2019) learn separate attention distributions for lemma characters and tags, and Anastasopoulos and Neubig (2019) use separate encoders for lemma characters and tags. However, decomposing feature bundles into structured vectors, such as one-hot encodings or binary features from feature geometry (Harley and Ritter, 2002), remains unexplored in the literature.

3. Methodology

3.1. Data

3.1.1. Real verbs

We use Spanish verbal paradigms in IPA transcription from Kakolu Ramarao et al. (2025). Each verb has a 12-cell present-tense paradigm crossing three persons (1, 2, 3), two numbers (singular, plural), and two moods (indicative, subjunctive). Verbs are classified as *L-shaped* if the 1SG.IND stem matches all subjunctive stems but differs from the remaining indicative stems, and as *NL-shaped* otherwise. The full dataset contains 299 L-shaped and 4,859 NL-shaped verbs.

To investigate the effect of input frequency on learning, we create three training conditions by varying the proportion of L-shaped verbs: 10%L-90%NL, 50%L-50%NL, and 90%L-10%NL. The 10%L condition most closely approximates the natural distribution of L-shaped verbs in the Spanish lexicon, where they constitute roughly 7% of verb types. For each condition, we sample 332 lemmas from the full pool and split them into training (70%, 232 lemmas), development (10%, 34 lemmas), and test (20%, 66 lemmas) sets. Following Goldman et al. (2022) and Kodner et al. (2023), we ensure that there is no lemma overlap between the training, development, and test sets, so that models must generalize to unseen lemmas. The L-shaped to NL-shaped ratio is maintained across all three splits.

Each training example consists of two source form-tag pairs and one target tag (see Section 3.2). From a 12-cell paradigm, we generate all 660 unique (unordered source pair, target) triples per lemma and subsample 25% per training seed. We repeat the procedure with three lemma splits and four data subsamples, yielding 12 independently trained models per condition per architecture.

3.1.2. Nonce verbs

To compare model behavior with human generalization, we use behavioral data from the wug test experiment by Nevins et al. (2015) on 107 native Spanish speakers. In a paradigm completion task, participants were presented two forms of 15 nonce verbs exhibiting novel stem-final consonant alternations between fricatives and stops (e.g., *bus-* vs. *but-*, *mif-* vs. *mip-*) and asked to produce a third form.

Participant responses were converted from Spanish orthography to IPA transcription, and only responses matching one of the expected paradigm cells were retained. For each nonce verb, we evaluate model predictions on the same three paradigm cells tested in the human study (1SG.IND, 2SG.SBJV, and 3SG.SBJV), using the source pairs in both orderings: two source pairs for 1SG.IND and one each for the two subjunctive cells, yielding 8 test items per verb (8 x 15 verbs = 120 test items per model run).

3.2. Task

To mirror the above wug test, where participants are given two forms and must produce a third, we frame the task as two-source morphological re-inflection (Kann et al., 2017): given two source form-tag pairs from a verb's paradigm and a target feature bundle, the model must produce the corresponding target form. For example, given (*pongo*, `V;IND;PRS;1;SG`) and (*ponga*, `V;SBJV;PRS;1;SG`) with target features `V;SBJV;PRS;2;SG`, the model should produce *pongas*. Following Wu et al. (2021), we model this

as character-level sequence-to-sequence transduction. During training, models observe all source-target combinations for each lemma; at test time, they must generalize to entirely unseen lemmas.

3.3. Model architectures

All five models share a common encoder-decoder transformer backbone with 4 layers, 4 attention heads, embedding dimension $d = 256$, and feedforward dimension 1024. The models differ in how the encoder handles morphosyntactic tags along two dimensions: (i) whether tags receive sequential positional encoding or are assigned a fixed position, and (ii) how tags are represented. We organize the models into two classes based on positional encoding: the first class assigns tags sequential positions as it is done with characters, while the second treats tags as position-invariant. The three architectures in the position-invariant class further differ in tag representation.

Training. We use the same hyperparameter settings as Kakolu Ramarao et al. (2025). All models are optimized using Adam (Kingma and Ba, 2015) with a learning rate of 0.001 and a linear warmup schedule over 4,000 steps. We use a batch size of 400, dropout of 0.3, and label smoothing of 0.1. Training proceeds for a maximum of 10,000 update steps. At inference, we use beam search with a beam size of 5. For each of the three frequency conditions (10%L, 50%L, 90%L), we train 12 models per architecture.

3.3.1. Class 1: Uniform positional encoding

Vanilla. The baseline concatenates form characters and feature tags into a single flat sequence. Characters are tokenized individually (e.g., *pongo* $\rightarrow$ `p o n g o`), while each morphosyntactic feature bundle is a single vocabulary entry (e.g., `V;IND;PRS;1;SG`). Both characters and tags are embedded through a shared learned embedding matrix. Both token types receive sequential sinusoidal positional encoding based on their position in the concatenated sequence. The model has no mechanism to distinguish tag tokens from character tokens (Kakolu Ramarao et al., 2025).

Character-separated. Identical to VANILLA, except that feature tags are decomposed into individual characters separated by whitespace (e.g., `V;IND;PRS;1;SG` $\rightarrow$ `V IND PRS 1 SG`). All tokens share the same embedding table and receive uniform sequential positional encoding.

3.3.2. Class 2: Position-invariant tags

The UniMorph schema (Sylak-Glassman, 2016) treats morphosyntactic features as unordered bundles, suggesting that sequential positional encoding may be inappropriate for morphosyntactic specifications. The following three architectures assign tags a fixed positional index of 0 while characters retain sequential positional encoding. They differ in how they represent tag content.

Feature-invariant. Following Wu et al. (2021), tags are embedded through the same shared matrix as in VANILLA but are assigned a fixed positional index of 0, while characters retain sequential positional encoding.

Feature-onehot. Tags are assigned a fixed positional index of 0 and represented not as atomic tokens but as structured feature vectors. Each tag is parsed into its component features and encoded as a one-hot vector over feature categories: mood (IND, SBJV), person (1, 2, 3), and number (SG, PL), yielding a sparse 7-dimensional binary vector. A learned linear projection maps this vector to the same d-dimensional embedding space used by character tokens. At each position, the model selects between the character embedding and the feature projection based on token type.

Feature-geometric. Like FEATURE-ONEHOT, tags are assigned a fixed positional index of 0 and represented as structured vectors. However, instead of one-hot category vectors, tags are encoded as binary feature vectors following the feature-geometric analysis of Harley and Ritter (2002): person is decomposed into [$\pm$participant, $\pm$author] (1st = [+,+], 2nd = [+,$-$], 3rd =[$-$,$-$]), number into [$\pm$plural], and mood into [$\pm$indicative], yielding a 4-dimensional binary vector. This vector is projected to the embedding space through a learned linear transformation.

3.4. Evaluation

3.4.1. Sequence accuracy

We evaluate overall sequence accuracy as the percentage of predictions that exactly match the target form, and additionally report it separately for L-shaped and NL-shaped verbs.

3.4.2. Stem accuracy by verb type

The distinction between L-shaped and NL-shaped verbs is defined by the stem-final consonant alternations: L-shaped verbs use one stem in the 1SG.IND and all subjunctive cells and a different stem elsewhere. Whether a model has learned the alternation pattern is therefore most directly

assessed at the stem level. We extract stems by stripping the conjugation suffixes (for *-ar*, *-er*, and *-ir* verb classes) from both predicted and target forms. A prediction is stem-correct if the extracted stem matches the target stem. We report stem accuracy separately for L-shaped and NL-shaped verbs.

3.4.3. Paradigm shape analysis

Beyond aggregate accuracy, we ask whether models treat the 12 paradigm cells in a way that reflects the L-shaped pattern, i.e., whether they achieve *structural generalization* (Hupkes et al., 2023). If a model has learned the L-shape, its stem predictions should systematically differ between L-cells (1SG.IND + all subjunctive) and the remaining indicative cells. We first report raw mean stem accuracy per paradigm cell across all test lemmas to examine whether models produce different stems across cells.

To quantify whether this per-cell pattern reflects the L-shape, we apply a scoring transformation to stem accuracies aggregated across all test lemmas, regardless of verb type: L-cell accuracies are kept as is, while the remaining indicative cell accuracies are inverted ($1 - accuracy$), since for L-shaped verbs an incorrect stem in a non-L indicative cell corresponds to the model producing the L-shaped stem where the regular stem is expected. After this transformation, a high score indicates that the model produced the L-shaped stem for that cell. If L-cells score high while non-L cells score low, this suggests L-shaped behavior. We then apply k-means clustering ($k = 2$) to group the 12 cells based on these transformed scores. A model that has acquired the L-shaped paradigm should produce two clusters: one containing 1SG.IND and all subjunctive cells, and another containing the five remaining indicative cells.

3.4.4. Nonce verb evaluation

To compare model generalization with human behavior, we evaluate all models on the nonce verb stimuli from the wug test experiment of Nevins et al. (2015). For each of the 15 nonce verbs, the model is tested on the same three paradigm cells as in the human study (1SG.IND, 2SG.SBJV, 3SG.SBJV). We report stem accuracy and per-cell paradigm shape for models alongside the human behavioral data.

4. Results

4.1. Real verbs

4.1.1. Sequence accuracy

Figure 1 compares sequence-level accuracy across the five architectures and three frequency conditions. The three position-invariant models (Class 2: FEATURE-INVARIANT, FEATURE-ONEHOT, FEATURE-GEOMETRIC) consistently outperform the two sequential-PE models (Class 1: VANILLA, CHARACTER-SEPARATED). In the 10%L condition, where L-shaped examples are scarce, the gap is largest: Class 2 models achieve mean accuracies of 68-70%, whereas VANILLA reaches 59% and CHARACTER-SEPARATED only 53%. As L-shaped training data increases, overall accuracy improves for all architectures and the gap narrows: in the 90%L condition, all five models converge to the 81-83% range. Class 1 models also exhibit higher variance in the 10%L and 50%L conditions, whereas Class 2 models maintain low variability even when L-shaped examples are scarce. This indicates that position-invariant tag encoding provides a stronger and more stable inductive bias when L-shaped examples are scarce, but with sufficient data even simpler architectures achieve comparable performance.

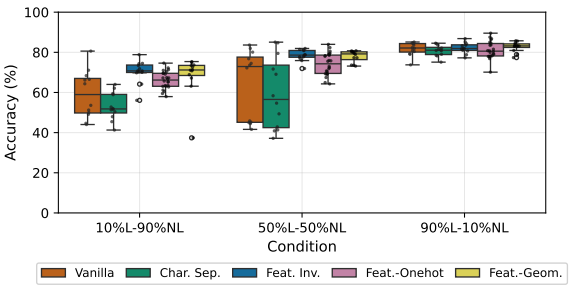


Figure 1: Overall sequence accuracy for all five architectures across the three frequency conditions.

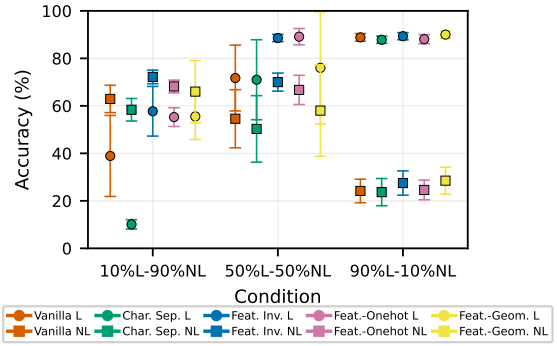


Figure 2: L-shape and NL-shape accuracy across five model architectures. Circles denote L-shape accuracy, squares denote NL-shape accuracy.

Furthermore, we examine how accurately models predict L-shaped forms compared to NL-shaped forms across the three test conditions in Figure 2. As L-shaped training frequency increases, L-shaped accuracy rises, reaching 88-90% in the

90%L condition across all models. Conversely, NL-shaped accuracy drops as NL-shaped verbs become scarcer, falling to 24-29% in the 90%L condition. This symmetric pattern indicates that all models, regardless of architecture, are more strongly driven by training frequency than by the regularity of the pattern: forms from the minority verb class are consistently predicted poorly, even when that class is the regular one.

In the critical 10%L condition, NL-shaped accuracy exceeds L-shaped accuracy for all models, reflecting the dominance of NL-shaped verbs in training. The gap is most extreme for CHARACTER-SEPARATED (L: 10%, NL: 58%) and VANILLA (L: 37%, NL: 62%), whereas Class 2 models maintain more balanced performance (FEATURE-INVARIANT: L: 58%, NL: 72%; FEATURE-ONEHOT: L: 55%, NL: 68%; FEATURE-GEOMETRIC: L: 56%, NL: 67%).

4.1.2. Stem accuracy by verb type

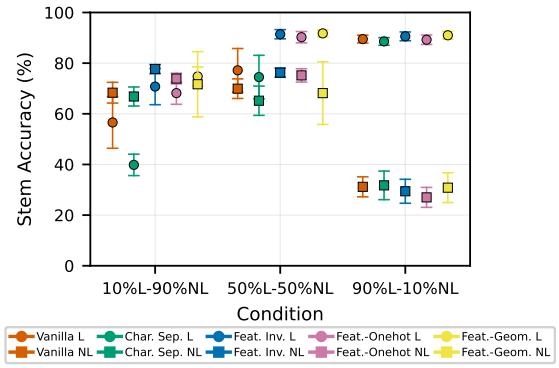


Figure 3: Stem accuracy separated by verb type. Circles denote L-shaped stem accuracy; squares denote NL-shaped stem accuracy.

Figure 3 evaluates stem accuracy, isolating the stem-final consonant alternation that defines the L-shaped pattern from suffix errors. Stem accuracy is consistently higher than sequence accuracy.

All models reach 89-92% L-shaped stem accuracy in the 90%L condition. NL-shaped stem accuracy decreases with decreasing NL training frequency, falling to 27-31% in the 90%L condition, mirroring the frequency dependence observed at the sequence level.

In the critical 10%L condition, Class 2 models show a clear advantage in L-shaped stem accuracy: FEATURE-GEOMETRIC (75%), FEATURE-INVARIANT (71%), and FEATURE-ONEHOT (68%) all outperform VANILLA (57%) and CHARACTER-SEPARATED (40%). This indicates that position-invariant tag encoding helps models learn the correct stem alternation even from limited L-shaped exposure.

4.1.3. Paradigm shape analysis

We apply the paradigm shape analysis described in Section 3.4.3. Figures 4-8 show raw (untransformed) per-cell mean stem accuracy ($\pm$ SD) across all test lemmas for the three conditions. Black-dashed borders indicate which cells belong to the L-shape for reference. Class 1 models show a clear accuracy contrast between L-cells and the remaining indicative cells even at 10%L, whereas Class 2 models show more uniform accuracy across cells.

To quantify whether this pattern reflects the L-shape, we apply the scoring transformation and k-means clustering described in Section 3.4.3. We report results across all three frequency conditions because the critical question is not merely *whether* models can recover the L-shaped clustering, but *how much L-shaped exposure they require to do so*. A perfect L-shape would yield high transformed scores for L-cells (1SG.IND + all subjunctive) and low scores for the remaining indicative cells, producing two distinct clusters. In the 50%L and 90%L conditions, all five architectures recover the correct L-shaped clustering. Table 2 shows the cluster assignments in the critical 10%L condition, where L-shaped verbs are rare in training.

Cell	Exp.	Van	C-Sep	F-Inv	F-1H	F-Geo
1SG.IND	L	L	L	L	L	L
2SG.IND	NL	NL	NL	NL	NL	NL
3SG.IND	NL	NL	NL	NL	NL	NL
1PL.IND	NL	NL	NL	NL	NL	NL
2PL.IND	NL	NL	NL	NL	NL	NL
3PL.IND	NL	NL	NL	NL	NL	NL
1SG.SBJV	L	L	L	L	L	L
2SG.SBJV	L	L	L	L	L	L
3SG.SBJV	L	L	L	L	L	L
1PL.SBJV	L	NL*	L	L	L	L
2PL.SBJV	L	NL*	L	L	L	L
3PL.SBJV	L	L	L	L	L	L

Table 2: K-means cluster assignments in the 10%L condition. Column abbreviations: Van = VANILLA, C-Sep = CHARACTER-SEPARATED, F-Inv = FEATURE-INVARIANT, F-1H = FEATURE-ONEHOT, F-Geo = FEATURE-GEOMETRIC, Exp. = expected cluster under a perfect L-shape. L = L-shape cluster, NL = non-L cluster. *Misclassified cells.

All three Class 2 models and CHARACTER-SEPARATED correctly cluster 1SG.IND with all six subjunctive cells. VANILLA achieves only a partial L-shape: 1PL.SBJV and 2PL.SBJV are misclassified with the non-L indicative cells. Although CHARACTER-SEPARATED achieves the correct grouping, its L-shape scores are lower than those of the Class 2 models, indicating weaker paradigm differentiation.

4.2. Nonce verb evaluation

4.2.1. Stem accuracies: models vs. humans

Figure 9 compares relaxed stem accuracy on the 15 nonce verbs from Nevins et al. (2015) across

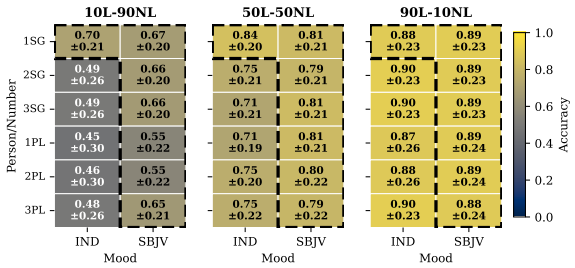


Figure 4: VANILLA: per-cell stem accuracy across the three frequency conditions.

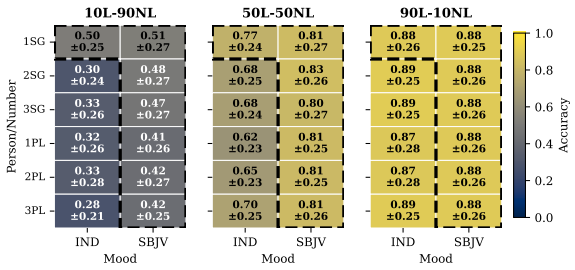


Figure 5: CHARACTER-SEPARATED: per-cell stem accuracy across the three frequency conditions.

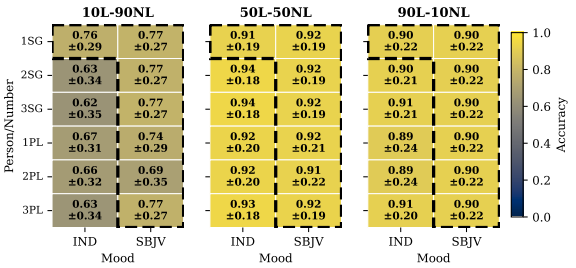


Figure 6: FEATURE-INVARIANT: per-cell stem accuracy across the three frequency conditions.

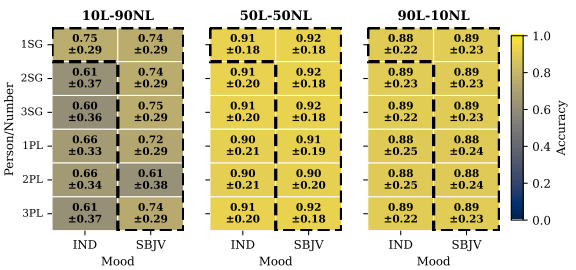


Figure 7: FEATURE-ONEHOT: per-cell stem accuracy across the three frequency conditions.

all five architectures and human participants. All models outperform the human mean of 33%: in the 10%L condition, CHARACTER-SEPARATED and VANILLA reach 49% and 48% respectively, while the three Class 2 models cluster around 43-44%.

Unlike the real-verb results, increasing L-shaped training frequency does not improve nonce-verb stem accuracy. All architectures decline from 10%L to 90%L, with Class 2 models dropping more sharply (from 43-44% to 32-37%) than Class 1 models (from 48-49% to 37-42%). This pattern, consistent with Kakolu Ramarao et al. (2025), suggests that models trained on higher proportions of L-shaped verbs generalize less effectively to novel stems.

4.2.2. Paradigm shape: models vs. humans

Figures 10 and 11 show per-cell nonce-verb stem accuracy for models and humans, respectively. Class 1 models, Class 2 models, and humans each exhibit distinct paradigm shapes. Human participants (Figure 11) show higher accuracy on 1SG.IND (37%) than on subjunctive cells (27-29%), suggesting that they preferentially generalize the stem alternation to 1SG.IND, the most frequent and salient L-shaped cell, rather than to subjunctive forms.

Class 2 models show the opposite pattern: subjunctive stem accuracy is high (79–88% at 10%L) while 1SG.IND accuracy is near zero (0-4%). These models have learned to produce the L-shaped stem in subjunctive cells but fail to extend it to 1SG.IND for novel items, yielding a partial L-shape that captures the mood distinction but misses the defining feature of the morphome, the inclusion of 1SG.IND. Class 1 models (VANILLA, CHARACTER-SEPARATED) show near-uniform accuracy across all cells (47-50% at 10%L), with no preference for indicative or subjunctive. These models do not exhibit any paradigm shape on nonce verbs.

5. Discussion

Positional encoding and tag representation (RQ1). Of the two architectural dimensions we manipulated, positional encoding proves decisive. When L-shaped verbs are scarce in training, position-invariant models recover higher L-shaped accuracy than sequential-PE models across all metrics. How tags are represented, whether as single indivisible vocabulary entries (e.g., `V;IND;PRS;1;SG`), one-hot feature vectors, or linguistically motivated feature geometries, matters far less: the three position-invariant architectures perform comparably despite substantial differences in how they encode morphosyntactic information. As L-shaped training frequency increases, all architectures converge, suggesting that position-invariant encoding acts as an inductive bias that compensates for data sparsity. Across all architectures, the minority verb class in training is consistently predicted poorly, confirming the frequency dependence shown by Kakolu Ramarao et al. (2025). However, position-invariant models maintain higher L-shaped accuracy than sequential-PE models un-

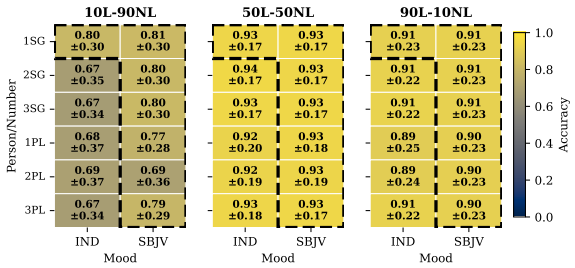


Figure 8: FEATURE-GEOMETRIC: per-cell stem accuracy across the three frequency conditions.

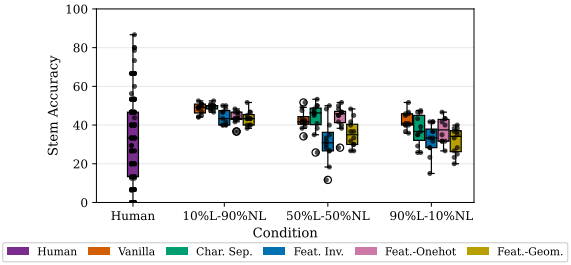


Figure 9: Stem accuracy on nonce verbs from Nevins et al. (2015). Human results (left) are shown alongside model results across the three frequency conditions.

der the same low-frequency conditions, suggesting that the reliance on input frequency that Kakolu Ramarao et al. (2025) observed is a property of the vanilla architecture rather than of neural inflection models in general.

Structural pattern acquisition (RQ2). Position-invariant models recover the correct L-shaped clustering, grouping 1SG.IND with all subjunctive cells, even from limited exposure. Among sequential-PE models, VANILLA only partially captures the pattern, misclassifying 1PL.SBJV and 2PL.SBJV, while CHARACTER-SEPARATED achieves the correct grouping but with lower L-shape scores, indicating weaker paradigm differentiation. The per-cell stem accuracy heatmaps corroborate this: Class 2 models achieve higher and more uniform accuracy across all cells. Class 1 models show lower overall accuracy with a wider gap between L-cells and the remaining indicative cells, reflecting greater difficulty with the pattern. This suggests that position-invariant models acquire the L-shape as a structural paradigmatic pattern.

Productive generalization and comparison with humans (RQ3). Yet none of the models extend the L-shaped pattern productively to novel stems. Position-invariant models generalize the L-shaped stem to subjunctive cells but fail almost entirely on 1SG.IND, while humans do the opposite (Nevins

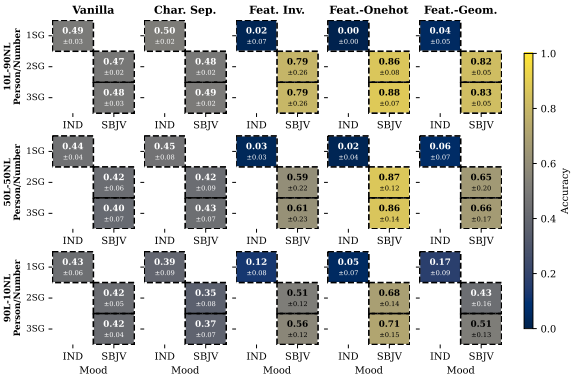


Figure 10: Per-cell mean stem accuracy on nonce verbs for all five architectures across the three frequency conditions.

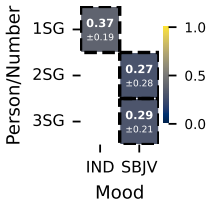


Figure 11: Per-cell mean stem accuracy on nonce verbs for human participants from Nevins et al. (2015).

et al., 2015). This likely reflects a difference between the two test sets: the real verb test set shares its stem alternation patterns with training, whereas the nonce verbs introduce consonant alternations absent from Spanish. Models succeed when alternation patterns are familiar but fail when both stems and alternations are novel, consistent with prior evidence that neural inflection models struggle to match human wug-test behavior (Corkery et al., 2019; Weissweiler et al., 2023). Neither model class reproduces the human paradigm shape, indicating that human morphological generalization differs qualitatively from that of the transformer architectures tested here.

6. Conclusion

We compared five encoder-decoder transformers on their ability to acquire and generalize the Spanish L-shaped morphome. The critical architectural distinction is between the two classes of models, not between individual architectures within each class: models within Class 1 (sequential PE) and within Class 2 (position-invariant PE) behave similarly regardless of how morphosyntactic tags are represented. Position-invariant positional encoding enables learning the L-shaped paradigm even when L-shaped verbs are scarce. However, no ar-

chitecture productively generalizes this pattern to novel stems, and all models diverge qualitatively from human generalization behavior. These findings contribute to the debate on the cognitive status of morphomic patterns: the L-shape is statistically learnable from distributional evidence, but the generalization mechanisms of character-aware transformers differ from those of human speakers. Future work should probe models' internal representations to determine whether the L-shape is encoded but not realized in production, extend the analysis to Italian and Portuguese where experimental data exists (Nevins et al., 2015; Cappellaro et al., 2024), and explore training regimes that better support generalization of minority patterns.

7. Limitations

Whether our findings extend to other Romance languages or to the complete paradigm remains untested. The nonce verb stimuli from Nevins et al. (2015) introduce fricative-stop alternations absent from Spanish, making it difficult to separate failures of paradigm generalization from failures on novel phonological alternations. All five models share the same hyperparameters from Kakolu Ramarao et al. (2025), which may not be optimal for each architecture individually. Additionally, our models are trained on paradigm-balanced data where each cell appears equally often, whereas humans encounter indicative forms, particularly first-person singular, far more frequently than subjunctive forms in natural language. Finally, we compare only transformer variants. Non-neural models such as linear discriminative learning have shown competitive performance on cognitively plausible morphological generalization tasks (Jeong et al., 2023) and may behave differently on the acquisition and generalization of morphomic patterns.

CRediT authorship contribution statement

We follow the CRediT taxonomy[2]. Conceptualization: [AKR]; Data curation: [AKR]; Formal Analysis: [AKR]; Investigation: [AKR, KT, DB]; Methodology: [AKR, KT, DB]; Supervision: [KT, DB]; Visualization: [AKR]; and Writing – original draft: [AKR] and Writing – review & editing: [KT, DB, AKR].

Acknowledgements

We gratefully acknowledge the support of the central HPC system "HILBERT" at Heinrich Heine University Düsseldorf. We thank Martin Krämer and Niklas Wiskandt for their valuable feedback. We also thank the reviewers for their constructive feedback, which improved our work.

[2] https://credit.niso.org/

8. Bibliographical References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Chiara Cappellaro, Nina Dumrukcic, Isabella Fritz, Francesca Franzon, and Martin Maiden. 2024. The cognitive reality of morphomes. evidence from Italian. *Morphology*, 34(1):33–71.

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? encoder-decoder neural networks as cognitive models of English past tense inflection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877, Florence, Italy. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra

Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2022. Position information in transformers: An overview. *Computational Linguistics*, 48(3):733–763.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 864–870. Association for Computational Linguistics.

Heidi Harley and Elizabeth Ritter. 2002. Person and number in pronouns: A feature-geometric analysis. *Language*, 78(3):482–526.

Adi Haviv, Ori Ram, Ofir Press, Peter Izsak, and Omer Levy. 2022. Transformer language models without positional encodings still learn positional information. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1382–1390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Cheonkam Jeong, Dominic Schmitz, Akhilesh Kakolu Ramarao, Anna Stein, and Kevin Tang. 2023. Linear discriminative learning: a competitive non-neural baseline for morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 138–150, Toronto, Canada. Association for Computational Linguistics.

Akhilesh Kakolu Ramarao, Kevin Tang, and Dinah Baer-Henney. 2025. Frequency matters: Modeling irregular morphological patterns in Spanish with transformers. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4474–4489, Vienna, Austria. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. Neural multi-source morphological reinflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 514–524. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Jordan Kodner and Salam Khalifa. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Modeling inflection in language acquisition. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 157–175, Seattle, Washington. Association for Computational Linguistics.

Jordan Kodner, Sarah Payne, Salam Khalifa, and Zoey Liu. 2023. Morphological inflection: A reality check. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6082–6101, Toronto, Canada. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Martin Maiden. 2011. Allomorphy, autonomous morphology and phonological conditioning in the history of the Daco-Romance present and subjunctive. *Transactions of the Philological Society*, 109(1):59–91.

Martin Maiden. 2018. *The Romance verb: Morphomic structure and diachrony*. Oxford University Press.

Martin Maiden. 2021. The morphome. *Annual Review of Linguistics*, 7:89–108.

P H Matthews. 1972. *Cambridge studies in linguistics: Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation series number 6*. Cambridge studies in linguistics. Cambridge University Press, Cambridge, England.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.

Andrew Nevins, Cilene Rodrigues, and Kevin Tang. 2015. The rise and fall of the L-shaped morphome: diachronic and experimental studies. *Probus: International Journal of Latin and Romance Linguistics*, 27(1):101–155.

Paul O'Neill. 2025. Defective verbs in portuguese: a morphomic approach. *Cognitive Linguistics*, 36(2):153–181.

Ben Peters and André F. T. Martins. 2019. IT–IST at the SIGMORPHON 2019 shared task: Sparse two-headed models for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 50–56, Florence, Italy. Association for Computational Linguistics.

David E. Rumelhart, James L. MollClelland, and the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, volume 1. The MIT Press.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew Spencer and Mark Aronoff. 1994. Morphology by itself: Stems and inflectional classes. *Language*, 70:811.

Gregory T Stump. 2001. *Inflectional morphology: A theory of paradigm structure*, volume 93. Cambridge University Press.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (uni-morph schema). Technical report, Center for Language and Speech Processing, Johns Hopkins.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON*

Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 1–39, Online. Association for Computational Linguistics.

Yu-An Wang and Yun-Nung Chen. 2020. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6840–6849, Online. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Comparing Transformer Model Interpretability with Human Cognition: A Dual Analysis of Attention and Attribution

Lingchen Kong, Jinnie Shin, Pavlo Antonenko
University of Florida
l.kong@ufl.edu, jinnie.shin@coe.ufl.edu, p.antonenko@coe.ufl.edu

Abstract

Understanding whether transformer-based language models exhibit processing patterns that correspond to human cognition remains a central question in cognitively motivated computational linguistics. This study examines the statistical correspondence between model interpretability measures and human cognitive signals in a sentence-level relation classification task. We analyze attention weights and attribution scores (Integrated Gradients, Leave-One-Out, and LIME) from six fine-tuned transformer models and compare them to gaze duration derived from the ZuCo 2.0 corpus and human-annotated word importance. Results show that early-layer attention exhibits stable, moderate correlations with gaze duration that are largely unaffected by fine-tuning. Attribution methods, by contrast, demonstrate stronger correlations with human word-importance judgments, particularly for correctly predicted instances, though higher predictive accuracy does not consistently imply stronger human alignment. These findings suggest that interpretability measures capture distinct aspects of model computation, only some of which overlap with human cognitive signals. Systematic, multi-level evaluation against human benchmarks is therefore essential for assessing cognitive plausibility and for developing more transparent, trustworthy, and human-aligned NLP systems.

Keywords: Transformer Models, Explainable AI, Cognitive Modeling, Eye-tracking, Cognitive Plausibility

1. Introduction

Transformer-based language models have substantially advanced natural language processing (NLP) and demonstrated strong performance across multiple downstream tasks, including machine translation (Wang et al., 2022), automatic essay scoring (Shin and Gierl, 2024), and question answering (Nassiri and Akhloufi, 2022). Despite these successes, open questions remain regarding whether the internal computations of these models correspond to patterns observed in human language processing (Gu et al., 2025; Schrimpf et al., 2021). Within cognitive science and computational linguistics, this issue extends beyond performance optimization and relates directly to whether artificial systems can serve as empirically informative models of behavioral patterns observed in human language processing (Goldstein et al., 2025). Recent interdisciplinary work has therefore shifted toward evaluating not only model accuracy but also the extent to which model behaviour corresponds to human attentional and reasoning patterns (Eberle et al., 2022; Gao et al., 2025).

Interpretability methods have become a primary tool for investigating model decision processes (Mi et al., 2020). Among these, attention weights in transformer architectures are often interpreted as indicators of token importance. Because attention distributes representational focus across input tokens, it has been proposed as a potential computational correlate of token salience that may relate to human perceptual attention during reading (Zou et al., 2023). However, whether attention reliably explains model decisions remains debated. Some studies demonstrate correlations between transformer attention distributions and human eye-tracking signals, suggesting partial correspondence (Bensemann et al., 2022; Sen et al., 2020), whereas others argue that attention weights may not faithfully represent causal feature importance (Jain and Wallace, 2019; Serrano and Smith, 2019).

Beyond attention, attribution-based interpretability methods aim to provide more direct explanations of model predictions (Schwalbe and Finzel, 2023). Techniques such as Integrated Gradients (IG, Sundararajan et al., 2017), Leave-One-Out (LOO, Li et al., 2016) feature perturbation, and Local Interpretable Model-Agnostic Explanations (LIME, Ribeiro et al., 2016) attempt to quantify the contribution of individual tokens to model outputs. While these approaches have been widely applied for model transparency, limited research has examined whether attribution scores correspond to human judgments about linguistic relevance or reasoning importance within controlled comprehension tasks. This gap limits the extent to which interpretability findings can be connected to cognitive modeling objectives.

Human cognitive signals provide a valuable external benchmark for evaluating model explanations. Eye-tracking research has demonstrated that fixation duration and gaze patterns reflect attention allocation and lexical processing during reading (Degno and Liversedge, 2020). Similarly, human annotation of word importance can approximate reasoning-level judgments about semantic rele-

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 86–98
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

vance (Lottridge et al., 2023). Integrating these signals enables a multimodal evaluation of whether model interpretability methods capture behavioral patterns associated with perceptual and decision-related processing.

This study investigates statistical correspondence between transformer model interpretability measures and human cognitive signals in a sentence relation classification task. Using six fine-tuned transformer models and the ZuCo 2.0 corpus (Hollenstein et al., 2023), we systematically compare attention weights and attribution-based explanation scores against human gaze duration and annotated word importance. We examine alignment across model layers, interpretability methods, and prediction correctness to determine when model explanations approximate human attentional and decision-related signals and when they diverge.

This work makes three primary contributions. First, it provides a comprehensive comparison between attention-based and attribution-based interpretability methods using both perceptual and reasoning-oriented human benchmarks. Second, it investigates how fine-tuning and model performance influence measured correlation strength with human cognitive signals. Third, it provides a multimodal evaluation framework combining eye-tracking, human annotation, and interpretability analysis to assess whether transformer explanations reflect human-like language processing. By identifying the conditions under which alignment emerges, this study provides empirical constraints on how transformer processing patterns relate to human cognition and informs the development of more transparent and cognitively grounded NLP systems.

2. Related Work

2.1. Neural Language Models as Cognitive Processing Systems

Neural network models have long served as computational approximations of human language processing via distributed representations learned from data (Mikolov et al., 2013; Rumelhart et al., 1986). Transformer-based language models extend this paradigm through self-attention mechanisms that enable global contextual integration and the formation of progressively abstract representations across layers (Vaswani et al., 2017). Emerging evidence suggests that transformer representations exhibit structural properties that parallel functional characteristics associated with multi-stage human language processing (Caucheteux et al., 2021). Empirical studies show that earlier layers tend to encode lexical and syntactic information, whereas deeper layers capture abstract semantic and task-level representations (Jawahar et al., 2019; Tenney et al., 2019). These findings are consistent with cognitive theories proposing hierarchical processing during reading, in which early perceptual and lexical decoding stages precede semantic interpretation and higher-level comprehension processes (Dien, 2009; Hauk et al., 2006). However, predictive success alone does not guarantee cognitive plausibility. As a result, interpretability methods have been used to examine whether internal model computations reflect human-like information processing.

2.2. Interpretability Metrics in Transformer Models

2.2.1. Attention Weights

In transformer architectures, attention weights regulate how information flows between tokens by determining the relative influence of each token when contextualizing others (Vaswani et al., 2017). Through this mechanism, the model learns to distribute focus across the input sequence, enabling it to capture syntactic and semantic dependencies during representation construction (Kobayashi et al., 2020). Because attention explicitly distributes representational focus, it has frequently been interpreted as an indicator of token importance and model reasoning. However, the explanatory validity of attention remains contested. Jain and Wallace (2019) demonstrated that alternative attention distributions can frequently produce identical model predictions, suggesting that attention may not uniquely determine decision outcomes. Similarly, Serrano and Smith (2019) showed that zeroing out high-attention inputs does not substantially change the model's output, and attention-based rankings often fail to identify the most decision-critical inputs relative to gradient-based rankings. These findings indicate that attention may primarily reflect internal representational routing rather than direct causal contribution to predictions. This limitation motivates investigating complementary interpretability methods that more directly capture decision-level influence.

2.2.2. Attribution Scores

Attribution scores quantify each input feature's contribution to the model's output by assigning an importance value relative to a baseline (Hao et al., 2021). Unlike attention, which reflects contextual encoding processes, attribution methods are intended to capture how much each token influences the final prediction, providing insight into model decision processes. In this study, we computed attribution scores in transformer models using three complementary methods. Integrated Gradients (IG)

estimates feature importance by integrating gradients along a continuous path from a baseline input (e.g., zero embedding) to the observed input, producing attributions that reflect the accumulated sensitivity of the model output to each token (Sundararajan et al., 2017). Leave-One-Out (LOO) representation erasure measures importance by removing or masking individual tokens and evaluating the resulting change in model predictions, providing a perturbation-based estimate of each token's contribution (Li et al., 2016). LIME approximates local model decision boundaries by fitting interpretable surrogate models to perturbed versions of the input, where samples closer to the original input are weighted more heavily, yielding locally faithful approximations of token importance (Ribeiro et al., 2016).

Together, these attribution techniques provide complementary perspectives on features that influence model predictions, capturing gradient sensitivity, perturbation-based effects, and local decision approximations. Despite their widespread application in explainable AI (Azad et al., 2025; Gaspar et al., 2024), relatively few studies have examined whether attribution scores correlate with human judgments during language comprehension. Understanding this relationship helps clarify whether attribution-based explanations reflect cognitively meaningful decision processes.

2.3. Evaluating Transformer Attention via Human Gaze Correlations

Eye-tracking research provides fine-grained measurements of attentional allocation during reading, with gaze duration, fixation frequency, and total reading time serving as established indicators of lexical processing difficulty and perceptual attention (Rayner, 2009, 1998; Staub et al., 2010). These gaze signals have therefore been adopted as external benchmarks for evaluating whether neural language models exhibit human-like attention patterns.

Multiple studies have reported moderate to strong correlations between transformer attention distributions and human gaze signals. Bensemann et al. (2022) found that early-layer attention weights in pretrained BERT models moderately align with human dwell time patterns, although they cautioned that better alignment does not necessarily improve model performance. Similarly, Wang et al. (2024) reported moderate to strong positive correlations $(0.39 - 0.78)$ between five different human gaze signals (e.g. gaze duration and the number of fixations) and the BERT model's attention patterns across different layers.

Cross-linguistic studies provide further support for the attention-gaze alignment. Morger et al. (2022) found correlations exceeding 0.5 across German, Dutch, English, and Russian reading tasks when comparing total reading time with attention patterns and the gradient-based saliency. Kozlova et al. (2024) observed positive early-layer correlations $(0.45 - 0.77)$ between attention patterns and gaze signals in Russian texts but found that task-specific fine-tuning did not significantly improve alignment. Consistent with these findings, Eberle et al. (2022) reported minimal changes in attention-gaze alignment following BERT fine-tuning. Together, these studies indicate that, in several reading-related settings, attention distributions from pretrained transformer models have shown statistically significant correlations with selected human gaze measures.

2.4. Integrating Human Gaze Signals into Transformer Training

Recent research has sought to actively integrate predicted or measured gaze signals into transformer architectures. Dong et al. (2022) introduced the GazBy, a joint model that integrates human gaze fixation estimation with BERT and showed improved performance in passage re-ranking tasks. Sood et al. (2020) proposed a hybrid text saliency model that mimics the human gaze patterns and demonstrated improved performance in paragraph generation and sentence compression when integrated into transformer attention layers. Wang et al. (2024) demonstrated that incorporating predicted gaze signals into BERT-based models can improve performance across sentiment analysis and semantic similarity tasks, while Zhang and Hollenstein (2024) reported improved question-answering performance using eye-tracking attention masks.

Parallel findings in vision tasks further highlight the potential of integrating human cognitive signals during training. Sharan et al. (2019) demonstrated improved visual question-answering performance by training models to attend to human-identified salient image regions, while Rong et al. (2021) showed similar improvements in visual classification using gaze-based saliency maps. In software engineering, EyeTrans incorporated human attention into transformer-based code summarization, yielding improved code summarization performance (Zhang et al., 2024).

However, integrating human gaze signals remains methodologically challenging. Kozlova et al. (2024) found inconsistent performance improvements when incorporating gaze signals as training objectives, while Dong et al. (2022) demonstrated that performance gains depend strongly on where gaze information is integrated within model architectures. These findings highlight the complexity of effective human-data integration and emphasize

that alignment with human cognition is beneficial only when operationalized with consideration for model architecture and task objectives.

3. Methods

3.1. Data Collection

This study utilized the ZuCo2.0 Corpus (Hollenstein et al., 2023), which provides synchronized eye-tracking and EEG data from 16 native English-speaking adults during sentence reading. The corpus includes both a normal reading and a task-specific relation identification paradigm. The present study focused exclusively on the task-specific paradigm, in which participants actively identify semantic relations between entities within sentences. Sentences in the task paradigm were collected from Wikipedia and annotated with relation categories such as "Political Affiliation", "Employment", and "Education". Sentences that did not express a target relation were labelled as "Control" and served as a reduced relational-demand baseline condition.

Two complementary human-derived measures were used as evaluation benchmarks. First, attentional allocation during reading was operationalized using word-level gaze duration, defined as the total fixation time on each word aggregated across fixations and averaged across participants. Gaze duration is a well-established indicator of lexical processing difficulty and reading-time allocation (Carpenter and Just, 2017; Strandberg et al., 2023).

Second, decision-relevant salience was measured through human annotation of word importance for relation classification. Two annotators independently ranked the top 20% of words they judged to be most critical for identifying each sentence relation, following interpretability thresholding practices established in prior attribution research (Ju et al., 2022). Rankings were converted into graded importance weights based on selection order, with unselected words assigned zero weight. Inter-rater agreement was measured using Spearman's rank correlation, as annotators produced ordinal word-importance rankings. Agreement was 0.77 across all sentences and 0.84 when Control sentences were excluded, indicating substantial consistency in relational word-importance judgments.

3.2. Transformer Models and Fine-Tuning

Six pretrained transformer models were evaluated to examine whether human-model alignment is consistent across architectural and representational differences. The selected models were grouped into three categories. Base models included BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019), representing standard transformer architectures. Compact models included ALBERT and DistilBERT (Lan et al., 2020; Sanh et al., 2019), which reduce parameters via weight sharing or compression. Alternative models included DeBERTa variants (He et al., 2021) with disentangled attention for improved relational encoding.

To fine-tune transformer models for sentence relations prediction, we collected a training set of 3,353 English sentences from Wikipedia, matching the source of the ZuCo 2.0 task-specific sentences in category distribution, sentence length, and lexical complexity (measured by Flesch Reading Ease). The dataset was split into 80% for model training and 20% for validation. Fine-tuning was conducted for one and three epochs, which were treated as controlled manipulations of task adaptation strength and performance optimization. Training used AdamW optimization, a learning rate of 2e-5, batch size 16, and 10% warmup steps. Model performance was evaluated using F1 score and accuracy.

3.3. Model Interpretability Measures

Model-derived interpretability measures were extracted to parallel the two human cognitive benchmarks. Attention weights were obtained from both the first and last transformer layers to capture early and late contextual integration. Within each layer, attention weights were averaged across attention heads and aggregated column-wise to estimate the total incoming attention weight assigned to each token. Sub-word attention values were summed to produce word-level attention scores compatible with gaze measurements. Token-level attribution scores were estimated using IG, LOO, and LIME. For IG and LOO, sub-token scores were aggregated to the word level through summation, while LIME produced word-level scores directly. Attribution scores were ranked within each sentence, and the top 20% of words were retained to match the human annotation protocol, enabling direct comparison between model-derived and human salience patterns.

3.4. Analysis Plan

The analyses followed the unified evaluation pipeline illustrated in Figure 1, which summarizes data preparation, model training, and comparison procedures across experiments. Experiment 1 evaluates the alignment between transformer attention weights and human gaze duration, assessing whether the distribution of attention across tokens corresponds to human gaze patterns during relational reasoning. Correlation magnitudes are compared across model architectures and fine-tuning

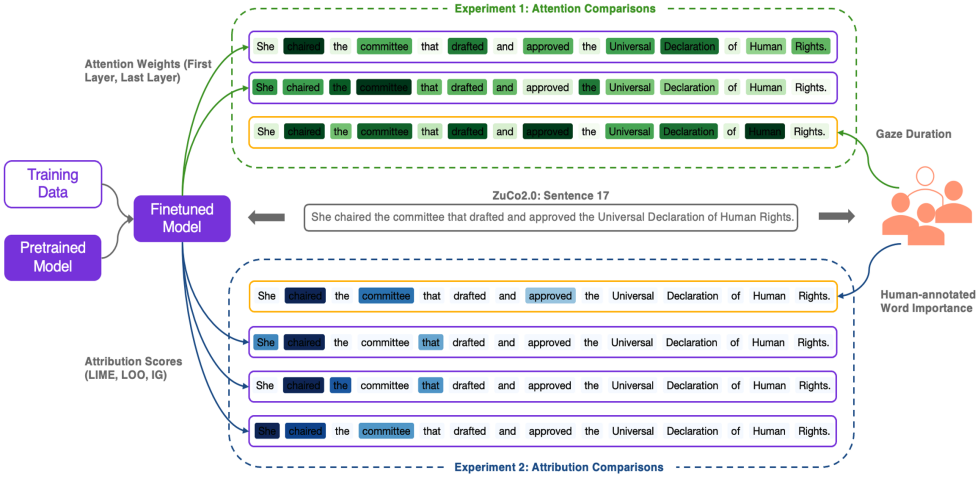


Figure 1: Conceptual overview of the analysis pipeline for evaluating human–model alignment in sentence-level relation classification. Darker shades indicate higher values (attention weight or attribution score). Blue shading corresponds to attribution scores, and green shading corresponds to attention weights.

stages to examine how differences in task performance relate to differences in human-model alignment.

Experiment 2 examines alignment between attribution scores and human-annotated word importance, evaluating whether words deemed influential by the model coincide with those selected by human annotators as critical for identifying relational meaning. The analysis further compares correlation strength across attribution methods and model correctness (correct vs. incorrect predictions) to determine whether stronger alignment is associated with higher prediction accuracy.

4. Results

4.1. Attention-Gaze Alignment

Table 1 summarizes average sentence-level correlations between transformer attention weights and human gaze duration across architectures, layers, and training epochs. The strongest attention-gaze alignment emerged in the first transformer layer, particularly for BERT and DistilBERT. BERT exhibited the highest correlation (0.657 across epochs), followed closely by DistilBERT (≈ 0.61). DeBERTa models showed weaker but consistently positive first-layer correlations ($\approx\ 0.28 - 0.34$), whereas RoBERTa demonstrated negligible alignment in both layers. ALBERT diverged from this pattern, showing higher correlations in the last layer than the first (last layer $= 0.298$ and 0.248 vs. first layer $= -0.044$ and 0.032), despite its parameter-sharing architecture.

Figure 2 illustrates the distribution of these correlations. First-layer distributions for BERT and DistilBERT are clearly right-shifted, with medians

Model	Epoch	F1	Corr. (Layer)	
			First	Last
ALBERT-base	1	0.73	-0.044	**0.298**
ALBERT-base	3	0.76	0.032	**0.248**
DistilBERT	1	0.76	**0.608**	0.068
DistilBERT	3	0.80	**0.613**	0.020
BERT-base	1	0.71	**0.657**	0.167
BERT-base	3	0.78	**0.657**	0.106
RoBERTa-base	1	0.75	**0.018**	0.054
RoBERTa-base	3	0.77	**0.017**	0.005
DeBERTa-base	1	0.74	**0.283**	0.183
DeBERTa-base	3	0.77	**0.287**	-0.011
DeBERTa-large	1	0.69	**0.337**	0.174
DeBERTa-large	3	0.77	**0.328**	0.157

Note. Strongest correlations per model are in bold.

Table 1: Average sentence-level correlations between human gaze duration and transformer attention weights by model, layer, and fine-tuning epoch

near the reported means and relatively tight dispersion, indicating relatively stable moderate alignment across many sentences. DeBERTa variants show weaker but still positive first-layer alignment, while RoBERTa's distributions cluster around zero in both layers, reflecting uniformly low correlations with human gaze duration. Last-layer distributions generally shift toward lower values or include broader negative tails, with ALBERT as the notable exception whose last-layer distribution centers in the positive range.

To understand whether better model performance in F1 translates into improvement in attention-gaze alignment from first layer, we tested the sentence-level paired differences in correla-

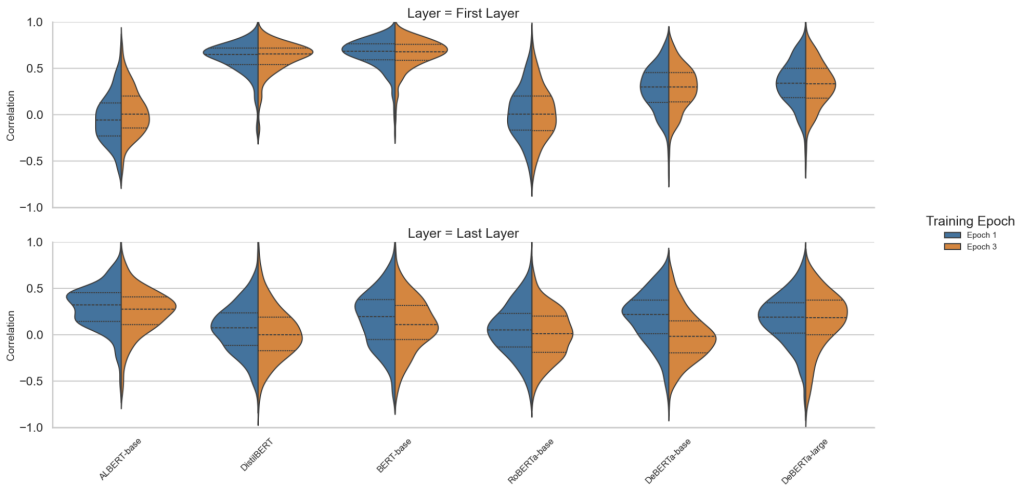


Figure 2: Distribution of sentence-level correlations between human gaze duration and transformer attention weights by model, layer, and fine-tuning epoch.

tion between epoch 1 and epoch 3 using Wilcoxon signed-rank tests. Results show no significant improvement in correlation between attention weights and human gaze duration after fine-tuning for any model except ALBERT ($Z = 11.95, p < .001, r = 0.66$). For all other models, the Wilcoxon signed-rank test did not support the hypothesis that fine-tuning improves alignment. Despite small increases in mean correlation for some models (e.g., DistilBERT: 0.608 to 0.613), the signed-rank distribution indicated that most sentences experienced reduced alignment ($Z = -3.83, p < .001, r = -0.21$). These findings suggest that attention-gaze alignment is more strongly associated with model architecture and pretraining conditions than with task-specific fine-tuning.

4.2. Attribution-Word Importance Alignment

Table 2 summarizes average correlations between human-annotated word importance and model attributions across methods (IG, LOO, LIME) and training epochs. IG consistently yields the strongest alignment, with RoBERTa achieving the highest overall mean correlation (0.374), followed by BERT (0.352) and DistilBERT (0.333). LOO generally performs mid-range and is the top method only for ALBERT, where it slightly exceeds IG (0.283 vs. 0.255 at epoch 3). LIME produced the weakest alignment overall, with correlations near zero for several models.

Table 3 reports correlations between human-annotated word importance and model attributions for sentences the models classified correctly versus incorrectly. The results showed a consistent pattern where the attribution scores from IG and LOO demonstrated higher correlations with human word

Model	Epoch	Attribution Scores		
		IG	LOO	LIME
ALBERT-base	1	0.244	**0.282**	0.150
ALBERT-base	3	0.255	**0.283**	0.117
DistilBERT	1	**0.333**	0.270	0.133
DistilBERT	3	**0.332**	0.277	0.135
BERT-base	1	**0.304**	0.251	0.094
BERT-base	3	**0.352**	0.272	0.124
RoBERTa-base	1	**0.345**	0.237	0.176
RoBERTa-base	3	**0.374**	0.253	0.176
DeBERTa-base	1	0.159	**0.227**	0.104
DeBERTa-base	3	**0.229**	0.222	0.069
DeBERTa-large	1	**0.333**	0.181	0.111
DeBERTa-large	3	**0.308**	0.247	0.142

Note. Strongest correlations per model are in bold.

Table 2: Average sentence-level correlations between human word-importance judgments and model attribution scores across attribution methods, models, and fine-tuning epochs

importance in correctly predicted sentences compared to incorrectly predicted ones. For example, BERT showed an IG correlation of 0.194 for incorrect predictions versus 0.400 for correct predictions at epoch 3. And ALBERT exhibited a notable LOO correlation increase from 0.137 to 0.336 at epoch 1. LIME exhibited smaller and more inconsistent differences across prediction outcomes. Group differences were evaluated using Mann-Whitney U tests, which confirmed significantly higher attribution–importance alignment for correct predictions across models and epochs for both IG and LOO ($p < .001$). No consistent significant differences were observed for LIME. These results indicate that when models make correct decisions, their attribu-

Model	Epoch	Attribution Scores					
		IG		LOO		LIME	
		IC	C	IC	C	IC	C
ALBERT-base	1	0.139	**0.283**	0.137	**0.336**	0.104	**0.166**
ALBERT-base	3	0.192	**0.274**	0.149	**0.323**	**0.156**	0.105
DistilBERT	1	0.167	**0.382**	0.157	**0.303**	0.133	**0.135**
DistilBERT	3	0.177	**0.372**	0.180	**0.302**	**0.136**	0.135
BERT-base	1	0.159	**0.361**	0.128	**0.300**	0.064	**0.105**
BERT-base	3	0.194	**0.400**	0.184	**0.297**	0.110	**0.179**
RoBERTa-base	1	0.170	**0.406**	0.123	**0.276**	**0.194**	0.169
RoBERTa-base	3	0.207	**0.426**	0.170	**0.280**	**0.211**	0.165
DeBERTa-base	1	0.019	**0.208**	0.104	**0.270**	0.082	**0.111**
DeBERTa-base	3	0.088	**0.271**	0.128	**0.249**	0.051	**0.075**
DeBERTa-large	1	0.215	**0.389**	0.078	**0.230**	0.101	**0.116**
DeBERTa-large	3	0.206	**0.339**	0.182	**0.267**	**0.183**	0.130

Note. Strongest correlations per condition are in bold.

Table 3: Average sentence-level correlations between human word-importance judgments and model attribution scores, separated by incorrect (IC) and correct (C) predictions

tion scores (particularly IG and LOO) more closely reflect human judgments of word importance, while LIME provides comparatively limited alignment.

5. Discussion

Evaluating whether neural language models approximate human cognitive processes during reading remains a challenge for cognitively grounded natural language processing (Kozlova et al., 2024). The present study contributes to this objective by systematically comparing transformer interpretability measures (i.e. attention weights, attribution scores) and two complementary human cognitive indicators (i.e., gaze duration, human-annotated word importance) during sentence-level relation classification. Across models, early-layer attention weights showed moderate and consistent alignment with gaze duration, whereas later-layer attention exhibited reduced alignment as the models progressed toward making final predictions. On the other hand, attributions scores, particularly Integrated Gradients (IG) and Leave-One-Out (LOO), showed more consistent and higher alignment with human-annotated word importance, especially for correctly predicted sentences. However, alignment patterns varied across model architectures and attribution methods, and perturbation-based LIME explanations showed comparatively unstable alignment.

These findings refine previous studies that early-layer attention aligns with human gaze behavior (Bensemann et al., 2022; Eberle et al., 2022; Sen et al., 2020; Brandl and Hollenstein, 2022) by showing that such alignment is stable and largely unaffected by fine-tuning, supporting the interpretation that early attention captures general perceptual or lexical features formed during pre-training (Kozlova et al., 2024). From a cognitive modeling perspective, this pattern supports the view that attention often reflects positional and distributional prominence, which aligns with human perceptual focus under the eye-mind hypothesis (Just and Carpenter, 1980), particularly when cognitive engagement is low. Our analysis of gaze duration reveals that participants tend to allocate more attention to visually distinct or complex words, such as numbers or uncommon terms, a pattern partially reflected in early transformer attention distributions, which is influenced by token frequency, orthographic complexity, and relative positioning. The weaker alignment observed in final-layer attention likely reflects increasing abstraction toward task-specific output representations (Clark et al., 2019; van Aken et al., 2019), which may rely on non-explicit or aggregate cues rather than human-salient information.

Our results extend prior work by demonstrating that attribution scores more closely approximate human decision-level reasoning. This finding is consistent with prior research emphasizing the utility of attribution scores in approximating human judgment in automated scoring studies. For example, Lottridge et al. (2023) found that IG-based saliency maps (i.e., word importance maps) aligned closely with human annotations in automated scoring tasks, though challenges persisted for longer responses and phrase-level distinctions. Similarly, Poulton and Eliens (2021) discovered that gradient-based techniques, especially Input X Gradient and Integrated Gradients, demonstrated the highest consistency in aligning with human-selected keywords when used on transformer-based Automated Short Answer Grading (ASAG) models trained with the

SQuAD 2.0 dataset.

We further observed that alignment between attribution scores and human word importance improves substantially for correctly predicted sentences. In these cases, both humans and models may rely on highly diagnostic semantic cues, which is consistent with principles of bounded rationality, where decision-makers prioritize the most informative elements when processing information under cognitive constraints (Gigerenzer and Goldstein, 1996). However, improved task performance did not uniformly translate into stronger human-model alignment. Notably, models such as DeBERTa-large often achieved higher F1 scores by emphasizing frequent but semantically trivial function words (e.g., was, of, the, a), instead of relying on meaning-bearing content. This divergence suggests that, while humans selectively attend to conceptually relevant cues to support semantic reasoning, models may instead exploit superficial statistical regularities to optimize classification. This pattern reflects the Clever Hans effect (Lapuschkin et al., 2019), in which performance gains arise from reliance on data set-specific artifacts rather than cognitive strategies relevant to the task, ultimately compromising human reasoning alignment and limiting generalizability.

Beyond empirical findings, this study contributes conceptually by integrating both process-level (gaze duration) and decision-level (attribution) indicators to evaluate transformer interpretability. Existing frameworks often compare explanations to post hoc human annotations (Doshi-Velez and Kim, 2017) without accounting for the cognitive processes underlying those judgments. Our dual-alignment approach addresses this gap by linking model attention to real-time human reading behavior and attribution scores to deliberative reasoning. This distinction reflects dual-process views of reading comprehension, where initial lexical processing precedes meaning integration (Rayner, 1998), and supports ongoing discussions that neural representations may align with cognitive operations at different architectural layers (Nonaka et al., 2021; Schirrmeister et al., 2017; Yamins et al., 2014). In addition, we show that higher task accuracy does not necessarily imply better cognitive alignment, reinforcing arguments that explainability must be evaluated on human-relevant criteria rather than performance alone (Lipton, 2018). By identifying where alignment emerges and where it breaks down, such as when models exploit shallow structural cues for prediction, our findings fill a critical gap in understanding how attribution methods approximate human reasoning (Garcia et al., 2024). This offers a pathway for refining explanation techniques to produce cognitively meaningful and trustworthy model interpretations.

From a practical perspective, these results provide actionable guidance for the development of cognitively informed transformer systems. The observed positive association between attribution-word importance alignment and model performance, particularly for IG and LOO, suggests that incorporating human-annotated importance signals into training or fine-tuning may promote models that reason in ways more consistent with human cognitive judgment (Bhatia and Richie, 2024; Hong et al., 2024). Such alignment is especially relevant in applications where justification quality matters as much as accuracy, such as automated essay scoring, clinical text assessment, and educational decision-support systems (Schneider et al., 2022). Our results also demonstrate that early-layer attention naturally aligns with human gaze patterns irrespective of task-specific tuning, implying that pre-trained models already encode perceptual regularities that approximate low-level human reading behavior. This suggests that approximate gaze signals could be simulated using early attention mechanisms in contexts where eye-tracking data is unavailable (Hollenstein et al., 2021), reducing data collection costs while preserving cognitive relevance. Importantly, these implications should be approached with caution: human-model alignment is not guaranteed by high accuracy but emerges more reliably in cases of correct, semantically grounded reasoning. This underscores the value of integrating human-centered interpretability evaluations into model development workflows to discourage reliance on superficial cues and promote trustworthy decision-making.

6. Conclusion

This study examined the extent to which transformer interpretability measures statistically correspond to human-derived indicators of reading and decision processes in a sentence relation classification task. By jointly evaluating attention–gaze and attribution–word importance correlations across architectures, layers, and fine-tuning stages, we provide a structured assessment of human–model alignment at both process and decision levels. Results show that early-layer attention weights exhibit stable, moderate correlations with gaze duration that are largely unaffected by task-specific fine-tuning, suggesting that pretrained representations encode perceptual or lexical regularities that overlap with human reading-time allocation. In contrast, attribution methods, particularly Integrated Gradients and Leave-One-Out, demonstrate stronger correlations with human-annotated word importance, especially for correctly predicted sentences. However, higher predictive performance does not uniformly imply stronger alignment, and

some high-performing models rely on superficial cues that diverge from human-identified relevance. Taken together, these findings indicate that different interpretability measures capture distinct aspects of model computation, only some of which overlap with human cognitive signals. Systematic evaluation against human-derived benchmarks is therefore necessary to assess the cognitive plausibility of transformer explanations and to inform the development of more transparent and human-aligned NLP systems.

7. Limitations and Future Work

Despite offering new insights into human-model alignment, this study is subject to several methodological and conceptual constraints. First, the operationalization of human cognition was limited to gaze duration and word-importance ratings from a single eye-tracking dataset (ZuCo2.0). While gaze duration is an established indicator of attentional allocation, it does not fully capture deeper reasoning or semantic integration processes. Similarly, binary top 20% annotations simplify human judgment into discrete relevance tiers. Future research should incorporate temporally dynamic and multivariate cognitive signals such as fixation regressions, saccade patterns, or EEG measures (Hollenstein et al., 2020), and expand to datasets featuring greater linguistic, demographic, and task diversity to strengthen generalizability.

Second, the analysis was restricted to encoder-based transformers within the BERT family. While this constraint enabled controlled architectural comparison, it limits conclusions about autoregressive decoder-only LLMs or hybrid architectures. Evaluating models trained under different objectives (e.g., causal language modeling) would clarify whether observed alignment patterns depend on training signals or architectural inductive biases.

Third, interpretability measures were evaluated primarily through correlation analysis. While correlation provides an interpretable and comparable summary metric, it compresses structured alignment into a single scalar statistic. This aggregation obscures distributional properties such as sparsity, positional concentration, or higher-order interaction patterns across tokens. Future work should complement correlation-based metrics with distribution-sensitive similarity measures, permutation controls, and intervention-based evaluations to more fully characterize explanatory correspondence.

Finally, the scope of investigation was limited to sentence-level relation classification. Extending this research to more complex, open-ended, or generative NLP tasks (e.g., reasoning, automated scoring, long-form comprehension), as well as non-text domains such as computer vision or multimodal learning, would enable broader evaluation of human-model alignment across cognitive demands and representational structures.

8. Acknowledgements

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

9. Bibliographical References

Mohammad Azad, Md Faraz Kabir Khan, and Sameh Abd El-Ghany. 2025. Xai-enhanced machine learning for obesity risk classification: A stacking approach with lime explanations. *IEEE Access*.

Joshua Bensemann, Alex Peng, Diana Benavides-Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and Michael Witbrock. 2022. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, Dublin, Ireland. Association for Computational Linguistics.

Sudeep Bhatia and Russell Richie. 2024. Transformer networks of human conceptual knowledge. *Psychological Review*, 131(1):271–306.

Stephanie Brandl and Nora Hollenstein. 2022. Every word counts: A multilingual analysis of individual human alignment with model attention. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 72–77, Online only. Association for Computational Linguistics.

Patricia A. Carpenter and Marcel Adam Just. 2017. *Cognitive Processes in Reading: Models Based on Readers' Eye Fixations*, page 177–213. Routledge.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *International conference on machine learning*, pages 1336–1348. PMLR.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Federica Degno and Simon P. Liversedge. 2020. Eye movements and fixation-related potentials in reading: A review. *Vision*, 4(1):11.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Joseph Dien. 2009. The neurocognitive basis of reading single words as seen through early latency erps: A model of converging pathways. *Biological Psychology*, 80(1):10–22.

Sibo Dong, Justin Goldstein, and Grace Hui Yang. 2022. Gazby: Gaze-based bert model to incorporate human attention in neural information retrieval. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 182–192.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. Increasing alignment of large language models with language processing in the human brain. *Nature Computational Science*, 5(11):1080–1090.

Basile Garcia, Crystal Qian, and Stefano Palminteri. 2024. The moral turing test: Evaluating human-llm alignment in moral decision-making.

Diogo Gaspar, Paulo Silva, and Catarina Silva. 2024. Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron. *IEEE Access*, 12:30164–30175.

Gerd Gigerenzer and Daniel G. Goldstein. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669.

Ariel Goldstein, Eric Ham, Mariano Schain, Samuel A. Nastase, Bobbi Aubrey, Zaid Zada, Avigail Grinstein-Dabush, Harshvardhan Gazula, Amir Feder, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Yossi Matias, Orrin Devinsky, Noam Siegelman, Adeen Flinker, Omer Levy, Roi Reichart, and Uri Hasson. 2025. Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, 16(1).

Chanyuan Gu, Samuel A. Nastase, Zaid Zada, and Ping Li. 2025. Reading comprehension in l1 and l2 readers: neurocomputational mechanisms revealed through large language models. *npj Science of Learning*, 10(1).

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12963–12971.

O. Hauk, M.H. Davis, M. Ford, F. Pulvermüller, and W.D. Marslen-Wilson. 2006. The time course of visual word recognition as revealed by linear regression analysis of erp data. *NeuroImage*, 30(4):1383–1400.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A. Jäger, and Nicolas Langer. 2023. The zuco benchmark on cross-subject

reading task classification with eeg and eye-tracking data. *Frontiers in Psychology*, 13.

Seung-Kyu Hong, Jae-Seok Jang, and Hyuk-Yoon Kwon. 2024. Enhancing performance of transformer-based models in natural language understanding through word importance embedding. *Knowledge-Based Systems*, 304:112404.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. 2022. Logic traps in evaluating attribution scores. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5911–5922, Dublin, Ireland. Association for Computational Linguistics.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Anastasia Kozlova, Albina Akhmetgareeva, Aigul Khanova, Semen Kudriavtsev, and Alena Fenogenova. 2024. Transformer attention vs human attention in anaphora resolution. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 109–122, Bangkok, Thailand. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1).

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Susan Lottridge, Sherri Woolf, Mackenzie Young, Amir Jafari, and Chris Ormerod. 2023. The use of annotations to explain labels: Comparing results from a human-rater approach to a deep learning approach. *Journal of Computer Assisted Learning*, 39(3):787–803.

Jian-Xun Mi, An-Di Li, and Li-Fang Zhou. 2020. Review study of interpretation methods for future interpretable machine learning. *IEEE Access*, 8:191969–191985.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Felix Morger, Stephanie Brandl, Lisa Beinborn, and Nora Hollenstein. 2022. A cross-lingual comparison of human and model relative word importance. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 11–23, Gothenburg, Sweden. Association for Computational Linguistics.

Khalid Nassiri and Moulay Akhloufi. 2022. Transformer models used for text-based question answering systems. *Applied Intelligence*, 53(9):10602–10635.

Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. 2021. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience*, 24(9):103013.

Andrew Poulton and Sebas Eliens. 2021. Explaining transformer-based models for automatic short answer grading. In *2021 5th International Conference on Digital Technology in Education*, ICDTE 2021, page 110–116. ACM.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Keith Rayner. 2009. Eye movements in reading: Models and data. *Journal of eye movement research*, 2(3):28.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. 2021. Human attention in fine-grained classification.

David E Rumelhart, James L McClelland, PDP Research Group, et al. 1986. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 2017. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420.

Johannes Schneider, Robin Richner, and Micha Riser. 2022. Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 33(1):88–118.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4596–4608, Online. Association for Computational Linguistics.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2931–2951.

Komal Sharan, Ashwinkumar Ganesan, and Tim Oates. 2019. *Improving Visual Reasoning with Attention Alignment*, page 219–230. Springer International Publishing.

Jinnie Shin and Mark J. Gierl. 2024. *Automated Short-Response Scoring for Automated Item Generation in Science Assessments*, page 504–534. Routledge.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341.

Adrian Staub, Sarah J White, Denis Drieghe, Elizabeth C Hollway, and Keith Rayner. 2010. Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5):1280.

Andrea Strandberg, Mattias Nilsson, Per Östberg, and Gustaf Öqvist Seimyr. 2023. Eye movements are stable predictors of word reading ability in young readers. *Frontiers in Education*, 8.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2019. How does bert answer questions?: A layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1823–1832. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Bingbing Wang, Bin Liang, Lanjun Zhou, and Ruifeng Xu. 2024. Gaze-infused bert: Do human gaze signals help pre-trained language models? *Neural Computing and Applications*, 36(20):12461–12482.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Leran Zhang and Nora Hollenstein. 2024. Eye-tracking features masking transformer attention in question-answering tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7057–7070, Torino, Italia. ELRA and ICCL.

Yifan Zhang, Jiliang Li, Zachary Karas, Aakash Bansal, Toby Jia-Jun Li, Collin McMillan, Kevin Leach, and Yu Huang. 2024. Eyetrans: Merging human and machine attention for neural code summarization. *Proceedings of the ACM on Software Engineering*, 1(FSE):115–136.

Jiajie Zou, Yuran Zhang, Jialu Li, Xing Tian, and Nai Ding. 2023. Human attention during goal-directed reading comprehension relies on task optimization. *eLife*, 12.

Correlating Language Model Surprisal With Cloze and Plausibility: Getting the Best of Both Measures

Kate Rebecca Belcher[1], **Matthew W. Crocker**[2]
[1]Leibniz-Institut für Wissensmedien (IWM), Tübingen, [2]Saarland University, Saarbrücken
k.belcher@iwm-tuebingen.de, crocker@coli.uni-saarland.de

Abstract

Prediction is central to both expectation-based theories of human language processing (such as Surprisal Theory), and the objective of neural network-based causal language models, where upcoming tokens are predicted based on their preceding context. With this similarity in mind, we investigated how language model predictions align with human linguistic prediction measures. We investigated the extent to which small-sized causal LLMs capture two common proxy measures of human surprisal – cloze probability and plausibility – in their predictive patterns. For this analysis, we created a new dataset of 660 sentence pair items with a minimal triplet design, in which target words vary across the full scale of word predictability, and calculate metric alignment by way of Pearson correlation. We find a stronger overall correlation of LM-surprisal with plausibility than with cloze, and, notably, the relationships between LM-surprisal and each of the two offline measures is found to vary depending on the relative predictability of the target word. We conclude that LM-surprisal offers a distinct perspective as a predictability measure than both offline behavioural measures, and that it may offer a useful tool in teasing apart nuances in predictability in certain instances which are not always captured by cloze probability and plausibility alone.

Keywords: Surprisal, expectancy, cloze probability, plausibility, large language models

1. Introduction

The expectation-based theories of language comprehension (Hale, 2001; Levy, 2008), state that the difficulty of processing a word is inversely proportional to its *surprisal*: the negative log probability of the word appearing given the preceding context. Words that are more expected in a given context have a lower surprisal and are typically read more quickly, whereas less expected words have a higher surprisal – conveying more information relative to more expected alternatives – and are read more slowly (Smith and Levy, 2013).

The equation for the calculation of surprisal:

$$Surprisal(x) = \log_2 \frac{1}{P(x|\text{context})} \quad (1)$$

can equally be reframed as:

$$Surprisal_{k+1} = -\log P(w_{k+1}|w_1 \dots w_k) \quad (2)$$

which intuitively expresses how surprisal – as an index of human processing effort – relates to the next word prediction language modeling objective used in causal language models. As (Levy, 2008) notes, surprisal creates a "causal bottleneck", such that whilst surprisal theory assumes a probabilistic model underlying linguistic comprehension, it makes no claims regarding specific mechanisms or linguistic representations. Being model-agnostic, surprisal measures can be derived from all types of probabilistic models, including transformer-based (Vaswani et al., 2017) large language models. Recently there has been considerable interest in determining the extent to which surprisal estimates derived from these language models (LM-surprisal) captures effects of predictability in human comprehension. One common approach is to correlate LM-surprisal with on-line measures of comprehension effort such as reading times, based on the hypothesis that these should index the cognitive effort due to human surprisal.

In this paper, however, we rather focus on the two off-line measures – namely cloze probability and plausibility judgements – that are commonly used in psycholinguistics to estimate the expectancy of a word in context. Cloze probabilities, determined by eliciting multiple human continuations for a given context, provide a direct measure of the predicted probability distribution for the upcoming word before it is encountered – something which cannot be directly revealed by reading times. While such cloze-based expectancy estimates have been shown to correlate well with reading times (e.g. Smith and Levy, 2011), they are limited in their ability to distinguish low probability targets. Plausibility jugdements, by contrast, provide an off-line measure of the integration difficulty of a word once it has been read which has been also argued to reflect surprisal (Brouwer et al., 2021), and can be obtained across the entire range of possible targets. In this paper, we explore the relationship between LM-surprisal derived from Transformer-based models and offline measures, across a broad range of expectancy values. Our findings shed light on both how to best estimate human surprisal, and how these proxy measures systematically align and diverge across the surprisal spectrum.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 99–109
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

1.1. Online vs Offline Measures

In studies of human language processing, online measures such as reading-times and eye-tracking measures are those which record information about how the comprehender processes language in real time as they incrementally read the words of a sentence or discourse, and provide a robust measure of comprehension effort. Such online measures of comprehension effort are thus indirectly related to surprisal, via the hypothesis that the two should be inversely proportional (Hale, 2001). Such online indices have been widely investigated, establishing a robust link of surprisal with both reading times (e.g. Shain et al., 2022; Smith and Levy, 2013; Aurnhammer and Frank, 2019; Merkx and Frank, 2021) and neurophysiological brain responses (Michaelov et al., 2024; Frank et al., 2015; Krieger et al., 2025). Offline measures, on the other hand, provide a complementary means to elicit more introspective responses to various dimensions of a linguistic stimulus – requiring the participant to make an active judgement or supply an answer – also allowing participants more time. In the present investigation, we focus on two human offline measures of word expectancy that are known to correlate with processing difficulty, namely cloze probability (introduced by Taylor (1953)) and plausibility judgements.

In a cloze task, participants are presented with a partial sentence prefix, and asked to provide what they judge to be the most likely word to follow, resulting over a sample of participants in a probability distribution of possible continuations. A widely acknowledged limitation of the cloze methodology is that despite providing a good representation of the probability of *likely* continuations, any word that is not mentioned in the study results in a cloze probability of zero. As cloze studies typically include less than 50 participants, this leads to a lack of nuance regarding the probabilities of less expected words, which, according to cloze, are all equally unexpected.

Plausibility is often linked to the idea of 'naturalness' (e.g. Amouyal et al., 2024), 'making sense' (e.g. Federmeier and Kutas, 1999), or how 'likely [it is] to occur in the real world' (Fedorenko et al., 2020). Kauf et al. (2024) link plausibility to world knowledge, concluding that humans judge sentences that align with their knowledge of the world as plausible, and those which violate this world knowledge as implausible. Whilst these definitions indicate a slightly broader scope than cloze, the central idea of expectation and predictability is common, and, as with cloze testing, plausibility judgements are obtained across a sample of people from which an overall judgement is determined for each experimental stimulus.

Though intuitively the relationship between plausibility and cloze probability may seem clear, such that high cloze probability items are typically also highly plausible (Brouwer et al., 2021), the relationship is not linear, and certain scenarios highlight how the concepts diverge, such that unexpected items in terms of cloze may still be judged as plausible (Delogu et al., 2021).

1.2. Comparing Offline Measures and LM-Surprisal

We define language model surprisal (or LM-surprisal) here as the inverse log likelihood of a token's occurrance given the preceding context, as derived from a language model trained with the objective to minimize the perplexity of next word prediction. Tokens refer to word-like or subword strings, 'tokenized' according to a tokenizer algorithm (such as Byte Pair Encoding (BPE) (Wang et al., 2020) or SentencePiece (Kudo and Richardson, 2018)). LM-surprisal is characterized by its data-rich basis and by the possibility to robustly estimate surprisal for any combination of context and continuation, as LM-surprisal estimates are derived from the output layer of the language model, with probabilities generated for all tokens across the entire vocabulary space of the model (Radford et al., 2019; Brown et al., 2020).

Both LM-surprisal and cloze probability serve as metrics to approximate linguistic statistics and word predictability (Smith and Levy, 2011). Rather than representing "true surprisal" values, these measures are better understood as distinct estimates of expectancy derived from different underlying models. A key distinction is that cloze probability is argued to incorporate human-centric factors like world knowledge and metacognitive strategies (Frank et al., 2015, p. 9), which are also often linked to plausibility (Kauf et al., 2024). However, the gap between these metrics is narrowing; Eisape et al. (2020) demonstrate that the correlation between LM-surprisal and cloze probability improves as model architectures become more sophisticated, with Transformer-based models outperforming simpler ones. This relationship is further supported by evidence of significant correlations between LM-surprisal, cloze probability, and plausibility across recent literature (e.g., Michaelov et al., 2024).

We therefore carry out a more systematic investigation into the ways in which LM-surprisal aligns with these two offline correlates of human surprisal, emphasizing the full range of target word predictability. Levy (2008) notes that despite the prominence of cloze probability in work on prediction, it is typically only higher cloze items (with a cloze probability >0.3) that are considered. We target this gap in our work, in which we build a dataset in which target words are designed to cover a broad distribution of word predictability, with a par-

ticular focus on the variation in lower expectation items, given that surprisal theory "relies on difficulty asymmetries between low-probability words" (Levy, 2008, p.42). Specifically, we adapt the design of Federmeier and Kutas (1999), who investigated the effect of target word expectancy based on cloze, and semantic category membership on the N400 ERP brain response. In their study, stimuli consisted of sentence pairs including a context sentence, followed by a continuation that contained the target word. The most expected single-word completion, as determined by cloze testing, determined the target in the high expectancy Condition A. The two target words for Conditions B and C are both unexpected, normed to have a cloze probability of <0.05. Conditions B and C differ crucially, however, in that Condition B is of the same semantic category as the most expected completion, while Condition C is semantically unrelated, belonging to a different basic semantic category. The consequence of this manipulation is that while B and C have similarly low cloze probability, targets in the B condition are generally rated as more plausible – a difference that was also attested in the N400 responses. Federmeier and Kutas also investigate the effect of contextual constraint on plausibility, that is, the extent to which the sentential context leads the comprehender towards a very specific completion of the second sentence, and confirm previous results from Schwanenflugel and LaCount (1988); Schwanenflugel and Shoben (1985) that less constraining contexts license a broader range of possible completions. This led to the finding that low cloze targets in Conditions B and C were judged to be more plausible in low constraint compared to high constraint items.

A subset of the Federmeier and Kutas dataset was later used by Ettinger (2020) to examine pragmatic inference and common sense reasoning in the BERT language model (Devlin et al., 2019). Specifically, Ettinger (2020) notes that common sense knowledge must be used to establish what is being spoken about in the first sentence of the sentence pair (the context), and pragmatic inference in determining how the second sentence relates to it. The experimental design of Federmeier and Kutas (1999) encourages a range of target word sentence completions from highly expected to highly unexpected, and ensures variation in the relative predictability levels within low cloze items, providing desirable characteristics for our experiment. In addition, the experimental set up allows us to investigate whether the same-category preference found in humans by Federmeier and Kutas (1999) is also reflected in patterns of LM-surprisal. Adapting the terminology used in Ettinger (2020), we refer to the variance of fit to the sentential context within zero-cloze items as the *pragmatic fit*. An example item from our stimuli is shown in Table 1.

1.3. Predictions

In light of the differing sensitivities of cloze probability and plausibility judgements, we explore how LM-surprisal aligns with these offline measures across the full spectrum of word predictability. Specifically, we address whether LM-surprisal correlates more strongly with the predictive nature of cloze tasks or the integrative nature of plausibility ratings. Looking beyond correlations with the entire dataset, we investigate the extent to which LM-surprisal may be differentially sensitive to these metrics depending on target expectancy. We predict that cloze probability will strongly correlate with LM-surprisal for expected words (in particular, Condition A), where human active prediction is most salient. Conversely, we examine whether plausibility is able to provide a more robust fit with LM-surprisal for less expected or "zero-cloze" items (Conditions B and C), as it captures the fine-grained pragmatic fit and world knowledge that cloze tasks fail to distinguish. Finally, we investigate if LM-surprisal captures the same interaction between contextual constraint and plausibility observed in humans, particularly in less constraining contexts where a broader range of completions is licensed.

2. Dataset Creation

Dataset creation followed four steps: stimuli generation; cloze testing for the generated stimuli; arrangement of target words for 'unexpected' conditions; and plausibility testing of all items. As in the original work by Federmeier and Kutas (1999), the language of the dataset is English.

2.1. Stimuli Generation

ChatGPT (version GPT 3.5) (Brown et al., 2020) was used for the initial generation of stimuli following the design of Federmeier and Kutas (1999). The model was prompted via the web interface to generate similar sentence pair examples, with examples from the original published stimuli by Federmeier and Kutas (1999) used as examples in a few-shot prompt. The final stimuli items were then compiled by manually reviewing and filtering the resulting synthetic stimuli according to the design criteria.

As in Federmeier and Kutas (1999), target word manipulations were centered around category-based rotations, following the concept of "basic category" as outlined by Rosch et al. (1976). A basic category is described as a category of concrete noun, which carries the most information to distinguish itself from other basic categories. An

Example stimulus	*The power went out during the storm, leaving the whole house in darkness. Susan rummaged in the drawer in search of a ...*				
Condition	Target	Cloze	Plausibility	Av. Cloze (Full dataset)	Av. Plaus. (Full dataset)
A	*torch*	0.8	6.72	$0.591_{\pm 0.223}$	$6.361_{\pm 0.465}$
B	*bulb*	0	3.11	$0.008_{\pm 0.032}$	$3.611_{\pm 1.529}$
C	*screw*	0	1.79	$0.0002_{\pm 0.002}$	$2.747_{\pm 1.326}$

Table 1: Average and example cloze and plausibility statistics

example would be "dog", for which a person can typically name several category members (e.g. "chihuahua", "dachshund", "corgi"), but all category members share the majority of features, such that the basic category "dog" is maximally distinguishable from other basic categories (e.g. "bird", "rodent" etc.). In total, target words stem from 58 distinct basic categories, belonging to 9 higher-level "super-categories" (meaning higher categorical levels such as "animals", "plants", or "tools and household objects").

2.2. Cloze Testing

Stimuli were divided into two lists, stratified such that all target semantic categories were equally represented across the lists. Stimuli were presented one at a time with the final word of the second sentence left blank. Participants, recruited via the crowdsourcing website *Prolific*, responded to each question by typing their completion into the empty text field. 30 participants were obtained for each list, and participants could not participate in both lists, leading to 60 unique participants. All participants were aged between 18 and 65 years old (mean age 39.5) and reported both their *primary language* and *first language* as English. After collection, data was minimally processed through spelling error correction, resolution of number agreement marking, and the unification of spelling variations. Cloze probabilities were then calculated from this data, and ranged from 0.033 (where only one person gave a specific response) to 1.0 (where all 30 participants gave the same response).

2.3. Target Word Selection

The response with the highest cloze probability (i.e. given with the highest frequency) determined the "best completion" target word for high predictability Condition A. Every word appeared as a target in each condition, helping to mitigate differences in conditions arising from lexically specific effects. For Condition B, target words were rotated within their basic semantic category to appear with another sentence pair (i.e. completions for the category 'dog' were rotated among other 'dog' stimuli, etc.). Condition C target words belonged to a different basic category, but the same super-category (i.e. at the next hierarchical level). This was accomplished by rotating target words from conditions A or B of another item which satisfy this criteria. For example, a target word from the basic category 'dog' would appear with a stimulus item from a different basic category of the super category 'animals'. Zero-cloze words were chosen as preference, and if this was not possible, the target word was as low-cloze as possible. For both Condition B and C, small morphosyntactic changes, such as resolving number agreement differences, were allowed to ensure that the resulting sentences were always grammatical, regardless of condition. This resulted in a fully counter-balanced design, such that each target word appeared with a sentence pair in each of the three conditions, resulting in a final dataset comprising of 220 sentence pairs, each with three possible endings.

2.4. Plausibility Ratings

Plausibility ratings were collected for all sentence pair contexts in all three conditions (660 items in total, split across six lists). A modal average of 18 individual responses (mean av. 19.67[1]) were collected for each item. The final plausibility rating for an item was the average of all responses. As with the cloze study, all participants were native English-speaking adults (mean age 37.9), recruited via *Prolific*. Judgements were made on a Likert scale with numbered points and labels at the extreme ends, from 1 (*"not plausible at all"*)' to 7 (*"very plausible"*). We asked participants to judge *"How plausible is the last sentence in the context of the whole text?"* Formulation in this way allowed participants to focus on the entire text and use all information available in their judgements. As the target word was always the final word in the stimulus, the plausibility judgement was always made directly after encountering the target word. In both the cloze and plausibility studies, six attention check items were included throughout the study. Correctly responding to at least four of the six items was a prerequisite for inclusion in the study. No par-

[1] For one list, 28 responses were collected due to a technical error. Post-hoc analysis showed no statistical difference in mean rating with 28 responses compared to a random 18 responses, so all participants were included in the final data.

ticipants were excluded on the basis of the attention checks, so all data was included in the study.

2.5. Dataset Characteristics

The resulting dataset contains 220 sentence pairs, each with three different target word conditions, to give a total of 660 experimental items. The average cloze probabilities and plausibility ratings for each condition are shown in Table 1. The averages by condition suggest a clear trend: Condition A target words are both the most expected in terms of cloze probability, and rated most plausible. While average cloze probability for Condition B is slightly higher than Condition C, both B and C were far less expected than A. Average plausibility similarly decreases from Condition A to Condition C. However, average plausibility for Condition B sits in the middle of the 1-7 scale at 3.611, and Condition C almost one point lower at 2.747, while neither are highly implausible. These more graded average plausibility ratings of Conditions B and C relative to A are consistent with our basic category membership manipulation – same category targets have plausibility ratings closer to A, than different category targets.

At the item-level, Condition A is rated as most plausible in nearly 95% of stimuli, demonstrating a strong human preference for the high expectation word in context. Between conditions B and C, Condition B is rated as more plausible than Condition C in 72.7% of items, leaving just over one quarter of items where Condition C, although semantically further from the best completion, is perceived as more plausible than the within-category alternative. The relationship between cloze and plausibility is shown in the swarm plots in Figure 1. Strikingly, we observe little variation in plausibility for higher cloze probability items, however, in low cloze items, the full range of the plausibility scale is utilized. This clearly demonstrates the divergence of the metrics, and highlights the aforementioned limitation of cloze in capturing the full extent of word expectancy. The full dataset with cloze and plausibility ratings can be found at: https://osf.io/x3wzs/.

3. Language Model Analyses

We correlate LM-surprisal against the human-derived cloze probabilities and plausibility ratings using Pearson correlations. Although simple correlational analyses do not allow us to control for other factors such as lexically specific effects, these potential issues are in part mitigated by the cross-condition counterbalancing design. LM-surprisal measures were obtained by calculating token-level output probabilities of target words, given the context of the sentence pair. Where a target word

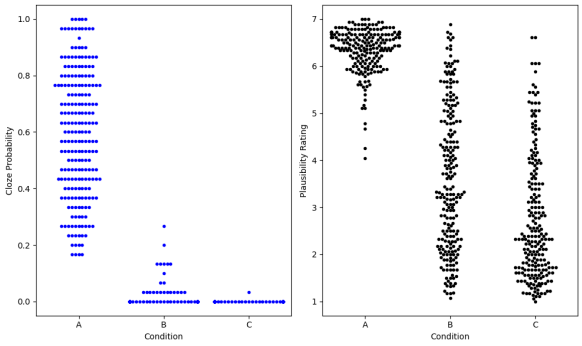


Figure 1: Swarm plots displaying the distribution of plausibility ratings and cloze probability across the dataset, by condition

consisted of multiple sub-word tokens, the respective sub-word token probabilities were summed, following the commonly used methodology for this scenario (Krieger et al., 2025; Kauf et al., 2024). We extracted LM-surprisal values from several small decoder-only language models. All models were open-source, so that token probabilities could be accessed, and were base models (rather than instruction-tuned), to ensure the same language modeling objective (next word prediction). We focus on four comparable models of between 1-2 billion parameters: GPT2-XL (Radford et al., 2019), GPT-Neo (Black et al., 2021), Falcon-RW-1B (Penedo et al., 2024) (henceforth, Falcon-1B), and Qwen3-1.7B-base (Yang et al., 2025) (henceforth, Qwen3-1B). Recent works suggest that models in this smaller size range are more appropriate for modeling human behavioural measures than larger models (Oh and Schuler, 2023; Oh et al., 2024), thus, they are the focus of our investigation. We compare the correlation of LM-surprisal with both cloze probability and plausibility, evaluating the entire dataset, individual conditions, and the modulation of contextual constraint level on the observed patterns. All reported correlations are significant to the 0.01 level unless otherwise indicated.

3.1. Alignment of Cloze and LM-Surprisal

Across the entire dataset, the relationship between cloze and surprisal is clearly observable, with all models achieving a moderately strong negative correlation. Falcon-1B shows the strongest correlation of -0.689. A result of the conditional design of the dataset is the clustering of items in the zero cloze area, which is clearly observable in Figure 2. Therefore, we look at how the correlation strength varies at the individual condition level, to give greater insight into where correlation stems from item level differences, as opposed to the contrast alone of zero-cloze and non-zero cloze items between con-

ditions. The by-condition correlations are reported in Table 2. We observe moderate correlations between LM-surprisal and cloze across the models for Condition A, and, despite the skew towards low-cloze items in Condition B, correlations are comparable to, or even exceed the strength of Condition A correlations across models. Given the intended lack of cloze variation for Condition C items, there is no observable correlation between cloze and surprisal for Condition C.

3.2. Alignment of Plausibility and LM-Surprisal

Table 3 reports the Pearson correlation between LM-surprisal and plausibility for the evaluated models. For the entire dataset, the overall correlation is strong across models, indicating that LM-surprisal accounts well for human-rated plausibility variation. The strongest correlation is observed for Qwen3-1B, slightly exceeding Falcon-1B. At the individual condition level, we observe significant correlations for all conditions, with the weakest correlation in Condition A, where -0.15 $< r <$ -0.25. Correlations in the unexpected (B/C) conditions are significantly stronger, ranging between -0.41 and -0.51, representing moderate negative correlations, which are stronger in Condition C than Condition B. In addition to strong overall alignment, Qwen3-1B shows the strongest correlation for the intermediary Condition B items, however is outperformed by Falcon-1B in Conditions A and C, suggesting that Falcon-1B is better able to approximate predictability at the more extreme ends of the scale. The correlation for Falcon-1B is plotted with all conditions together in Figure 3.

3.3. Constraint Level Analysis

In an additional analysis, we evaluate the effect of contextual constraint in modulating target word plausibility and corresponding LM-surprisal. The data are split into two subsets according to the constraint level of the context. Following Federmeier and Kutas (1999), we use the median cloze probability of the Condition A target word as the threshold. Items where the top cloze word has a cloze probability of 0.6 or higher form the 'high constraint' group.

Table 4 shows average human plausibility by condition and constraint level, with the highest of each constraint pair marked in bold. Higher constraining items are judged as more plausible than less constraining items in Condition A, however in Conditions B and C, we observe the opposite effect, such that less constraining items are viewed as *more* plausible than those with more highly constraining contexts. These between-constraint differences were significant (Mann-Whitney-U test), and reflect the same patterns found by Federmeier and Kutas (1999), further verifying the similarity in characteristics between the two datasets, and indicating that less constraining contexts license a broader range of possible completions (Schwanenflugel and Shoben, 1985; Schwanenflugel and LaCount, 1988). The corresponding constraint-modulated correlations between surprisal and plausibility are also reported for the two best performing models of the four models tested. In Condition A, we observe no significant correlation when broken down by constraint level. In Condition B, for three of the models, we observe a pattern whereby the correlation with high constraint items exceeds the equivalent low-constraint correlation. Falcon-1B is an exception to this, where a stronger correlation is observed for low constraint items. In Condition C, a stronger correlation between LM-surprisal and plausibility is

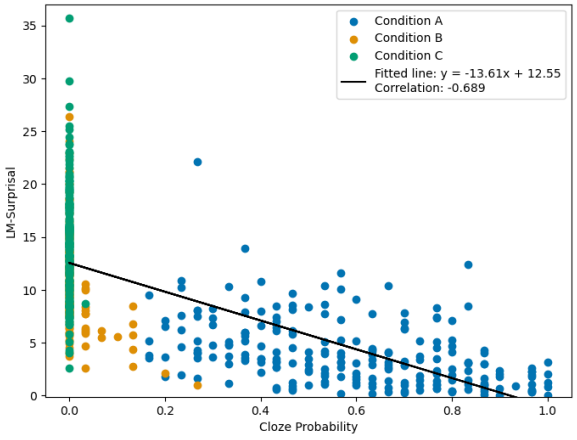


Figure 2: Correlation between Falcon-1B surprisal and cloze probability with conditions colour-coded

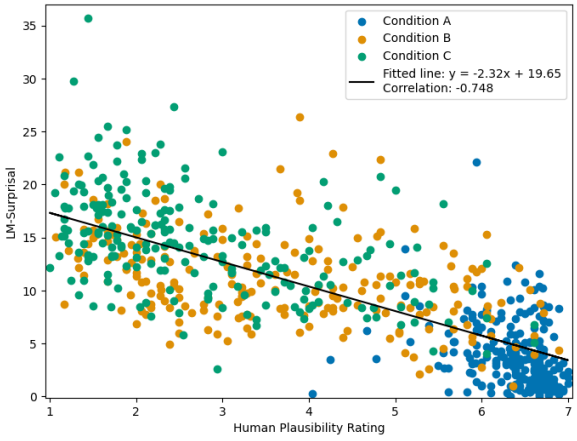


Figure 3: Correlation between Falcon-1B surprisal and plausibility with conditions colour-coded

LM-surprisal vs Cloze	GPT2-XL	GPT-Neo	Falcon-1B	Qwen3-1B
Full data (ABC)	-0.62	-0.625	**-0.689**	-0.671
Condition A	-0.307	-0.296	**-0.386**	-0.279
Condition B	-0.344	-0.358	**-0.375**	-0.35
Condition C	-0.058^{ns}	-0.109^{ns}	-0.078^{ns}	-0.062^{ns}

Table 2: Pearson correlation with cloze for all data and by condition (^{ns} indicates correlation is **not** significant, **boldface** indicates the strongest correlation across models)

LM-surprisal vs Plausibility	GPT2-XL	GPT-Neo	Falcon-1B	Qwen3-1B
Full dataset (ABC)	-0.693	-0.696	-0.748	**-0.753**
Condition A	-0.208	-0.207	**-0.245**	-0.164
Condition B	-0.43	-0.434	-0.415	**-0.466**
Condition C	-0.494	-0.468	**-0.504**	-0.502

Table 3: Correlations of LM-surprisal with plausibility across all data and by condition, **boldface** indicates the strongest correlation across models)

found for *low* constraint items consistently across models. This mirrors the interaction between constraint and plausibility found in Conditions B and C in the human plausibility ratings. The between-constraint differences in the surprisal-plausibility correlations for Conditions B and C did not reach significance, however, the qualitative pattern suggests that LM-surprisal captures plausibility best in the case of *less* expected, and *less* contextually constrained items.

3.4. Comparing Cloze and Plausibility Correlations

We exemplify the side-by-side correlations with cloze and plausibility for Falcon-1B and Qwen3-1B in Table 5. We observe that, in both cases, surprisal provides a statistically significantly better correlation for human plausibility than cloze ($p<0.05$). Comparing the correlations for each condition side-by-side, we find that in Condition A, for expected items, surprisal provides a qualitatively better fit for cloze than plausibility ($p=0.505$ for Falcon-1B). However, for the less expected items in Conditions B and C, surprisal provides a better fit for plausibility than for cloze for both conditions, which is statistically significant in Condition C. Thus, we find that whilst surprisal shows correlations with both cloze probability and plausibility ratings, the strength of correlation varies with both the expectancy of the final word, and the pragmatic expectancy of this word within its sentential context.

Furthermore, for the Falcon-1B estimates, which are overall in greatest alignment with the human measures, a pattern arises whereby, with cloze, the strength of the correlation decreases over the conditions with relative expectancy. The strongest correlation is found in Condition A, with a slightly weaker correlation in Condition B, and no significant correlation at all in Condition C. Conversely, with plausibility, surprisal offers the weakest correlation in Condition A, with an increase in strength for Condition B, and further increase to the strongest correlation in Condition C. Thus, we see an inverse relationship between surprisal and cloze, and surprisal and plausibility as a function of a word's expectancy (whether measured by cloze or plausibility). This across-condition pattern is stable across models for the surprisal-plausibility correlation, and, whilst the corresponding pattern with cloze is less defined for Qwen3-1B, such that the strongest correlation with cloze is in Condition B rather than Condition A, the inverse relationship observed at the extreme ends of the predictability scale still holds, with Conditions A and C showing contrasting patterns for both variables.

4. Discussion

In the present paper we investigated the nature of the relationship between LM-suprisal and offline behavioural measures of expectancy. Our results suggest that LM-surprisal is able to capture expectancy variations at both ends of the scale: In high expectation items, which are more common in active prediction, surprisal reflects some of the nuance in expectation captured by cloze. Moreover, for less expected completions, LM-surprisal still provides a graded picture of word expectation in context, which corresponds to the pragmatic fit of that word in its context. For the highly expected words in Condition A, we found that surprisal was more strongly correlated with cloze probability than plausibility. However, in the low expectation conditions, correlation strength was significantly higher between surprisal and plausibility. Thus, for predictions of a language model, the whole spectrum of relative predictability can be captured more effectively than by either one of cloze or plausibility taken alone. Returning to the predictions in Section 1.3,

	Condition A		Condition B		Condition C	
	H	L	H	L	H	L
Human plausibility	**6.484**	6.232	3.394	**3.837**	2.506	**2.996**
Falcon-1B LM-surp. vs plaus.	-0.11^{ns}	-0.187^{ns}	-0.365	**-0.449**	-0.476	**-0.504**
Qwen3-1B LM-surp. vs plaus.	-0.139^{ns}	-0.064^{ns}	**-0.479**	-0.435	-0.489	**-0.523**

Table 4: Constraint effects on human plausibility and LM-surprisal-plausibility correlation strength. (^{ns} indicates a non-significant correlation, **boldface** indicates the strongest correlation for each H/L constraint pair)

	Falcon-1B		**Qwen3-1B**	
	Surp. vs Cloze	Surp. vs Plaus.	Surp. vs Cloze	Surp. vs Plaus.
Entire dataset	-0.689	**-0.748**	-0.671	**-0.753**
Condition A	**-0.386**	-0.245	**-0.279**	-0.164
Condition B	-0.375	**-0.415**	-0.35	**-0.466**
Condition C	-0.078^{ns}	**-0.504**	-0.062^{ns}	**-0.502**

Table 5: Relationships of LM-surprisal (Falcon-1B/Qwen3-1B) with cloze probability and plausibility (^{ns} indicates lacks of significant correlation, **boldface** indicates the strongest correlation for each pair)

these findings support the idea that LM-surprisal is differentially sensitive to the two measures in accordance with relative expectancy. This pattern proved fairly robust across the models tested.

Our results highlight how the differences between the human offline prediction metrics are heightened when considering words from the extreme ends of the scale in terms of word probability. Cloze probability allows for a more fine-grained distinction in expectancy among high-expectation targets than plausibility ratings, which instead demonstrate ceiling effects such that the probability of high-expectation words in lower constraining contexts is overestimated, relative to those in higher constraining contexts. Conversely, plausibility ratings offer a more graded differentiation of less expected targets, possibly reflecting the importance of pragmatic and world knowledge over purely linguistic constraints in estimating surprisal for these conditions. The constraint analysis offered further support to this analysis, demonstrating a clear trend across all models that for the least expected completions (Condition C), the correlation of surprisal and plausibility is stronger for less constraining items than highly constraining items. Our results pattern with those of both Federmeier and Kutas (1999) and Schwanenflugel and LaCount (1988); Schwanenflugel and Shoben (1985) for human-rated plausibility, with the additional insight that LM-surprisal patterns appear also to follow this pattern for the least expected items.

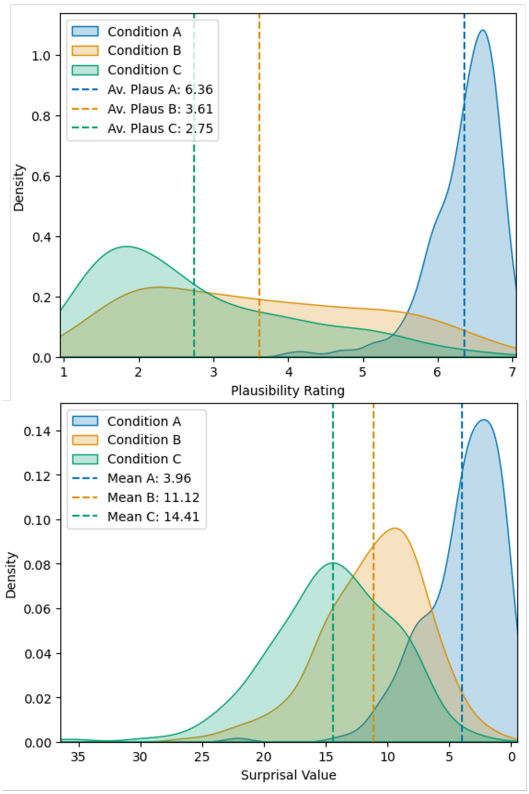


Figure 4: Density plots for human plausibility (upper) and Falcon-1B surprisal (lower) by condition. Surprisal is plotted with lowest surprisal on the right of the scale.

We also observe where LM-surprisal deviates from the two offline metrics. The density plots in Figure 4 illustrate that the categorical design of the dataset resulted in clearly distinguishable peaks in surprisal across conditions. The categorical membership appeared to be less influential for participants judging the plausibility of these continuations, as demonstrated by the wide, evenly spread distribution across the plausibility scale for Condition B. Additionally, although the by-constraint correlation of surprisal and plausibility reflected the same patterns as human plausibility for Condition C, the equivalent result was only found with one model for Condition B, appearing to be less consistent with

the human plausibility results.

5. Conclusion

In this study, we investigated the extent to which LM-surprisal aligned with two common offline human predictability metrics, with a particular focus on low expectation items, which, while crucial for surprisal theory, are often lesser explored in the literature (Levy, 2008). Our dataset design deliberately encouraged the full use of the probability scale for target items, confirming the fact that low expectation in terms of cloze probability does not always correspond to low plausibility. We empirically showed that LM-surprisal is *differentially sensitive* to the two metrics, providing a closer fit for cloze probability than plausibility when words are highly expected, and a closer fit for plausibility than cloze when words are less expected in context. In the lowest expectancy items, the correlation between surprisal and plausibility also displayed similar constraint-based patterns as human plausibility data. These results suggest that the two behavioural metrics provide complementary estimates of human surprisal, with LM-surprisal capturing the strengths of each. Our findings thus emphasize that, while both offline measures offer a valuable means for estimating human surprisal – and thus evaluating language models – neither measure fully captures the entire expectancy spectrum. Our results pave the way for the more nuanced use of these measures both in future evaluations of more sophisticated models, as well as in experimental psycholinguistic investigations. Future experimental work may wish to combine both offline human measures, especially depending on the range of expectancies in the data. Further language model analyses may seek to investigate how LM-surprisal correlates with plausibility as model size increases, and whether the patterns found by Oh et al. (2024) with naturalistic corpora are also applicable for low-expectancy stimuli such as those in our study. Overall, our results confirm that while cloze is a well-established measure for distinguishing more and less expected words, it fails to capture the relative expectancy of less expected and implausible items. Plausibility offers a potential alternative human measure to quantify expectancy for these items. Importantly, however, language models demonstrate the potential to provide a robust estimate of human surprisal across the whole distribution of expectancy, modeling the best elements of both measures.

6. Acknowledgements

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Germany —Project-ID 232722074—SFB 1102. The first author carried out this work while at Saarland University and the University of Malta, and gratefully acknowledges the support of the Erasmus Mundus European Masters Programme in Language and Communication Technologies (EU grant no. 2019-1508), and funding from the Joachim Herz Stiftung through the ALEE project. We kindly thank the anonymous reviewers for their comments.

7. Limitations

We acknowledge several limitations in the study. The conclusions drawn about LM-surprisal can only be drawn in relation to the specific models tested, acknowledging the nature of the models as causal language models trained on next word prediction, and as small LLMs, the largest of which has 1.7 billion parameters. This is notably smaller than the largest and most state of the art language models now available. However, larger models have been shown to offer a reduced ability to model human behavioural measures compared to those in the range evaluated here (Oh and Schuler, 2023), and, such models are typically closed source, which hinders the analysis of output probabilities, and may be trained with additional training objectives, such as instruction-tuning, reducing the link between the original motivation for the study. While language models were used both in the stimuli generation process and in the evaluation of the stimuli, the types of LMs used for each task have fundamentally distinct training objectives, with the instruction-tuned chatbot model ChatGPT used in stimuli generation, and base language models – pretrained on next-token prediction only – for the evaluation. Although we use a counter-balanced dataset design, we acknowledge that using Pearson correlation as our primary analysis method may not fully reveal the relationship of the predictors we consider here. Further analyses, such as regression analyses, could help to tease apart these factors.

8. Ethics Statement

All participants in the data collection studies for the creation of the dataset used in this research gave their informed consent and received remuneration for their participation. We believe any broader ethical concerns of the conducted analyses to be minimal.

9. Bibliographical References

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In

Findings of the Association for Computational Linguistics: EACL 2024, pages 166–181, St. Julian's, Malta. Association for Computational Linguistics.

Christoph Aurnhammer and Stefan Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with meshTensorFlow.

Harm Brouwer, Francesca Delogu, Noortje J. Venhuizen, and Matthew W. Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Francesca Delogu, Harm Brouwer, and Matthew W. Crocker. 2021. When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, 1766:147514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiwalayo Eisape, Noga Zaslavsky, and Roger Levy. 2020. Cloze distillation: Improving neural language models with human next-word prediction. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 609–619, Online. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kara D. Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 263–277, Miami, Florida, US. Association for Computational Linguistics.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W. Crocker. 2025. On the limits of llm surprisal as a functional explanation of the n400 and p600. *Brain Research*, 1865:149841.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, page 12–22, Online. Association for Computational Linguistics.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics:*

EMNLP 2023, pages 1915–1921, Singapore. Association for Computational Linguistics.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2024. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data only. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439.

P. J. Schwanenflugel and K. L. LaCount. 1988. Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:344–354.

Paula J Schwanenflugel and Edward J Shoben. 1985. The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, 24(2):232–252.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020. Neural machine translation with byte-level subwords. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9154–9160.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Is Cross-Lingual Transfer in Bilingual Models Human-Like? A Study with Overlapping Word Forms in Dutch and English

Iza Škrjanec[1,2], **Irene Elisabeth Winther**[3], **Vera Demberg**[1], **Stefan L. Frank**[3]
[1]Saarland University, Germany [2]Zuse School ELIZA, Germany [3]Radboud University, the Netherlands
{skrjanec, vera}@lst.uni-saarland.de, {irene.winther, stefan.frank}@ru.nl

Abstract

Bilingual speakers show cross-lingual activation during reading, especially for words with shared surface form. Cognates (friends) typically lead to facilitation, whereas interlingual homographs (false friends) cause interference or no effect. We examine whether cross-lingual activation in bilingual language models mirrors these patterns. We train Dutch-English causal Transformers under four vocabulary-sharing conditions that manipulate whether (false) friends receive shared or language-specific embeddings. Using psycholinguistic stimuli from bilingual reading studies, we evaluate the models through surprisal and embedding similarity analyses. The models largely maintain language separation, and cross-lingual effects arise primarily when embeddings are shared. In these cases, both friends and false friends show facilitation relative to controls. Regression analyses reveal that these effects are mainly driven by frequency rather than consistency in form-meaning mapping. Only when just friends share embeddings are the qualitative patterns of bilinguals reproduced. Overall, bilingual language models capture some cross-linguistic activation effects. However, their alignment with human processing seems to critically depend on how lexical overlap is encoded, possibly limiting their explanatory adequacy as models of bilingual reading.

Keywords: bilingualism, cognate, interlingual homograph, surprisal, semantic similarity

1. Introduction

Compared to speakers of a single language, bilingual speakers process some linguistic phenomena differently during reading, a notable example being words with the same surface form in multiple languages. For cognate words, the surface form as well as the meaning is equivalent or very similar across languages (e.g. *winter* in Dutch and English). For interlingual homographs (henceforth, false friends) the surface form is the same, but the meaning is not (e.g. *brand* means 'fire' in Dutch).

Empirical studies of human reading find that bilinguals process cognate words faster than non-cognate control words (Libben and Titone, 2009; Bultena et al., 2014; Lauro and Schwartz, 2017). This cognate facilitation effect has not been observed in monolinguals, suggesting that it arises specifically from the coexistence of two languages within a single speaker.

For reading of false friends, studies report either slower reading of false friends than control words or no difference in bilingual speakers (Libben and Titone, 2009; Titone et al., 2011; Pivneva et al., 2014; Hoversten and Traxler, 2016). Slower reading of false friends in bilinguals is thought to reflect interference caused by diverging form-meaning mappings across languages.

Both facilitation and interference effects have been attributed to cross-language activation (e.g. Dijkstra and van Heuven, 2002; Kroll et al., 2012). Explaining these cross-lingual interactions requires explicit and testable computational models of how the two languages interact within a single cognitive system (Frank, 2021). Neural language models

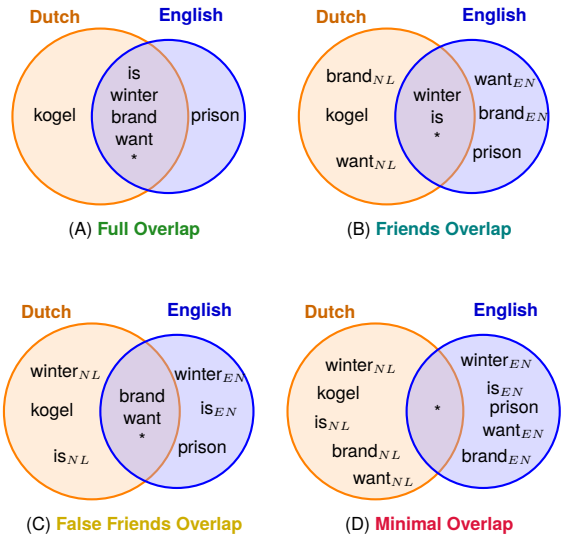


Figure 1: Vocabulary conditions. **Full overlap (A)**: All identical surface forms that appear in Dutch and English are shared between the two languages, each represented by a single embedding. **Friends Overlap (B)**: Only cognates and loan words are shared. **False Friends Overlap (C)**: Only false friends are shared. **Minimal Overlap (D)**: Only punctuation and named entities are shared, while other tokens have language-specific embeddings. Across all conditions, punctuation and named entities are shared (denoted as $*$).

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 110–123
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(LMs) are strong candidates as computational models of second language learning and bilingual reading (Yadavalli et al., 2023; Oba et al., 2023; Aoyama and Schneider, 2024; Constantinescu et al., 2025).

In Natural Language Processing, cross-lingual transfer is often desired when structural or lexical similarities or simply lexical overlap between languages can be leveraged to improve downstream performance (e.g., Kallini et al., 2025; K et al., 2020; Wu and Dredze, 2019; Pires et al., 2019). In contrast, this work evaluates bilingual LMs for cross-lingual transfer between Dutch and English and explores under what conditions this transfer seems human-like (if at all). We test LMs for lexical processing via vocabulary manipulation in four conditions (Figure 1), controlling how overlapping word forms are treated in the vocabulary. In the Friends Overlap condition, each word that is a *friend* (a cognate or a loan word) is assigned a single, language-unspecific embedding allowing both Dutch and English to inform it. In the False Friends Overlap setting, the same holds for words that are false friends (but not friends). The other two conditions enforce either sharing of each form-overlapping word (Full Overlap) or a nearly complete separation of languages in the embedding space (Minimal Overlap).

We evaluate each vocabulary condition through the lens of two signals important in LM training as well as cross-lingual activation in humans: word frequency and language context (i.e. the language of the sentence).

Our results show that bilingual models generally separate the two languages, except when they are explicitly tied by a shared embedding of a word. In that case, the models show cross-lingual effects, specifically facilitation effects for friends as well as false friends compared to their respective control words. We find that this facilitation is driven by word frequency, and is not influenced by form-meaning mapping across languages. The only condition that can explain human data is Friends Overlap (i.e. facilitation for friends, but not for false friends). This suggests that while LMs can reproduce certain cross-lingual activation patterns, their behavior aligns with human bilingual reading only under specific vocabulary conditions[1].

2. Background and Related Work

2.1. Vocabulary Design

A large number of related studies use an equivalent of the Full Overlap condition when training a bilingual LM, so each overlapping surface form is shared between languages (Winther et al., 2021; Roslund and Matusevych, 2022; Oba et al., 2023; Constantinescu et al., 2025). Our vocabulary conditions are inspired by Kallini et al. (2025), who experiment with vocabulary manipulations to find that, for downstream tasks such as natural language inference and question answering, any sharing is beneficial (even that of false friends) in contrast to no overlap. Aoyama and Schneider (2024) reset the embedding layer before starting L2 acquisition, which essentially does not allow cross-lingual transfer between overlapping word forms.

In terms of the share of first (L1) and second (L2) language, some studies used a balanced proportion where each language has the same budget of training tokens (Oba et al., 2023; Kallini et al., 2025; Constantinescu et al., 2025), while others simulated larger L1 exposure compared to L2 (Roslund and Matusevych, 2022; Constantinescu et al., 2025). Winther et al. (2021) observe that the cognate facilitation effect in LMs trained on Dutch-English or Norwegian-English depends on the portion of L1/L2 and the presentation order of the languages, whereby an LM that is first trained on the L1 (75% training samples), and then on the L2 at the end of each epoch exhibits lower surprisal for cognates relative to non-cognates, mirroring the facilitation effect observed in bilinguals.

2.2. Role of Frequency and Context

Language models learn more frequent words earlier and better (Chang and Bergen, 2022; Razeghi et al., 2022). This provides a relevant link to cross-lingual transfer in bilinguals. According to the cumulative frequency hypothesis (Voga and Grainger, 2007; Strijkers et al., 2010; Midgley et al., 2011), the cognate facilitation effect in bilinguals stems from exposure to cognates regardless of the language: exposure strengthens the word's lexical representation due to the same form-meaning mapping. It is unclear how this fares with false friends, where the form-meaning mapping diverges across languages, i.e. whether we can expect cross-lingual transfer to be influenced by frequency.

Aside from word frequency, context is an important cue in LM training and inference. LMs learn how to use context information to disambiguate lexically ambiguous words within one language (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). One question that arises is whether a bilingual LM can use the language of the preceding context as a cue to process an overlapping word form without influence/transfer from the other language, or whether the fact that an overlapping form occurs in two language contexts makes it more ambiguous and less predictable.

[1]Our code is made available at https://github.com/izaskr/cross_lingual_transfer_dutch_english_forms.

3. Manipulation of Vocabulary Sharing

We design 4 conditions for the LM vocabulary to test the role of sharing/separation of embeddings of words with the same surface form across Dutch and English. When a word form is *shared* (see intersections in Figure 1), it has a single word embedding. If this form happens to appear in both Dutch and English, then samples from both languages contribute to the embedding during training. In case of *separation* (see the areas outside of the intersection in Figure 1), each word form is encoded for the language of the sentence it is in. Their embeddings are informed only by the samples in the respective language.

In condition Full overlap (A), there is no active intervention: each word form that appears in both languages has a single, language-unspecific embedding. In other conditions, we manipulate what is shared and what is separated.

In condition Friends Overlap (B), only friends (cognate and loan words) across the two languages are in the intersection. This condition is motivated by hypotheses that cognate words might share orthographic and (partially) semantic representations in a bilingual's mental lexicon, while false friends are represented by two different orthographic representations (Dijkstra and van Heuven, 2002; Lemhöfer and Dijkstra, 2004). All other words have language-specific embeddings.

Only false friends are placed in the intersection in condition False Friends Overlap (C). Each false friend has a single entry in the vocabulary despite different meanings across languages (e.g., *brand* means 'fire' in Dutch).

Finally, in Minimal Overlap (D), the word forms from the two languages are completely separated given the sentence language. A language-specific processing is assumed, regardless of overlapping form/meaning across languages.

Note that punctuation and named entities are placed in the intersection across all conditions. They are often not language-specific, refer to the same extra-linguistic entity and are thus not a part of our manipulation. Named entities were identified by spaCy's xx_ent_wiki_sm model[2] prior to tokenizer and LM training.

4. Implementation

4.1. Training Data

To imitate human language exposure, we train our LMs on a corpus of diverse genres: 49% of tokens

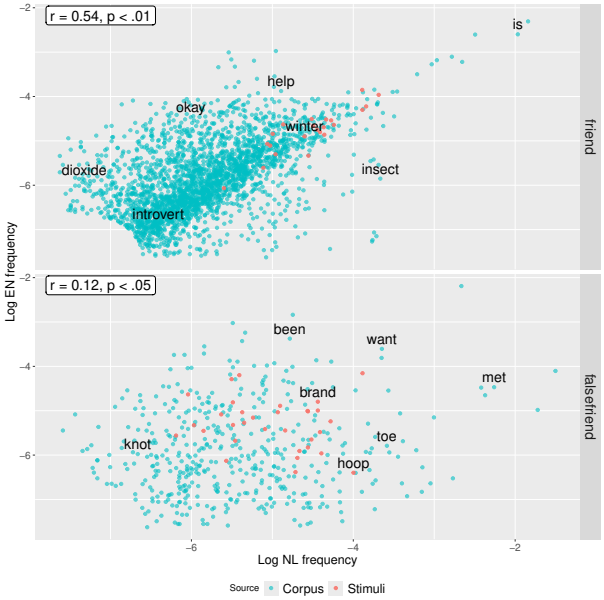


Figure 2: The frequency of each word in the Dutch and English portion of the LM training data. The frequencies are generally positively correlated: a word form that is frequent in Dutch is frequent in English, but more so for *friends* than *false friends*. The plot includes (false) friend words from psycholinguistic stimuli (Bultena et al., 2014; Huisman, 2025) as well as other compiled lists and our training data (here labeled as Corpus).

in the data come from non-fiction text (Wikipedia[3]), 26% from transcribed scripted speech (OpenSubtitles[4] and TedTalks[5]), and 25% from web-crawled data (CC100, Conneau et al., 2020)[6]. In total, each training corpus had about 400 million tokens. To simulate an unbalanced late Dutch-English bilingual, the training data consists mostly of Dutch (75% tokens), while the rest is English. In each epoch, Dutch samples were presented first, followed by English ones.

4.2. Overlapping Word Forms

To obtain a broad coverage of words sharing only form, or both form and meaning between Dutch and English, we annotated a list of words that appear in both the Dutch and English corpora as friends (cognates or loan words) and false friends (interlingual homographs). We joined our list with existing ones (Bultena et al., 2014; Poort and Rodd, 2019;

[2] https://spacy.io/models/xx

[3] https://dumps.wikimedia.org/nlwiki/latest

[4] https://opus.nlpl.eu/OpenSubtitles/nl&en/v2024/OpenSubtitles

[5] https://object.pouta.csc.fi/OPUS-NeuLab-TedTalks/v1/tmx/en-nl.tmx.gz

[6] https://data.statmt.org/cc-100

Lefever et al., 2020; Huisman, 2025) to obtain a set of 2,806 friends and 511 false friends, both referring to word types without named entities. The Dutch and English frequencies of each word have a significant positive correlation (Figure 2).

Some words are homographs within one language and can be both a friend and a false friend (e.g. *monster* in Dutch can have the English meaning as well as 'sample'). These words are excluded from our vocabulary manipulation.

Class annotations are available for a subset of the items, comprising of 1,598 friends and 379 false friends: all cognate and loan words are annotated as content words in both languages. Among false friends, 4% have different classes in Dutch and English (e.g. *toe* and *met* are content words in English, but function words in Dutch).

4.3. Psycholinguistic Stimuli

We evaluate the LMs on stimuli from two reading studies, in which the participants were late bilinguals with Dutch as their first language and English as their second. In the *friend study* (Bultena et al., 2014), participants read English sentences with target words that were either Dutch-English friends (cognates) (1a) or non-cognate controls (1b)[7].

1a. The residents dislike the *winter* for the ...

1b. The residents dislike the *prison* for the ...

The friend and control targets were matched with respect to word length and English word frequency. In total, the stimuli consisted of 22 item sentences (with two conditions per item). The context preceding the target words was designed to not be semantically constraining and biased towards the meanings of target words. The analysis of reading times by Dutch-English bilinguals revealed a cognate facilitation effect, with faster reading for cognates than the corresponding control words.

The stimuli in the *false-friend study* (Huisman, 2025) include word pairs of false friends and controls that were presented to participants in Dutch sentences, that is, in their first language, in a self-paced reading paradigm. Again, each sentence had two conditions: with a false friend (2a) or with a language-unique control word (2b):

2a. De beelden van de *brand* zullen hen ...

2b. De beelden van de *kogel* zullen hen ...

All target words were nouns and all false friends were form-identical between Dutch and English. Control words were paired with false friends to match them on word length and Dutch and English word frequency. The stimuli by Huisman (2025) consists of 32 item sentences, all designed to be semantically non-constraining in the context preceding the target words. While this study did not find significant reading time differences between false friends and control words in bilinguals (and thus no cross-language activation), the manipulation provides a valuable evaluation of the LMs tested here, particularly due to the lack of prior work on false-friend reading in Dutch-English bilinguals.

[7]We only use items where the cognate is a form-identical noun.

4.4. Tokenizers

The tokenization process included multiple components. All words annotated as *friends* or *false-friends* (Section 4.2) as well as the language-unique control words (Section 4.3) were tokenized as single subwords across all conditions to avoid biases from multi-subword segmentation (Lesci et al., 2025). We train a byte-level BPE tokenizer on combined Dutch-English train set (64k vocabulary, minimum frequency of 2) and used it to tokenize all remaining words. We also train a byte-level BPE tokenizer on named entities (10k vocabulary, minimum frequency of 2).

4.5. Transformer Language Models

For each vocabulary manipulation, we train a separate Transformer model with the causal language modeling task, following the GPT2-small configuration. In a single epoch, the model was first presented Dutch samples (300 million tokens), followed by English ones (100 million tokens) creating a sequential regime. Each LM was trained for 2 epochs. See Appendix C for vocabulary sizes and parameter counts under different conditions. Appendix D shows loss curves on the training and test sets.

5. Vocabulary Interventions in Bilingual LMs

5.1. Language Context

We first explore how (dis)similar the preceding contexts of the same surface form are across Dutch and English. If the LM "understands" that a friend word carries the same meaning in Dutch and English, this should be reflected in the similarity of its contextual and word embeddings across languages. In the same vein, if the LM represents the language-specific meanings of false friends, their context and word embeddings should be dissimilar as well. In conditions that assign each overlapping word form a shared embedding (Friends Overlap for friends; False Friends Overlap for false friends; Full Overlap for both), it might be easier for the LM

to learn friends as cross-lingual meaning equivalents, but it might be more difficult to tease apart the language-specific meanings of false friends.

Method. For each word t in the stimuli set of friends and false friends T, we sample a set of 500 sentences containing that word from the training data from each language (yielding 1,000 sentences per word in total). For each sentence in the sample, we calculate the mean-pooled embedding of the preceding context of word t (without the word itself). For each language-specific subcorpus, we average across these embeddings, obtaining $\mu_C^{NL}(t)$ and $\mu_C^{EN}(t)$.

For each sentence in the sample, we also obtain the contextualized embedding of the target word t itself. We then average across the sentences per language, yielding $\mu_W^{NL}(t)$ and $\mu_W^{EN}(t)$ for Dutch and English sentences, respectively.

We measure the similarity of these embeddings between languages using cosine similarity[8]. The similarity between $\mu_C^{NL}(t)$ and $\mu_C^{EN}(t)$ tells us how similar the preceding contexts of the same word are across languages. In contrast, cosine similarity between $\mu_W^{NL}(t)$ and $\mu_W^{EN}(t)$ estimates the similarity of the target word embedding across languages. See Appendix E for illustrative diagrams. We use the embeddings from layer 12. Results from layer 5 show very similar patterns.

Results. Figure 3 shows that the contexts for the same word in Dutch and English are generally estimated as dissimilar and vary little across vocabulary conditions, i.e. manipulating what is shared in the embedding space does not lead to more similarities or differences for the context preceding target words (upper panel). Focusing on the embeddings of the target words themselves (lower panel), the results indicate that separating the languages for overlapping word forms leads to lower similarity for the word affected by separation: in the Friends Overlap condition, false friends have language-specific vocabulary entries. The distance between Dutch and English sentences containing these words is larger than when each false friend has a single embedding. Friends, likewise, show a lower similarity when their embeddings are language-specific (False Friends Overlap condition). In the Minimal Overlap condition, the contexts as well as word embeddings stay apart and dissimilar across Dutch and English. Generally,

[8]Due to known issues with anisotropy in contextualized representations (Ethayarajh, 2019), we standardize the embeddings following Timkey and van Schijndel (2021). We estimate a mean and standard deviation embedding from a sample of 30k instances and standardize the embeddings before pooling.

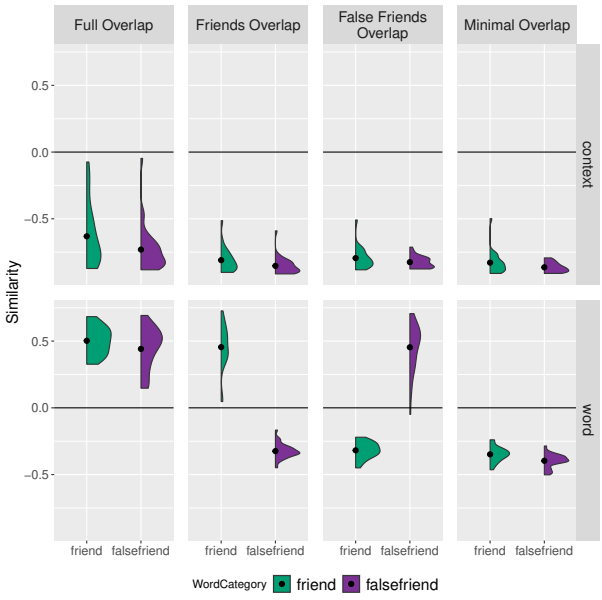


Figure 3: Cosine similarity between Dutch and English across the context and word-only embeddings for training data samples for target words (friends or false friends). A larger similarity means the LM considers Dutch and English embeddings more similar.

this result shows that the contexts are coded as dissimilar between Dutch and English, and the target word embeddings are similar when each is represented by a single, language-unspecific embedding, no matter the form-meaning match in Dutch and English.

5.2. Processing Effort via Surprisal

We further focus on the predictability of overlapping word forms in comparison to language-unique words given the same preceding sentence context by estimating their surprisal. We use word surprisal, $surp(w) = -\log_2 p(w|context)$, as a computational correlate of processing effort as observed in humans with lower surprisal corresponding to easier processing (e.g. Hale, 2001; Levy, 2008; Shain et al., 2024).

Bilingual speakers tend to show different behavior for language-unique words and words with overlapping surface form. Specifically, reading a word with form overlap across languages might show patterns of cross-language activation: facilitation for friends (faster reading in comparison to controls), and interference for false friends (slower reading in comparison to controls). We expect lower surprisal for friends than controls, reflecting the ease of processing of friends over controls. In contrast, we expect lower surprisal for controls than false friends.

With respect to the vocabulary configurations, we examine how sharing/separation of embeddings affects the form-meaning correspondence of overlapping word forms. On the one hand, in conditions where each overlapping surface forms has a single embedding, this word might be more ambiguous as it appears across Dutch and English contexts, possibly reducing its predictability in a given sentence. If so, shared forms should have higher surprisal than language-unique controls. On the other hand, a shared embedding receives more training signal than two single ones and more than a frequency-matched control, which could result in reduced surprisal compared to controls. We use $\alpha = .05$ in all statistical tests.

Method. Using sentence stimuli from the *friend study* and the *false-friend* study (see Section 4.3), we estimate surprisal for each target word. Importantly, in each item, the target word (either overlapping or control) is preceded by the same context. We test for effects of processing effort by comparing the surprisal of overlapping words to that of controls.

We use mixed-effects regression in `lme4` (Bates et al., 2015) with surprisal as the outcome, and word category and word position in sentence (index) as fixed effects, and a random intercept for sentence pair (item)[9]. We fit a regression model for each vocabulary condition separately for the two studies, resulting in 8 regression models in total.

Results. Figure 4 shows surprisal estimates across the two word categories and vocabulary conditions. For *friends*, we find that they have significantly lower surprisal than control words in two conditions: Full Overlap, and Friends Overlap. In both conditions, each friend surface form has a single embedding, which increases the word's predictability relative to controls. For *false friends*, we find that in the Full Overlap and False Friends Overlap conditions, their surprisal is significantly lower than that of controls. Again, in these conditions, the false friend's surface form has a single embedding.

Taken together, both types of overlapping word forms (friends and false friends) show smaller surprisal relative to language-unique controls when their surface forms are represented by a single shared embedding. These results indicate that embedding sharing increases the predictability of overlapping surface forms regardless of their form-meaning correspondence. Surprisingly, neither word category show lower predictability, either because of language context ambiguity or form-meaning divergence.

[9]Model structure: $surp \sim WordCategory + WordIndex + (1|Item)$. WordCategory is sum-coded (1 friend/falsefriend, −1 control).

5.3. Frequency Benefit across Languages

In bilingual training, overlapping surface forms receive additional exposure relative to language-unique controls due to their occurrence in both languages. This raises the question whether increased frequency drives improved predictability of words with overlapping surface forms across languages (i.e. as per the cumulative frequency hypothesis) or whether their predictability might depend on the language of the sentence. This analysis builds on the facilitation effects observed in Section 5.2 focusing on the effect of a word's frequency in Dutch vs. English on its surprisal.

The two frequency measures may play different roles. In the *friend study*, the target words appear in English sentences, making Dutch the "other" language. In the *false-friend study*, the sentences are in Dutch, so English serves as the "other" language. This analysis examines whether both the sentence-language frequency ($\text{freq}_S(w)$) and the other-language frequency ($\text{freq}_O(w)$) influence surprisal of word w.

Method. Both friends and false friends exhibited facilitation relative to controls in vocabulary conditions where their embeddings were shared between Dutch and English. Thus, to compare the two frequency measures ($\text{freq}_S(w)$ and $\text{freq}_O(w)$), we consider the Full Overlap and Friends Overlap conditions for the *friend study*, and Full Overlap and False Friends Overlap for the *false-friend study*. The frequency estimates of the target words are based on the Dutch and English portions of the LM training data; they are log-transformed and normalized. For each study, we start by fitting a regression model with surprisal estimated by word category, normalized word index and vocabulary condition[10]. In the next step, we add freq_S to the model[11], and finally freq_O[12]. We compare these regression models using likelihood ratio tests to determine whether the addition of the frequency terms provides a better fit.

Results. The results for the *friend study* show a main effect of word category as facilitation for friends over controls ($\beta = -.5$, $p < .01$). Upon adding word frequency in English (freq_S), the word category effect was still negative, but not a reliable effect ($\beta = -.2$, $p = .06$). There is a main

[10]The maximal model that would converge was $surp \sim WordCategory + Vocabulary + WordIndex + (1|Item)$.

[11]$surp \sim freq_S + WordCategory + Vocabulary + WordIndex + (1|Item)$

[12]$surp \sim freq_O + freq_S + WordCategory + Vocabulary + WordIndex + (1|Item)$. Interaction terms did not improve the model fit for any regression models.

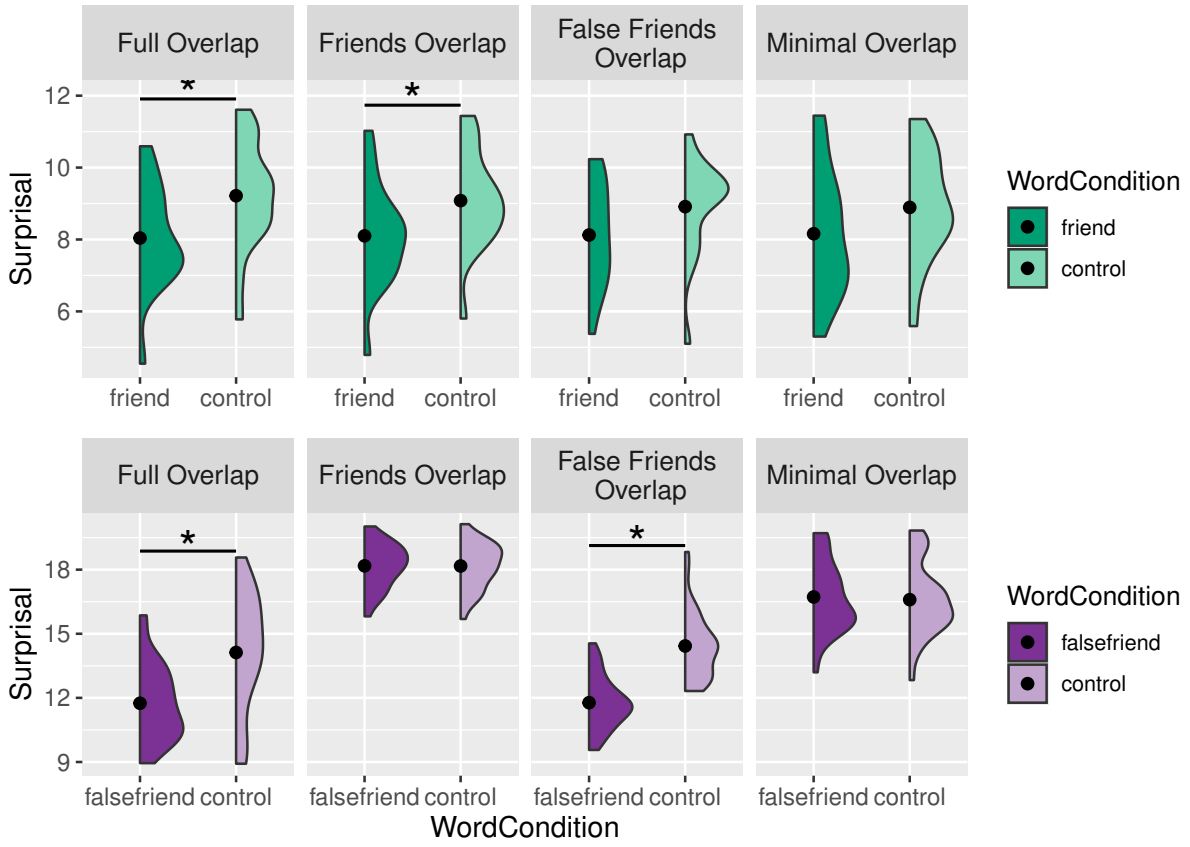


Figure 4: Surprisal estimates for friends/controls and false friends/controls from LMs trained under different vocabulary conditions. The symbol $*$ indicates a statistically significant difference ($p < .05$) between the two word types.

effect of English frequency ($\beta = -1.1$, $p < .01$), indicating that higher frequency leads to smaller surprisal. Finally, including Dutch word frequency (freq_O), we observe that English frequency is still a significant predictor ($\beta = -1.2$, $p < .01$), but Dutch frequency is not ($p = .9$) and neither is word category ($p = .4$). The model comparison revealed that including Dutch word frequency in the model does not contribute to the fit ($\chi^2(1) = .002, p = .97$).

For the *false-friend study*, we find a facilitation effect for false friends compared to controls ($\beta = -1.3$, $p < .01$). When Dutch frequency (freq_S) is included in the model, there is still a significant facilitation effect for false friends ($\beta = -1.2$, $p < .01$) as well as a significant main effect of Dutch frequency ($\beta = -.5$, $p < .01$). Finally, when English frequency (freq_O) is added to the model, Dutch frequency remains a significant predictor ($\beta = -.5$, $p < .05$), and English frequency is significant as well ($\beta = -.7$, $p < .05$), but the difference between false friends and controls is no longer significant ($\beta = -.6$, $p = .06$). Model comparison further indicates that including English word frequency significantly improves model fit ($\chi^2(1) = 4.49, p = .034$).

Overall, these results indicate that effects of bilingual exposure differ between L1 and L2. In the *friend study*, the facilitation effect for friends over controls is accounted for by sentence-language frequency (English), with seemingly no contribution from the other-language frequency (Dutch). However, correlation of fixed effects sheds light on this: there is a large correlation between word category and Dutch frequency ($r = -.9$), essentially signaling to the regression model that words that do not appear in Dutch (i.e. have a Dutch frequency of 0) are English control words. For friends, the English and Dutch frequencies are highly correlated in the model ($r = -.7$), as well as in the data (see Figure 2). This means that English frequency sufficiently predicts variance between control and friend words. The variance that could be explained by Dutch frequency is already explained by the English one, making Dutch frequency redundant.

In contrast, the facilitation effect found for the *false-friend study* seems to be driven by both the frequency in the language of the sentence and frequency in the "other" language. In this model, word category and English frequency are correlated as

well ($r = -.9$), again pertaining to Dutch control words that have a 0 frequency in English. However, the model correlation between Dutch and English false friends frequency is weak ($r = -.15$, see also Figure 2 for data correlation). This is not surprising since in the case of false friends, the same surface form is mapped to different meanings and possibly to different parts of speech, with different distributional patterns across Dutch and English. We hypothesize that this is the reason both Dutch and English frequencies remain significant in the *false-friend study*.

Generally, our results provide support for the cumulative frequency hypothesis for both friend and false friend words. Both languages contribute to the facilitation effect, but for friends it is harder to disentangle the effects of Dutch and English as frequency sources due to their high correlation, especially across highly similar languages (Schepens et al., 2013).

6. Discussion

This work asked whether cross-lingual transfer in bilingual Transformer LMs produces patterns that resemble cross-lingual activation in bilingual readers, focusing on words with overlapping surface forms across Dutch and English. To test this, we trained LMs on L1 Dutch and L2 English, and evaluated the models on experimental stimuli from studies of Dutch-English bilinguals, which included words with overlapping surface forms between the two languages. Designing a cognitively plausible vocabulary in bilingual LMs is an open question. We addressed it by manipulating lexical sharing directly by controlling which form-overlapping words received an embedding shared between languages versus language-specific embeddings.

We find that the bilingual LM keeps the two languages apart in the embedding space, but cross-lingual effects do occur via embedding sharing: friends (i.e. cognate and loan words) show lower surprisal than language-unique control words in the conditions where either each form overlapping word is shared between Dutch and English, or only friends are shared between the two languages. Similarly, false friends have lower surprisal than controls either in conditions where sharing holds for all form overlapping words or just for false friends. These facilitation effects are largely explained by frequency-based signals.

The vocabulary condition that matches human data is one where only friends are shared between Dutch and English: the LM in this condition exhibits facilitation effects for friends without the corresponding facilitation for false friends. Other conditions fail to resemble human bilinguals with respect to processing the different word types: 1) the conditions with full overlap or just for false friends show facilitation for false friends, 2) the condition with minimal overlap (named entities and punctuation) does not show cross-lingual effects. Crucially, simulations of bilingual reading with LMs typically assume full surface overlap (see Section 2.1), which may bias conclusions about cross-lingual transfer and L2 learning.

Among the conditions tested, the one sharing only friends has the best *descriptive adequacy* with respect to empirical findings of bilingual reading of overlapping words. However, it lacks *explanatory adequacy* (Jacobs and Grainger, 1994; Frank, 2021): the tokenizer and LM are based on explicitly coded rules about which forms are friends (and should each have a single embedding) and which are not. We cannot assume this is the mechanism behind cognate facilitation effect found in bilinguals. Importantly, humans learn meaning from grounded language learning, sensorimotor and social cues, and not (just) streams of text (Warstadt and Bowman, 2022; Chang and Bergen, 2022). Yet, when the LM was left to train without explicit coding of (false) friends, it resulted in behavior that does not reflect that of bilingual humans.

One of the open questions is why shared embeddings result in facilitation for false friends, rather than interference. One possibility is that surprisal in unconstrained sentence contexts mostly reflects the LM's ability to predict the surface form given the few semantic and lexical cues, and is as such mostly informed by frequency and distributional patterns rather than semantic competition between meanings. Another possibility might lie in the stimuli language (L1, so Dutch) and exposure. The LMs were largely trained on Dutch, while only a quarter of the data was in English. Perhaps the LM did not have a chance to properly learn the English meaning, and therefore does not exhibit interference effects when processing a sentence in Dutch. Generally, bilinguals tend to show cross-lingual effects in L2 reading and less so in L1 (Titone et al., 2011).

One of the main findings confirms the role of frequency in overlapping word forms specifically in the conditions with shared embeddings. When an embedding is shared, it is trained on data from both languages, which increases its effective frequency and typically reduces surprisal. This frequency benefit appears regardless of whether the word is a friend or a false friend. This suggests that overlap primarily provides a learning advantage through shared counts, not through semantic consistency.

The present study uses the GPT2-small Transformer architecture. Cross-lingual activation patterns in cognate facilitation have been observed across a range of architectures, including LSTM models (Winther et al., 2021), Transformers (albeit

shallower than GPT2-small), and simple recurrent networks (Roslund and Matusevych, 2022). This suggests that cognate facilitation is driven less by architectural choice than by training data and procedure: specifically, unbalanced exposure (with greater L1 than L2 input) consistently gives rise to a cognate facilitation effect (Winther et al., 2021; Roslund and Matusevych, 2022).

Furthermore, Winther et al. (2021) report that the order of L1 and L2 presentation plays a crucial role: the cognate effect emerges either when the model is first pretrained on L1 and subsequently trained on a mix of L1 and L2, or (in the absence of pretraining) when L1 data comprises the first 75% of the epoch and L2 data the remainder. We follow the latter training procedure, but it would be valuable to test the former with different vocabulary designs as well.

Our results can be situated within the framework of the BIA+ model (Dijkstra and van Heuven, 2002), which is one of the most influential computational accounts of bilingual word recognition. The model includes orthographic, phonological and semantic lexical representations with bidirectional connections between the representation types. In contrast, in our LMs, word representations are encoded at the embedding layer, where a word form either has a single (shared) or two language-specific embeddings. With this difference in mind, the vocabulary condition that best matches human bilingual data is broadly consistent with the BIA+ assumptions that cognates have a special shared representation due to their overlap in both form and meaning, while interlingual homographs (false friends) are represented by two separate orthographic entries. The role of form frequency in our results relates to the BIA+ model as well, where frequency is implemented as resting activation levels. Importantly, in BIA+, frequency and form overlap are separable sources of facilitation, while they are harder to disentangle in our LMs, where sharing an embedding both increases a word's effective frequency and enables cross-lingual transfer at the same time. Further comparisons between BIA+ and bilingual LMs, for instance using tasks that more directly probe semantic competition or that vary sentence context constraints, could shed more light on how well these two types of models align in explaining cross-lingual activation.

7. Conclusion

We presented a controlled study of vocabulary sharing in bilingual Dutch-English Transformer language models, targeting the question of whether bilingual LMs exhibit human-like cross-language activation. The results show that bilingual LMs generally keep Dutch and English contexts distinct, and cross-lingual effects arise primarily when the model shares embeddings. Under embedding sharing, overlapping forms become more predictable than language-unique controls, resulting in facilitation for both friends and false friends. This facilitation is largely explained by frequency and does not depend on whether the form-meaning mapping is shared across languages.

Among the evaluated vocabulary conditions, only the condition with shared embeddings for cognate and loan words reproduces the qualitative human pattern of facilitation for these words. Overall, these findings suggest that neural language models can reproduce some cross-lingual activation patterns, however, their alignment with bilingual reading behavior depends critically on how lexical overlap is treated in the encoding step.

8. Limitations

Our findings are currently based on a small-scale manipulation, which affected between 2.3% and 4.3% of the vocabulary depending on the condition. We did not explore how lexical interventions affect other aspects of cross-lingual transfer, e.g. syntactic processing.

We tested the models on two sets of psycholinguistic stimuli, but only the *friend study* (Bultena et al., 2014) has previously shown effects in bilinguals. The stimuli from the *false-friend* experiment (Huisman, 2025) did not produce cross-lingual effects, limiting the extent to which model behavior can be compared to behavioral data. This points to a general challenge of scarcity of bilingual human responses to test the human-likeness of LMs.

To measure cross-lingual effects in LMs, we use surprisal and similarity between representations, but other measures might capture patterns that correspond to human behavior, for example attention or the activation from the feed-forward networks (Oh and Schuler, 2022; Kuribayashi et al., 2025).

9. Acknowledgements

We thank Martin van Harmelen and Mayank Jobanputra for their help in implementing tokenizer and language model training. We also thank two anonymous reviewers for their constructive feedback.

Iza Škrjanec is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

Iza Škrjanec benefited from a Short-Term Scientific Mission funded by COST Action MultiplEYE (CA21131), supported by COST (European Cooperation in Science and Technology).

10. Bibliographical References

Tatsuya Aoyama and Nathan Schneider. 2024. Modeling nonnative sentence processing with L2 language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4927–4940, Miami, Florida, USA. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Sybrine Bultena, Ton Dijkstra, and Janet G. van Hell. 2014. Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, 67:1214–1241.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating critical period effects in language acquisition through neural language models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Ton Dijkstra and Walter van Heuven. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5:175 – 197.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Stefan L. Frank. 2021. Toward computational models of multilingual sentence processing. *Language Learning*, 71:193–218.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Liv J. Hoversten and Matthew J. Traxler. 2016. A time course analysis of interlingual homograph processing: Evidence from eye movements*. *Bilingualism: Language and Cognition*, 19(2):347 – 360.

Merit Huisman. 2025. The role of phonology on recognising interlingual homographs in L1 sentence context. Master's thesis, Radboud University, Nijmegen, the Netherlands.

Arthur M. Jacobs and Jonathan Grainger. 1994. Models of visual word recognition: Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, 20(6):1311–1334.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Julie Kallini, Dan Jurafsky, Christopher Potts, and Martijn Bartelds. 2025. False Friends are not foes: Investigating vocabulary overlap in multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21138–21154, Suzhou, China. Association for Computational Linguistics.

Judith F. Kroll, Paola E. Dussias, Cari A. Bogulski, and Jorge R. Valdes Kroff. 2012. Juggling two languages in one mind: What bilinguals tell us about language processing and its consequences for cognition. volume 56 of *Psychology of Learning and Motivation*, pages 229–262. Academic Press.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Justin Lauro and Ana I. Schwartz. 2017. Bilingual non-selective lexical access in sentence contexts: A meta-analytic review. *Journal of Memory and Language*, 92:217–233.

Els Lefever, Sofie Labat, and Pranaydeep Singh. 2020. Identifying cognates in English-Dutch and

French-Dutch by means of orthographic information and cross-lingual word embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4096–4101, Marseille, France. European Language Resources Association.

Kristin Lemhöfer and Ton Dijkstra. 2004. Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition*, 32:533–550.

Pietro Lesci, Clara Meister, Thomas Hofmann, Andreas Vlachos, and Tiago Pimentel. 2025. Causal estimation of tokenisation bias. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28325–28340, Vienna, Austria. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177.

Maya Libben and Debra A. Titone. 2009. Bilingual lexical access in context: evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(2):381–90.

Katherine J. Midgley, Phillip J. Holcomb, and Jonathan Grainger. 2011. Effects of cognate status on word comprehension in second language learners: An erp investigation. *Journal of Cognitive Neuroscience*, 23(7):1634–1647.

Miyu Oba, Tatsuki Kuribayashi, Hiroki Ouchi, and Taro Watanabe. 2023. Second language acquisition of neural language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Irina Pivneva, Julie Mercier, and Debra Titone. 2014. Executive control modulates cross-language lexical activation during l2 reading: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(3):787–796.

Eva D. Poort and Jennifer M. Rodd. 2019. A Database of Dutch-English Cognates, Interlingual Homographs and Translation Equivalents. *Journal of Cognition*, 2(1):15.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rasmus Roslund and Yevgen Matusevych. 2022. Modeling sentence processing effects in bilingual speakers: A comparison of neural architectures. In *Annual Meeting of the Cognitive Science Society*.

Job Schepens, Ton Dijkstra, Franc Grootjen, and Walter J. B. van Heuven. 2013. Cross-language distributions of high frequency and phonetically similar cognates. *PLOS ONE*, 8(5):1–15.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Kristof Strijkers, Albert Costa, and Guillaume Thierry. 2010. Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. *Cerebral Cortex*, 20(4):912–928.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Debra A. Titone, Maya Libben, Julien Mercier, Veronica Whitford, and Irina Pivneva. 2011. Bilingual lexical access during l1 sentence reading: The effects of l2 knowledge, semantic constraint, and l1-l2 intermixing. *Journal of experimental psychology. Learning, memory, and cognition*, 37 6:1412–31.

Madeleine Voga and Jonathan Grainger. 2007. Cognate status and cross-script translation priming. *Memory & Cognition*, 35(5):938–952.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*.

Irene Elisabeth Winther, Yevgen Matusevych, and Martin J. Pickering. 2021. Cumulative frequency can explain cognate facilitation in language models. In *Annual Meeting of the Cognitive Science Society*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Aditya Yadavalli, Alekhya Yadavalli, and Vera Tobin. 2023. SLABERT talk pretty one day: Modeling second language acquisition with BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11763–11777, Toronto, Canada. Association for Computational Linguistics.

Appendix

A. Translations

English translations of Dutch sentences in Section 4.3.

2a The images of the fire ...

2b The images of the bullet ...

B. Language Model Hyperparameters

The context window size was reduced to 256. We used gradient accumulation with the effective batch size of 512. Weight decay was set to 0.1, the learning rate to 5e-4 with a cosine learning rate scheduler. The initial 1k steps of training were used as warm-up. All models were trained under the same random seed.

C. Vocabulary and LM Sizes

Condition	Vocabulary size	Num. parameters
A	77,369	144,672,000
B	141,877	194,214,144
C	144,170	195,975,168
D	144,681	196,367,616

Table 1: Vocabulary size and number of trainable parameters for the language model trained in each condition.

D. Training and Evaluation Loss

Figure 5 shows the loss values on the training and test sets at the end of each of 6 epochs in each condition. The training loss combines the values for the Dutch and English portion of the training set. We show separate values for the test set of each language. At the very end of each epoch, the LMs have been exposed to English more recently, so the test loss for English is quite low in all conditions, but high for Dutch, which indicates some level of attrition of Dutch. Compared to the other conditions, condition A (Full Overlap) shows a relatively higher training loss, but also lower test loss for Dutch.

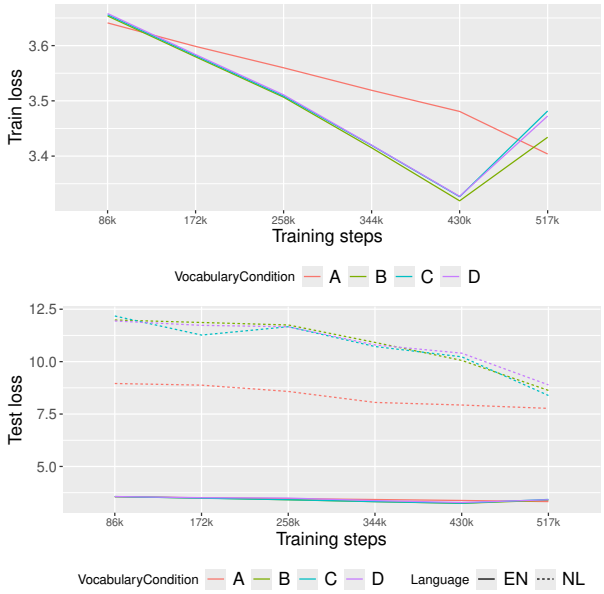


Figure 5: Training and evaluation loss.

E. Diagrams for 5.1

Figure 6 illustrates mean pooling over the preceding context of target words to obtain μ_C. Figure 7 shows how we obtained the mean-pooled token representation μ_W.

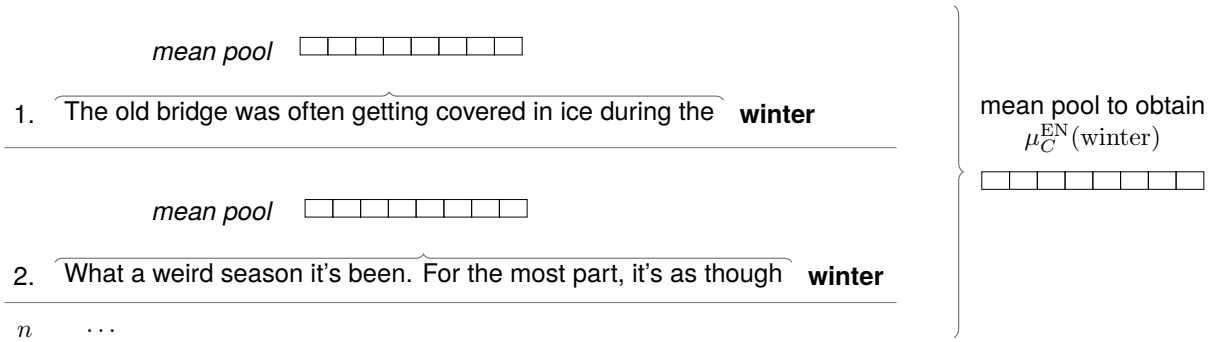


Figure 6: Mean pooling over the embeddings in the preceding tokens withing each sentence ($n = 500$). Mean pooling over these to obtain a context embeddings from English samples μ_C^{EN} for the token *winter*. The same was performed over Dutch sentences to obtain $\mu_C^{NL}(winter)$.

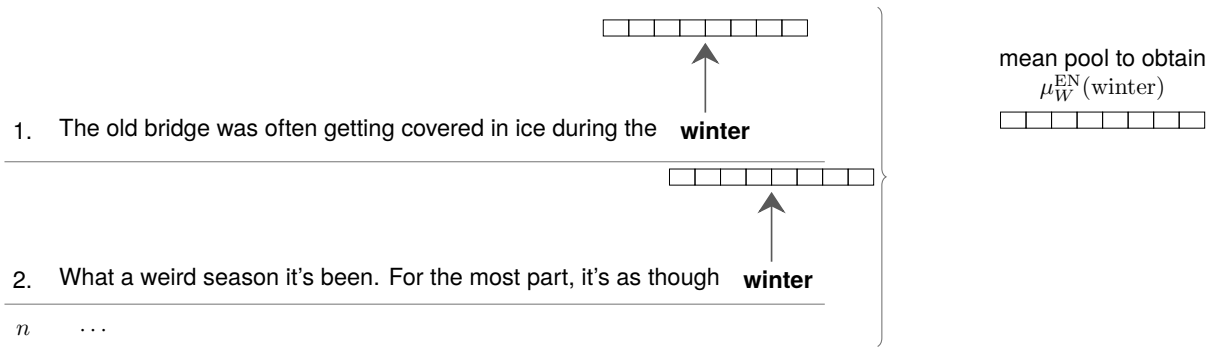


Figure 7: Mean pooling over the token representations of the target word in n English sentences ($n = 500$) to obtain $\mu_W^{EN}(winter)$. Similarly, n Dutch sentences are used to calculate $\mu_W^{NL}(winter)$.

Brain-to-Text Decoding with Brain Atlases and Brain Foundation Models

Haruka Akama[1], Ryo Yoshida[1], Max Müller-Eberstein[1,2], Yohei Oseki[1]
[1]The University of Tokyo, Japan
[2]IT University of Copenhagen, Denmark
{haruka-akama, yoshiryo0617, iijimax, oseki}@g.ecc.u-tokyo.ac.jp

Abstract

Brain-to-text decoding using pre-trained Large Language Models (LLMs) as decoders has recently begun to enable open-ended, sentence-level generation directly from neural recordings. However, despite these decoder-side advances, a critical bottleneck still lies in the encoder which compresses extremely high-dimensional fMRI data into compact brain representations. This study investigates the contrast between three primary encoding strategies: an *expert-driven mapping*, **Brain Atlas**, of fMRI signals to 424 brain regions; *data-driven mappings*, namely **PCA**-reductions, as typically used in prior work; and a *hybrid* **Brain Foundation Model** (**BFM**), combining an expert-driven atlas encoding with large-scale data-driven pretraining. Experiments on the Narratives fMRI dataset demonstrate that including an expert-driven mapping significantly outperforms the purely data-driven PCA-configurations across all evaluation metrics. Additionally, we find that adding the computationally expensive BFM on top of the heuristic Brain Atlas encoding yields no statistically significant gains. Our ablation analyses reveal that this divide is driven by Brain Atlas features smoothing cross-subject and temporal variance, while retaining a sparse importance profile which prioritizes signals from brain subnetworks related to language processing. These findings highlight expert-driven, spatially-aware feature aggregation as a key direction for future decoding performance gains.

Keywords: brain-to-text decoding, fMRI, brain atlas, brain foundation model, large language model

1. Introduction

Brain-to-text decoding aims to reconstruct natural language directly from neural activity. As such, it holds significant promise for Brain–Machine Interface (BMI) applications, particularly for individuals with severe motor impairments such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome, who are unable to express their intended language through speech or movement. While early work focused on word-level recognition (Pereira et al., 2018), the recent adoption of Large Language Models (LLMs) as decoders has begun to enable open-ended, sentence-level generation from fMRI recordings (Ye et al., 2025).

Despite this progress on the decoder side, a critical bottleneck on the encoder side has gone largely unexplored. Prior work has primarily relied on Principal Component Analysis (PCA), or custom down-projections, to compress the extremely high-dimensional brain signal (often hundreds of thousands of voxels) into a compact representation that the decoder can process. As a result, it remains an open question as to how *encoder design* affects decoder performance, as well as how the *utilization* of the raw brain signals differs across each encoding method.

To shed light on these questions, this study contrasts three general strategies for encoding raw brain data from functional Magnetic Resonance Imaging (fMRI; Figure 1): an *expert-driven mapping*, **Brain Atlas**, based on brain atlases that define functional boundaries in the brain based on neuroscience literature; *data-driven mappings*, namely **PCA**-reductions of the fMRI data following prior work (Ye et al., 2025); and a *hybrid mapping*, namely the **Brain Foundation Model** (**BFM**; Caro et al., 2024), which further encodes the expert-driven Brain Atlas representation through a highly

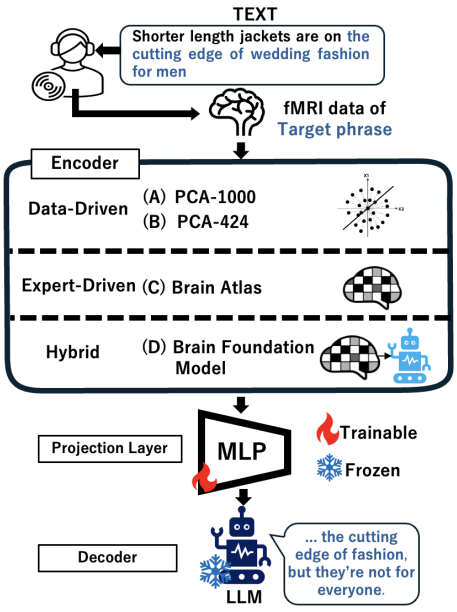


Figure 1: Overview of the brain-to-text decoding pipeline, within which we study four encoder designs: Purely data-driven (A) PCA-1000, (B) PCA-424; purely expert-driven (C) Brain Atlas; and a hybrid (D) Brain Foundation Model (Section 3).

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 124–144
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

data-driven contextualized embedding model, pre-trained on 6,700 hours of fMRI recordings.

Towards a better understanding of how brain signal encoder design affects language decoder performance, this work contributes:

- Encoder design ablations covering expert-driven Brain Atlas mappings, data-driven PCA reductions, hybrid BFM embeddings, and two control settings (Section 3);
- Brain-to-text decoding experiments, which show that expert-driven Brain Atlas mappings outperform PCA and control baselines, as well as BFM on efficiency (Section 5);
- A regional ablation analysis, which shows that Brain Atlas dimensionality reduction exhibits a sparse contribution structure focused on language processing areas, whereas BFM and PCA features are broadly distributed and include negative contributions that impede performance (Section 6);

Our results establish expert-driven compression as a more effective design choice for brain-to-text decoding, and highlight the issues underlying current, purely data-driven encoder designs.

2. Related Work

Using non-invasive neural recordings, such as fMRI, to decode different types of information has been a formidable challenge across modalities, owing to the inherent trade-off between temporal and spatial resolution in neuroimaging and the significant modality gap between neural activity and externalized concepts. Due to the high information density of brain data, the direct reconstruction of concrete concepts has only recently been enabled with the advent of more advanced statistical machine learning methods. This includes the classification of mental states (Haynes and Rees, 2006), as well as early formulations of how to potentially decode image information from the visual cortex (Thirion et al., 2006).

For the reconstruction of linguistic information, initial studies similarly focused on the closed-set problem of word-level classification (Mitchell et al., 2008), with later studies introducing phoneme-level reconstruction, in order to enable the decoding of unseen words (Pei et al., 2011). For an even higher level of flexibility, Pereira et al. (2018) further introduced universal decoders generating semantic vectors via ridge regression. At the sentence level, Tang et al. (2023) added a selective decoding framework to a pre-trained language model to enable higher-quality decoding. In the end, all of these methods nonetheless rely on candidate sets to delimit the decoding space, and do not perform fully free-form text generation.

With the advent of more capable decoder models, the open-endedness, granularity, and quality of reconstructed outputs has increased dramatically. The introduction of diffusion models (Ho et al., 2020) in computer vision has lead to full brain-to-image decoding pipelines with high fidelity (Takagi and Nishimoto, 2023; Chen et al., 2023; Wang et al., 2024). Meanwhile, the introduction of LLMs has begun to enable open-ended brain-to-text decoding at the sentence level: Ye et al. (2025) proposed one of the first open-ended generation pipelines, which map PCA-reduced fMRI data to the embedding space of an LLM, in order to decode brain activity from reading and listening to stories back to full sentences. To verify that the model exploits brain-specific information, they additionally introduced a permuted control model with randomly mismatched signals. The unpermuted model significantly outperformed this control with a win rate of 66.5%, confirming that the decoder leverages genuine neural information. In our experiments, we adopt this pipeline as our direct baseline.

While advancements on the decoder side have driven continual improvements to generation quality, the encoder side has remained fairly constant. Due to the extremely high dimensionality of fMRI data (e.g., 250k voxels for a resolution of $3\times3\times4$ mm), typically, some form of dimensionality reduction (e.g., PCA) is applied, before the resulting compressed representation is passed through a learned projection (e.g., small MLP), which maps the brain activation vector to the decoder's embedding space. This configuration has largely remained unchanged across all aforementioned configurations, yet presents a critical and understudied bottleneck to the quality of the decoded outputs. As such, this work explicitly studies the effects of different encoder design choices on the task of open-ended, sentence-level brain-to-text decoding.

3. Methodology

We first outline the brain-to-text decoding pipeline of Ye et al. (2025), and define four encoder designs that are either data-driven, expert-driven, or both, to evaluate in the subsequent experiments.

Task Formulation. A raw fMRI scan is represented as a four-dimensional array $\mathcal{X} \in \mathbb{R}^{X\times Y\times Z\times T}$, where (X, Y, Z) denote the spatial dimensions and T the number of time steps. The spatial dimensions are first flattened to yield a matrix $M \in \mathbb{R}^{T\times V}$ ($V = X \cdot Y \cdot Z$); taking a single row gives $\mathbf{x} \in \mathbb{R}^V$, the fMRI signal at one time step. Given $\mathbf{x}$ and a text context $y_{<t} = (y_1, \ldots, y_{t-1})$, the goal of brain-to-text decoding is to generate the

subsequent tokens $y_t, y_{t+1}, \ldots$ that approximate the speech perceived by the subject.

As illustrated in Figure 1, the decoding pipeline consists of three stages:

1. **Encoder** $f_{\text{enc}} : \mathbb{R}^V \to \mathbb{R}^d$ reduces the high-dimensional fMRI signal to a compact d-dimensional representation $\mathbf{z} = f_{\text{enc}}(\mathbf{x})$. In this work we compare four encoder designs that differ in their dimensionality reduction strategy, i.e., whether they are purely data-driven, expert-driven, or hybrid.

2. **Projection** $g : \mathbb{R}^d \to \mathbb{R}^{d_{\text{LLM}}}$ maps the encoder output into the LLM's embedding space, yielding the brain embedding $\mathbf{z}_{\text{brain}} = g(\mathbf{z})$.

3. **LLM Decoder** generates text autoregressively, conditioned on $\mathbf{z}_{\text{brain}}$ and the text context. Added special tokens $\langle$`brain/`$\rangle$ and $\langle$`/brain`$\rangle$ mark the boundaries of the brain-derived embedding in the LLM input sequence.

Since neural activity during language comprehension is a continuous, distributed process across multiple cortical regions (Heald et al., 2023), as well as due to high amounts of noise during the recording process, information loss during the encoding phase can become a critical bottleneck for decoding performance. Therefore, the central design question of this work is the choice of encoder f_{enc}. Specifically, how to most effectively compress the high-dimensional fMRI signal while preserving the linguistically relevant neural information needed for language generation.

Encoder Designs. We organize our experimental configurations along two primary axes: **data-driven** dimensionality reduction, and **expert-driven** anatomical mapping. Within this space, we evaluate the following four encoder configurations:

- **(A) PCA-1000 (Data-driven):** fMRI signals are reduced to 1,000 dimensions using PCA, then mapped by the projection layer. This configuration is used in Ye et al. (2025) (Section 3.1).

- **(B) PCA-424 (Data-driven):** fMRI signals are reduced to 424 dimensions using PCA, then mapped by the projection layer. This configuration serves as a matched-dimension PCA baseline to (C) Brain Atlas (Section 3.1).

- **(C) Brain Atlas (Expert-driven):** fMRI signals are converted into 424 mean activations grouped by parcellating the brain into the corresponding regions of interest (ROIs) using the AAL-424 atlas (Nemati et al., 2020), then mapped by the projection layer (Section 3.2).

- **(D) Brain Foundation Model; BFM (Hybrid):** The same 424 ROI-wise signals as in (C) are fed to a contextual embedding model, pre-trained on fMRI data (Caro et al., 2024), followed by the projection layer (Section 3.3).

In terms of purely data-driven approaches, PCA-based dimensionality reduction is often used in prior work, including Ye et al. (2025). In contrast, brain atlases provide an expert-driven approach that parcellates the brain into anatomically informed ROIs, thereby preserving functional and anatomical organization. Additionally, the recent introduction of pre-trained brain foundation models such as Caro et al. (2024)'s BFM offers a hybrid encoding approach, combining an initial expert-driven atlas mapping with large-scale data-driven pretraining.[1] Despite these fundamental differences in design principles, the impact of the encoding strategy on decoding performance has not yet been investigated in prior work. We argue that one key to improving brain-to-text decoding lies in the design of this encoding stage.

Control Settings. To additionally evaluate the effectiveness of using any brain signal, we further employ two control designs based on Ye et al. (2025):

- **(E) LLM-only:** The LLM receives text embeddings only, with no fMRI input, i.e., special tokens with no intermediate brain embeddings. This condition isolates the contribution of brain signals by providing a text-only upper bound.

- **(F) Random:** fMRI signals are replaced with pre-generated random 424-dimensional vectors, which are passed through the projection layer. Thus, the model architecture is identical to (C), but the brain input contains only noise. This condition tests whether the model exploits any meaningful brain signal.

3.1. PCA Baselines

Principal Component Analysis (PCA) is a parameter-free dimensionality reduction method, that is driven purely by the variance characteristics of its input data. As it is used in Ye et al. (2025), it further serves as our direct baseline.

Before PCA, fMRI data are normalized: for each voxel its temporal mean is subtracted, and the result is divided by a per-voxel global standard deviation estimated by pooling all subjects' recording per dataset. Given the normalized time series $\tilde{X} \in \mathbb{R}^{T_{\text{all}} \times V}$ (all subjects' timesteps for one dataset

[1] While it would be of interest to investigate BFM without the initial brain atlas mapping, this is the only input format accepted by the pre-trained model.

concatenated), PCA finds $d+1$ orthogonal directions $W = [\mathbf{w}_1, \ldots, \mathbf{w}_{d+1}] \in \mathbb{R}^{V\times(d+1)}$ that maximize explained variance:

$$\mathbf{w}_k = \underset{\|\mathbf{w}\|=1,\ \mathbf{w}\perp\mathbf{w}_{1:k-1}}{\arg\max} \ \mathrm{Var}\left(\mathbf{w}^\top \tilde{X}^\top\right).$$

The first principal component (PC_1), which predominantly captures globally correlated physiological noise (e.g., heartbeat, respiration; see Carbonell et al., 2011; Birn, 2012; Murphy and Fox, 2017), is discarded. Each normalized volume is then projected onto a d-dimensional representation using PC_2 through PC_{d+1}:

$$\mathbf{z} = W_{2:d+1}^\top(\tilde{\mathbf{x}} - \boldsymbol{\mu}) \in \mathbb{R}^d,$$

where $\boldsymbol{\mu} \in \mathbb{R}^V$ is the column mean of $\tilde{X}$ and $W_{2:d+1}$ collects columns 2 through $d+1$ of W. We evaluate two output dimensionalities: (A) $d = 1{,}000$, replicating Ye et al. (2025); and (B) $d = 424$, matching the Brain Atlas feature dimensionality.

A defining property of PCA is that its components are determined purely by variance maximization across the full voxel space, without reference to any anatomical parcellation. This means that it can be applied in a purely data-driven manner, without any expert annotation. However, it may also capture high-variance components that reflect non-neural sources—such as head-motion artifacts and physiological noise—rather than functionally organized neural signals.

3.2. Brain Atlas

Generally, a brain atlas provides a static mapping of normalized voxel positions to anatomically informed brain regions. For our experiments, we adopt the Automated Anatomical Labeling 424 atlas (AAL-424; Nemati et al., 2020), which parcellates the brain into $K = 424$ non-overlapping ROIs within the normalized Montreal Neurological Institute (MNI) coordinate system (Figure 2). Since AAL-424 is also used in the initial encoding step of the BFM (Caro et al., 2024), this configuration allows us to compare the performance of an expert-driven mapping with and without an additional data-driven encoder model. Unlike PCA, whose partition of the voxel space is determined entirely by data-driven variance maximization, the atlas partition is fixed *a priori* based on anatomical boundaries (major sulci) grounded in neuroscience knowledge.

Formally, a brain atlas defines a partition $\mathcal{A} = \{R_1, R_2, \ldots, R_K\}$ of the voxel index set, where $R_k \subseteq \{1, \ldots, V\}$ is the set of indices assigned to the k-th ROI. The R_k are disjoint, and $\bigcup_k R_k \subseteq \{1, \ldots, V\}$. For each ROI, the encoder computes the mean activation over its constituent voxels:

$$z_k = \frac{1}{|R_k|} \sum_{v\in R_k} x_v, \quad k = 1, \ldots, K.$$

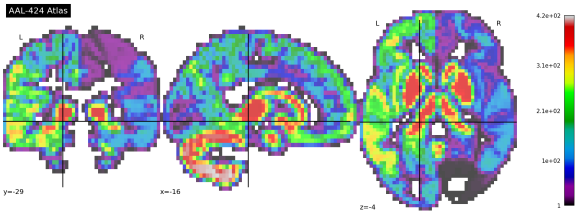


Figure 2: The AAL-424 brain atlas in MNI space (coronal, sagittal, and axial slices). Each color denotes a distinct ROI indexed from 1–424. The fine-grained parcellation covers both cortical and subcortical structures.

The full encoder output is the concatenation of all ROI signals:

$$\mathbf{z} = [z_1, z_2, \ldots, z_K]^\top \in \mathbb{R}^K, \quad K = 424.$$

As with PCA, the resulting ROI signals are normalized per dataset and per subject: for each ROI, its temporal mean (within the recording) is subtracted and the result is divided by a per-ROI global standard deviation estimated from that subject's recordings within the dataset. Crucially, each dimension z_k corresponds to a specific, anatomically labeled brain region, making the resulting representation directly interpretable. Additionally, the normalization to MNI space corrects for cross-subject morphological variability, ensuring that ROI R_k refers to the same anatomical structure across individuals and enabling meaningful cross-subject aggregation.

Beyond the ROI-based AAL-424 parcellation, the Akiki-Abdallah hierarchy groups the same 424 ROIs into 24 modules (AA-24; Akiki and Abdallah, 2019) nested within 7 canonical parent networks (Visual, Ventral Salience, Central Executive, Sensorimotor, Default Mode, Dorsal Salience, and Subcortical), which we use for our region-level ablation analysis in Section 6.

3.3. Brain Foundation Model

In order to encode brain recordings in a denser manner, recent work has proposed mirroring masked language models (Devlin et al., 2019) and vision transformers (Dosovitskiy et al., 2021) to train brain foundation models (BFMs). In this work, we leverage a 13M-parameter BFM by Caro et al. (2024), pre-trained on approximately 6,700 hours of non-linguistic fMRI data through self-supervised learning. This particular BFM offers a hybrid expert and data-driven configuration for our experiments, as it first maps the raw fMRI data to the AAL-424 atlas, before embedding it further. In the pretraining phase, activation patches from random brain regions are masked, and the model learns to reconstruct them from the remaining context, thereby capturing inter-regional functional relationships and

spatiotemporal co-activation patterns without labeled data.

Following the intended design of BFM, we discard the reconstruction head used in pretraining and use only the frozen Transformer backbone as a feature extractor over AAL-424 parcel-level fMRI signals. The Brain foundation model processes each recording by dividing it into non-overlapping chunks of 200 timepoints. Within each chunk, the time series is further segmented into non-overlapping temporal patches of 20 timepoints. Let $\mathbf{H}_{v,p} \in \mathbb{R}^{d_b}$ denote the final-layer hidden vector for voxel (parcel) index $v = 1, \ldots, V$ at patch index p, after removing the CLS token. We then apply mean pooling over the spatial (voxel/parcel) dimension to obtain a fixed-length patch representation $\mathbf{z}_p \in \mathbb{R}^{d_b}$:

$$\mathbf{z}_p = \frac{1}{V} \sum_{v=1}^{V} \mathbf{H}_{v,p}.$$

For each sample, the patch representation corresponding to that sample is used as a single brain token and finally passed to the projection layer. Through this configuration, we examine the effects of leveraging external data-driven knowledge for the informativeness of brain activation encoding.

4. Experimental configuration

Next, we describe the datasets, models, training procedure, and evaluation metrics used to compare the effect of the four encoder designs (A–D) defined in Section 3 on brain-to-text decoding performance.

Models. Following the sentence-level brain-to-text decoding pipeline of Ye et al. (2025), the projection consists of two hidden layers with ReLU activations. Similarly, the final LLM decoder uses `GPT-2 small` (Radford et al., 2019) (124M parameters) across all experimental settings. The brain foundation model (Caro et al., 2024) comprises approximately 13M frozen parameters and was used exclusively in configuration (D).

Training Data. For a direct comparison with Ye et al. (2025), we utilize data from 27 subjects in the *Narratives* dataset (Nastase et al., 2021), consisting of fMRI recordings of individuals listening to approximately 4.6 hours of auditory stories. Specifically, we used five stories: *pieman*, *lucy*, *notthefallintact*, *slumlordreach*, and *tunnel*. Recordings were acquired on a Siemens Skyra 3T scanner with a spatial resolution of 3×3×4 mm and a repetition time of 1.5 s. For each subject, data were split into training (60%), validation (20%), and test (20%) sets based on temporal order. The dataset underwent standard preprocessing via `fMRIPrep 20.0.5` (Esteban et al., 2019), including motion correction and normalization to `MNI152NLin2009cAsym` space.

Training Procedure. The parameters of the LLM decoder are kept frozen, with only the projection layer and the embeddings of the special tokens ⟨`brain/`⟩ and ⟨`/brain`⟩ around $\mathbf{z}_{\text{brain}}$ being subject to optimization. In configuration (D), the parameters of the brain foundation model are also kept frozen.

The input embedding sequence $\mathbf{X}$ to the LLM is defined as:

$$\mathbf{X} = \left[\langle \texttt{brain/} \rangle,\ \mathbf{z}_{\text{brain}},\ \langle \texttt{/brain} \rangle,\ \mathbf{e}_1, \ldots, \mathbf{e}_{t-1} \right]$$

where $\mathbf{z}_{\text{brain}}$ denotes the encoder output used as the brain embedding (i.e., PCA features in configurations (A) and (B), brain atlas-based features in (C), and BFM-derived features in configuration (D)). Here, $\mathbf{e}_1, \ldots, \mathbf{e}_{t-1}$ are the token embeddings of the preceding context provided to the decoder. We minimize the following cross-entropy loss to maximize the likelihood of the ground-truth sentence continuation:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{M} \sum_{t=1}^{M} \log\, p(y_t \mid y_{<t}, \mathbf{X})$$

In addition to encoding linguistically relevant neural information, $\mathbf{z}_{\text{brain}}$ needs to conform to the decoder model's embedding space. While this alignment can be learned jointly with the main $\mathcal{L}_{\text{CE}}$ objective, Ye et al. (2025) find that some of their configurations benefit from an additional warm-up phase, in which the projection layer is first trained to reduce the distance to the target embedding space. As such we report additional ablations including this warm-up phase in Appendix A.2.

Hyperparameters. Once again following Ye et al. (2025), the mini-batch size was set to 1, and the maximum number of epochs was set to 100. We used the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 10^{-3}. To prevent overfitting, the dropout rate was set to 0.5. Training was terminated if no improvement in validation performance was observed for 10 consecutive epochs.

Evaluation Metrics. We employ the following standard metrics for text generation: **BLEU-1** (Papineni et al., 2002) to evaluate n-gram overlap based on "Precision" with a brevity penalty; **ROUGE-1/L** (Lin, 2004) to evaluate "Recall" at the levels of unigrams and the longest common subsequence; and **WER** (Morris et al., 2004) to measure the word-level error rate based on substitutions, deletions, and insertions. For testing the statistical significance of metric differences, we employ the two-sided Wilcoxon signed-rank test with Pratt handling

of zeros (Pratt, 1959), continuity correction, and Benjamini-Hochberg FDR correction for multiple comparisons (Benjamini and Hochberg, 1995).

5. Results

Table 5 presents the results for our four encoder configurations. Across all metrics, the two encoder designs incorporating expert-driven brain atlas inputs (C and D) consistently outperform the purely data-driven PCA approaches (A and B). Surprisingly, the hybrid BFM approach yields no statistically significant gains over using only the brain atlas mapping. Both Brain Atlas and BFM further significantly outperform the control settings (E and F), implying that information from the brain signal is being used to improve decoding performance.

We note that the absolute metric levels (e.g., WER ≈ 0.94) are consistent with the current state of the art in non-invasive fMRI-to-text decoding (Tang et al., 2023; Ye et al., 2025), and reflect inherent limitations of the recording modality: fMRI integrates neural activity via an indirect hemodynamic proxy, which is limited to a time-resolution of 1.5s per volume. In stark contrast, upper bound performance on brain-to-text decoding can reach down to 3% WER (Makin et al., 2020), but only if invasive electrodes are directly implanted in the brain (iEEG with a time-resolution of 200ms), the task is speaking the text (and not listening), and the output vocabulary stems from a closed set.

Brain Atlas significantly outperforms both PCA baselines, and control settings. The Brain Atlas encoding approach achieves BLEU-1 = 0.1276, ROUGE-1 = 0.1186, ROUGE-L = 0.1115, and WER = 0.9439. Compared to PCA-1000, which achieved the highest performance in Ye et al. (2025), Brain Atlas sees improvements of $+16.9\%$ in BLEU-1, $+14.0\%$ in ROUGE-1, and $+13.5\%$ in ROUGE-L. Similarly, the relative improvement to PCA-424 is $+11.2\%$, $+8.1\%$, and $+7.6\%$ respectively. At the same time, Brain Atlas also outperforms both control settings, showing that the brain signal it encodes contributes to improvements in decoding performance. All improvements remained significant after FDR correction at $q < 0.05$ (see * in Table 5). To better understand why the Brain Atlas encoding yields these gains, we conduct multiple ablation studies in Section 6.1.

BFM shows no significant improvement over Brain Atlas. BFM, which utilizes an initial brain atlas mapping before further embedding the data using a pre-trained Transformer, achieved BLEU-1 = 0.1282, ROUGE-1 = 0.1195, ROUGE-L = 0.1121, and WER = 0.9435. Compared to Brain Atlas alone, this leads to marginal numerical increases of $+0.5\%$ (BLEU-1), $+0.8\%$ (ROUGE-1), and $+0.5\%$ (ROUGE-L). However, none reached significance after FDR correction. Validation loss was also comparable at 4.870 (C) vs. 4.884 (D). As such, despite being the most data-intensive method, appending a pre-trained BFM does not appear to improve the quality of neural representations for language tasks. We further analyze potential reasons behind this discrepancy in Section 6.2.

PCA-424 significantly outperforms PCA-1000. Counter-intuitively, reducing the number of PCA components from 1,000 to 424 also yielded significant improvements across all metrics after FDR correction at $q < 0.05$ (see † in Table 5). This result further helps disentangle the contributions of dimensionality and encoding strategy to the overall A $\rightarrow$ C improvement. Since both B and C use $d = 424$ dimensions, the B $\rightarrow$ C gap directly measures the benefit of expert-driven encoding. In practice, the effect of changing the encoding strategy ($\Delta\text{Loss}_{\text{B}\rightarrow\text{C}} = -0.623$) is roughly twice as large as the dimensionality-reduction effect alone ($\Delta\text{Loss}_{\text{A}\rightarrow\text{B}} = -0.319$), with both differences remaining independently significant after FDR correction. In Section 6.1, we further examine the mechanistic basis behind the difference in performance between PCA-424 and PCA-1000, as well as between PCA and Brain Atlas.

Qualitative Examples. Table 2 presents two representative decoded outputs from Brain Atlas. Additional qualitative examples are provided in Appendix Table 5. In Example 1, the model correctly recovers *heart* and *pounding*; in Example 2, it correctly recovers *know*. In both cases, generation then drifts into repetitive continuation, a known degeneration pattern of autoregressive decoding (Holtzman et al., 2020; Welleck et al., 2020), which here likely indicates limited additional neural evidence beyond the initially matched tokens. While the overall quality of the results reflects the general difficulty of open-ended sentence recovery from neural signals (Tang et al., 2023; Ye et al., 2025), these examples, together with the quantitative metric improvements, suggest that expert-driven encoding supports the recovery of key lexical items from non-invasive neural signals.

6. Analysis and Discussion

Our results indicate that expert-driven encoder designs significantly outperform purely data-driven approaches. Simultaneously, we find that adding the hybrid BFM on top of the Brain Atlas encodings does not significantly improve decoding performance, despite the added computational complexity. To better understand the reasons underlying

	Cond.	BLEU-1 ↑	ROUGE-1 ↑	ROUGE-L ↑	Loss ↓	WER ↓
Encoders	(A) PCA-1000	0.1092	0.1040	0.0982	5.812	0.9513
	(B) PCA-424	0.1148†	0.1097†	0.1036†	5.493†	0.9486†
	(C) Brain Atlas	0.1276*	0.1186*	0.1115*	4.870*	0.9439*
	(D) BFM	0.1282*	0.1195*	0.1121*	4.884*	0.9435*
Controls	(E) LLM-only	0.1154	0.1067	0.1045	5.079	0.9481
	(F) Random	0.1148	0.1094	0.1029	5.502	0.9507

Table 1: Brain-to-text decoding performance across encoder designs and control conditions. Conditions (E) and (F) serve as controls that establish, respectively, the language-model-only baseline and the effect of an uninformative brain embedding. The gap between (C) and (E) quantifies the information gain attributable to genuine neural signals (ΔLoss $= -0.209$ nats ≈ 0.30 bits/token). For statistical significance among encoder designs (A–D), * indicates significantly better than both (A),(B),(E),(F); † indicates significantly better than (A); all after FDR correction at $q < 0.05$.

Context	Prediction	Ground Truth
…earth in dead silence it is really quiet so i can really hear my	**heart pounding** and then i hear the sound of my **heart pounding** and then i hear the sound of my **heart pounding** and then i hear the sound of my **heart pounding**	**heart pounding**
…you the straps and you	**know** what i'm talking about you know what i'm talking about you know what i'm talking about you know what i'm talking about you know what i'm	**know** through the

Table 2: Qualitative output examples from the Brain Atlas encoder design. **Blue bold** indicates words matching the ground truth.

these findings, we compare the stability of each encoding strategy across subjects and time. Additionally, we analyze where in the brain the signal used by each encoder originates from.

6.1. How Does Expert Knowledge Improve Encoding Quality?

Inter-Subject Correlation (ISC) measures the Pearson correlation of brain encodings between pairs of subjects listening to the same auditory stimulus. High ISC indicates that the representation captures stimulus-driven, reproducible neural responses rather than subject-specific noise. For each story, we computed ISC across every subject pair, averaging across all dimensions with non-zero standard deviation. The mean ISC of Brain Atlas was 0.1090, approximately $10.8\times$ higher than PCA-424 (0.0101) and $22.7\times$ higher than PCA-1000 (0.0048). This large gap indicates that ROI-based averaging in Brain Atlas preserves neural responses that are reliably evoked across subjects by the same stimulus. Consequently, the projection layer receives a consistent, cross-subject signal that is easier to map to language representations.

Temporal Autocorrelation. Because fMRI measures neural activity indirectly via the BOLD (Blood-Oxygenation-Level-Dependent) signal, brain activations are intrinsically temporally smooth due to the hemodynamic response function (i.e., a gradual increase and decrease in oxygen levels). As such, a high lag-1 autocorrelation corr(x_t, x_{t+1}) for each active dimension indicates a more physiologically plausible signal. The mean autocorrelation of Brain Atlas was 0.5753, markedly higher than PCA-424 (0.4870) and PCA-1000 (0.3016). The decrease in correlation with increasing PCA dimensionality points to the progressive incorporation of high-frequency noise components into higher-order principal components—a factor which we analyze in greater detail next.

Additional Noise in PCA-1000 versus PCA-424. Despite $2.36\times$ as many dimensions, PCA-1000 gains only 7.3 percentage points of explained variance (the fraction of voxel variance captured by PCA), and even underperforms the random baseline. Per-voxel contribution maps in Appendix Figure 6, showing how much each voxel drives the components, further reveal that this marginal gain is concentrated at the extra-brain periphery. This indicates that head-surface artifacts and physiological noise rather than genuine neural signals are being captured by the additional feature dimensions. These noise-derived components elevate the overall contribution floor in PCA-1000, suppressing the contrast of truly informative regions (cerebellum, basal ganglia, occipital lobe) and make the decoder's learning problem harder (see Appendix A.3).

Expert versus Data-driven. Taken together, these two metrics demonstrate that dimensionality reduction preserving the anatomical spatial structure of the brain substantially contributes to brain-to-text decoding performance. PCA extracts information based on variance maximization, without explicitly accounting for anatomical topography or local spatial structure, allowing noise to dominate higher-order components. In contrast, Brain Atlas compresses signals to 424 dimensions via ROI-based averaging, preserving functional and anatomical organization. This delivers more consistent cross-subject neural responses and more physiologically plausible temporal dynamics than either PCA baseline. This anatomically coherent representation likely contributes to the stabilization of the projection into the LLM embedding space and enables higher-quality decoding.

6.2. Why Does BFM Not Improve over Brain Atlas?

Since Brain Atlas substantially outperforms both PCA baselines, a hybrid approach combining an expert-driven encoding with a data-driven brain foundation model pre-trained on large-scale neuroimaging data should intuitively yield further performance improvements. However, BFM provides no significant improvement over Brain Atlas.

Inter-Subject and Temporal Correlation. The mean ISC of BFM is 0.2718—approximately $2.5\times$ higher than Brain Atlas (0.1090)—yet this apparent advantage does not translate into better decoding performance. Similarly, BFM shows an extremely high lag-1 autocorrelation of 0.8987—far above the physiologically plausible BOLD range of Brain Atlas (0.5753) and the highest value among all configurations—with its temporal correlation uniquely narrow and concentrated near 1.0, with near-zero variance across dimensions (Appendix Figure 4). This indicates that BFM output features are uniformly over-smoothed and show essentially no dimension-wise variation in temporal regularity. Together with the ISC finding, this pattern suggests that BFM's representation collapses into a highly redundant signal. We hypothesize this behavior to be an artifact of BFM's attention design, which pools activations across both the temporal, and spatial dimensions (i.e., different brain regions). The high ISC and autocorrelation thus reflect over-smoothing rather than genuine neural consistency, directly limiting the projection layer's ability to map the representation to language space (see Appendix A).

Regional Contributions. To identify which brain regions drive decoding performance in Brain Atlas versus BFM representations, we conducted an ablation analysis, selectively masking activations from different brain regions. We group the original 424 ROIs into 24 functional subnetwork groups (e.g., Language Default Mode, Auditory Somatomotor, Subcortical Cerebellum) according to the AA-24 hierarchy (Akiki and Abdallah, 2019). Next, we mask each of these 24 groups and measured the resulting change in test loss ($\Delta\text{Loss}_g = \mathcal{L}_{\text{ablated}}(g) - \mathcal{L}_{\text{baseline}}$). Positive ΔLoss indicates a positive contribution to decoding, whereas negative ΔLoss indicates interference.

Brain Atlas exhibits a strikingly sparse importance profile, with only 5 of 24 groups showing non-zero ΔLoss (Figure 3). This means these top-5 groups contribute positively, while all other groups have no effect on decoding performance. Notably, the *Language Default Mode Network* (lang DM; superior temporal and inferior frontal gyri) ranks second at $\Delta\text{Loss} = 0.023 \pm 0.004$, confirming that Brain Atlas selectively captures signals from established neural correlates of language comprehension. This sparse, reliably positive profile simplifies the downstream projection task by providing a clean, discriminative input with no conflicting signals to suppress.

In contrast, BFM engages all 24 groups, with 8 showing *negative* ΔLoss, indicating that a substantial portion of its representation actively impedes language decoding (see Appendix A.4). Furthermore, while 3 out of 5 top contributing groups overlap across Brain Atlas and BFM, the latter's groups have a maximum contribution that is 46.5% smaller, while specifically not including the lang DM.

Future Directions for Hybrid Encoders. The success of Brain Atlas highlights expert-driven, anatomically grounded feature aggregation as a simple, performant, and robust encoder design. For hybrid encoders to match this, BFM representations would need equivalent spatial selectivity and linguistic focus. Pretraining on linguistic brain data, fine-tuning, and task-adapted brain atlases that more directly emphasize language-processing regions may all improve future hybrid encoder design, similarly to advances in the text-only modality (Devlin et al., 2019). Finally, a more granular encoding of the time-dimension—e.g., via a learned aggregation such as attention pooling—may also help retain important information for language decoding.

7. Conclusion

This study investigated the effect of different encoder design strategies on brain-to-text decoding, contrasting data-driven PCA baselines against an expert-driven, anatomically informed mapping via the AAL-424 brain atlas. We additionally evaluated a hybrid encoder that supplements the same atlas-

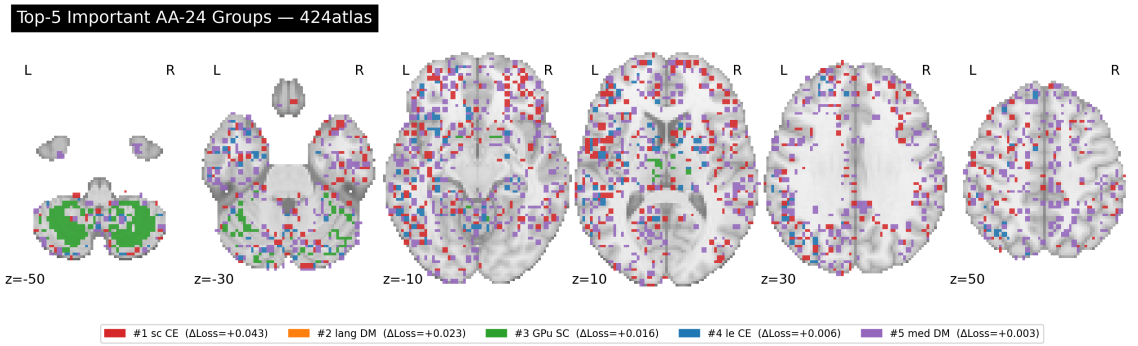


Figure 3: Top-5 contributing AA-24 groups for Brain Atlas (axial slices, $z = -50$ to $+50$ mm): (1) Subcortical Cerebellum (sc CE, $\Delta\text{Loss} = +0.043$, red); (2) Language Default Mode Network (lang DM, $\Delta\text{Loss} = +0.023$, orange); (3) Globus Pallidus/Putamen (GPu SC, $\Delta\text{Loss} = +0.016$, green); (4) Left Cerebellum (le CE, $\Delta\text{Loss} = +0.006$, blue); (5) Medial Default Mode Network (med DM, $\Delta\text{Loss} = +0.003$, purple). The spatial distribution confirms anatomically coherent localization of the most predictive regions.

based representations with a pre-trained brain foundation model, combining expert-driven and large-scale data-driven encoding.

The Brain Atlas approach significantly outperformed both control settings, as well as both PCA baselines on every evaluation metric, demonstrating that its expert-driven, anatomically grounded dimensionality reduction actively encodes brain signal information which substantially benefits decoding performance. Inter-subject correlation and temporal autocorrelation analyses further confirm that ROI-based averaging preserves more stimulus-driven, cross-subject reproducible neural responses and more BOLD-like temporal dynamics than either PCA baseline.

Introducing a BFM on top of the Brain Atlas encoding, however, yielded no statistically significant improvements. Regional ablation analysis reveals that BFM features exhibit a dense, mixed-sign contribution profile that impedes the projection layer, in contrast to the sparse, reliably positive profile of Brain Atlas. This highlights developing language-specialized BFMs as an important future challenge.

Taken together, our findings suggest that encoder-side representation quality—particularly expert-driven, anatomically grounded dimensionality reduction—plays a more decisive role in brain-to-text decoding than model scale alone. More broadly, the results suggest that the choice of dimensionality-reduction strategy is a key factor to improving brain-to-text decoding performance.

8. Limitations

First, due to implementation constraints, our experiments were conducted on five stories sourced from the Narratives dataset (Nastase et al., 2021), using a single decoder (`GPT-2 small`). While we were able to closely reproduce, and thus compare our methods with, the results of (Ye et al., 2025), generalizability to other fMRI datasets, stories with a different narrative structure, as well as additional languages, or decoder architectures remains to be established. Similarly, we employ only one type of brain atlas (AAL-424). This allows us to compare the effect of adding a BFM, which uses the same initial anatomical mapping (Caro et al., 2024). Given AAL-424's effectiveness, future work would likely benefit from investigating other expert-driven mappings of functionally grouped brain regions.

Second, the task of brain-to-text decoding currently lacks standard training procedures. As such, we make certain design decisions based on prior work, but do not exhaust all hyperparameter possibilities. Specifically, for extracting a brain signal embedding from BFM, the mean pooling strategy used to aggregate its spatial token representations is a simple but potentially suboptimal design choice. As proposed in Section 6.2, alternative aggregation strategies such as attention pooling thus warrant investigation in future work. Additionally, we use BFM in a frozen state. Depending on the downstream task, this is the intended use of the model (Caro et al., 2024), however, due to its non-linguistic pretraining, full fine-tuning, may still yield different results. Similarly, the projection layer has been found to benefit from an additional warm-up phase to explicitly align it with the embedding space of the decoder (Ye et al., 2025). In Appendix Table 3, we find that adding this representational alignment pretraining stage retains the performance ordering from our main results in Section 5 sans statistical significance. This suggests that the MSE alignment objective can compensate for encoder-side representation quality to some degree. Nonetheless, the fact that Brain Atlas outperforms the other encoder designs, even without additional pretraining, highlights the cost-effectiveness of an informative encoding strategy. More broadly, an important open

question is whether a fully learnable encoder could further improve performance beyond the frozen configurations studied here. Our frozen-encoder design isolates the effect of the encoding *strategy* itself from that of task-specific optimization, but end-to-end fine-tuning on language tasks may unlock complementary gains, particularly as larger linguistic fMRI datasets become available.

Third, all primary metrics are reported as averages across 27 subjects, and subject-level variance is not explicitly characterized. Although the Wilcoxon signed-rank test operates on per-subject scores and is therefore sensitive to the distribution of individual differences, it is possible that the observed group-level advantages are driven by a subset of subjects with particularly strong neural signal quality. Future work should examine per-subject performance distributions to assess the consistency of the reported effects.

Fourth, the BFM encoder processes brain recordings in non-overlapping chunks of 200 timepoints (300 s at repetition time (TR) = 1.5 s) via global self-attention. Consequently, the BFM representation associated with a given word w_t can incorporate neural signals from later timepoints within the same chunk, potentially spanning several minutes of subsequent story content. This introduces substantially greater temporal leakage than in the Brain Atlas and PCA baselines, which are computed independently at each TR. Strictly speaking, this means that BFM and Brain Atlas are not compared under equally causal configurations. However, our goal is not real-time decoding, but to test whether large-scale data-driven pretraining improves the quality of brain-to-text mapping. From that perspective, the null result is even more notable: despite enjoying a temporal information advantage, BFM still does not outperform Brain Atlas.

9. Ethics Statement

All fMRI data used in this study are drawn from the publicly available Narratives dataset (Nastase et al., 2021), which was collected under institutional ethical approval with informed consent from participants. No new human subjects experiments were conducted in this work. The dataset does not contain personally identifiable information.

10. Acknowledgements

We thank the anonymous reviewers for their helpful feedback. This work was supported by JSPS KAKENHI Grant Number JP24H00087, JST PRESTO Grant Number JPMJPR21C2, JST CREST Grant Number JPMJCR2565, JST BOOST Grant Number JPMJBY24B2, and Carlsberg Foundation Grant Number CF-25-0624.

11. Bibliographical References

Teddy J Akiki and Chadi G Abdallah. 2019. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific reports*, 9(1):19290.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Rasmus M Birn. 2012. The role of physiological noise in resting-state functional connectivity. *NeuroImage*, 62(2):864–870.

Felix Carbonell, Pierre Bellec, and Amir Shmuel. 2011. Global and system-specific resting-state fMRI fluctuations are uncorrelated: principal component analysis reveals anti-correlated networks. *Brain Connectivity*, 1(6):496–510.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed A Rizvi, Matteo Rosati, Christopher Averill, James L Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van Dijk. 2024. BrainLM: A foundation model for brain activity recordings. In *The Twelfth International Conference on Learning Representations*.

Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 2023. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22710–22720.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

John-Dylan Haynes and Geraint Rees. 2006. Decoding mental states from brain activity in humans. *Nature reviews neuroscience*, 7(7):523–534.

James B Heald, Daniel M Wolpert, and Máté Lengyel. 2023. The computational and neural bases of context-dependent learning. *Annual Review of Neuroscience*, 46(1):233–258.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Joseph G Makin, David A Moses, and Edward F Chang. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582.

Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. In *Proceedings of Interspeech 2004*, pages 2765–2768. ISCA.

Kevin Murphy and Michael D Fox. 2017. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *NeuroImage*, 154:169–173.

Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. 2020. A unique brain connectome fingerprint predates and predicts response to antidepressants. *IScience*, 23(1).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Xiaomei Pei, Dennis L Barbour, Eric C Leuthardt, and Gerwin Schalk. 2011. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. *Journal of neural engineering*, 8(4):046028.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

John W Pratt. 1959. Remarks on zeros and ties in the wilcoxon signed rank procedures. *Journal of the American Statistical Association*, 54(287):655–667.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. `https://cdn.openai.com/better-language-models/language-models.pdf`. OpenAI technical report.

Yu Takagi and Shinji Nishimoto. 2023. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14463.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866.

Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116.

Shizun Wang, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. 2024. Mindbridge: A cross-subject brain decoding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11333–11342.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020.

Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. 2025. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1):346.

12. Language Resource References

Esteban, Oscar and Markiewicz, Christopher J and Blair, Ross W and Moodie, Craig A and Isik, A Ilkay and Erramuzpe, Aliaga and Kent, James D and Goncalves, Mathias and DuPre, Elizabeth and Snyder, Madeleine and others. 2019. *fMRIPrep: a robust preprocessing pipeline for functional MRI*. Nature Methods. PID https://doi.org/10.1038/s41592-018-0235-4. Software resource.

Nastase, Samuel A and Liu, Yun-Fei and Hillman, Hanna and Zadbood, Asieh and Hasenfratz, Liat and Keshavarzian, Neggin and Chen, Janice and Honey, Christopher J and Yeshurun, Yaara and Regev, Mor and others. 2021. *The Narratives fMRI dataset for evaluating models of naturalistic language comprehension*. Scientific Data. PID https://doi.org/10.1038/s41597-021-01033-3. Dataset resource.

Appendix

A. Additional Ablation Analyses of Encoder Signal Quality

A.1. Details on Inter-Subject and Temporal Correlation

The performance gap between Brain Atlas and the PCA baselines in Table 5 motivates a closer look at the statistical quality of each encoder's input representation. We measure two complementary properties—inter-subject correlation (ISC) and lag-1 temporal autocorrelation—that respectively reflect stimulus-driven reliability across subjects and physiological plausibility of the BOLD signal profile (Figure 4). Brain Atlas achieves the highest ISC (0.1090) and autocorrelation (0.5753), confirming that expert-driven parcellation yields cleaner, more reproducible signals than either PCA baseline. Notably, BFM's extreme autocorrelation (0.8987) and elevated ISC (0.2718) indicate over-smoothing rather than genuine neural fidelity, consistent with its failure to improve over Brain Atlas in decoding performance (Section 6.2).

A.2. Projection pretraining

We additionally test whether warm-up pretraining of the projection layer—an MSE-based phase that first aligns the projection layer's outputs to the LLM embedding space (Ye et al., 2025)—can compensate for encoder-side signal quality differences (Table 3). With warm-up, the performance rank of all four encoders remains the same as without, however, they lose statistical significance. This suggests that representational alignment can partially substitute for encoder quality. The fact that the Brain Atlas can achieve these performance gains without added training complexity nonetheless highlights the importance of including expert knowledge in the encoder's design.

A.3. PCA-424 versus PCA-1000

Signal-to-Noise Ratio. The story-averaged cumulative explained variance ratio (EVR) of PCA-424 was 63.9%, compared to 71.2% for PCA-1000 (Figure 5). Despite a 2.36-fold increase in the number of components, the gain in EVR is only 7.3 percentage points, i.e., 576 additional components account for merely 7.3% of total variance. To determine what this additional variance represents, we analyzed per-voxel contributions to PCA components. Mean per-voxel contributions were comparable between PCA-424 and PCA-1000 for both brain voxels (5.6 vs. 5.4×10^{-4}) and non-brain voxels (3.6 vs. 3.7×10^{-4}). However, the contribution difference map (PCA-1000 $-$ PCA-424; Figure 6, bottom; red = PCA-1000 dominant, blue = PCA-424 dominant) reveals that red predominates at the extra-brain periphery. This indicates that the additional components in PCA-1000 are largely derived from non-brain noise: by the variance-maximization principle, PCA assigns components to extra-brain voxels whose large variance arises from head-surface artifacts and physiological noise. The minimal information gain (7.3% EVR) is therefore further devalued by being concentrated in non-brain noise rather than genuine neural signals.

Contrast within the Brain. Voxel contribution maps (Figure 7) show that both PCA-424 and PCA-1000 concentrate high-contribution voxels in the same regions—the cerebellum, basal ganglia, and occipital lobe. The key difference lies not in which regions are targeted, but in the *contrast* between high and low-contribution regions. PCA-424 exhibits sharper contrast, with high-contribution voxels clearly prominent against surrounding areas. PCA-1000 shows a more homogeneous distribution in which the salience of high-contribution voxels

is relatively suppressed. This is consistent with the noise allocation finding: noise-derived components in PCA-1000 elevate the overall contribution floor within the brain, obscuring truly informative voxels. The decoder trained on PCA-424 inputs therefore receives clearer signals about which voxels to prioritize, contributing to more stable learning and improved decoding performance.

A.4. Details on Regional Ablations

To identify which brain regions contribute to decoding performance and to explain why BFM does not improve over Brain Atlas, we conducted a region-level ablation analysis. Activations in each of the 24 functional subnetwork groups (AA-24 hierarchy; Akiki and Abdallah, 2019) were selectively masked, and the resulting change in validation loss ($\Delta\text{Loss} = \mathcal{L}_{\text{ablated}} - \mathcal{L}_{\text{baseline}}$) was measured (Table 4, Figure 9). Positive ΔLoss indicates that the masked group is relied upon for decoding; negative values indicate that it actively impedes decoding. Brain Atlas exhibits a sparse, positive-only profile in which only 5 of 24 groups carry non-zero importance. Its second highest contributing group is the Language Default Mode Network (superior temporal and inferior frontal gyri), directly reflecting established neural correlates of language comprehension. In contrast, BFM distributes importance across all 24 groups with 8 showing negative ΔLoss, suggesting that its dense, entangled representation introduces irrelevant or conflicting signals that hinder the projection into language space. The brain maps below confirm the anatomically coherent localization of the most predictive regions for each encoder.

B. Qualitative Decoding Examples

To complement Table 5's quantitative evaluation, Tables 5–8 present randomly selected decoding outputs from all four encoder configurations alongside the corresponding ground-truth transcripts. Table 5 shows 20 examples from (C) Brain Atlas, while Tables 6, 7, and 8 each show 10 examples from (A) PCA-1000, (B) PCA-424, and (D) BFM, respectively. Examples are drawn from the held-out test portions of all five stories to provide a representative sample of decoder behavior across diverse narrative contexts. While BLEU-1 and ROUGE scores are modest in absolute terms (but equivalent to prior work, such as Ye et al., 2025), the qualitative outputs show partial thematic alignment with the reference texts, particularly in content words and broad topic area. A notable qualitative difference is that Brain Atlas outputs tend to exhibit less degenerate repetition than the other encoders: PCA-424, PCA-1000, and BFM predictions frequently collapse into repetitive token or phrase loops (e.g., wait wait wait …", it's not your fault it's not your fault …"), whereas Brain Atlas outputs more often form grammatically coherent sentences with varied vocabulary, suggesting greater output diversity and naturalness.

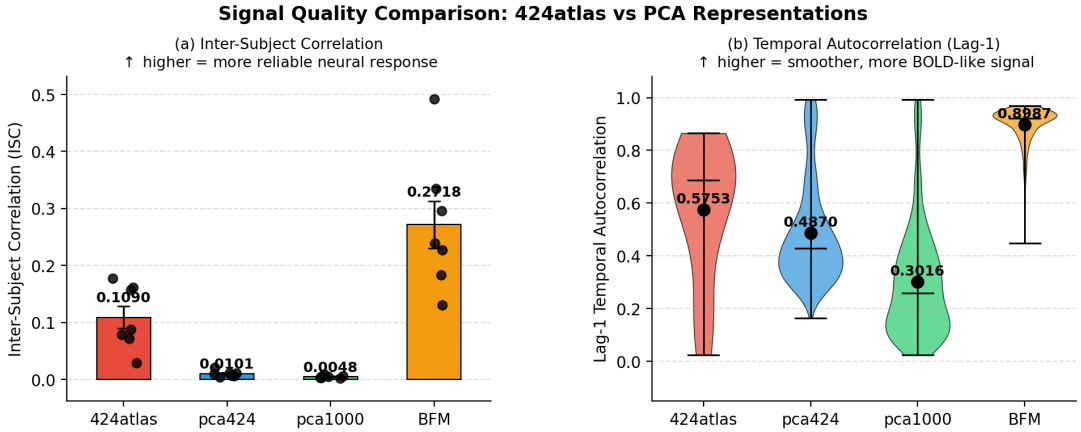


Figure 4: Signal quality comparison. **(a)** *Inter-Subject Correlation (ISC)*: ISC means are C = 0.1090, B = 0.0101, A = 0.0048, and D = 0.2718. Each dot represents the per-story ISC; error bars indicate standard error margin across stories. Higher ISC indicates more stimulus-driven, cross-subject reproducible neural responses. **(b)** *Lag-1 Temporal Autocorrelation*: means are C = 0.5753, B = 0.4870, A = 0.3016, and D = 0.8987. Violin plots reflect per-dimension distributions, with dots denoting means. Section 6 discusses how the mid-level correlations of Brain Atlas translate into better performance compared to PCA and BFM.

configuration	BLEU-1	ROUGE-1	ROUGE-L	Val. Loss	WER (↓)
(A) PCA-1000	0.1243	0.1170	0.1101	5.088	0.9444
(B) PCA-424	0.1252	0.1177	0.1106	5.019	0.9441
(C) Brain Atlas	0.1265	0.1180	0.1109	4.844	0.9441
(D) BFM	0.1255	0.1165	0.1093	4.868	0.9449

Table 3: Evaluation results with warm-up training (MSE alignment + cross-entropy) according to Ye et al. (2025). While the performance ordering of Table 5 is roughly maintained, no statistically significant pairwise differences are observed. This pattern suggests that warm-up absorbs signal-quality differences introduced by dimensionality reduction strategies.

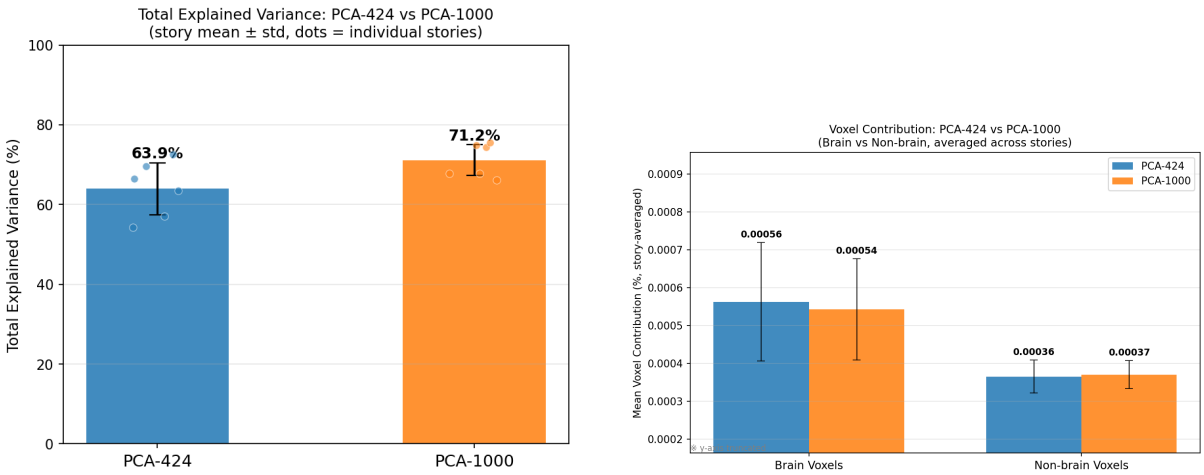


Figure 5: Left: Story-averaged total explained variance ratio (EVR) for PCA-424 (63.9%) and PCA-1000 (71.2%). Despite a 2.36-fold increase in dimensionality, the EVR gain is only 7.3 percentage points. Right: Mean per-voxel contribution for brain vs. non-brain voxels (PCA-424 and PCA-1000).

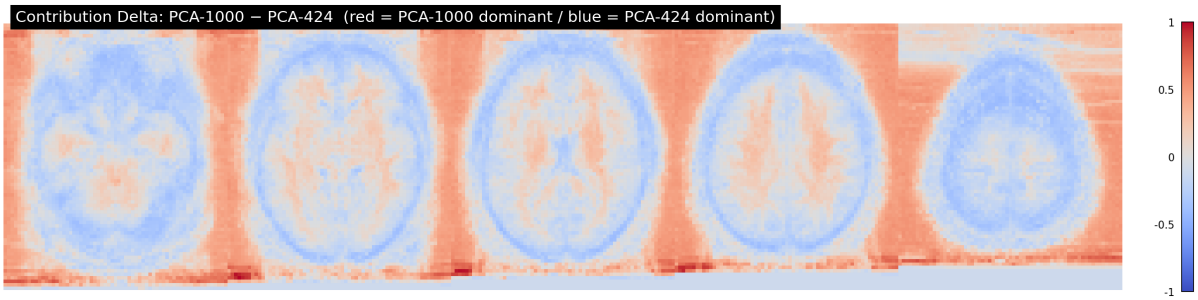
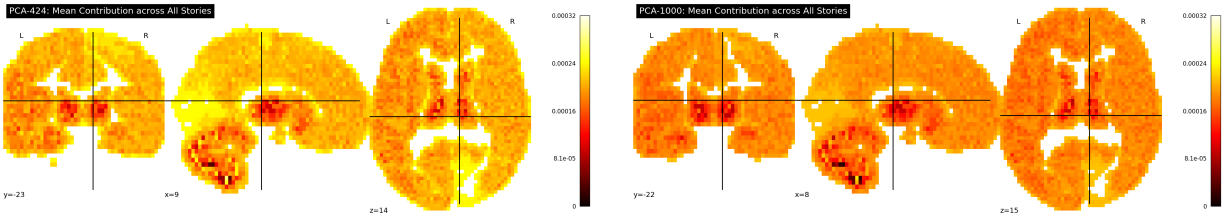


Figure 6: Contribution difference map (PCA-1000 − PCA-424). Red indicates PCA-1000-dominant regions; blue indicates PCA-424-dominant regions. Red concentration at the brain periphery shows that the additional PCA-1000 components largely reflect extra-brain noise.



(a) PCA-424 (b) PCA-1000

Figure 7: Mean voxel contribution maps across all stories for (a) PCA-424 and (b) PCA-1000. High-contribution voxels are concentrated in similar regions (cerebellum, basal ganglia, occipital lobe) in both configurations. PCA-424 shows sharper contrast between high- and low-contribution regions; PCA-1000 shows a more uniform, noise-elevated distribution.

Table 4: Top-5 AA-24 network groups ranked by ablation importance ($\Delta\text{Loss} = \mathcal{L}_{\text{ablated}} - \mathcal{L}_{\text{baseline}}$; mean $\pm$ standard error margin across 27 subjects). † indicates groups with positive ΔLoss in *both* models.

424atlas					**BFM**				
Rank	Group	ΔLoss			Rank	Group	ΔLoss		
1	sc CE†	+0.043	$\pm$	0.009	1	med DM†	+0.023	$\pm$	0.012
2	lang DM†	+0.023	$\pm$	0.004	2	sc CE†	+0.016	$\pm$	0.009
3	GPu SC†	+0.016	$\pm$	0.002	3	cent SM	+0.011	$\pm$	0.006
4	le CE†	+0.006	$\pm$	0.001	4	GPu SC†	+0.010	$\pm$	0.007
5	med DM†	+0.003	$\pm$	0.001	5	ri CE	+0.006	$\pm$	0.006

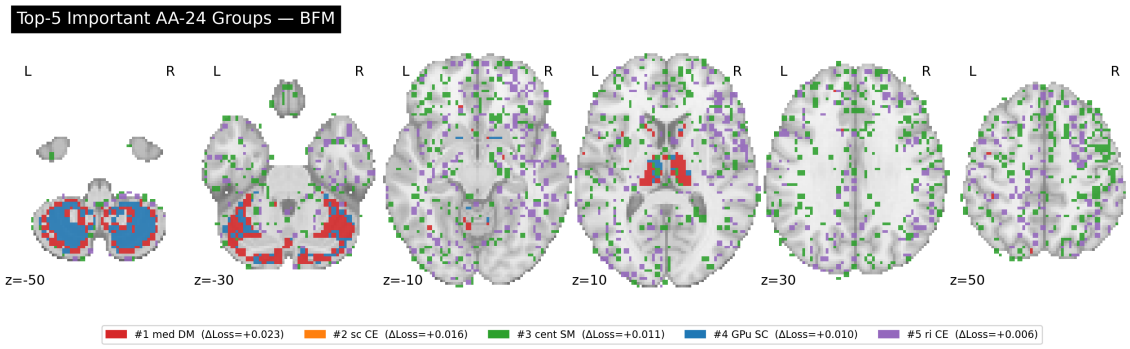


Figure 8: Top-5 contributing AA-24 groups for BFM projected onto MNI space (axial slices, $z = -50$ to $+50$ mm). Color indicates group identity (rank order): red = #1 med DM (Medial Default Mode Network, $\Delta\text{Loss} = +0.023$), orange = #2 sc CE (Subcortical Cerebellum, $+0.016$), green = #3 cent SM (Central Sensorimotor, $+0.011$), blue = #4 GPu SC (Globus Pallidus/Putamen, $+0.010$), purple = #5 ri CE (Right Cerebellum, $+0.006$). The spatial distribution confirms anatomically coherent localization of the most predictive regions.

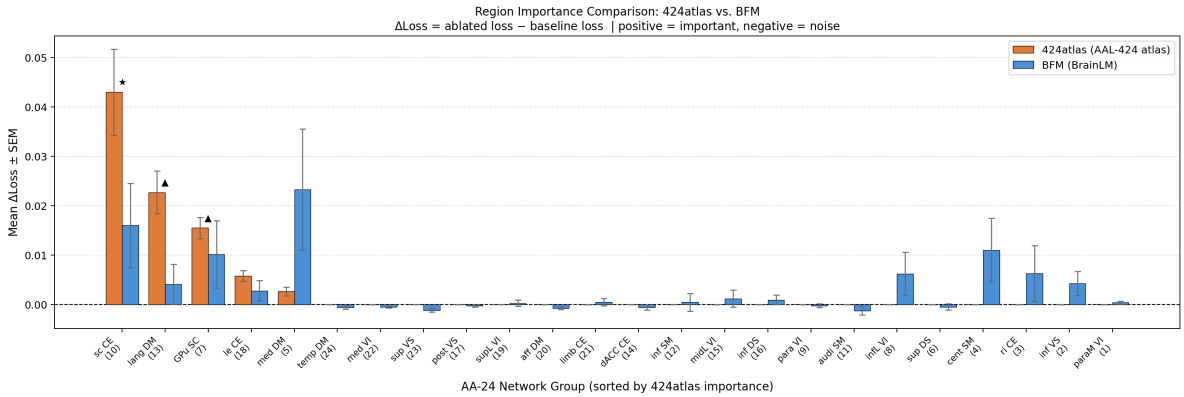


Figure 9: Ablation importance of AA-24 groups for Brain Atlas (orange, left bars in each pair) and BFM (blue, right bars in each pair). ΔLoss = ablated loss $-$ baseline loss; mean $\pm$ SEM across 27 subjects. Groups are sorted by Brain Atlas importance (left = most important). Positive values indicate reliance; negative values indicate noise contribution. Brain Atlas exhibits a sparse profile (5 of 24 groups non-zero), whereas BFM distributes importance broadly with 8 groups showing negative contributions. † Groups that are positive in both models.

Table 5: Random qualitative examples (1–20) from (C) Anatomical Atlas (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…date today is june	i'm going to look at the wall and see if there's a hole in the wall we're going to find out what it's all about. oh yeah	fifteenth yet i could swear
Narratives_060	…to be having a lot of nightmares	when i woke up in the morning and i woke up in the morning and i woke up in the morning and i woke up in the morning and i woke up	lately really
Narratives_065	…get in bed and relax tell the truth is it going to do you ...	if you don't tell the truth **and** i'm going **to** get out of here so i'll go back **to** my room **and** wait for you **to** come in **and**	**to** sit around **and** stew
Narratives_016	…alan did not want to	be in the room so he said you know what i'm going to do and i'm going to go out there and i'm going to tell you this story	part with one of his
Narratives_144	…what he did in	the war **he** killed people **and** you know i was living in the city of alan you know i lived in the city of alan you know i went to	brazil **and he** said
Narratives_041	…no i don't believe it	's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible it's impossible	swanson
Narratives_052	…with a clever imitation guy mary i'm going look around a little more well ...	**right** i'm going to look around a little more well your dinner will be ready all **right** i'm going to look around a little more well your dinner will be	**right** save it me there are
Narratives_066	…she learned some two hundred fifty signs	that she had learned from her father **and** he was **the** one who taught her how to teach her how to do it. i'm not sure if you've ever	**and the** big question is
Narratives_134	…well uh you know but	**uh** you know um i'm not going to be able to go out and **uh** um um um um um um um um um um um um um um um um um um	**uh** the stress of suddenly i was
Narratives_058	…morning guy woke up	in the middle of the night and he was lying on the floor with his hands behind his back and he had no idea what was going on so he decided to	screaming
Narratives_026	…i must warn you that it's useless to buck us	**out** of this place we've got to get away from here you're going to have to be careful what you're doing there's no way we're going to	come **out** peacefully
Narratives_141	…new york rules he would get g corp that actually has a seven twenty ...	**that** he would have to **go** to and i was the first one to come out of the door and it was a really cool thing to be able to get in	which were worth even though you can **go** do **that** great
Narratives_137	…thing you know a southern guy with an	old lady and i was like oh yeah you know what i'm going to tell you this is gonna be the story of my life and i'm going to tell	eight year perfect attendance pin from
Narratives_111	…bob mean so all your friends are a why member of and are you ...	'm going to the hospital and i'm going to a hospital and i'm going to an emergency surgery and i'm going to be in the hospital and i'm	then probably would've never said you know when i grow
Narratives_052	…these new products the dream guy	i've been dreaming about for a long time and i've been dreaming about all the new things i've been dreaming about since i've been asleep i've been	went home that night
Narratives_144	…of this building and i went normal life and my	dad **and i** went **to** school **and** my dad was a teacher **and i** was in college **and i** was the first guy **to** go **to** university **and i** was going **to**	**to** bob **and i** said bob you know those funny little sheltered ...
Narratives_140	…of an apartment	building and **i** was living **in** the same apartment building as bob and **i** had been living there for a long time and **i** knew that bob was going to be living	**in** new york **i** guess you weren't

Subject	Context	Prediction	Ground Truth
Narratives_041	…i think it's some sort of mass hypnosis or something	like that but it's not something that we've been able to find out what **it** means they're trying to get rid of **us** so i think it's a	well why doesn't **it** work for **us**
Narratives_084	…you know just talking made now to go to space i said have you ...	-saving **surgery** and you've been in space and you've been **to** space and you've been **to** space and you've been **to**	financial crime no no i used **to** threatening **surgery** so what do ...
Narratives_061	…recommend against crossing that line but i had just	found out that i was a student at **the** university so i decided to give it a shot and i went to **the** dean's office where i learned that i had	seen **the** line

Table 6: Random qualitative examples (1–10) from (A) PCA-424 (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…excuse me sir it's quite alright	oh wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait wait	do remember the phone call
Narratives_060	…mr in eh no he had an appointment with mr	mr he had an appointment with m mr he had an appointment with mr he had an appointment with mr he had an appointment with mr	the human research institute
Narratives_065	…no i didn't arthur he said his eyes on the	wall and he said yeah i don't know where **the** arthur dead man is they haven<brain/> oh yeah i've been told by **the** police that there's a	far dim end of **the** room
Narratives_016	…alan did not want to	get hit by the landlord so he died and that's the story **of** the tragedy **of** the tragedy **of** the tragedy **of** the tragedy **of** the tragedy **of** the tragedy **of**	part with one **of** his
Narratives_144	…connecticut and i work for the new york times	you know **i** work for the new ynew york times you know **i** work for the new york times you know **i** work for the new york	and **i** was working a story one
Narratives_041	…the power and wash out the memory of the day you	have to go back to the beginning of the day you have to go back to the beginning of the day you have to go back to the beginning of the day	your friend swanson
Narratives_052	…concrete more	concrete you see concrete more concrete underneath but underneath you see concrete more concrete underneath you see concrete more concrete underneath but underneath you ...	metal and here on the
Narratives_066	…year old she was eating at the	**table** and she was sitting down **with** her hands behind her head and she was eating a cup of coffee and she had just gone to sit down **with** her hands behind	**table with** us forks
Narratives_134	…trouble and um and bob realized he says you can't do ...	you can't do that and bob realized he realized he realized he realized he realized he realized he realized he realized he realized he realized he realized ...	be the first suspect
Narratives_058	…mary	look what is it you want to know what is it here i'm looking out the door and there is a huge pile of strange materials that has been put together	mary

Table 7: Random qualitative examples (1–10) from (B) PCA-1000 (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…excuse me sir it's quite alright	what are you going to **do** it's okay I'm going to get out of here hurry up and get out of here let's get out of here hurry up	**do** remember the phone call
Narratives_060	…mr in eh no he had an appointment with mr	mr mr mr mr mr mr mr mr mr mr mr mr mr mr mr	the human research institute
Narratives_065	…no i didn't arthur he said his eyes on the	wall and he was leaning against **the** wall looking up at **the** ceiling where **the** wall had fallen onto **the** wall there were no arthur's eyes on **the** wall but	far dim end of **the** room
Narratives_016	…alan did not want to	be the **one** that was on the roof **of** the building at the time, so he said he had to go up there and they were all over the place and	part with **one of** his
Narratives_144	…connecticut and i work for the new york times	**and** now **i** am **working** for the new y you know **i** work for the new y times **and** now **i** am **working** for the new york times **and**	**and i** was **working** a story one
Narratives_041	…the power and wash out the memory of the day you	have to be very careful you don day you have to be very careful you don day you have to be very careful you don day you have	your friend swanson
Narratives_052	…concrete more	than concrete more than concrete more<brain/> cement more than concrete more concrete more than concrete more **metal** more than concrete more concrete more than concrete ...	**metal** and here on the
Narratives_066	…year old she was eating at the	**table** she was eating at the **table** she was eating at the **table** she was eating at<brain/> **table** she was eating at the **table** she was eating at the **table**	**table** with us forks
Narratives_134	…trouble and um and bob realized he says you can't do ...	you can't do that and bob realized he realized he realized he realized he realized he realized he realized he realized he realized he realized he realized ...	be the first suspect
Narratives_058	…mary	what is it look what is it look what is it look what is it look what is it look what is it look what is it look what is it look what is it look	mary

Table 8: Random qualitative examples (1–10) from (D) Brain Foundation Model (seed=20260223). Matching words are highlighted in **blue bold**; text is truncated for readability.

Subject	Context	Prediction	Ground Truth
Narratives_052	…excuse me sir it's quite alright	you've got to be careful what you're going to **do** it's not your fault it's not your fault it's not your fault it's not your	**do** remember the phone call
Narratives_060	…mr in eh no he had an appointment with mr	he had an appointment with mr he had an appointment with m guess what he was going to do it's been a long time since you've been here	the human research institute
Narratives_065	…no i didn't arthur he said his eyes on the	girl and he was like oh yeah i don't know what you're going to do but i'm going to go to **the** hospital and i'm going to get	far dim end of **the** room
Narratives_016	…alan did not want to	go to the hospital but he was going to the hospital and he said i'm going to the hospital and i'm going to the hospital and i'm going to	part with one of his
Narratives_144	…connecticut and i work for the new york times	**and** i've been **working** for the new y you know it's **a** lot of money to be able to work in the city **and** uh yeah so I'm	**and** i was **working a** story one
Narratives_041	…the power and wash out the memory of the day you	've been waiting for the day you've been waiting for the day you've been waiting for the day you've been waiting for the day you've been waiting for	your friend swanson
Narratives_052	…concrete more	than anything else you've ever seen in **the** history of **the** world it's been a long time since we've had a chance to have an event like this one	metal and here on **the**
Narratives_066	…year old she was eating at the	kitchen **table** she was sitting on the couch and she was asleep she was asleep she was asleep she was asleep she was asleep she was asleep she was	**table** with us forks
Narratives_134	…trouble and um and bob realized he says you can't do ...	you can't do that and he says yeah i'm going to **be the first** one to go and um and uh and um and um and um and um and	**be the first** suspect
Narratives_058	…mary	**mary mary mary mary mary mary mary mary mary** m oh my god oh my god oh my god oh my	**mary**

Floating or *Suggesting Ideas*? A Large-Scale Contrastive Analysis of Metaphorical and Literal Verb–Object Constructions

Prisca Piccirilli[1] and **Alexander Fraser**[2,3] and **Sabine Schulte im Walde**[1]
[1]Institute for Natural Language Processing (IMS), University of Stuttgart, Germany
[2]Technical University of Munich (TUM), Germany
[3]Munich Center for Machine Learning (MCML), Germany
{prisca.piccirilli, schulte}@ims.uni-stuttgart.de

Abstract

Metaphor is a pervasive feature of everyday language, enabling speakers to express abstract concepts in terms of more concrete domains. While prior work has extensively examined metaphors from cognitive and psycholinguistic perspectives, large-scale empirical comparisons between metaphorical and literal language remain limited, especially when they compete in conveying the same meaning. We address this gap through a systematic contrastive analysis of 297 near-synonymous English verb–object (VO) pairs (e.g., *float idea* vs. *suggest idea*) in approximately 2 million extracted corpus sentences, allowing us to examine their contextual usage. Using five established NLP tools – SEANCE, TAALES, TAASSC, TAACO and TAALED – we derive 2,293 metaphor-related cognitive and linguistic features capturing affective, lexical, syntactic, and discourse-level properties. We address two research questions: (i) whether these features consistently differ between metaphorical and literal language (**cross-pair analysis**), and (ii) whether individual VO pairs show strong internal divergence between their metaphorical and literal variants (**within-pair analysis**). Cross-pair results reveal robust global tendencies: literal contexts are associated with higher lexical frequency, stronger cohesion, and greater structural regularity, whereas metaphorical contexts show increased affective loading, imageability, lexical diversity, and constructional specificity. In contrast, within-pair analyses reveal substantial heterogeneity and most pairs display non-uniform directional effects across cognitive or linguistic dimensions. Our findings suggest that there is no single, consistent distributional pattern that consistently distinguishes metaphorical from literal language. Instead, the differences we observe are largely construction-specific. Overall, by combining large-scale data with a broad range of cognitive and linguistic features, this study offers a fine-grained understanding of the contrast between metaphorical and literal VO usages.

Keywords: metaphorical language, cognitive and linguistic empirical features, verb-object constructions

1. Introduction

Metaphor is a "necessary" language feature of everyday thought and communication that allows speakers to conceptualize abstract ideas in terms of more concrete domains (Ortony, 1975; Lakoff and Johnson, 1980; van den Broek, 1981; Schäffner, 2004, i.a.). While metaphor has been extensively studied in cognitive linguistics and psycholinguistics (Gibbs, 1989; Blasko, 1999; Giora, 2002; Glucksberg, 2003; Steen, 2004, i.a.), its empirical analysis in contrast to literal language use remains limited. For instance, *float the idea* and *suggest the idea* both convey the act of proposing an idea, but they differ in register, connotation, and metaphorical framing: *float the idea* employs a conceptual metaphor (IDEAS ARE OBJECTS IN MOTION), projecting an abstract act into a concrete physical domain (Lakoff and Johnson, 1980). In contrast, *suggest the idea* is literal and denotes the speech act directly. Pragmatically, the verb *float* conveys tentativeness or strategic intent, while *suggest* is more neutral and direct, and commonly used in formal and informal contexts (Cameron and Deignan, 2003; Shutova, 2010).

Understanding how metaphorical and literal expressions differ at scale is crucial not only for theoretical insights but also for practical applications such as machine translation, sentiment analysis, and information retrieval. In this paper, we fill this gap by conducting a large-scale, systematic analysis of metaphorical and literal language in contrast. We curate and analyze a dataset of approximately 2 million sentences containing semantically equivalent verb-object (VO) pairs, where one variant is metaphorical, e.g., *float idea*, and the other is literal, e.g., *suggest idea*. This paired design enables controlled comparison across usage contexts while holding the underlying meaning constant. Using a variety of NLP tools[1], we automatically extract a wide range of over 2,200 features, allowing us to explore cognitive, distributional and structural differences between metaphorical and literal usages. We conduct a large-scale and thorough descriptive analysis of these contrasts along two axes: we identify (i) which cognitive and linguistic features systematically distinguish metaphorical from literal language across VO pairs (**cross-pair**), and (ii) which VO pairs are “most strongly metaphorical” or “most strongly literal” across features but **within-**

[1]https://www.linguisticanalysistools.org/

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 145–161
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

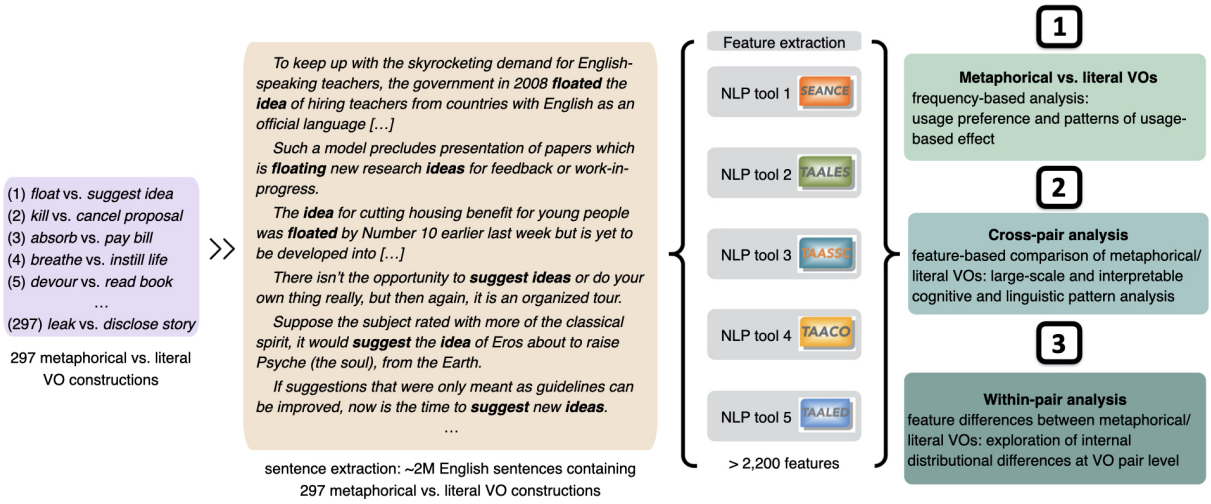


Figure 1: Pipeline of our approach: based on an existing set of metaphorival vs. literal VO pairs, we extract sentences containing these VOs. We then run five NLP tools on all sentences (masking the VO, see Section 3) to extract more than 2,200 cognitive and linguistic features in order to analyze the contrast between metaphorical and literal language usage from (i) a cross-pair and (ii) a within-pair perspective.

pair. Figure 1 provides a visual pipeline of our work study. In sum, our contributions are threefold:

- We present a large-scale, contrastive dataset of metaphorical and literal verb-object expressions in naturalistic English text.
- We conduct an empirical analysis of contextual feature-based differences between metaphorical and literal verb-object expressions.
- We shed light on the internal metaphorical-literal differences at the verb-object level.

By focusing on metaphorical and literal paraphrases with shared meaning, this work advances the computational study of figurative language beyond detection, offering a scalable and interpretable approach to metaphor-literal contrast. Our findings open new avenues for metaphor modeling and the injection of cognitive and linguistic knowledge, and it underlines the importance of metaphorical variation in language understanding tasks[2].

2. Datasets and Linguistic Features

VO Pairs We downloaded the VOs from Piccirilli et al. (2024). They collected a set of 47 VOs from previous work (Mohammad et al., 2016; Shutova, 2010; Piccirilli and Schulte im Walde, 2021; Stowe et al., 2022), which they semi-automatically extended by collecting the most frequently observed shared direct objects of the corresponding verbs, resulting in a total of 297 metaphorical VOs and their corresponding literal paraphrases.

VO Sentences Piccirilli et al. (2024) extracted 1,691 English sentences containing this set of VO pairs. We thus use this publicly available dataset, VOLIMET[3]. However, since our goal is to analyze large amounts of natural language, we enlarge this dataset by automatically extracting all available English sentences containing the VOs from the ENCOW corpus (ENglish COrpora from the Web, Schäfer and Bildhauer, 2012; Schäfer, 2015), where the object is the direct object of the verb. We extract the entire sentence regardless of its length or the distance between the verb and object, which allows us to capture contextual variations that are directly relevant to our features. ENCOW's coverage of multiple writing styles, registers, domains and genres (e.g., blogs, news articles, academic texts), ensures that our analyses reflect natural language in general. Our combined datasets therefore contain close to 2M English sentences covering the whole set of 297 metaphorical vs. literal VOs. Appendix A.1 presents the full list of metaphorical and literal VOs and their sentence-level frequencies, reflecting how often each VO appears in context.

Features Critically, we enriched our dataset with a range of cognitive and linguistic features: they serve as a basis for quantitative analyses of relationships between metaphorical vs. literal language and these properties. We extracted a total of 2,293 features using publicly-available NLP tools[4] which provide measures that are at the core of contrasting metaphorical vs. literal language usage, i.e., these features are related to lexical sophistication

[2]All data and analysis files, including the full set of detailed outputs, are available in our repository at https://github.com/priscapiccirilli/Met-Lit-Contrast.

[3]https://github.com/priscapiccirilli/VOLIMET

[4]SEANCE, TAALES, TAASSC, TAACO, TAALED

and diversity, text cohesion, syntactic complexity, and sentiment analysis (with integers in brackets referring to corresponding numbers of features):

SEANCE (1,271) assesses sentiment, emotion, and social cognition in texts using dictionaries and categories (Crossley et al., 2017).

TAALES (461) analyzes word frequency, range, concreteness, and psycholinguistic properties to assess lexical sophistication (Kyle and Crossley, 2015; Kyle et al., 2018).

TAASSC (355) examines syntactic structures to measure complexity and variety in sentence construction (Kyle, 2016).

TAACO (168) measures text cohesion using lexical overlap, connectives, and semantic similarity (Crossley et al., 2016, 2019).

TAALED (38) computes lexical diversity metrics like type-token ratio, MTLD (Measure of Textual Lexical Diversity), and HD-D (Hypergeometric Distribution D) to evaluate vocabulary variation (Kyle et al., 2021).

3. Comparison of Metaphorical and Literal Verb-Objects: Approach

For comparing metaphorical vs. literal VOs, this work addresses two research questions:

RQ1 Are there consistent differences in **cognitive and linguistic feature patterns** between literal and metaphorical language (= **cross-pair**)?

RQ2 Are there **specific VO pairs** showing a particularly strong contrast in cognitive and linguistic features between literal and metaphorical uses (= **within-pair**), and do we find systematic patterns for literal- vs. metaphor-dominant VO pairs?

RQ1 To investigate how metaphorical and literal expressions differ in their contexts, we conducted pairwise comparisons of automatically extracted features across all sentences containing our 297 VOs. For each VO, we extracted contextual features using the NLP tools presented in Section 2. To avoid the introduction of lexical bias, we masked both the verb and the object in all sentences by replacing them with MASK placeholders. All sentences containing the same VO were grouped and processed together as a single input to each tool. This allowed us to derive aggregated feature representations at the VO level, based solely on the surrounding context across all occurrences of the VO, but independent of the lexical identity of the verb and object themselves. We then compared the feature scores between the metaphorical and literal variants of each pair, to investigate whether metaphorical sentences consistently differ from their literal counterparts along certain dimensions, e.g., whether they tend to exhibit higher emotional valence, lower cohesion, or greater syntactic complexity. We standardized all feature scores using z-score normalization to ensure comparability across scales.

We conducted two levels of analysis. First, to assess (1) **cross-pair feature distinctions**, we computed the mean difference and conducted a Wilcoxon signed-rank test for each feature across all aligned metaphorical-literal VO pairs. Features were classified based on both statistical significance and effect size. Specifically, we considered a feature **important** if it showed a statistically significant difference $p < 0.05$ *and* a substantial effect size defined as a mean difference of ≥ 1 standard deviations following z-score normalization. Features that were statistically significant but had smaller effect sizes were labeled as **significant but small**, indicating a consistent trend without a strong magnitude. Features with non-significant differences $p \geq 0.5$ were not considered further. We present the analyses and discussion in Sec. 4.2.

RQ2 Second, we computed (2) **within-pair feature differences** to determine which VO pairs showed the largest divergence in their metaphorical vs. literal usages. For each aligned VO pair, we calculated both signed and absolute differences across all normalized feature scores. We then computed the mean absolute difference across features per pair, using this value as an indicator of overall metaphoric–literal divergence. VO pairs with a mean absolute difference ≥ 1 standard deviations were interpreted as **linguistically distinctive**; that is, their metaphorical and literal forms showed consistent contextual differences across multiple linguistic features. For these pairs, we also identified which features contributed most to the divergence, and whether the feature was higher in the literal or metaphorical variant. This analysis (Section 4.3) allows us to explore the extent to which individual VO pairs show systematic variation in their surrounding contexts, as captured by the tools, thereby complementing our global feature-level comparisons.

4. Analyses and Discussion

This section provides our analyses and discussions regarding our two research questionsThese two main studies in Sections 4.2 and 4.3 are preceded by a frequency analysis (Section 4.1).

4.1. Metaphorical vs. Literal Preferences

Frequency Table 1 presents an overview of the distribution of literal and metaphorical sentences across our 297 VO pairs; in total, our dataset contains 2,058,787 sentences. In order to analyze

metaphorical versus literal usage preferences, we defined a metaphorical ratio MET_{ratio} for each VO pair and also for each verb:

$$MET_{ratio} = \frac{|MET|}{|MET|+|LIT|}$$

where $|MET|$ and $|LIT|$ denote the number of corpus occurrences of metaphorical and literal variants, respectively. We included only pairs and verbs with a total frequency > 10 to avoid instability from rare items. Figures 2 and 3 display scatter plots of metaphorical vs. literal preferences against total verb frequencies on a logarithmic scale. These visualizations reveal both the overall distribution across our verbs and (VO) pairs along with their strong literal or metaphorical preferences, thus contextualizing the literal dominance observed in raw counts as a consequence of skewed frequency distributions (Table 1).

Category	Count
Total number of sentences	2,058,787
Sentences with literal VOs	1,724,892
Sentences with metaphorical VOs	333,895
Total number of VO pairs	297
VO pairs literal > metaphorical	222
VO pairs metaphorical > literal	75
VO pairs with equal frequency	0

Table 1: Distribution of literal and metaphorical sentences across VO pairs in the dataset. The table reports the total number of sentences per category and the number of VO pairs for which one usage type predominates.

Due to the visual density of Figure 2, we report in Table 2 the 10 most (i) frequency-balanced, (ii) literal-dominant and (iii) metaphor-dominant VO pairs. Most targets (Figure 2) are concentrated towards the left side ($ratio < 0.5$), indicating that literal usages (in green) tend to be more frequent overall for many VO pairs. As shown in Table 1, 222 of our VOs are literally dominant, such as *read article* (vs. *devour article*), *address need* (vs. *attack need*), *pay bill* (vs. *absorb bill*) (middle part of Table 2), compared to 75 VO pairs which are predominantly used metaphorically, i.e., cluster of orange points with a $ratio > 0.5$, such as *cast doubt* (vs. *cause doubt*), *kill bill* (vs. *cancel bill*), *break pattern* (vs. *end pattern*) (middle part of Table 2). The metaphorically-dominant VOs might represent expressions that are lexicalized in natural language (highly-conventionalized metaphors or *dead* metaphors) and behave similarly to idiomatic expressions (Lakoff and Johnson, 1980; Bowdle and Gentner, 2005). Table 2 further shows that literal-dominant VOs tend to exhibit larger absolute Log_2 values, indicating a more pronounced frequency imbalance than metaphor-dominant VOs, suggesting that literal usages often overwhelmingly outnumber their metaphorical counterparts, whereas metaphorical preferences are typically less extreme.

Points near the center ($0.35 < ratio < 0.65$) indicate pairs where metaphorical and literal uses occur with roughly equal frequencies, e.g., *deflate* vs. *reduce currency*, *frame* vs. *pose idea*, *abuse* vs. *misuse product* (top part of Table 2). These balanced cases are particularly interesting for our further qualitative analysis, as they may indicate more flexible, context-dependent interpretation (Section 4.3).

Also, we observe a clear U-shaped pattern in the distribution of VO pairs in Figure 2: pairs that are strongly literal or strongly metaphorical are substantially more frequent than pairs with a balanced literal–metaphorical ratio (see lit. *read book* (upper left) and met. *cast doubt* (upper right) vs. lit. *practise profession* and met. *absorb experience* (middle). This pattern reflects a usage-based effect: highly frequent verb–object constructions become lexicalized, stabilizing either toward predominantly literal meanings or toward conventional metaphorical meanings. In contrast, low-frequency pairs show less semantic stabilization and therefore exhibit more variable usage, clustering near the center of the ratio scale.

Finally, an important insight from the verb-only plot (Figure 3) is that examining verbs independently of their objects can be misleading. While the verb–object *reduce currency* appears as a balanced phrase ($ratio \geq 0.5$ and relatively low frequency) the verb *reduce* on its own, aggregated across all objects, is strongly literal (ratio close to 0) with a very high frequency. This finding highlights how crucial it is to consider the verb and its object as a unit, as verbs behave very differently in terms of their frequency and metaphoricity when considered with or without their objects.

4.2. Cross-Pair Analysis: Feature-based Comparison of Metaphorical and Literal VOs

In this section, we address our first research question: **Are there consistent differences in linguistic feature patterns between literal and metaphorical language (= across VOs)**?

Affective Semantics (SEANCE) We observed 550 features that are statistically significant (with small effect). Features more prominent in literal uses (309) reflect positive affect, institutional structure, cognitive competence, and goal-directedness (valence, work and role-related lexicons, means–end procedural frames), as in *"This*

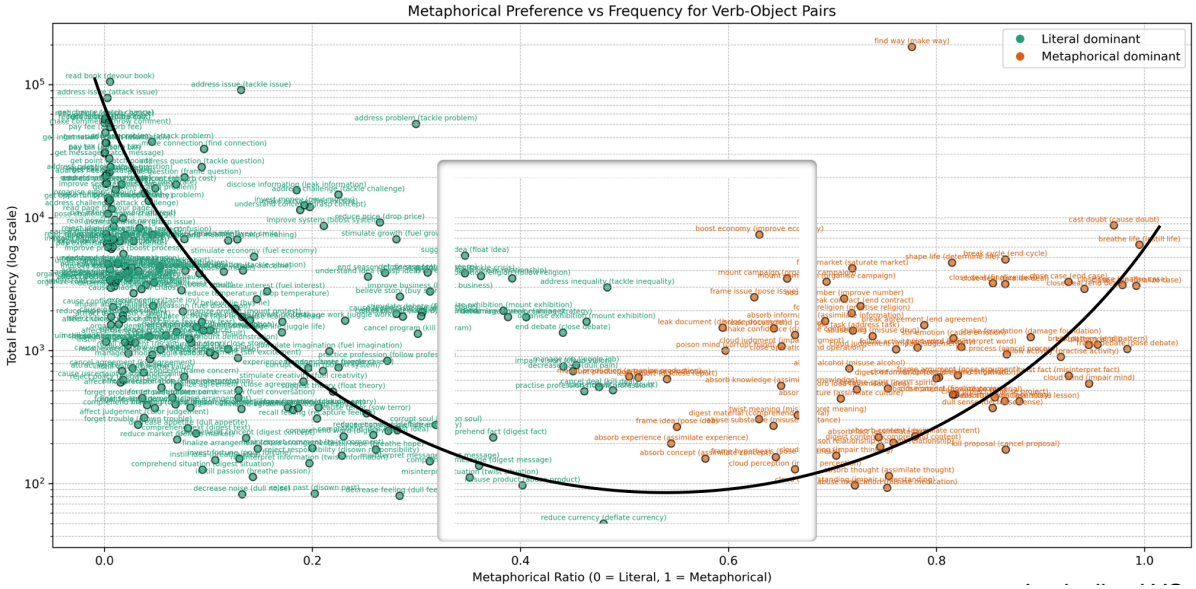


Figure 2: Metaphorical vs. literal VO preferences against total frequencies on a logarithmic scale, with points color-coded by dominance (green for literal-dominant, orange for metaphorical-dominant). The ratio ranges from 0 (entirely literal usage) to 1 (entirely metaphorical usage). Labels indicate the dominant VO followed by its corresponding counterpart in parenthesis, For example, met. VO *breathe life* (lit. VO *instill life*) is metaphorical dominant with a ratio of 1.0. For better readability, we report in Table 2 the 10 most (i) frequency-balanced, (ii) literal-dominant and (iii) metaphor-dominant VO pairs.

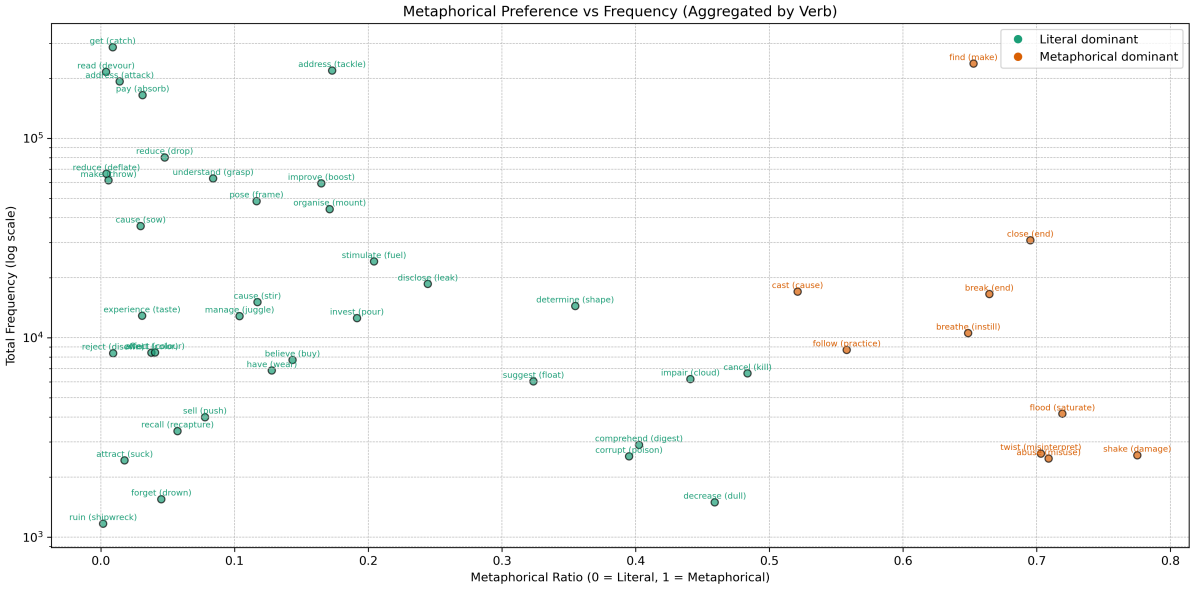


Figure 3: Metaphorical vs. literal verb-only preferences (independently of their direct objects) against total frequencies on a logarithmic scale, with points color-coded by dominance (green for literal-dominant, orange for metaphorical-dominant). The ratio ranges from 0 (entirely literal usage) to 1 (entirely metaphorical usage). Labels indicate the dominant verb followed by its corresponding counterpart in parenthesis.

***bill** was **paid** from an advance on a book whose unauthorised manuscript was published by another ally [. . .]"* (lit. VO). In contrast, metaphorical VO pairs were associated with 241 features related to negative affect, sensorimotor grounding, and social/personal vulnerability. These included emotion-laden categories (e.g., Fear, Disgust), embodied experience (e.g., Fall, Pain), and social identity markers (e.g., Race, RcEthic_Lasswell – a dictionary-based measure capturing the relative frequency of moral and ethical evaluation terms) as in *"Fears, suspicions, resentments and hatred have **poisoned relationships** across that divide in ways that threaten us all"* (met. VO). This pat-

Metaphorical	Literal	Met Count	Lit Count	Ratio	Log_2
Top 10 frequency-balanced VO Pairs					
absorb concept	assimilate concept	89	65	0.58	0.45
absorb experience	assimilate experience	109	91	0.55	0.26
frame idea	pose idea	147	120	0.55	0.29
mount production	organise production	322	305	0.51	0.07
mount production	organize production	322	316	0.51	0.03
sow doubt	cause doubt	246	257	0.49	-0.06
kill deal	cancel deal	253	279	0.48	-0.14
deflate currency	reduce currency	24	50	0.48	-1.06
abuse product	misuse product	39	58	0.40	-0.57
twist situation	misinterpret situation	39	72	0.35	-0.88
Top 10 VO pairs literal $<$ metaphor					
breathe life	instill life	6,227	30	1.00	7.68
close case	finalize case	3,047	24	0.99	6.94
frame debate	pose debate	1,011	17	0.98	5.85
close case	finalise case	3,047	66	0.98	5.52
cast doubt	cause doubt	8,494	257	0.97	5.04
break pattern	end pattern	1,060	50	0.95	4.38
cloud mind	impair mind	534	30	0.95	4.11
kill bill	cancel bill	1,043	59	0.95	4.13
close deal	end deal	2,744	168	0.94	4.02
twist fact	misinterpret fact	599	47	0.93	3.66
Top 10 VO pairs literal $>$ metaphor					
deflate cost	reduce cost	12	51,650	0.00	-12.07
absorb bill	pay bill	10	30,364	0.00	-11.57
devour article	read article	22	55,121	0.00	-11.29
absorb tax	pay tax	20	30,911	0.00	-10.60
attack need	address need	20	20,487	0.00	-10.00
absorb fee	pay fee	45	43,490	0.00	-9.93
catch chance	get chance	67	55,978	0.00	-9.71
devour story	read story	23	17,990	0.00	-9.61
catch result	get result	47	36,573	0.00	-9.60
shipwreck career	ruin career	2	1,165	0.00	-9.19

Table 2: Top VO pairs across distributional profiles: (i) the 10 most frequency-balanced pairs, (ii) the 10 pairs where metaphorical usage is more frequent than their literal counterparts, and (iii) the 10 pairs where literal usage is more frequent. **Met Count** and **Lit Count** indicate the number of sentences in which the metaphorical and literal VO occur, respectively. **Ratio** corresponds to the proportion of metaphorical usage (cf. Figure 2). **Log_2** represents the base-2 logarithm of the ratio between metaphorical and literal frequencies, capturing the direction and magnitude of the imbalance.

tern aligns with theories of metaphor as grounded in bodily and affective experience, often used to frame complex or disruptive states (Lakoff and Johnson, 1980; Gibbs, 2006). The contrast supports a broad cognitive distinction: literal language tends to encode structured and abstract functionality (*address question, reduce cost*, *suggest idea, understand meaning*) and institutional roles (*pay bill, organize conference*) (Turner, 1996; Boroditsky, 2000), while metaphor draws on vivid, sensorimotor grounding (*digest idea, grasp risk)* and emotionally charged semantics (*twist meaning, poison relationship)*, often serving to make abstract concepts experientially accessible (Kövecses, 2010; Mohammad et al., 2016).

Lexical Sophistication (TAALES) Data containing literal VOs scored higher across a broad set of 147 frequency-, range-, and corpus-based features They show higher word and n-gram frequencies across major corpora (SUBTLEXus, BNC, COCA), broader contextual range, greater overlap with academic wordlists, suggesting that literal expressions occur in more common, general-purpose, and infor-

mative contexts, rather than in emotive ones. We also observed higher scores on features for lexical decision times (e.g., LD_Mean_RT_CW – average lexical decision reaction time), indicating that literal language is easier to recognize and process. In contrast, metaphorical expressions showed 73 features with higher scores on psycholinguistic dimensions such as imageability, concreteness, and meaningfulness (MRC norms), along with greater semantic richness (broader and more salient association strength, high LSA-based similarity). Notice, for example, the concrete, technical, and lexically familiar language used in the sentence *"It's the modifications to database permissions that **causes issues** with bbPress 2.2"* (lit. VO), whereas the sentence *"Mikey's stereotypes are holding him back and **coloring** his **judgement**"* (met. VO) contains more abstract, highly imageable language, plus another conventionalized metaphorical expression (*smth. holding sb. back*). This contrast suggests that while literal language tends to rely on familiar, high-frequency lexical items optimized for ease of processing (Van Petten and Kutas, 1990; Dufau et al., 2015), metaphorical language engages more semantically rich, imageable, and conceptually salient terms (Bowdle and Gentner, 2005; Giora, 2002), potentially facilitating deeper conceptual associations and more nuanced interpretation.

Syntactic Complexity (TAASSC) Quite a few features (33) are statistically significant and with a large effect, especially in contexts containing metaphorical VOs (26 for metaphorical vs. 7 for literal contexts). Contexts of literal VOs showed higher variability in syntactic roles such as passive subjects, agents, and indirect objects, as in *"The care planning **system** in the home is currently being **improved**"* (lit. VO), suggesting flexible role-filling within relatively stable constructions. In contrast, metaphorical contexts scored higher on a larger set of features, including type-token ratios and lemma/construction frequency across multiple registers (academic, news, magazine, fiction), reflecting broader lexical and constructional diversity. The pattern holds with 171 significant but small-effect features, where literal contexts show more structural variability (e.g., in syntactic dependencies and construction frequency dispersion), reinforcing their flexibility in grammatical realization (Fazly et al., 2009; Wierzba et al., 2023). Metaphorical contexts show a greater affinity for modifiers (e.g., possession, adverbs), and more conventionalized or lexically cohesive patterns (e.g., higher collexeme ratios, more frequent lemma-construction associations), suggesting semantic specificity and idiomatic usage (Gries and Stefanowitsch, 2004).

Lexical Cohesion (TAACO) Metaphorical and literal language is also contrastive with regards to cohesion patterns. Among features showing significant and large effects, literal contexts score higher on measures such as lexical density, repetition of content and pronoun lemmas, suggesting stronger local cohesion and referential continuity. In contrast, metaphorical contexts show higher type-token ratios across a wide range of grammatical categories (e.g., nouns, verbs, adjectives, adverbs) and n-gram windows (bigram/trigram lemma TTR), indicating greater lexical diversity and a broader distribution of lexical items across the discourse. For features with significant but small effects, contexts of literal VOs again show more consistent adjacent overlap across a variety of part-of-speech categories, as well as higher semantic similarity (e.g., Word2Vec, LSA) between adjacent segments. These suggest stronger sequential cohesion and thematic persistence. In contrast, metaphorical expressions scored higher on MATTR (Moving-Average Type-Token Ratio variant) and showed slightly more binary overlap for content and function words, as well as increased use of temporal and oppositional discourse markers and a higher pronoun-to-noun ratio, pointing to a more fragmented but stylistically varied cohesion profile. This pattern aligns with previous findings from TAASSC and TAALES: literal language favors consistency, repetition, and higher cohesion at both the lexical and discourse levels, while metaphorical language tends to exhibit greater lexical and constructional variety (Halliday and Hasan, 1976).

Lexical Diversity (TAALED) Similarly to the other tools, we observed significantly higher counts of tokens, types, and lexical density measures for literal contexts, suggesting a denser and more informationally rich lexical structure. Also, literal contexts present moderate variation in word types (i.e., more varied vocabulary) within a relatively stable lexical range (i.e., within a coherent and semantically cohesive set of words). In contrast, contexts of metaphorical VOs scored higher on alternative type-token ratio measures, indicating a greater variety of word types, especially for function words. These findings, supported by smaller effects in measure like MATTR, MSTTR (Mean Segmental Type–Token Ratio), and MTLD (Measure of Textual Lexical Diversity), which provide length-controlled estimates of vocabulary variation, suggest that metaphorical language uses a more diverse and flexible vocabulary, particularly through varied use of function words that enforce cohesion and style (Kimmel, 2010; Piccirilli and Schulte Im Walde, 2022).

Figure 4: Tool-level aggregation of distributional dominance for 93 clearly-distinctive VO pairs (mean absolute difference ≥ 1 standard deviations in at least three tools, detailed methodology explained in Sections 3 and 4.3). Feature-level contrasts were collapsed into tool-level directional judgments (literal-dominant in green vs. metaphor-dominant in orange) for VO pairs meeting the distinctiveness criterion in at least three tools. VO pairs are sorted top to bottom from mostly literal-dominant to mostly metaphor-dominant. For example, the VO pair *instill idea vs. breathe idea* behaves significantly more "metaphorical-like" across features of four tools (SEANCE, TAACO, TAALES and TAASSC), i.e., the metaphorical variant *breathe idea* introduces robust distributional differences detectable across linguistic features (structurally, lexically, discourse-wise, etc.) Sections 3 and 4.2 describe the linguistic aspects captures by the tools.

4.3. Within-Pair Analysis: Feature Differences between Metaphorical and Literal VOs

In Section 4.2, we took a **cross-pair** approach and looked at the overall contrast between metaphorical vs. literal VOs (= all metaphorical VOs vs. all literal VOs). In contrast, we now zoom into the contrast **within-pair**, that is to say we focus on each individual VO pair (e.g., *float idea* vs. *suggest idea*) and all the respective sentences the VOs appear in. Our goal is to determine whether our near-synonymous literal and metaphorical VOs show substantial internal differences, and whether such differences are systematic across pairs or rather pair-specific. We are therefore addressing our second research question: **Among the most distinctive VO pairs, which ones – if any – stand out for showing especially strong contrasts in linguistic features between their literal and metaphorical uses?**

Feature-level differences For each VO pair and each feature, we computed the **signed standardized difference**:

$$signed_diff = z_{metaphor} - z_{literal}$$

where z-scores were obtained using a single StandardScaler fitted on the combined literal and metaphorical data to ensure a shared reference distribution. Positive values indicate higher feature values for the metaphorical variant, while negative values indicate higher values for the literal variant. We also computed the **absolute difference** $|signed_diff|$ capturing magnitude differences irrespective of direction.

Aggregation across features: For each VO pair, we then computed the **mean absolute difference** across all features to quantify the overall distinctiveness of each pair, as well as the **mean signed difference** across features, capturing the overall directional bias (literal-dominant vs. metaphor-dominant). VO pairs were then classified as **clearly distinctive** if their mean absolute difference equalled or exceeded a threshold of 1 standard deviations. This threshold identifies pairs whose literal and metaphorical realizations differ substantially across all features. For each pair, we report: the (i) mean absolute difference, (ii) features (across tools) exceeding the defined threshold, and (iii) direction of dominance (literal vs. metaphorical)This step does not assume that all features move in the same direction; rather, it quantifies the overall degree of distributional divergence between the two variants. Finally, we can compute the directional classification. We define as **literal-dominant** the pairs with a negative mean signed difference and **metaphor-dominant** the ones with a positive

mean signed difference. Again, this classification reflects aggregate tendencies across all features. A metaphor-dominant pair does not imply that every feature is higher in the metaphorical variant, but that the balance of standardized feature differences favors that variant.

Given the high dimensionality of the overall feature space ($> 2,000$ features across the five NLP tools we use), visualizing feature-level contrasts per distinctive VO pair separately for each tool proved to be difficult to interpret. As a result, to improve readability and interpretability, while assessing robustness of our approach, we identified VO pairs that met our two distinctiveness criterion: (1) a mean absolute difference ≥ 1 standard deviations (2) in at least three of the five tools. For each tool, VO pairs were further classified as literal-dominant or metaphor-dominant based on the mean signed standardized difference across features.

We present in Figure 4 the aggregations of feature-level contrasts into tool-level directional judgments (literal-dominant in green vs. metaphor-dominant in orange), allowing us to assess whether metaphor–literal asymmetries persist across our independent linguistic feature set. We observe a lot of heterogeneity: according to our strict criteria, 93 VO pairs proved to be clearly distinctive, with only a subset of VO pairs actually showing consistent dominance patterns across tools. (i) Seven VOs are literal-dominant across tools (green rows), such as *forget* vs. *drown problem*, *organize* vs. *mount event*, *attract* vs. *suck worker*. (ii) Six VOs are mostly metaphor-dominant (orange rows), such as *instill* vs. *breathe idea*, *understand* vs. *grasp meaning*, *disclose* vs. *leak document*. (iii) The rest of them – 80 VOs – actually show mixed patterns, where no consistent directional bias emerges. These mixed VOs are neither literal nor metaphor dominant, e.g., *manage* vs. *juggle project*, *address* vs. *attack topic*, meaning that different linguistic dimensions pull in different directions. For example, consider the VO pair *pay bill* vs. *absorb bill*: the metaphorical variant *absorb bill* is on average dominant across TAACO and TAASSC, for which features related to lexical cohesion and syntactic complexity, respectively, scored higher than for its literal counterpart *pay bill*. On the contrary, features measuring affective semantics (SEANCE), lexical sophistication (TAALES) and lexical diversity (TAALED) are pulling towards the literal variant *pay bill*.

Our analysis suggests that metaphor–literal differences do not follow stable, recurring linguistic patterns. Instead, distributional asymmetries are pair-specific, with each VO pair showing its own pattern of divergence. Moreover, directional patterns are not consistent even across contrasts involving the same literal–metaphorical verb pair. For instance, the contrast between *instill* and *breathe* is literal-dominant in combination with the direct object *spirit*, but mostly metaphor-dominant with *idea*. This finding indicates that distributional asymmetries are once again (as mentioned in Section 4.1) not verb-driven but construction-specific: the unit of divergence is the verb–object pairing rather than the lexical verb itself.

5. Conclusion

This study provides a large-scale, feature-based comparison of metaphorical and literal verb–object constructions in natural English sentences. Leveraging nearly two million corpus instances and over 2,200 linguistic features derived from five complementary NLP tools, we investigated both cross-pair and within-pair patterns of divergence.

At the **cross-pair** level, metaphorical and literal language show consistent distributional tendencies. Literal contexts are characterized by higher lexical frequency, stronger local cohesion, and more structurally regular patterns, suggesting stability, informativeness, and easier processing. In contrast, metaphorical contexts display greater affective intensity, sensorimotor grounding, lexical diversity, and constructional specificity, aligning with theoretical accounts that view metaphor as experientially grounded and semantically enriching.

In contrast, our **within-pair** analyses reveal that the cross-pair tendencies do not translate into uniform pair-level behavior. Although a subset of VO pairs shows strong and robust divergence (literal or metaphorical dominant), most pairs present mixed dominance patterns across linguistic dimensions. Moreover, contrasts are construction-specific rather than verb-driven: the same verb may pattern differently depending on its object, indicating that metaphoricity emerges at the level of the verb–object construction pairing rather than the lexical verb alone.

Our findings have both theoretical and methodological implications. Theoretically, they challenge the assumption of stable linguistic signatures of metaphoricity and instead support a usage-based, construction-sensitive view of figurative language. Methodologically, our results demonstrate the value of aggregating multidimensional feature profiles while preserving pair-level granularity. Future work may extend this approach to cross-linguistic settings, translation studies, and predictive modeling, further integrating linguistic theory with large-scale computational analysis. As a matter of fact, our follow-up study makes use of the extracted linguistic features and the knowledge learned from this work as input to machine learning models in order to observe the important features for predicting metaphorical vs. literal language.

6. Limitations

Several limitations should be acknowledged. First, our analysis is restricted to the English language, and the patterns observed in this work may not generalize to other languages with different morphological, syntactic, or metaphorical conventions, especially given the fact that the NLP tools we used are built on and for the English language. Second, although the corpus covers multiple genres (e.g., fiction, academic, web), it represents a single source, and corpus-specific distributional biases cannot be ruled out. Third, our dataset comprises 297 verb–object pairs, which, even though systematically constructed, cannot be considered fully representative of metaphorical or literal language as a whole. Finally, the very large number of extracted features ($> 2,200$) made it necessary to aggregate and enforce thresholding procedures to ensure readability and interpretability: while this multidimensional approach enables broad coverage, it may also obscure more fine-grained effects at the level of individual features or verb-object features.

7. Acknowledgments

This research was supported by the DFG Research Grants SCHU 2580/4-1 (*MUDCAT: Multimodal Dimensions and Computational Applications of Abstractness*) and SCHU 2580/7-1 and FR 2829/8-1 (*MeTRapher: Learning to Translate Metaphors*). Prisca Piccirilli is also supported by the Studienstiftung des deutschen Volkes. We are grateful to Annerose Eichel, Neele Falk and the IMS SemRel research group for helpful discussions, suggestions and feedback regarding versions of this work. We thank Sven Naber for his assistance with data extraction and processing. We would also like to thank the anonymous reviewers for their constructive feedback.

8. Bibliographical References

Dawn G. Blasko. 1999. "Only the Tip of the Iceberg": Who Understands what about Metaphor? *Journal of Pragmatics*, 31:1675–1683.

Lera Boroditsky. 2000. Metaphoric Structuring: Understanding Time Through Spatial Metaphors. *Cognition*, 75(1):1–28.

Brian Bowdle and Dedre Gentner. 2005. The Career of Metaphor. *Psychological Review*, 112:193–216.

Lynne Cameron and Alice Deignan. 2003. Combining Large and Small Corpora to Investigate Tuning Devices Around Metaphor in Spoken Discourse. *Metaphor and Symbol*, 18.

Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating Semantic Similarity and Text Overlap. *Behavior Research Methods*, 51(1):14–27.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2016. The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion. *Behavior Research Methods*, 48(4):1227–1237.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2017. Sentiment Analysis and Social Cognition Engine (SEANCE): An Automatic Tool for Sentiment, Social Cognition, and Social Order Analysis. *Behavior Research Methods*, 49(3):803–821.

Stephane Dufau, Jonathan Grainger, Katherine Midgley, and Phillip Holcomb. 2015. A Thousand Words Are Worth a Picture: Snapshots of Printed-Word Processing in an Event-Related Potential Megastudy. *Psychological Science*, 26.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103.

Raymond W. Gibbs. 1989. Understanding and Literal Meaning. *Cognitive Science*, 13(2):243–251.

Raymond W. Gibbs. 2006. Metaphor Interpretation as Embodied Simulation. *Mind & Language*, 21(3):434–458.

Rachel Giora. 2002. Literal vs. Figurative Language: Different or Equal. *Journal of Pragmatics*, 34:487–506.

Sam Glucksberg. 2003. The Psycholinguistics of Metaphor. *TRENDS in Cognitive Science*, 7(2):92–96.

Stefan Gries and Anatol Stefanowitsch. 2004. Extending Collostructional Analysis: A Corpus-Based Approach to Alternations. *International Journal of Corpus Linguistics*, 9:97–129.

Michael A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Michael Kimmel. 2010. Why We Mix Metaphors (and Mix Them Well): Discourse Coherence, Conceptual Metaphor, and Beyond. *Journal of Pragmatics*, 42:97–115.

Zoltan Kövecses. 2010. *Metaphor: A Practical Introduction*, 2nd edition. Oxford University Press, New York.

Kristopher Kyle. 2016. *Measuring Syntactic Development in L2 Writing: Fine-Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication*. Ph.d. dissertation, Georgia State University.

Kristopher Kyle and Scott A. Crossley. 2015. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, 49(4):757–786.

Kristopher Kyle, Scott A. Crossley, and Charles Berger. 2018. The Tool for the Analysis of Lexical Sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3):1030–1046.

Kristopher Kyle, Scott A. Crossley, and Scott Jarvis. 2021. Assessing the Validity of Lexical Diversity Using Direct Judgements. *Language Assessment Quarterly*, 18(2):154–170.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a Medium for Emotion: An Empirical Study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Andrew Ortony. 1975. Why Metaphors Are Necessary and Not Just Nice. *Educational Theory*, 25(1):45–53.

Prisca Piccirilli, Alexander Fraser, and Sabine Schulte im Walde. 2024. VOLIMET: A Parallel Corpus of Literal and Metaphorical Verb-Object Pairs for English–German and English–French. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 222–237, Mexico City, Mexico. Association for Computational Linguistics.

Prisca Piccirilli and Sabine Schulte im Walde. 2021. Synonymous Pairs of Metaphorical and Literal Expressions in Context: An Empirical Study and Dataset *to tackle* or *to address the question*. In *Proceedings of the Workshop DiscAnn*, Tübingen, Germany.

Prisca Piccirilli and Sabine Schulte Im Walde. 2022. What Drives the Use of Metaphorical Language? Negative Insights from Abstractness, Affect, Discourse Coherence and Contextualized Word Representations. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 299–310, Seattle, Washington. Association for Computational Linguistics.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany. Institut für Deutsche Sprache.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey. European Language Resources Association.

Christina Schäffner. 2004. Metaphor and Translation: Some Implications of a Cognitive Approach. *Journal of Pragmatics*, 36(7):1253–1269.

Ekaterina Shutova. 2010. Automatic Metaphor Interpretation as a Paraphrasing Task. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies:*, pages 1029–1037, Los Angeles, California. Association for Computational Linguistics.

Gerard J. Steen. 2004. Can Discourse Properties of Metaphor Affect Metaphor Recognition? *Journal of Pragmatics*, 36(7):1295–1313.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Mark Turner. 1996. *The Literary Mind: The Origins of Thought and Language*. Oxford University Press, Oxford.

Raymond van den Broek. 1981. The Limits of Translatability Exemplified by Metaphor Translation. *Poetics Today*, 1(4):73–87.

Cyma Van Petten and Marta Kutas. 1990. Interactions Between Sentence Context and Word Frequency in Event-related Potentials. *Memory & cognition*, 18:380–93.

Marta Wierzba, J. Brown, and Gisbert Fanselow. 2023. The Syntactic Flexibility of German and English Idioms: Evidence from Acceptability Rating Experiments. *Journal of Linguistics*, 60:1–38.

A. Supplementary Materials

A.1. Met vs. Lit Verb-Object Pairs

Table 3: Metaphorical vs. Literal Verb–Object Pairs with Corpus Frequencies (N=297)

Metaphorical VO	Literal VO	Met Count	Lit Count
absorb knowledge	assimilate knowledge	354	190
absorb information	assimilate information	1,154	511
absorb idea	assimilate idea	369	141
absorb culture	assimilate culture	308	127
absorb material	assimilate material	382	85
absorb lesson	assimilate lesson	366	50
absorb fact	assimilate fact	166	57
absorb content	assimilate content	177	50
absorb experience	assimilate experience	109	91
absorb concept	assimilate concept	89	65
absorb thought	assimilate thought	86	28
absorb cost	pay cost	1,226	16,540
absorb fee	pay fee	45	43,490
absorb bill	pay bill	10	30,364
absorb tax	pay tax	20	30,911
absorb debt	pay debt	61	21,344
absorb interest	pay interest	178	9,713
absorb expense	pay expense	40	5,774
abuse alcohol	misuse alcohol	525	208
abuse drug	misuse drug	947	335
abuse substance	misuse substance	175	97
abuse medication	misuse medication	70	23
abuse product	misuse product	39	58
attack problem	address problem	1,707	35,477
attack issue	address issue	239	79,175
attack need	address need	20	20,487
attack question	address question	114	21,778
attack challenge	address challenge	83	11,541
attack point	address point	232	4,098
attack situation	address situation	45	3,465
attack topic	address topic	22	4,066
attack change	address change	54	3,579
attack cause	address cause	155	3,094
attack matter	address matter	49	3,451
boost economy	improve economy	4,687	2,755
boost service	improve service	177	16,003
boost system	improve system	1,829	6,837
boost situation	improve situation	16	6,316
boost process	improve process	98	5,235
boost education	improve education	87	3,837
boost work	improve work	80	3,424
boost business	improve business	871	1,910
boost result	improve result	211	2,630
boost number	improve number	1,749	710
break agreement	end agreement	1,225	329
break cycle	end cycle	4,193	647
break relationship	end relationship	1,318	2,321
break contract	end contract	1,571	590
break marriage	end marriage	694	1,094
break process	end process	945	523
break pattern	end pattern	1,060	50
breathe life	instill life	6,227	30
breathe sense	instill sense	25	1,074
breathe confidence	instill confidence	30	1,144

Metaphorical VO	Literal VO	Met Count	Lit Count
breathe value	instill value	26	659
breathe spirit	instill spirit	391	128
breathe love	instill love	66	308
breathe hope	instill hope	47	133
breathe idea	instill idea	16	134
breathe passion	instill passion	16	96
buy story	believe story	724	1,823
buy word	believe word	125	2,997
buy lie	believe lie	260	1,805
cast doubt	cause doubt	8,494	257
cast issue	cause issue	98	6,530
cast fear	cause fear	282	1,365
catch disease	get disease	1,201	2,662
catch idea	get idea	201	52,090
catch chance	get chance	67	55,978
catch information	get information	72	36,240
catch result	get result	47	36,573
catch message	get message	134	27,705
catch point	get point	147	23,962
catch call	get call	43	20,692
catch problem	get problem	594	14,694
catch opportunity	get opportunity	54	13,177
close investigation	end investigation	504	124
close season	end season	1,037	2,805
close deal	end deal	2,744	168
close case	end case	3,047	240
close debate	end debate	601	762
close operation	end operation	643	309
close story	end story	122	903
close deal	finalise deal	2,744	424
close case	finalise case	3,047	66
close plan	finalise plan	98	825
close arrangement	finalise arrangement	27	367
close agreement	finalise agreement	65	293
close project	finalise project	381	64
close deal	finalize deal	2,744	468
close case	finalize case	3,047	24
close plan	finalize plan	98	867
close arrangement	finalize arrangement	27	156
close agreement	finalize agreement	65	438
close project	finalize project	381	86
cloud memory	impair memory	77	297
cloud ability	impair ability	39	1,977
cloud judgement	impair judgement	778	244
cloud judgment	impair judgment	702	376
cloud mind	impair mind	534	30
cloud vision	impair vision	335	424
cloud thinking	impair thinking	114	48
cloud perception	impair perception	85	43
cloud understanding	impair understanding	70	27
color judgement	affect judgement	12	301
color decision	affect decision	36	3,236
color choice	affect choice	10	1,555
color perception	affect perception	145	1,279
color experience	affect experience	77	1,222
color interpretation	affect interpretation	38	498

Metaphorical VO	Literal VO	Met Count	Lit Count
colour judgement	affect judgement	69	301
colour decision	affect decision	32	3,236
colour choice	affect choice	22	1,555
colour perception	affect perception	135	1,279
colour experience	affect experience	55	1,222
colour interpretation	affect interpretation	29	498
deflate economy	reduce economy	71	181
deflate cost	reduce cost	12	51,650
deflate price	reduce price	63	6,766
deflate value	reduce value	58	4,430
deflate supply	reduce supply	12	1,784
deflate wage	reduce wage	23	1,130
deflate market	reduce market	15	199
deflate currency	reduce currency	24	26
devour book	read book	587	104,728
devour article	read article	22	55,121
devour story	read story	23	17,990
devour page	read page	37	10,667
devour novel	read novel	77	8,517
devour information	read information	58	6,787
devour chapter	read chapter	21	6,345
devour news	read news	15	4,480
digest information	comprehend information	496	124
digest meaning	comprehend meaning	18	352
digest material	comprehend material	192	113
digest fact	comprehend fact	83	139
digest text	comprehend text	21	217
digest concept	comprehend concept	30	190
digest idea	comprehend idea	60	172
digest word	comprehend word	52	174
digest content	comprehend content	154	49
digest situation	comprehend situation	12	115
digest message	comprehend message	49	87
disown past	reject past	17	67
disown idea	reject idea	14	7,519
disown policy	reject policy	9	567
disown responsibility	reject responsibility	37	125
drop price	reduce price	2,439	6,766
drop cost	reduce cost	218	51,650
drop rate	reduce rate	817	15,826
drop temperature	reduce temperature	357	2,077
drown trouble	forget trouble	9	268
drown pain	forget pain	29	414
drown problem	forget problem	16	423
drown feeling	forget feeling	16	373
dull appetite	decrease appetite	21	240
dull pain	decrease pain	315	384
dull sense	decrease sense	316	54
dull noise	decrease noise	11	72
dull feeling	decrease feeling	23	58
find excuse	make excuse	2,374	9,643
find way	make way	149,262	42,972
find connection	make connection	3,155	29,720
float idea	suggest idea	1,798	3,386
float theory	suggest theory	108	383
float concept	suggest concept	48	316

Metaphorical VO	Literal VO	Met Count	Lit Count
flood market	saturate market	2,989	1,167
follow profession	practice profession	228	610
follow religion	practice religion	1,371	2,125
follow activity	practice activity	824	230
follow profession	practise profession	228	266
follow religion	practise religion	1,371	537
follow activity	practise activity	824	72
frame question	pose question	1,553	18,561
frame problem	pose problem	570	12,804
frame challenge	pose challenge	97	9,544
frame issue	pose issue	1,574	945
frame debate	pose debate	1,011	17
frame concern	pose concern	49	588
frame argument	pose argument	535	117
frame idea	pose idea	147	120
frame hypothesis	pose hypothesis	102	56
fuel debate	stimulate debate	595	1,351
fuel growth	stimulate growth	1,929	4,940
fuel economy	stimulate economy	733	4,359
fuel interest	stimulate interest	437	2,354
fuel discussion	stimulate discussion	149	1,821
fuel demand	stimulate demand	553	1,262
fuel activity	stimulate activity	154	1,487
fuel imagination	stimulate imagination	214	775
fuel creativity	stimulate creativity	112	473
fuel conversation	stimulate conversation	57	385
grasp meaning	understand meaning	802	5,927
grasp concept	understand concept	2,149	9,262
grasp issue	understand issue	284	8,076
grasp point	understand point	531	6,359
grasp problem	understand problem	188	6,464
grasp situation	understand situation	220	3,935
grasp reason	understand reason	59	4,449
grasp language	understand language	87	4,220
grasp idea	understand idea	913	2,688
grasp risk	understand risk	23	3,262
grasp question	understand question	30	3,016
juggle job	manage job	353	425
juggle project	manage project	149	7,242
juggle work	manage work	389	1,291
juggle life	manage life	234	1,136
juggle career	manage career	168	578
juggle school	manage school	38	825
kill proposal	cancel proposal	156	24
kill project	cancel project	517	1,284
kill bill	cancel bill	1043	59
kill program	cancel program	403	934
kill process	cancel process	813	124
kill agreement	cancel agreement	20	717
kill deal	cancel deal	253	279
leak report	disclose report	330	280
leak information	disclose information	2,977	13,122
leak document	disclose document	884	603
leak story	disclose story	361	56
mount production	organise production	322	305
mount event	organise event	49	14,000

Metaphorical VO	Literal VO	Met Count	Lit Count
mount campaign	organise campaign	2,288	1,007
mount conference	organise conference	19	3,384
mount exhibition	organise exhibition	765	885
mount demonstration	organise demonstration	96	1,286
mount protest	organise protest	237	1,455
mount production	organize production	322	316
mount event	organize event	49	5,955
mount campaign	organize campaign	2,288	1,197
mount conference	organize conference	19	2,947
mount exhibition	organize exhibition	765	1,229
mount demonstration	organize demonstration	96	1,050
mount protest	organize protest	237	1,552
poison mind	corrupt mind	596	402
poison system	corrupt system	148	556
poison process	corrupt process	31	343
poison soul	corrupt soul	88	188
poison relationship	corrupt relationship	141	48
pour money	invest money	2,383	10,003
pour fortune	invest fortune	20	134
push drug	sell drug	311	3,680
recapture feeling	recall feeling	63	245
recapture memory	recall memory	34	1,112
recapture moment	recall moment	55	961
recapture experience	recall experience	43	887
shake confidence	damage confidence	870	440
shake foundation	damage foundation	1,124	138
shape result	determine result	100	1,125
shape life	determine life	3,737	848
shape outcome	determine outcome	451	3,472
shape success	determine success	104	2,795
shape strategy	determine strategy	724	1,061
shipwreck career	ruin career	2	1,165
sow doubt	cause doubt	246	257
sow death	cause death	44	18,023
sow confusion	cause confusion	345	7,077
sow chaos	cause chaos	76	2,622
sow conflict	cause conflict	18	2,110
sow panic	cause panic	33	1,808
sow fear	cause fear	133	1,365
sow violence	cause violence	22	930
sow uncertainty	cause uncertainty	11	591
sow terror	cause terror	85	252
sow hatred	cause hatred	68	181
stir excitement	cause excitement	113	766
stir confusion	cause confusion	22	7,077
stir reaction	cause reaction	94	3,894
stir feeling	cause feeling	530	1,357
stir emotion	cause emotion	1,009	227
suck worker	attract worker	12	690
suck talent	attract talent	31	1,693
tackle question	address question	2,250	21,778
tackle issue	address issue	11,995	79,175
tackle problem	address problem	15,180	35,477
tackle concern	address concern	299	17,506
tackle challenge	address challenge	3,352	11,541
tackle situation	address situation	533	3,465

Metaphorical VO	Literal VO	Met Count	Lit Count
tackle point	address point	111	4,098
tackle crisis	address crisis	1,322	2,491
tackle matter	address matter	347	3,451
tackle inequality	address inequality	1,445	1543
tackle task	address task	1,017	396
taste freedom	experience freedom	145	594
taste pain	experience pain	21	5,984
taste life	experience life	130	3,837
taste joy	experience joy	102	2,071
throw remark	make remark	72	13,467
throw comment	make comment	282	47,768
twist word	misinterpret word	876	187
twist fact	misinterpret fact	599	47
twist meaning	misinterpret meaning	218	109
twist comment	misinterpret comment	37	148
twist situation	misinterpret situation	39	72
twist information	misinterpret information	28	114
twist message	misinterpret message	46	101
wear smile	have smile	876	5,975

Social Meaning in Large Language Models: Structure, Magnitude, and Pragmatic Prompting

Roland Mühlenbernd
Leibniz-Centre General Linguistics, Berlin, Germany
muehlenbernd@leibniz-zas.de

Abstract

Large language models (LLMs) increasingly exhibit human-like patterns of pragmatic and social reasoning. This paper addresses two related questions: do LLMs approximate human social meaning not only qualitatively but also quantitatively, and can prompting strategies informed by pragmatic theory improve this approximation? To address the first, we introduce two calibration-focused metrics distinguishing structural fidelity from magnitude calibration: the Effect Size Ratio (ESR) and the Calibration Deviation Score (CDS). To address the second, we derive prompting conditions from two pragmatic assumptions: that social meaning arises from reasoning over linguistic alternatives, and that listeners infer speaker knowledge states and communicative motives. Applied to a case study on numerical (im)precision across three frontier LLMs, we find that all models reliably reproduce the qualitative structure of human social inferences but differ substantially in magnitude calibration. Prompting models to reason about speaker knowledge and motives most consistently reduces magnitude deviation, while prompting for alternative-awareness tends to amplify exaggeration. Combining both components is the only intervention that improves all calibration-sensitive metrics across all models, though fine-grained magnitude calibration remains only partially resolved. LLMs thus capture inferential structure while variably distorting inferential strength, and pragmatic theory provides a useful but incomplete handle for improving that approximation.

Keywords: large language models, social inference, pragmatics, social meaning, magnitude calibration, pragmatic prompting, evaluation, numerical (im)precision

1. Introduction

Large language models (LLMs) increasingly exhibit sophisticated forms of pragmatic and social reasoning. Recent work has shown that they can recover conversational implicatures (Ruis et al., 2023; Sravanthi et al., 2024; Scherrer et al., 2024), reason pragmatically about scalar expressions (Cho and Kim, 2024), and produce context-sensitive social judgments that align with expert human evaluations (Mittelstädt et al., 2024). A growing body of work further suggests that LLMs can simulate human samples in social science experiments, reproducing population-level patterns of social judgment (Argyle et al., 2023; Santurkar et al., 2023). This paper pursues two related but distinct questions about the quality of this reasoning.

The first concerns measurement. Most existing evaluations of LLM social reasoning focus on directional or categorical agreement: whether a model identifies the correct implication or ranks alternatives similarly to humans. Yet many aspects of human social evaluation are inherently graded. The strength of inferred traits (e.g., competence, friendliness) depends on subtle interactions between linguistic form and context. A model may reproduce the qualitative direction of an effect while systematically exaggerating or attenuating its magnitude. This structure–magnitude dissociation has been documented in broad social science domains, where LLMs have been shown to reproduce effect directions while overestimating magnitudes by factors of 2–10 (Hewitt et al., 2024; Cui et al., 2025; Argyle et al., 2023). Crucially, however, most existing work reports this discrepancy as a descriptive side finding, without metrics designed to quantify it in a principled way (Hullman et al., 2025). We address this gap by introducing two magnitude-sensitive metrics, the *Effect Size Ratio (ESR)* and the *Calibration Deviation Score (CDS)*, that operationalize the distinction between *structural fidelity* and *magnitude calibration*.

The second question concerns explanation and intervention. Can prompting strategies informed by pragmatic theory improve how well LLMs approximate human social meaning? We ground our prompting conditions in two well-established assumptions from pragmatics: that social meaning arises from reasoning over linguistic alternatives in context, and that listeners evaluate speakers by inferring their knowledge states and communicative motives. If LLMs engage in genuinely pragmatic social reasoning, then prompts that explicitly activate these reasoning processes should modulate model behavior in theoretically predictable ways, which allows us to test not only how well LLMs approximate human social meaning, but whether pragmatic theory provides a useful handle for improving that approximation.

We apply both contributions to a case study on numerical (im)precision, a domain in which the interplay between linguistic form, context, and social

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 162–171
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

inference is well documented (Beltrama et al., 2022; Solt et al., 2025), and in which LLM pragmatic behavior has already attracted attention (Tsvilodub et al., 2025). Across three frontier LLMs and four theory-motivated prompting conditions, we find a consistent dissociation: all models achieve high structural alignment but differ markedly in magnitude calibration. Knowledge-and-Motives-Aware prompting partially restores human-like calibration in overconfident models, while combined prompting yields architecture-dependent trade-offs rather than uniform improvement.

2. Theoretical Background

Many instances of social meaning are not directly encoded in linguistic form but emerge from inferential processes listeners apply when interpreting speakers' utterances (Acton, 2019; Beltrama, 2020). These inferences often concern social attributes of the speaker, including competence, knowledgeability, and communicative intent, that listeners update based on the speaker's linguistic choices and the context in which they occur (Beltrama and Papafragou, 2023; Beltrama et al., 2022; Solt et al., 2025). Two well-established assumptions from pragmatics ground our evaluation framework and motivate our prompting conditions.

Reasoning over Alternatives. Listeners evaluate a speaker's linguistic choice against alternatives they could have produced. In Gricean pragmatics (Grice, 1975), a speaker's selection of a weaker or less precise expression where a stronger one was available licenses inferences about their epistemic state or intent (Levinson, 2000). Alternative-sensitive reasoning produces context-dependent social evaluations: the social meaning of numerical precision is modulated by contextual demands, with precise forms enhancing perceived status more strongly in high-precision contexts (e.g., formal testimony) than in casual ones (Beltrama et al., 2022; Solt et al., 2025). Social inference is thus not triggered by form alone, but by the relationship between form, available alternatives, and context.

Speaker Knowledge and Motives. Listeners also infer social attributes by reasoning about *why* a speaker chose a particular expression: what knowledge states and communicative motives plausibly explain the observed choice. This is central to Grice's (1957) account of meaning as intention recognition, and is grounded in the notion that utterances are interpreted against a shared communicative context (Stalnaker, 1999). Empirically, Beltrama and Papafragou (2023) showed that violations of Gricean norms of relevance and informativeness systematically reduce social evaluations of competence and warmth, mediated by listeners' inferences about speaker motives.

RSA as a Unifying Framework. The Rational Speech Act framework (Frank and Goodman, 2012; Goodman and Frank, 2016) is the most prominent formal account within the broader tradition of probabilistic pragmatics (Franke and Jäger, 2016), and integrates both assumptions above. In RSA, a pragmatic listener interprets an utterance by reasoning jointly over the space of alternative utterances a rational speaker could have produced *and* over the speaker's latent beliefs and communicative goals. Social meaning emerges from this joint inference: the same form (e.g., an approximate number) can warrant different social evaluations depending on which alternatives were available and what knowledge state or motive best explains the speaker's choice. This integration motivates treating the two assumptions not as independent factors but as complementary components of a single inferential process, a structure directly reflected in our Combined prompting condition.

Implications for Evaluation and Prompting. These two assumptions have direct methodological consequences. First, they imply that evaluating social meaning in LLMs requires going beyond directional agreement: a model may reproduce the *direction* of a social inference while failing to capture its graded *strength*, which depends on how competing alternatives and inferred speaker states are weighted. This motivates our distinction between structural fidelity and magnitude calibration, and the metrics we introduce to operationalize it.

Second, the assumptions motivate our prompting conditions directly. If social inference involves reasoning over alternatives and epistemic uncertainty about potential knowledge states and motives of the speaker, then prompts that explicitly activate these two aspects should modulate model behavior in theoretically predictable ways — allowing us to test not only how well LLMs approximate human social meaning, but whether pragmatic theory provides a useful handle for improving that approximation.

3. Behavioral Baseline: Social Inferences from (Im)Precision

We ground our LLM evaluation in Experiment 1 of Solt et al. (2025), which investigates how the choice of numerical precision level conveys social meaning about the speaker, and how this meaning is modulated by the pragmatic requirements of the utterance context. The study's central question is whether the degree to which the level of precision in an expression impacts attributions of competence, knowledgeability, and related traits depends on the

	Precise	Approximate
HP (insurance claim)	*"The bicycle cost $500."*	*"The bicycle cost about $500."*
LP (friend inquiry)	*"The bicycle cost $500."*	*"The bicycle cost about $500."*

Table 1: Example stimuli from the *bicycle* scenario across the four experimental conditions. HP context: Jamie reports the cost to an insurance agent. LP context: Jamie answers a friend who is casually considering buying a bicycle. The utterance form (precise vs. approximate) is identical across contexts; only the pragmatic demands differ.

contextual demands for precision, or whether it operates uniformly regardless of context. This tests the core pragmatic prediction that social meaning is not a fixed property of linguistic form but arises from the relationship between form and context. Numerical (im)precision is a particularly well-suited test case for LLM evaluation: it offers a clearly defined set of linguistic alternatives (precise vs. approximate forms), experimentally validated human effect sizes as a quantitative benchmark, and prior evidence that LLMs engage in precision-related pragmatic reasoning (Tsvilodub et al., 2025).

Design and materials. Participants ($N = 371$) were recruited online and randomly assigned to one of six everyday dialog scenarios involving a numerical expression, in one of four conditions, crossing utterance form (precise vs. approximate numerical expression) with contextual precision requirements (high-precision [HP] vs. low-precision [LP] needs). Scenarios were pretested to ensure that their two contextual versions differed reliably in required precision level. For each scenario, participants rated the speaker on six social dimensions using 7-point Likert scales: *competent*, *knowledgeable*, *well-prepared* (competence-related); *helpful*, *likeable* (likeability-related); and *pedantic*. Table 1 illustrates the 2×2 design using the *bicycle* scenario. The HP and LP contexts establish different pragmatic demands for precision, such that the social cost of using an approximate form is predicted to be higher when precision is situationally required.

Results. The study yielded ten statistically significant effects that constitute the directional structure of the human data. Five are *main effects of form*: precise speakers were rated significantly higher than approximate speakers on *competent*, *knowledgeable*, *well-prepared*, *helpful* and pedantic (all $p < .001$). Five additional effects are *form × context interactions*: the rating advantage of precise over approximate was significantly larger in HP than LP for *competent*, *well-prepared*, *helpful*, *likeable* ($p < .001$) and *knowledgeable* ($p < .05$). Together, these effects reflect the context-sensitivity of social meaning: the social cost of imprecision is more pronounced when precision is situationally required, while the social benefit of approximation emerges most clearly when high precision is not called for.

These ten effects (five main effects and five interactions) define the benchmark against which LLM outputs are evaluated in Section 6.

4. LLM Evaluation

Models and protocol. We evaluated three frontier LLMs accessed via API:

- GPT (`gpt-4o-mini`)
- Claude (`claude-sonnet-4-20250514`)
- Gemini (`gemini-2.5-pro`)

For each combination of scenario, context, utterance form, and social attribute, models were prompted to rate the speaker on the given attribute using the identical 7-point scale as in the human experiment. Each query was run $n = 10$ times at temperature $\tau = 1.0$, and model outputs were averaged to compute mean ratings per attribute × context × form condition, matching the structure of the human dataset.

Prompting conditions. To probe the role of pragmatic reasoning in LLM social inference, we implemented four prompting regimes grounded in the theoretical distinctions introduced in Section 2.

- *Minimal (MIN):* Reflects the exact instructions of the human experiment, serving as the baseline for default inference behavior. An example of a full prompt is provided in Appendix A.1.
- *Alternative-Aware (ALT):* Extends the minimal prompt with a one-shot chain-of-thought exemplar (Wei et al., 2022) to elicit explicit reasoning over alternative utterances and their contextual appropriateness, operationalizing the principle of Reasoning over Alternatives (Section 2). The addition to the minimal prompt is provided in Appendix A.1.
- *Knowledge-and-Motives-Aware (KMA):* Extends the minimal prompt with an instruction to consider multiple plausible speaker knowledge states and communicative motives before rating, operationalizing the principle of Speaker Knowledge and Motives (Section 2). The addition to the minimal prompt is provided in Appendix A.1.
- *Combined (COM):* Integrates both extensions above to test whether jointly activating both

pragmatic reasoning components yields improved alignment with human ratings.

5. Evaluation Metrics

Let H and M denote the human and model mean ratings, respectively, for a given attribute, context, and utterance form. We assess alignment at three levels.

Global pattern similarity. For each model–prompting condition pair, we measure overall alignment across all H–M pairs using three complementary metrics. The *Spearman rank correlation* (ρ) captures whether the model preserves the relative ordering of human ratings across conditions, without assuming a linear relationship. The *Concordance Correlation Coefficient* (CCC; Lin, 1989) jointly assesses co-variation and mean-level agreement. The *Root Mean Square Error* (RMSE) quantifies the average absolute deviation between H and M on the original 7-point scale, providing an interpretable measure of magnitude discrepancy.

Structural alignment. We assess whether models reproduce the direction of the ten significant effects established in the human experiment (Section 3). The *Directional Agreement Score* (DAS) checks, for each of the five significant main effects of form, whether the sign of the mean difference $\Delta = M_{\text{precise}} - M_{\text{approximate}}$ matches the human direction: $\text{sign}(\Delta_M) = \text{sign}(\Delta_H)$. The *Interaction Sensitivity Score* (ISS) applies the analogous check to the five significant form $\times$ context interactions, asking for each attribute whether the difference in Δ between HP and LP conditions has the correct sign: $\text{sign}(\Delta_M^{\text{HP}} - \Delta_M^{\text{LP}}) = \text{sign}(\Delta_H^{\text{HP}} - \Delta_H^{\text{LP}})$. Both scores range from 0 to 1, where 1 indicates perfect directional agreement with the human benchmark across all relevant effects.

Magnitude calibration. Beyond directional agreement, we assess whether models reproduce the *magnitude* of the ten significant human effects. The *Effect Size Ratio* (ESR) is computed for each significant main effect and form $\times$ context interaction separately:

$$\text{ESR} = \frac{|\Delta_M|}{|\Delta_H|} \tag{1}$$

where $\Delta = \bar{x}_{\text{precise}} - \bar{x}_{\text{approx}}$ for main effects, and $\Delta = (\Delta^{\text{HP}} - \Delta^{\text{LP}})$ for interactions, with $\Delta^c = \bar{x}^c_{\text{precise}} - \bar{x}^c_{\text{approx}}$ for context c. $\text{ESR} = 1$ indicates perfect magnitude match; $\text{ESR} > 1$ indicates exaggeration; $\text{ESR} < 1$ indicates attenuation. To summarize across all ten effects, the *Calibration Deviation Score* (CDS) is:

$$\text{CDS} = \frac{1}{n}\sum_{i=1}^{n} |\,\text{ESR}_i - 1\,| \tag{2}$$

where i indexes the n significant human effects (main effects and interactions) with non-zero $|\Delta_H|$. Lower CDS indicates closer alignment to human effect magnitudes overall.

By separating structural metrics (DAS, ISS) from magnitude metrics (ESR, CDS), this framework enables principled assessment of both *which* inferences LLMs make and *how strongly* they make them.

Model	Prompt	Spearman ρ	CCC	RMSE
GPT	MIN	0.856	0.703	0.811
	ALT	**0.880**	**0.812**	0.565
	KMA	0.829	0.754	0.601
	COM	0.832	0.786	**0.547**
Claude	MIN	0.849	0.730	0.766
	ALT	0.847	0.754	0.724
	KMA	0.920	0.739	0.701
	COM	**0.946**	**0.804**	**0.576**
Gemini	MIN	0.910	0.622	1.262
	ALT	0.908	0.576	1.422
	KMA	0.923	**0.680**	**1.066**
	COM	**0.932**	0.659	1.125

Table 2: Global pattern similarity metrics per model and prompting condition. Spearman ρ measures rank-order correspondence between model and human mean ratings across all conditions; CCC (Concordance Correlation Coefficient) jointly assesses co-variation and mean-level agreement; RMSE reports average deviation on the 7-point scale. Prompting conditions: MIN = Minimal; ALT = Alternative-Aware; KMA = Knowledge-and-Motives-Aware; COM = Combined. Bold indicates the best value per model per metric.

6. Results

Universal Structure, Variable Calibration. Structural alignment is uniformly high across all models and conditions: DAS and ISS equal 1.0 for all attributes with non-zero human effects, indicating perfect reproduction of both main effect polarity and form $\times$ context interaction directions. Spearman ρ values range from 0.829 to 0.946, confirming strong rank-order correspondence between model and human ratings across conditions.

However, the CCC and RMSE values in Table 2 reveal systematic calibration failures beneath this structural agreement. CCC penalizes not only unsystematic noise but also systematic deviations from the identity line $H = M$; the consistently

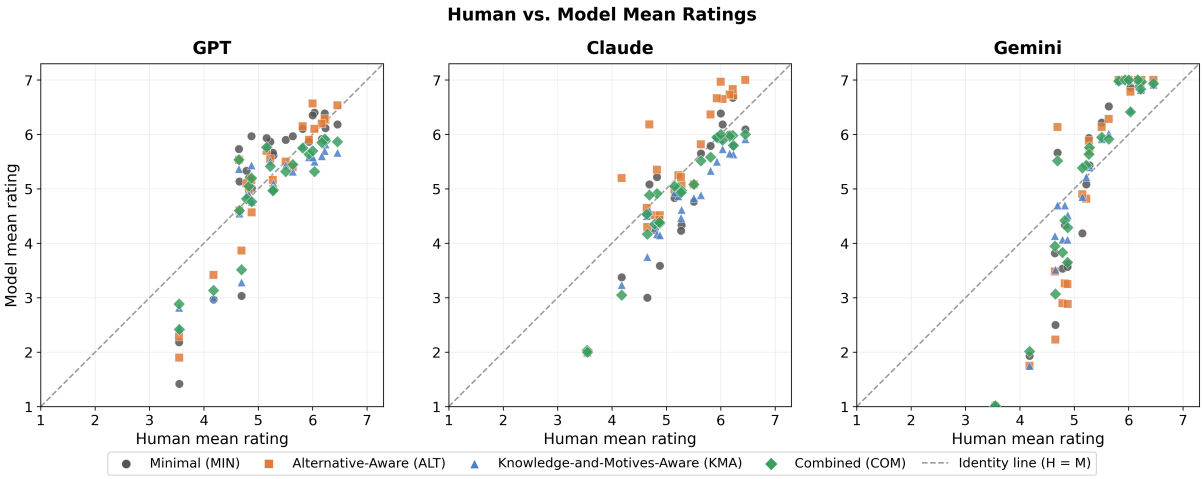


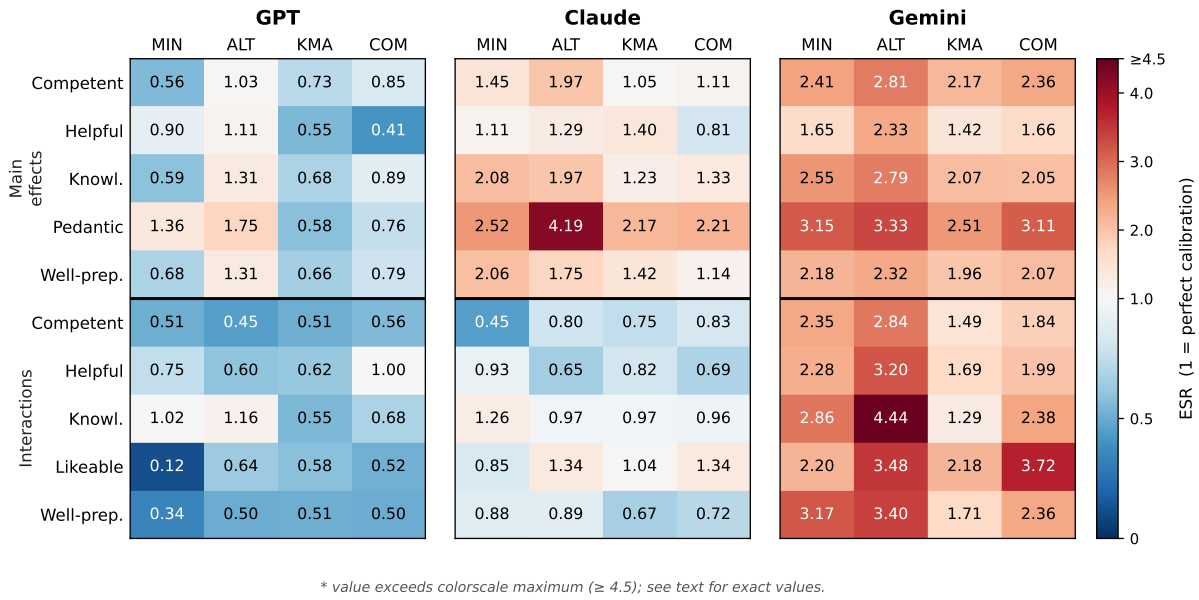
Figure 1: Human vs. model mean ratings across all conditions (scenarios, contexts, utterance forms, and social attributes) for each model and prompting condition. Each point represents one human–model mean rating pair; the dashed identity line ($H = M$) indicates perfect calibration. Points above the line indicate model overestimation; points below indicate underestimation. GPT clusters closely around the identity line across all conditions, reflecting near-calibrated magnitude alignment. Claude shows greater spread, with sensitivity to prompting condition visible in the vertical displacement of individual condition clusters. Gemini displays a characteristic compression along the x-axis with strong vertical spread, reflecting the severe magnitude inflation reported in Table 3. Prompting conditions: MIN = Minimal (gray circles); ALT = Alternative-Aware (orange squares); KMA = Knowledge-and-Motives-Aware (blue triangles); COM = Combined (green diamonds).

lower CCC values relative to Spearman ρ therefore directly operationalize the structure–magnitude dissociation: models preserve the *ordering* of human ratings while distorting their *scale*. This is further reflected in the RMSE values, which quantify the average deviation from human ratings on the 7-point scale. Gemini shows the most severe miscalibration (RMSE: 1.07–1.42), followed by Claude (RMSE: 0.58–0.77) and GPT (RMSE: 0.55–0.81). Figure 1 provides a global visualization of this structure–magnitude dissociation: while all models track the relative ordering of human ratings, systematic vertical displacement from the identity line reveals the degree of magnitude miscalibration for each model and prompting condition.

Magnitude Calibration Across Models. Table 3 reveals systematic architecture-dependent differences in magnitude alignment, with a consistent dissociation between main-effect and interaction calibration. GPT shows the best overall calibration (CDS: 0.3–0.4), with relatively modest deviations on both main effects and interactions across all prompting conditions. Claude exhibits moderate but uneven miscalibration: main-effect CDS varies substantially across conditions (0.4–1.24), suggesting high sensitivity to prompting, while interaction calibration is more stable (0.17–0.23) and consistently lower than main-effect deviation — a pattern not shared by the other models. Gemini displays severe magnitude inflation throughout, with interaction effects particularly affected (CDS_i: 0.67–2.47), frequently exceeding 2–$3\times$ the human effect magnitude, while main-effect miscalibration, though substantial (CDS_m: 1.03–1.72), is comparatively less extreme.

Model	Prompt	CDS_m	CDS_i	CDS
GPT	MIN	0.325	0.460	0.393
	ALT	0.303	0.392	0.348
	KMA	0.360	0.444	0.402
	COM	**0.259**	**0.349**	**0.304**
Claude	MIN	0.845	0.231	0.538
	ALT	1.235	0.205	0.720
	KMA	0.454	**0.167**	**0.310**
	COM	**0.395**	0.228	0.312
Gemini	MIN	1.389	1.572	1.480
	ALT	1.717	2.470	2.094
	KMA	**1.026**	**0.671**	**0.848**
	COM	1.241	1.434	1.338

Table 3: Calibration Deviation Scores for main effects (CDS_m), interaction effects (CDS_i), and their aggregate (CDS). Lower values indicate better magnitude alignment with the human benchmark. Prompting conditions: MIN = Minimal; ALT = Alternative-Aware; KMA = Knowledge-and-Motives-Aware; COM = Combined. Bold indicates the best value per model per metric.



** value exceeds colorscale maximum (≥ 4.5); see text for exact values.*

Figure 2: Effect Size Ratios (ESR) per model, prompting condition, and benchmark effect. Rows are grouped into main effects (top) and form × context interactions (bottom); columns correspond to prompting conditions (MIN, ALT, KMA, COM). Color encodes deviation from perfect calibration (ESR $= 1$, white): blue indicates attenuation, red indicates exaggeration. Values exceeding the colorscale maximum of 4.5 are marked with an asterisk.

Prompting Effects on Calibration. The KMA condition produces the most consistent calibration improvements across models prone to magnitude exaggeration. For Claude, CDS decreases from 0.538 (MIN) to 0.310 (KMA), while for Gemini, CDS decreases from 1.480 (MIN) to 0.848 (KMA), the largest absolute reduction observed across any model–condition pair. GPT, already well-calibrated at baseline, shows moderate sensitivity to prompting: COM achieves the best overall CDS from 0.393 (MIN) to 0.304 (COM).

Alternative-awareness prompting (ALT) produces mixed and sometimes adverse effects. While it reduces main-effect deviation for GPT (CDS_m: 0.325 $\rightarrow$ 0.303) and interaction calibration for Claude (CDS_i: 0.231 $\rightarrow$ 0.205), it substantially amplifies magnitude inflation for Gemini (CDS_i: 1.572 $\rightarrow$ 2.470), suggesting that explicitly foregrounding alternative utterances may exacerbate exaggeration in already poorly calibrated models.

Combined prompting (COM) stands out as the only condition that improves all calibration-sensitive metrics (CCC, RMSE, CDS_m, CDS_i) relative to minimal prompting for every model. For GPT, COM achieves the best overall CDS (0.304) and lowest RMSE (0.547). For Claude, COM produces the strongest global alignment across all three metrics (Spearman ρ: 0.946, CCC: 0.804, RMSE: 0.576) and the best CDS_m (0.395), though KMA yields better interaction calibration (CDS_i: 0.167 vs. 0.228 under COM). For Gemini, COM slightly reduces miscalibration relative to MIN (CDS: 1.480 $\rightarrow$ 1.338) while remaining less effective than KMA on CDS. The consistent cross-model improvement under COM, even for GPT, which shows little sensitivity to individual prompting components, suggests that jointly activating both pragmatic reasoning processes produces reliable alignment gains, even when the individual components yield mixed results. However, the substantial gap between COM and KMA for Gemini's calibration indicates that the two components are not fully additive, and that the chain-of-thought exemplar may partially interfere with the epistemic uncertainty instructions for fine-grained context sensitivity.

Figure 2 provides a detailed view of these patterns across all models, prompting conditions, and benchmark effects. For Claude and Gemini, main effects are systematically exaggerated (ESR > 1), while GPT shows near-calibrated or attenuated main effects throughout. Interaction effects show more model-specific variation across all three models. Gemini displays the strongest and most consistent exaggeration overall, with several cells exceeding the colorscale maximum. A left-to-right reduction in deviation is visible for Claude and Gemini, reflecting the positive effect of the KMA condition.

7. Discussion

Structure Without Calibration. Our results demonstrate a systematic dissociation between structural and quantitative alignment. All models

achieve perfect directional agreement (DAS = ISS = 1.0) and high rank correlations across all prompting conditions, yet CCC values fall consistently below Spearman ρ, and CDS reveals substantial magnitude deviations. This confirms that LLMs reliably learn *which* inferences arise from linguistic form and context, while variably failing to reproduce *how strongly* those inferences operate. The pattern suggests that models acquire directional pragmatic knowledge from training data, but do not faithfully encode the graded, probabilistic character of human social inference.

The architecture-dependent nature of calibration failure is noteworthy. GPT approximates human effect magnitudes closely across all conditions, while Gemini systematically inflates both main effects and interactions — sometimes by factors of 2–3. Claude occupies an intermediate position but is highly sensitive to prompting, suggesting that its default inference behavior is less stable. What drives these between-architecture differences remains an open question: since all three models are closed-source, their training objectives, fine-tuning procedures, and response normalization strategies are not publicly available, and we refrain from drawing strong causal conclusions from behavioral differences alone.

The cross-model consistency of COM has a further implication beyond model evaluation. If explicitly prompting for joint reasoning over alternatives *and* speaker knowledge states is the only intervention that reliably improves calibration across all architectures, this suggests that human-like pragmatic inference may itself require both components to operate simultaneously, consistent with the RSA view that listeners engage in joint inference over utterance alternatives and latent speaker states (Frank and Goodman, 2012; Goodman and Frank, 2016). Conversely, the adverse effects of ALT in isolation suggest that alternative-awareness without epistemic grounding may amplify rather than moderate social inferences. This dissociation could be investigated directly in human participants through paradigms that selectively manipulate access to alternative utterances and speaker context information.

Pragmatic Prompting and Its Limits. The prompting manipulations reveal a partial and asymmetric benefit of pragmatically informed instructions. Explicitly prompting for speaker knowledge states and motives (KMA) consistently reduces magnitude deviation in overestimating models suggesting that directing attention to epistemic uncertainty moderates the exaggeration of social inferences. This is consistent with the theoretical view that pragmatic meaning arises from reasoning over latent speaker states (Goodman and Frank, 2016; Bergen et al., 2016), and that models benefit from having this reasoning made explicit.

Alternative-awareness prompting (ALT), by contrast, produces mixed and sometimes adverse effects. For GPT, it yields modest improvements on both main-effect and interaction calibration. For Claude, it reduces interaction deviation but simultaneously inflates main-effect deviation to its highest value across all conditions, resulting in a net worse overall calibration than minimal prompting. For Gemini, ALT produces the worst calibration observed across any model–condition combination, severely amplifying magnitude inflation relative to baseline. This pattern suggests that explicitly foregrounding utterance alternatives, without anchoring the reasoning in uncertainty about speaker states, amplifies contrast effects rather than moderating them, most severely in models with stronger baseline calibration deficits.

A notable finding is that combined prompting (COM) is the only condition that improves all calibration-sensitive metrics relative to minimal prompting across all three models simultaneously. This cross-model consistency suggests that jointly activating reasoning over alternatives and over speaker knowledge states produces a more robust pragmatic inference process than either component alone. At the same time, COM clearly underperforms KMA on magnitude calibration for both models prone to exaggeration: for Claude, KMA yields better interaction calibration, and for Gemini the advantage of KMA over COM is even more pronounced, affecting both main-effect and interaction calibration. This consistent pattern suggests that the chain-of-thought exemplar introduced by ALT partially interferes with the epistemic uncertainty instructions when both are combined, and that this interference is more severe in models with stronger baseline calibration deficits. The overall picture is one of reliable directional improvement under COM, with remaining architecture-specific trade-offs at the level of individual calibration components.

Limitations and Future Directions. The evaluation is grounded in a single experimental paradigm involving numerical expressions across six scenarios and six social attributes. Generalization to other pragmatic domains, such as scalar implicature, politeness, or register variation, remains to be established. The human benchmark consists of condition means from a published behavioral study (Solt et al., 2025); by-participant variance is accounted for in the original study's statistical analysis, and our evaluation framework follows standard practice in LLM-as-participant work in operating at the level of condition means (Argyle et al., 2023; Santurkar et al., 2023). We evaluate three proprietary frontier models; conclusions should not be generalized to

open-weight architectures or smaller models, which may differ substantially in their pragmatic calibration.

Prompting manipulations approximate the relevant pragmatic reasoning mechanisms without constituting direct implementations. The prompts activate reasoning processes that are theoretically motivated but not formally equivalent to RSA inference; future work could examine whether more explicit computational instantiations of alternative-based or epistemic reasoning yield stronger calibration gains.

Comparing LLM condition means against individual human ratings rather than condition means yields consistently higher RMSE across all models and conditions (by 0.6–0.9 points on the 7-point scale), confirming that mean-level benchmarking is the more conservative measure and that reported calibration deviations are not an artifact of aggregation (see Appendix A.2). Future work could explore whether fine-tuning on calibrated human judgment data yields more robust alignment (Ouyang et al., 2022). The present study focuses on inference-time interventions; training-time calibration objectives remain an important direction for closing the structure–magnitude gap identified here.

8. Conclusion

We investigated whether frontier LLMs approximate human social meaning not only qualitatively but also quantitatively, grounding evaluation in experimentally measured human effect sizes. Across three models and four prompting conditions, all models reliably reproduce the directional structure of human social inference, a finding that is robust across architectures and prompting manipulations, while diverging substantially in magnitude calibration. Pragmatically informed prompting partially reduces these deviations, but its effects are architecture-dependent and not uniformly beneficial. The ESR and CDS metrics introduced here provide principled tools for diagnosing the structure–magnitude dissociation, and we argue that separating directional fidelity from magnitude alignment is a necessary step toward evaluating genuinely human-like social reasoning in LLMs.

Ethics Statement

This study evaluates proprietary LLMs via API under standard access conditions. The human behavioral data was collected in a previously published study (Solt et al., 2025) following standard ethical procedures for online behavioral research. Our findings concern model-level tendencies in social attribute inference; we caution against using automated social judgments of this kind in consequential decision-making contexts without careful human oversight.

Data Availability

The human behavioral data is reported in Solt et al. (2025) and is publicly available at `https://osf.io/m4rhn` (experiment1.csv). The LLM rating data generated in this study is publicly available at `https://github.com/muehlenbernd/llm-social-calibration`.

Acknowledgements

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1412, 416591334.

Bibliographical References

Eric K. Acton. 2019. Pragmatics and the social life of the English definite article. *Language*, 95(1):37–65.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Andrea Beltrama. 2020. Social meaning in semantics and pragmatics. *Language and Linguistics Compass*, 14(3):e12370.

Andrea Beltrama and Anna Papafragou. 2023. We are what we say: Pragmatic violations inform speaker inferences. *Glossa Psycholinguistics*, 2(1).

Andrea Beltrama, Stephanie Solt, and Heather Burnett. 2022. Context, precision, and social perception: A sociopragmatic study. *Language in Society*, pages 1–31.

Leon Bergen, Roger Levy, and Noah D. Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20, Bangkok, Thailand. Association for Computational Linguistics.

Ziyan Cui, Ning Li, Huaikang Zhou, et al. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1):3–44.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

H. Paul Grice. 1957. Meaning. *The Philosophical Review*, 66(3):377–388.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. Working paper, New York University.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. 2025. Validating LLM simulations as behavioral evidence. Preprint.

Stephen C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, MA.

Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

Justin M. Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. *Scientific Reports*, 14:27449.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The Goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 20827–20905. NeurIPS 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cindy (Cynthia) He, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Stephanie Solt, Roland Mühlenbernd, and Mariya Burbelko. 2025. Social meaning and pragmatic reasoning: The case of (im)precision. In *Proceedings of the Experiments in Linguistic Meaning (ELM 3)*, pages 371–382. Linguistic Society of America.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Robert Stalnaker. 1999. Context and content: Essays on intentionality in speech and thought. *Mind*.

Polina Tsvilodub, Kanishk Gandhi, Haoran Zhao, Jan-Philipp Fränken, Michael Franke, and Noah D. Goodman. 2025. Non-literal understanding of number words by language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

A. Supplementary Materials: Appendices, Software, and Data

A.1. Prompt Texts

The following example shows the minimal prompt of the scenario 'bicycle' for the high precision context, the approximate numerical expression, and the social attribute competent.

Minimal Prompt:

Task description:
In this experiment, you'll read a brief description of a situation involving two people. One of these people asks a question, and the second person answers it. Your task will be to answer some questions about the second person.

Task situation:
Jamie's new bicycle was stolen. Fortunately it was insured, so Jamie has called the insurance company.
Insurance agent: "How much did the bicycle cost? I'll start the paperwork right away."
Jamie: "The bicycle cost about $500."

Task:
Based on what Jamie says, how confident does Jamie sound?

Use a 7-point Likert scale:
1 = not at all confident
7 = very confident

Answer with a single number between 1 and 7. Give only the number, no other text.

For the Alternative-Aware condition, the minimal prompt is extended by a one-shot chain-of-thought exemplar, inserted between the *Task Description* and *Task Situation* blocks.

Alternative-Aware Prompt Extension:

Example situation:
Jordan and Sam are planning a work meeting.
Jordan: "Do you know what time the meeting starts?"
Sam: "It starts at around 9."

Example task:
Based on what Sam says, how polite does Sam sound?
Use a 7-point Likert scale:
1 = not at all polite
7 = very polite

Example reasoning (for illustration only):
Sam gives an approximate answer rather than an exact time. In this context, an approximate answer can be appropriate and polite, since it provides useful information without unnecessary detail. Nothing in Sam's response suggests rudeness or disrespect.

Example answer: 6

Now the actual task
You will now see a new situation.
Please answer the question based only on the information given.

For the Knowledge-and-Motives-Aware condition, the *Task* block of the minimal prompt is replaced by an extended version that includes explicit instructions to consider speaker knowledge states and communicative motives prior to rating.

Knowledge-and-Motives-Aware Prompt:

Task:
Based on what Jamie says, how competent does Jamie sound?

Before you answer, please note:
The same utterance can arise from different speaker knowledge states and motivations. You should therefore avoid assuming a single motive or level of knowledge unless the context clearly supports it.

- Step 1: Briefly list two or three plausible reasons why the speaker might have chosen this wording, considering both their possible knowledge state and their communicative goals.
- Step 2: Based on this uncertainty, provide a balanced social evaluation of the speaker.

Now it's your turn: How competent does Jamie sound?

Use a 7-point Likert scale:
1 = not at all competent
7 = very competent

Answer with a single number between 1 and 7. Give only the number, no other text.

The Combined condition integrates both extensions into the minimal prompt. Since they target different blocks of the prompt structure, the two additions can be inserted independently.

A.2. Mean- vs. Individual-Level RMSE

Table 4 compares RMSE computed against human condition means vs individual human ratings. Individual-level RMSE is substantially higher across all models and conditions (by 0.5–0.9 points), confirming that mean-level benchmarking is the more conservative measure and that reported calibration deviations are not an artifact of aggregation.

Model	Prompt	$\mathbf{RMSE}_{\mathbf{mean}}$	$\mathbf{RMSE}_{\mathbf{indiv}}$
GPT	MIN	0.811	1.542
	ALT	0.565	1.429
	KMA	0.601	1.441
	COM	0.547	1.420
Claude	MIN	0.766	1.521
	ALT	0.724	1.497
	KMA	0.701	1.489
	COM	0.576	1.433
Gemini	MIN	1.262	1.822
	ALT	1.422	1.937
	KMA	1.066	1.691
	COM	1.125	1.729

Table 4: RMSE computed against human condition means ($\mathbf{RMSE}_{\mathbf{mean}}$) vs. individual human ratings ($\mathbf{RMSE}_{\mathbf{indiv}}$) per model and prompting condition.

Headlines You Won't Forget: Can Pronoun Insertion Increase Memorability?

Selina Meyer[1] **Magdalena Abel**[2] **Michael Roth**[1]
[1]Natural Language Understanding Lab [2]Cognitive Psychology Lab
University of Technology Nuremberg
{firstname.lastname}@utn.de

Abstract

For news headlines to influence beliefs and drive action, relevant information needs to be retained and retrievable from memory. In this probing study we draw on experiment designs from cognitive psychology to examine how a specific linguistic feature, namely direct address through first- and second-person pronouns, affects memorability and to what extent it is feasible to use large language models for the targeted insertion of such a feature into existing text without changing its core meaning. Across three controlled memorization experiments with a total of 240 participants, yielding 7,680 unique memory judgments, we show that pronoun insertion has mixed effects on memorability. Exploratory analyses indicate that effects differ based on headline topic, how pronouns are inserted and their immediate contexts. Additional data and fine-grained analysis is needed to draw definitive conclusions on these mediating factors. We further show that automatic revisions by LLMs are not always appropriate: Crowdsourced evaluations find many of them to be lacking in content accuracy and emotion retention or resulting in unnatural writing style. We make our collected data available for future work.

Keywords: News Memorability, LLM-based Text Editing, Cognitive Psychology

1. Introduction

News research in NLP is often related to boosting engagement of news articles, as formalized by behavioural data such as article dwell time (Davoudi et al., 2019), likes, retweets, quotes, or replies (Gopalakrishna Pillai et al., 2025; Park et al., 2021). This includes work on generating or editing suitable headlines or social media posts and mitigating the impact of misinformation (Srba et al., 2024).

However, far less is known about how users process and retain news content, an equally critical factor in shaping belief and behaviour. Memorability plays a key role here: what users remember can influence what they believe and share. This is especially relevant in the age of generative AI, which has the potential to accelerate the production and spread of persuasive, yet misleading content (Spitale et al., 2023; Bashardoust et al., 2024; Garry et al., 2024). Cognitive psychology, particularly the illusory truth effect, suggests that repetition alone can enhance perceived truthfulness and increase the likelihood of information being shared (Pennycook et al., 2018; Vellani et al., 2023), highlighting the importance of understanding other factors, such as linguistic characteristics, that can shape memory.

While psychological drivers of belief in fake news have received considerable attention in cognitive psychology (e.g. see Pennycook and Rand, 2021), linguistic aspects that drive memorability of true or reputable news have largely been overlooked (cf. §2). In this paper, we address this gap through a preliminary set of experiments that focus on the memorability of news headlines. In the course of this, we also explore LLMs' capabilities to manipulate news headlines, in the form of directly addressing readers, to make them more memorable. Our experiments test whether minor, targeted edits affect memory in terms of *recognizing* and *recalling* headlines, and whether LLMs can reliably imple-

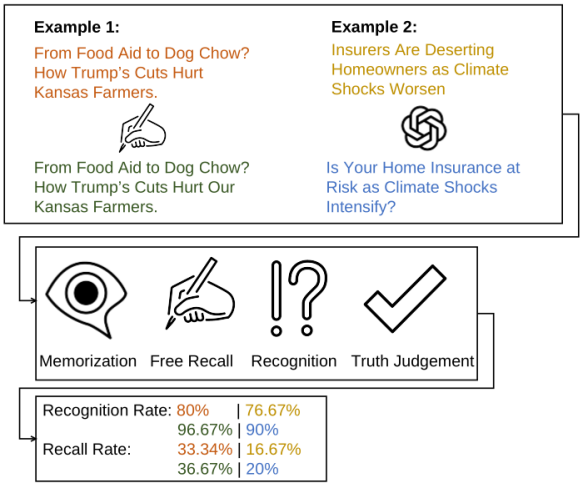


Figure 1: An overview of our experiment design. We ask humans and LLMs to insert first and second person pronouns in pre-existing news headlines. Participants are then shown headlines for a short time with the goal of memorizing them. In the examples shown, pronoun insertion considerably boosted recognition and recall.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 172–187
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

ment relevant edits without distorting the original meaning (§3). Our results show that pronoun insertion has mixed effects on memorability and LLM revisions are not fully reliable (§4).

In summary, our contributions are two-fold: we test LLMs on a linguistically motivated paraphrasing task and we measure downstream effects in memorization studies using experimental methods from cognitive psychology.

2. Related Work

Our work is related to text style transfer in that we manipulate one dimension of a text while preserving its core meaning (Mukherjee et al., 2024), but differs in that we focus on a targeted manipulation rather than changing the overall style of a piece of text. Prior research has shown that despite LLMs' generally impressive capabilities across many NLP tasks (Wei et al., 2022), even large LLMs often fail at simple tasks on which humans achieve perfect performance, such as writing sentences that contain a specific word, word unscrambling, or sentence editing (Efrat et al., 2023; Zhang and He, 2024). Fine-tuned models, even small ones, have been shown to outperform much larger base models on narrow text editing tasks, such as grammar correction (Raheja et al., 2023), whereas zero- and few-shot prompting has been shown to lead to inconsistent performance in text style transfer tasks, including language detoxification and sentiment transfer (Mukherjee et al., 2024), highlighting the continued importance of training data in such tasks.

Specifically focusing on news rewriting and headline generation, Gopalakrishna Pillai et al. (2025) explore different prompting strategies to rewrite news tweets to be more formal, casual, or factual, focusing on increasing predicted engagement, Ao et al. (2021) introduce a dataset of personalized headlines based on user preferences and candidate articles, and Chen et al. (2023) work on methods to leverage clickbaiting techniques, while keeping content faithful to increase reading interest and promote real information.

Beyond NLP-focused work, our approach is informed by findings from psychology, psycholinguistics and marketing showing that direct address and pronoun choice can influence memory even when propositional content is unchanged. For instance, Symons and Johnson (1997) discuss how information framed in relation to the self is more memorable, Brunyé et al. (2011) show that second-person constructions induce stronger reader involvement, and Cruz et al. (2017) suggest a robust effect of second person pronouns on consumer outcomes. Related to news memorability, Lutz et al. (2024) previously found different linguistic cues to affect cognitive and affective processing and Peña et al. (2023) show that tweet-style texts are generally more memorable than news headlines. Also related to our work are studies by Clark et al. (2026) and others on sentence recognition, which however do not take into account recall (i.e., the accessibility in memory in the absence of any retrieval cues). In contrast to this, we follow a common approach in cognitive psychology that provides a broader picture on memory by including measures for both recognition and recall (e.g. MacLeod and Kampe, 1996; Unsworth and Brewer, 2009).

3. Methods

We performed a linguistic analysis on Peña et al.'s data. The findings suggest that personal pronouns help distinguish highly memorable content from less memorable items. To test whether this effect holds with headlines alone, we conducted a pilot study using topic-balanced headlines with and without pronouns. The results indicated that headlines with first and second person pronouns tend to be more memorable.[1] Building on this insight, we explore the capabilities of a range of LLMs to insert such pronouns into real news headlines, without changing the content of the original headline or resulting in an unnatural writing style. Upon asserting the quality of the manipulated headlines, we conduct between-subject user studies to identify the effect of this specific linguistic change on memorability, in the absence of other discriminating factors. Overall, we run three memory studies, each informed by results of the preceding study.

3.1. Memory Studies

Our memory study design is based on established study structures from the field of cognitive psychology (e.g. see Peña et al., 2023; Abel and Bäuml, 2023) and consists of five phases:

- **Presentation Phase.** After reading and agreeing to the informed consent, participants view a fixed number of news headlines for 10 seconds each in random order, with no additional content shown. They are instructed to memorize them.
- **Distraction Phase.** Participants view and react to unrelated images for 60 seconds to reduce potential recency effects.
- **Recall Phase.** Participants freely recall and write down as many headlines as possible, aiming for exact wording. They are encouraged to spend at least 5 minutes on this task. If participants try to move to the next phase

[1] See Appendix A for details on our analysis of Peña et al.'s data and pilot study.

early, the system prompts them to take more time up to two times. After that, they may proceed even if less than 5 minutes have passed.

- **Recognition Phase.** All headlines of the presentation phase plus an equal number of unseen distractor headlines are shown in random order. For each headline, participants are asked to indicate whether they have seen the headline in the presentation phase.
- **Truth Judgement Phase.** In addition to recognition and recall, we also measured perceived truthfulness. To this end, all headlines shown in the recognition phase are presented again in random order. Participants indicate how false or true they personally believe the headline to be on a 7-point likert-scale scale ranging from "definitely false" to "definitely true".

Study Material and Procedure While all three studies share the same underlying design, they differ in study material, group assignments, and number of participants. We describe each study separately below. Across all experiments, each participant group has 30 participants, resulting in a total of 240 participants. The corresponding headline revision procedures are described in §3.2.

Study I We collect 32 headlines from 32 major news outlets, with eight headlines for each of four topics: entertainment, politics, environment, and health, excluding those that originally include pronouns. Using various LLMs, we insert at least one first- or second-person pronoun into 16 headlines, collecting quality judgements by 8 annotators for each revision. We only include LLM revisions judged as both accurate and appropriate by at least 62.5% of annotators.

Participants are randomly assigned to one of two groups. In the presentation phase, each group sees 16 headlines, balanced across topics: 8 original and 8 LLM-revised with pronouns inserted. The assignment is counterbalanced: Group A sees one set revised and the other original, while Group B sees the reverse. In the recognition phase, participants view the 16 headlines they have previously seen plus 16 held-out new ones which serve as distractor items (identical across groups). Of these new headlines, 7 are LLM-revised, ensuring that revised items are not recognized simply due to being the only ones containing pronouns. This design allows us to isolate the effect of pronoun use on memorability while controlling for content.

Study II Qualitative insights from study I results indicated that recognition improved when pronouns were organically integrated in the headline. We hypothesized that humans might achieve this more naturally than LLMs, which often relied on the addition of sentence fragments and clickbaity phrasing (see Table 1 for examples). To investigate this, we include human revisions into our study material: in study II, half of the headlines with pronouns presented in the presentation phase are revised by prolific workers, while the other half is LLM-revised. Again, participants are randomly assigned to two counterbalanced groups.

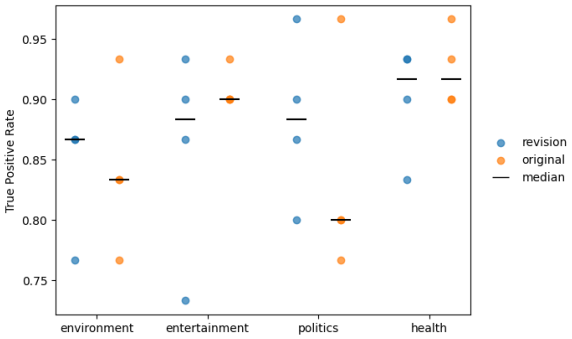


Figure 2: Mean true positive rates of original and revised headlines by topic (environment, entertainment, politics, health) in study II.

Study III Based on results obtained in study II, which indicate strong differences between effects of pronoun insertion across headline topics (see Figure 2), we narrow down the selection of headlines to only one topic for study III. The differences between original headlines and revisions seemed to be strongest for headlines related to politics in study II, leading us to collect 32 new headlines from this topic. All 32 headlines are paired with a revision with pronouns inserted. We only use human revisions for this study and all revisions are written by the same person.

Participants are assigned to four groups with counterbalanced headline sets across presentation and recognition phases. Groups 1 and 3 view opposite versions of original and revised headlines during presentation, while Groups 2 and 4 see the held-out distractor sets from Groups 1 and 3. This design doubles the amount of evaluated headlines and tests whether pronoun use increases false recognition of previously unseen headlines.

3.2. Pronoun Insertion

Models We use 8 LLMs of various sizes, including open-weight and proprietary models, to introduce first or second person pronouns to the collected headlines: `gpt-4o-mini-2024-07-18`, `gpt-4o-2024-08-06`, `Llama-3.1-8B-Instruct`, `Mistral-7B-Instruct-v0.3`, `Mixtral-8x7B-Instruct-v0.1`, `Qwen3-32B`

Headline	Version	Recognition Rate	Recall Rate
LLM Revisions which increased likelihood of recall and recognition			
As the World Warms, Extreme Rain Is Becoming Even More Extreme	Original	73.34	26.67
Are **You** Prepared for the Dramatic Increase in Extreme Rain as Earth Warms?	LLM Revision	86.67	46.67
As **Our** World Warms, Extreme Rain is Becoming Even More Extreme	Human revision	76.67	33.34
Study finds no link between aluminum in vaccines and autism, asthma	Original	76.67	43.34
Autism and Asthma: How a New Study Confirms No Connection to Aluminum in **Your** Vaccines	LLM Revision	90.00	43.34
Insurers Are Deserting Homeowners as Climate Shocks Worsen	Original	80.00	20.00
Is **Your** Home Insurance at Risk as Climate Shocks Intensify?	LLM Revision	86.67	30.00
LLM-Revisions which decreased likelihood of recall and recognition			
From Food Aid to Dog Chow? How Trump's Cuts Hurt Kansas Farmers.	Original	83.34	36.67
Your Kansas Farmers Are Suffering: Trump's Cuts Lead from Food Aid to Dog Chow	LLM Revision	63.34	26.67
From Food Aid to Dog Chow? How Trump's Cuts Hurt **Our** Kansas Farmers.	Human Revision	96.67	36.67
Téa Leoni and Tim Daly Marry in Intimate New York Wedding	Original	83.34	46.67
Your Inside Look at Téa Leoni and Tim Daly's Intimate New York Wedding	LLM Revision	80.00	33.34
Kennedy Family Reunites for Massive Fourth of July Celebration	Original	90.00	36.67
You Won't Believe the Kennedy Family's Massive Fourth of July Reunion!	LLM Revision	83.34	36.67

Table 1: Examples of headlines and LLM-revisions with their recognition and recall rates. Where available, equivalent human revisions are included for comparison.

(thinking mode enabled), `DeepSeek-V3-0324`, and `DeepSeek-R1-0528`. We set the temperature to 0.3 to allow for a limited amount of creativity and pass the same prompt to each model (see Appendix B).

Human Evaluation We collect annotations that assess the accuracy and stylistic appropriateness of revised headlines compared to the originals. Based on the multidimensional quality metrics framework (Lommel et al., 2013), we classify accuracy errors as misrepresentations, additions, or omissions, and style issues as grammar errors, awkwardness, or inconsistency.[2] 128 annotators recruited through Prolific review each original assigned to them alongside a revision and mark it as inaccurate or inappropriate only if at least one subcategory applies. They can also note shifts in tone or emotion. Before participating in annotation, annotators are required to pass a qualification test consisting of four original-revision pairs. Each original-revision pair receives 8 annotations. Annotators see 2–3 revisions per model and never see two revisions for the same headline. To compute inter-annotator agreement (IAA), we calculate the average raw agreement across annotators and annotation groups, as well as the mean of Krippendorff's α across annotator groups. Acceptance rates for accuracy, style, and emotion retention are determined by the proportion of annotators who rated the accuracy and style as acceptable and did not report any shift in emotion or tone relative to the original headline. We collect annotations for 232 original–revision pairs derived from 29 seed headlines, yielding 1,856 judgments.

Human Rewriting In addition to generating LLM revisions, we ask 10 Prolific workers who work in journalism, copywriting, or creative writing, to revise the original headlines. Participants are provided

[2]See Appendix C for instructions provided to annotators.

Model	Accuracy	Style	Emotion
GPT-4o	*62.9*$^{\pm 24.4}$	*65.5*$^{\pm 23.3}$	59.1$^{\pm 20.6}$
DeepSeek-chat	53.9$^{\pm 27.2}$	63.4$^{\pm 19.2}$	*61.2*$^{\pm 25.3}$
DeepSeek-reasoning	**65.5**$^{\pm 24.0}$	**67.2**$^{\pm 21.8}$	**62.1**$^{\pm 19.6}$
GPT-4o-mini	57.3$^{\pm 27.9}$	50.0$^{\pm 22.9}$	52.6$^{\pm 24.0}$
Mixtral	49.6$^{\pm 23.7}$	61.6$^{\pm 22.1}$	54.7$^{\pm 24.4}$
Qwen	47.4$^{\pm 27.2}$	51.7$^{\pm 22.6}$	56.5$^{\pm 20.2}$
Llama3-8b	50.0$^{\pm 24.5}$	47.8$^{\pm 25.5}$	53.5$^{\pm 23.8}$
Mistral	61.2$^{\pm 30.9}$	51.7$^{\pm 22.3}$	60.3$^{\pm 23.2}$

Table 2: Mean acceptance rates in % per model. Highest values per category are **bolded**, second highest values are *italicized*, lowest values are displayed in red.

with instructions that are slightly modified from the prompt given to the LLMs (see Appendix D) and must pass a shortened version of the qualification test used in the LLM revision annotations in order to take part. Each participant rewrites 15 headlines and may skip headlines they feel are not suitable for pronoun insertion. We receive between 7 and 10 revisions per original headline. If two or more participants insert a pronoun in a headline in the same way, it is included in study II (8 headlines overall).

For study III, revisions for all 32 original headlines are obtained from a graduate student enrolled at the university of this work, who is an English native speaker. The student received the same instructions as the prolific workers and revisions were checked for appropriateness and faithfulness to the original meaning by the first author of this work.

4. Results and Discussion

We begin by outlining results regarding different LLMs' performances at the pronoun insertion task defined above as judged by human annotators. After this, we elaborate on the results obtained across the three user studies and offer exploratory analyses to identify potential mediating effects between pronoun insertion and headline memorability.

4.1. LLM Revisions

As a result of annotation, we observe IAA scores for accuracy, style and emotion retention of α = 0.19 (60.32% raw agreement), α = 0.08 (56.64%) and α = -0.03 (55.33%), respectively, indicating that shifts in emotion between original headlines and revisions are especially subjective or difficult to judge for human annotators.

Differences between LLMs Headlines rewritten by DeepSeek-reasoning have the highest acceptance rates across all three annotation categories, whereas Qwen displays the worst performance on accuracy, Llama on style, and GPT-4o-mini on emotion retention (see Table 2). While revisions by larger LLMs generally seem to garner higher acceptance rates, this advantage is not consistent. For instance, the average accuracy acceptance rate for Mistral revisions is 11.64 and 7.32 percentage points higher than for Mixtral and DeepSeek-chat, and only 1.72 percentage points shy of GPT-4o's acceptance rate. This mirrors existing research on LLM's failures at solving simple text editing tasks out-of-the-box (Efrat et al., 2023; Zhang and He, 2024).

For the memorization experiment, we select LLM revisions with minimum acceptance rates of 62.5% (corresponding to at least 5 out of 8 annotators) for style and accuracy. Mean acceptance rates of selected revisions across experiments lie at 81.9% for style, 80.6% for accuracy and 65.13% for emotion retention.

Common Errors in LLM Revisions We qualitatively examine the LLM revisions with style and accuracy acceptance rates of 50% or less to identify common error types. Examples for each identified error type are presented in Table 3. Revisions with low acceptance rates commonly include forms of inappropriate role attribution, which incorrectly frame the reader or writer as an active participant in the headline content. Other common error types include the addition of hallucinated details not present in the original headline, the insertion of evaluative statements, which introduce an author stance not grounded in the original, and the omission of details from the original. In some cases, combinations of multiple error types are displayed at the same time.

4.2. Memory Experiment

As evaluation measures, we compute the true positive rate for each presented headline and the false positive rate for each distractor based on user inputs collected in the recognition phase. We additionally calculate recall rate as the frequency of a headline's appearance in free recall divided by its presentation frequency.

Recall Matching Recalled items are manually matched to their corresponding headlines and ambiguous cases (e.g., items matching multiple headlines) are not counted. For instance, all items containing the word *NASA* and no reference to another headline were matched to the headline *NASA Website Will Not Provide Previous National Climate Reports* or its revision, depending on experimental group. We provide some examples for recalled items and their respective original headlines in Table 5. This resulted in some items displaying a

Original Headline	LLM Revision
Inappropriate Role Attribution	
Bishop of major Catholic diocese exempts parishioners from Mass amid ICE raids	*Why I'm Exempting You* from Mass During ICE Raids
A couple tried for 18 years to get pregnant. AI made it happen	*You and your partner* tried for 18 years. AI helped *you* achieve it.
Hallucinated Details	
Earth is as far away from the sun as it ever gets. So why is it so hot?	You're far from the sun *in July*, but why does it still feel so hot?
Bishop of major Catholic diocese exempts parishioners from Mass amid ICE raids	Amid ICE raids, the bishop says you can skip Mass *for your safety*.
Insertion of Evaluative Statements	
Earth is as far away from the sun as it ever gets. So why is it so hot?	*I'm Baffled:* Earth Is At Its Greatest Distance From The Sun, But Why Am I Still So Hot?
Study finds no link between aluminum in vaccines and autism, asthma	*Your concerns about aluminum in vaccines and autism/asthma are valid*—but new research finds no link.
Ommission of Details	
Measles cases surge to record high *since disease was declared eliminated in the US*	Are You Aware That Measles Cases Have Hit a Record High?
Cierra Leaves 'Love Island USA' Due to a 'Personal Situation' Amid *Backlash Over Resurfaced Post*	Cierra Leaves 'Love Island USA': Here's What Her 'Personal Situation' Means for Fans Like You

Table 3: Examples of commonly found error types in LLM revisions of news headlines. Indicators for each error category are *italicized*.

Recognition TP / FP rates (%)			Recall (%) (cos. sim.)
Study I			
original	$85.4^{\pm 7.2}$	$9.0^{\pm 2.5}$	$34.2^{\pm 12.8}$ ($80.2^{\pm 5.8}$)
revision	$84.0^{\pm 8.1}$	$8.1^{\pm 3.2}$	$32.9^{\pm 10.5}$ ($80.5^{\pm 6.9}$)
Study II			
original	$87.7^{\pm 6.7}$	$9.4^{\pm 2.5}$	$34.4^{\pm 12.1}$ ($79.1^{\pm 7.1}$)
revision	$87.3^{\pm 6.4}$	$10.5^{\pm 5.5}$	$34.2^{\pm 14.3}$ ($75.4^{\pm 5.9}$)
Study III			
original	$80.9^{\pm 8.6}$	$6.2^{\pm 4.2}$	$30.5^{\pm 10.2}$ ($76.6^{\pm 9.7}$)
revision	$80.4^{\pm 9.2}$	$7.2^{\pm 4.9}$	$28.9^{\pm 10.7}$ ($76.0^{\pm 8.7}$)

Table 4: Mean rates of true positive (TP) hits and false alarms (FP) for *recognition* and average *recall* rates.

considerate amount of distortion or lack of detail compared to the headlines seen by participants in the presentation phase. To account for this, we measure recall distortion using the mean cosine similarity, based on S-BERT embeddings (Reimers and Gurevych, 2019), between recalled items and their respective ground truth headline. High similarity with the original means that participants remembered a headline in detail and correctly, whereas low similarity is an indicator for a participant remembering merely the gist of a headline or even misremembering the content of the headline.

Effects of Pronoun Insertion on Memorability Results for all three studies are summarized in Table 4. We run significance tests for the main measures on each study separately. We use two-tailed independent t-tests for normally distributed data and Mann-Whitney U tests were parametric assumptions are violated (see Table 6). Overall, the effects of pronoun insertion on the evaluation measures considered across the three studies are not significant, indicating that pronouns alone do not systematically affect news headline memorability.

Effects of Pronoun Insertion on Perceived Truthfulness Mean results for perceived truthfulness given headline version (original or revision) and whether it had been presented in the presentation phase or not are provided in Table 7. Although all original headlines included in the study were collected from reputable news venues (NYT, NPR, CNN, Yahoo news, CBS, CNBC, NBC, Washington Post, USA Today, and Forbes) and only revisions with a high accuracy acceptance rate were included in the study, we reasoned that the introduction of first and second person pronouns might affect truth judgements. Past work has also found evidence of the illusory truth effect: repetition affects perceived truthfulness of information (Pennycook et al., 2018; Vellani et al., 2023). Like for the main measures,

Recalled Text	Presented Headline	Cosine Similarity
New study shows no link between aluminum in vaccines and autism, asthma	Study finds no link between aluminum in vaccines and autism, asthma	1.00
Beyond the Lights: How Fireworks Affect you and your Animals	Beyond the Light Show: How Fireworks Affect You and Your Animals	0.99
John Goodman shows 200 lb weight loss	John Goodman Shows You His 200-Lb. Weight Loss Transformation.	0.88
JD Vance is pushing trumps agenda	How JD Vance shapes and sells the 'Trump doctrine' on foreign policy	0.72
jd vance forign affairs	What You Need to Know About JD Vance Shaping the 'Trump Doctrine' on Foreign Policy	0.65
Brad Pitt and Angelina Jolie	Brad Pitt Takes a Bold Step in His Legal Battle with Ex-Wife Angelina Jolie Over Their French Winery: Here's What You Need to Know	0.51
NASA will no longer monitor rain sometime	NASA Website Will Not Provide Previous National Climate Reports	0.47
Trump economy.	From Food Aid to Dog Chow? How Trump's Cuts Hurt Kansas Farmers.	0.31

Table 5: Examples for items recalled in the free recall phase and their cosine similarity with the corresponding headlines. While some recalled items are mostly faithful to the original headline shown during the presentation phase, others merely reproduce individual words or a general gist.

Study I	
Recognition TP	$t(30) = 0.54, p = 0.59$
Recognition FP	$t(14) = 0.66, p = 0.52$
Recall	$t(30) = 0.3, p = 0.76$
cos.-sim.	$t(30) = -0.13, p = 0.9$
Study II	
Recognition TP	$U = 136.5, p = 0.76$
Recognition FP	$t(14) = -0.5, p = 0.62$
Recall	$t(30) = 0.04, p = 0.96$
cos.-sim.	$t(30) = 1.57, p = 0.13$
Study III	
Recognition TP	$U = 538, p = 0.73$
Recognition FP	$U = 456.5, p = 0.46$
Recall	$t(62) = 0.64, p = 0.53$
cos.-sim.	$U = 539, p = 0.72$

Table 6: Results of statistical tests for main measures across the three studies.

no difference in truth judgements between original and revised headlines was found across studies. We do observe slight, though statistically insignificant indications of illusory truth effect, meaning that headlines encountered in the presentation phase

		Truth Judgement		
	Seen	*Study I*	*Study II*	*Study III*
Original	False	$4.55^{\pm 1.6}$	$4.42^{\pm 1.6}$	$4.60^{\pm 1.7}$
	True	$5.22^{\pm 1.6}$	$5.11^{\pm 1.7}$	$4.81^{\pm 1.7}$
Revised	False	$5.04^{\pm 1.5}$	$4.76^{\pm 1.8}$	$4.41^{\pm 1.8}$
	True	$5.26^{\pm 1.6}$	$5.11^{\pm 1.7}$	$4.74^{\pm 1.8}$

Table 7: Mean truth judgements of original and pronoun-inserted presentation headlines (seen) and distractor items (unseen).

were considered slightly more truthful on average than headlines first seen in the recognition phase.

Exploratory Analysis and Interpretation On average, revised headlines were longer and had shorter words (average character count: $81.09^{\pm 21.37}$, average word length: $4.84^{\pm 0.58}$) compared to original headlines ($67.78^{\pm 15.07}$, $5.18^{\pm 0.7}$). To identify to what extent this might impact memorability, we pool all headlines across the three studies and perform a correlation analysis, taking into account recognition and recall rates, cosine similarities of recalled items, headline lengths, and average word lengths (see Figure 3). We find a significant negative correlation between recall similarity and headline length based on both word and

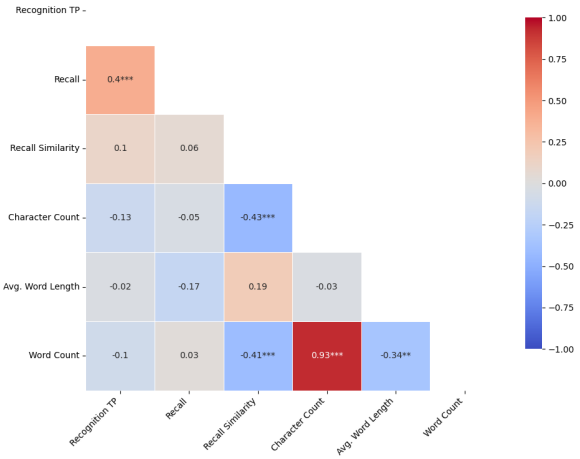


Figure 3: Bonferroni-corrected Pearson correlations between memory measures and headline length features. * denotes significant correlations at $p<0.05$, ** at $p<0.01$, and *** at $p<0.001$.

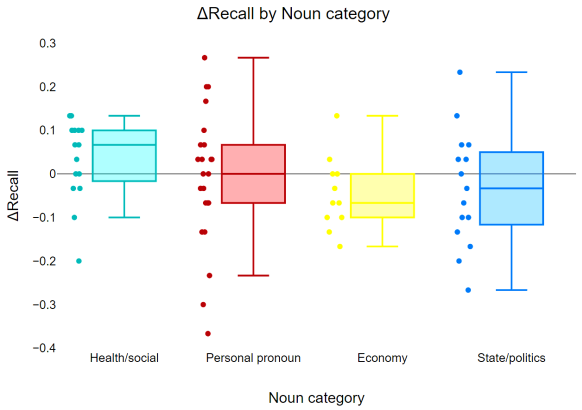


Figure 4: Increases/decreases in recall rate between original and revised headlines for insertions by nominal category, i.e. personal pronouns ("we", "you") and possessives with different nominal heads: Health/social (e.g. "your insurance", "our babies"), Economy (e.g. "your dollars", "our farmers") and State/politics (e.g. "your country", "our election").

character counts, but recall and recognition rates show no clear correlation with length features. This indicates that longer headlines tend to be remembered in less detail, but are retrievable and recognizable at similar rates as shorter headlines. We also find a significant positive correlation between recognition and recall rates, indicating that if a headline can be recognized correctly, it also tends to be accessible in the absence of any retrieval cues (and vice versa).

Qualitative inspection of recognition and recall rates further reveals individual headlines for which the addition of a pronoun clearly increases or decreases recognition. As shown in the examples provided in Table 1, it becomes apparent that revisions that naturally incorporate pronouns—either through restructuring or simple pronoun insertion—show a tendency to improve memorability, while clickbaity or unnatural edits seem to reduce it.

Moreover, effects on recall are influenced by the immediate context of the inserted pronoun: in particular, possessive pronouns with social/health-related nominal heads seem to improve recall, but economic ones do not (e.g. "our babies" vs. "our farmers", see Figure 4). A potential reason for this might be that information related to health and social factors has a higher potential to be personally relevant and contain actionable information, compared to political and economic information which is less directly controllable by consumers of news. Future work could incorporate judgements of personal relevance to verify this.

5. Conclusion

In this paper, we present computational experiments targeting a specific linguistic change, namely the insertion of first- and second-person pronouns into news headlines, along with user studies examining their effects on memorability. We show that LLMs do not always insert pronouns appropriately, as indicated by crowdsourcing evaluations. Collectively, our memory studies lead to the conclusion that pronoun insertion in itself has no consistent effect on memorability. A closer look revealed individual cases of memorability impairment and enhancement, with substantial variation across contexts suggesting a need for more fine-grained analyses and additional data. Moving forward, we plan to investigate other linguistic features that may more strongly influence memorability and to expand our evaluation of LLM capabilities in this context. Ultimately, our goal is to develop computational methods for making news and other information encountered online more memorable while preserving its original content. We believe that this approach has the potential to boost true, high-quality information over misinformation, thus complementing other efforts towards the mitigation of the impact of misinformation on society. To support further research, we release our data, including revision annotations and 7,680 unique memory and truth judgments from three experiments involving 240 participants.[3] We encourage NLP researchers to consider memorability as a modelling feature alongside engagement and related factors.

[3] https://zenodo.org/records/19254945

6. Limitations

We identify several limitations in this work, which we describe below:

We are aware of potential interaction effects of other aspects related to pronoun insertion, such as increases in headline length or long words, changes in syntactic structure, or use of loaded language, which might be responsible for some headlines gaining a boost in memorability, while others saw deterioration when pronouns were inserted. In the future, we plan to develop computational methods to make targeted changes, such as the ones presented, while changing as little as possible of the remaining text. We also plan to explore more impactful changes to news texts, including headlines, to explore a greater variety of linguistic features which can impact memorability in this context, and run studies on larger scales, including more study items in the process, to increase the robustness of experimental results. While we are confident that the set of experiments presented here is suitable to conclude that pronoun insertion alone does not consistently boost memorability of news headlines, we are aware that the comparably small number of included headline items limits our ability to explain why some headlines benefit from pronoun insertion, whereas for others it leads to a decrease in memorability. Consequently, in future studies, we will focus on creating a database large enough to uncover linguistic patterns that interact with changes to increase or decrease memorability.

In cognitive psychology, experiments as the one described in this paper are often conducted in lab settings. While using Prolific for data collection yields many advantages, it also increases the likelihood of participants using external tools to help them remember headlines or be exposed to external distractors. To counteract this, we specifically instructed participants not to use external tools and make sure they are undisturbed for the duration of the study. We also embedded the news headlines as images instead of text, to prevent participants from copy-pasting content for later study phases and made participants aware that their compensation does not depend on their performance during the memory tasks at the beginning of the study.

7. Ethical Considerations

Although previous research has found generative AI less likely to change content and tone of the original message when paraphrasing in formal contexts (e.g. academic, news) than in informal contexts (Tripto et al., 2024), LLM revisions in this context can potentially introduce misinformation and inaccuracies. To prevent showing misinformation to participants during the study, we only included headlines which were judged as accurate by at least 62.5% of annotators, with a mean accuracy acceptance rate of 80.6% over all presented LLM revisions.

For all data collected on Prolific, participants were compensated in GBP at a hourly rate equivalent to the current minimum wage in the country of this work (approx. 11 GBP). Participants who did not pass the qualification test were compensated for their time at the same rate. This corresponds to more than twice the federal minimum wage in the US, where all study participants and annotators were based. The student who rewrote the headlines for study III was employed at the university and compensated at an hourly rate in accordance to the official salary scale for research assistants in the country of this work.

All data was collected anonymously and does not allow conclusions about participants' identities. Participants in the memory study explicitly agreed to provided personal information (e.g. age, political orientation, summarized in Appendix E) being published in the informed consent before beginning the study.

Acknowledgements

We thank Noas Shaalan for support in preparing the data for study III. We also thank the anonymous CMCL reviewers for their valuable and constructive feedback.

8. Bibliographical References

Magdalena Abel and Karl-Heinz T Bäuml. 2023. Item-method directed forgetting and perceived truth of news headlines. *Memory*, 31(10):1371–1386.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Amirsiavosh Bashardoust, Stefan Feuerriegel, and Yash Raj Shrestha. 2024. Comparing the will-

ingness to share for human-generated vs. ai-generated fake news. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW2).

Tad T Brunyé, Tali Ditman, Caroline R Mahoney, and Holly A Taylor. 2011. Better you than i: Perspectives and emotion simulation during narrative comprehension. *Journal of Cognitive Psychology*, 23(5):659–666.

Chih Yao Chen, Dennis Wu, and Lun-Wei Ku. 2023. HonestBait: Forward references for attractive but faithful headline generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4810–4824, Toronto, Canada. Association for Computational Linguistics.

Thomas Hikaru Clark, Greta Tuckute, Bryan Medina, and Evelina Fedorenko. 2026. A distinctive meaning makes a sentence memorable. *Journal of Memory and Language*, 146:104700.

Ryan E. Cruz, James M. Leonhardt, and Todd Pezzuti. 2017. Second person pronouns enhance consumer involvement and brand attitude. *Journal of Interactive Marketing*, 39(1):104–116.

Heidar Davoudi, Aijun An, and Gordon Edall. 2019. Content-based dwell time engagement prediction model for news articles. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 226–233, Minneapolis, Minnesota. Association for Computational Linguistics.

Avia Efrat, Or Honovich, and Omer Levy. 2023. LMentry: A language model benchmark of elementary language tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10476–10501, Toronto, Canada. Association for Computational Linguistics.

Maryanne Garry, Way Ming Chan, Jeffrey Foster, and Linda A Henkel. 2024. Large language models (llms) and the institutionalization of misinformation. *Trends in cognigarry2024largetive sciences*.

Reshmi Gopalakrishna Pillai, Antske Fokkens, and Wouter van Atteveldt. 2025. Engagement-driven persona prompting for rewriting news tweets. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8612–8622, Abu Dhabi, UAE. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sha-

ran Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru,

Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Bernhard Lutz, Marc Adam, Stefan Feuerriegel, Nicolas Pröllochs, and Dirk Neumann. 2024. Which linguistic cues make people fall for fake news? a comparison of cognitive and affective processing. *Proc. ACM Hum. Comput. Interact.*, 8(CSCW1):1–22.

Colin M MacLeod and Kristina E Kampe. 1996. Word frequency effects on recall, recognition, and word fragment completion tests. *Journal of experimental psychology: Learning, memory, and cognition*, 22(1):132.

Sourabrata Mukherjee, Atul Kr. Ojha, and Ondrej Dusek. 2024. Are large language models actually good at text style transfer? In *Proceedings of the 17th International Natural Language Generation Conference*, pages 523–539, Tokyo, Japan. Association for Computational Linguistics.

Kunwoo Park, Haewoon Kwak, Jisun An, and Sanjay Chawla. 2021. How-to present news on social media: A causal analysis of editing news headlines for boosting user engagement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 491–502.

Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.*, 147(12):1865–1880.

Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402.

Tori Peña, Raeya Maswood, Melissa Chen, and Suparna Rajaram. 2023. Memory for tweets versus headlines: Does message consistency matter? *Appl. Cogn. Psychol.*, 37(4):768–784.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text editing by task-specific instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Ivan Srba, Olesya Razuvayevskaya, João A Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, Carolina Scarton, Kalina Bontcheva, and Maria Bielikova. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *arXiv [cs.CL]*.

Cynthia S Symons and Blair T Johnson. 1997. The self-reference effect in memory: a meta-analysis. *Psychological bulletin*, 121(3):371.

Qwen Team. 2025. Qwen3 technical report.

Nafis Irtiza Tripto, Saranya Venkatraman, Dominik Macko, Robert Moro, Ivan Srba, Adaku Uchendu, Thai Le, and Dongwon Lee. 2024. A ship of theseus: Curious cases of paraphrasing in LLM-generated texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6608–6625, Bangkok, Thailand. Association for Computational Linguistics.

Nash Unsworth and Gene A Brewer. 2009. Examining the relationships among item recognition, source recognition, and recall from an individual differences perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6):1578.

Valentina Vellani, Sarah Zheng, Dilay Ercelik, and Tali Sharot. 2023. The illusory truth effect leads to the spread of misinformation. *Cognition*, 236(105421):105421.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. In *Proceedings of the 30th Conference on Pattern Languages of Programs*, pages 1–31.

Yidan Zhang and Zhenan He. 2024. Large language models can not perform well in understanding and manipulating natural language at both character and word levels? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11826–11842, Miami, Florida, USA. Association for Computational Linguistics.

A. Pilot Study

Posthoc Analysis of Peña et al.'s data We obtain Peña et al.'s study data and perform analyses using spacy's (Honnibal et al., 2020) English transformer pipeline to identify linguistic features which may impact memorability of news headlines and tweets in their data. This post hoc analysis reveals significant correlations between memorability and pronoun use (Spearman Rank Coefficient ρ=0.31, p<0.001). At the same time, we also find that these characteristics are used significantly more in tweets than in news headlines (p<0.001) in the study items selected by Peña et al.. We were thus interested in whether these effects would persist, when only news headlines are used. In other words, we wondered if the increase in memorability of tweets compared to news headlines might stem from the language that is used, rather than the content or text type. To address this question, we performed a pilot user study using a within-subject design to reproduce findings from this analysis, using only naturalistic news headlines found in the wild and no tweets.

Pilot Study For our pilot study, we collected 32 news headlines from popular US news outlets with eight headlines for each of four topics: entertainment, politics, environment, and health. Within each topic, half the headlines contain pronouns and half do not. 60 participants were randomly split into two groups: each saw a different set of 16 headlines during the presentation phase, with the other 16 appearing first during the recognition phase. For both groups, headlines were balanced by topic and pronoun use. The study followed the same format as the main study.

We conducted a mixed-effects linear regression to investigate whether the presence of pronouns in headlines influenced their recognition, taking into account participants' experimental group. While neither the main effect of pronoun presence nor the main effect of experimental group reached statistical significance, their interaction did: headlines containing pronouns were recognized at different rates depending on the experimental group ($\beta = 0.73$, SE = 0.31, z = 2.39, p = 0.017).

Follow-up analysis of individual headlines revealed that Group 2—where recognition rates for pronoun-containing headlines were higher—had a greater number of headlines using first- and second-person pronouns compared to Group 1. Across all participants, we also observed a consistent pattern: headlines featuring first- and second-person pronouns were more likely to be recognized than those without any pronouns, reinforcing the importance of pronoun type in headline recognition (see Figure 5).

B. LLM Prompts and Setup

Our prompt is based on a combination of best-practice strategies suggested by White et al. (2023). We provide the LLM with a persona in the system prompt.

We also make use of the alternative approaches pattern and the reflection pattern described by White et al. (2023). After providing a description of

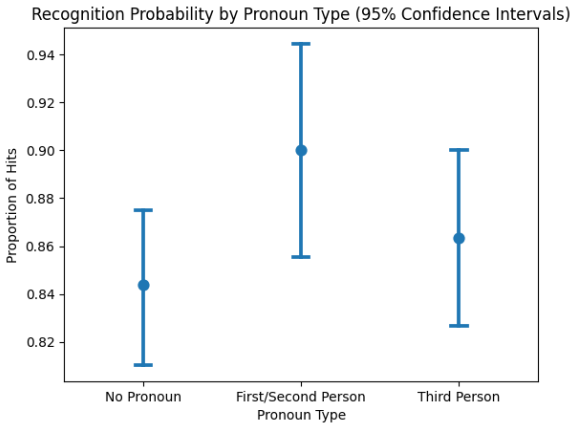


Figure 5: headline recognition likelihood given the contained pronoun type

the task and the constraints, LLMs are prompted to provide five revisions of a given headline. They are then asked to reflect on the quality of the headlines and choose the best headline based on specified criteria. The output is constrained by providing a response template and forcing JSON-format. If a model revision for a headline did not contain a first or second person pronoun on the first try (this happened for gpt-4o-mini, mistral, mixtral, and qwen), the generation process was repeated for the specific headline until the requirement was fulfilled.

System Prompt: *You are an editor at a high-quality newspaper. Your task is to subtly modify article headlines to make them more engaging, without altering the core message.*
Specifically, when given a headline, you will rewrite it by incorporating second-person (e.g., "you", "your") and/or first-person (e.g., "I", "my") pronouns. This will make the headline more relatable and attention-grabbing for the reader. Ensure that the revised headline remains true to the original tone and meaning.
Your goal is to make each headline more compelling and conversational, while maintaining clarity and relevance to the reader's experience. After generating multiple options, choose the one that fits best in terms of engagement, clarity, and relevance for the target audience.

Prompt: *You are given a news article headline. Your task is to rewrite it using first-person ("I", "my") and/or second-person ("you", "your") pronouns to make it more engaging and personally relevant to readers.*
Generate exactly five alternative versions of the headline. Each version should:
- Preserve the original tone and core message as closely as possible.
- Use first-person and/or second-person pronouns to create a direct, conversational appeal.
After generating the five rewrites, analyze which one is the most effective. Your analysis should consider:
- Reader engagement
- Clarity
- Faithfulness to the original meaning
Finally, select the single best version based on your reasoning.
Use the following json-format to return your output (no additional explanation or text):
$rewrite_1$*: First rewritten headline,*
$rewrite_2$*: Second rewritten headline,*
$rewrite_3$*: Third rewritten headline,*
$rewrite_4$*: Fourth rewritten headline,*
$rewrite_5$*: Fifth rewritten headline,*
$reasoning$*: Explain why one version stands out in terms of engagement, clarity, and preservation of the original message.,*
$best_headline$*: The best headline from above*
Original headline: "$row['headline']$"

Setup and Parameters We used the default settings for all LLM calls, with the temperature parameter set to 0.3. GPT (Achiam et al., 2023) and DeepSeek (Liu et al., 2024) models were accessed via the OpenAI API, while Mistral (Jiang et al., 2023) and LLaMA (Grattafiori et al., 2024) ran on a single A40 GPU. Mixtral (Jiang et al., 2024) and Qwen (Team, 2025) models were run on two A40 GPUs. Given the small number of headlines to process, generating all LLM outputs took less than an hour.

C. Instructions for Human Annotation of Rewritten Headlines

The following instructions, adapted from the multidimensional quality metrics framework Lommel et al. (2013) were provided to annotators for the evaluation of LLM-revised headlines. The instructions were followed by a list of 6 more examples. Headlines used for examples and the qualification test did not appear in the main annotation task.

Overview
You will see two versions of a news headline, one marked as "Original", the other as "Revision". A revision is a rewriting of the original headline, so that it includes one or more first or second person pronouns. Your task is to annotate whether the revised version retains the original content and style and to what extent it reflects a news headline you are likely to read on typical news outlets.
You are asked to judge the following categories:
1. Accuracy: Does the content in the revised version accurately reflect the content of the source text?

There are three ways, in which accuracy is commonly violated, which can also occur together:

- ***Misrepresentation**: The revision misrepresents information provided in the original headline*
- ***Addition**: The revision includes content not present in the original headline*
- ***Omission**: The revision is missing content present in the original headline*

***2. Style: Is the language style of the revised headline appropriate?** Inappropriate style can manifest in various forms, that can also occur together:*

- ***Grammar**: The revision contains grammar or language errors*
- ***Awkward Style**: The revision is grammatical, but unnatural as a news headline or awkward (e.g it involves excessive wordiness or overly embedded clauses)*
- ***Inconsistent Style**: The style or tone is inconsistent within the revision (e.g. factual, dry information is paired with sensationalism)*

If a revision differs in tone or emotion compared to the original version, you can indicate this in a separate checkbox. Also, feel free to add comments in the comment field. You can even suggest revision improvements, if you can think of a better phrasing (but this is not the main goal of this annotation task)

Examples:

***Original:** Jonathan Majors reportedly admits to being 'aggressive' with ex-girlfriend in newly released audio clip*
***Revision:** You need to hear Jonathan Majors' disturbing admission about being 'aggressive' with an ex-girlfriend*
1. Does the content in the revised version accurately reflect the content of the source text?
[x] Yes [ ] No
2. Is the language style of the revised headline appropriate?
[x] Yes [ ] No
[x] The revision differs in tone or emotion compared to the original
***Explanation**: The headline is appropriate as a news headline and accurately reflects the content of the original headline. The addition of the word "disturbing" changes the emotional tone of the revision compared to the original.*

D. Instructions for Human Revisions

The following instructions were given to participants when collecting human revisions of news headlines. The same instructions were also given to the graduate student who revised headlines for study III.

You are given a set of news article headlines, one after the other. Your task is to rewrite each headline using first-person (e.g. "I", "my", "our") and/or second-person (e.g. "you", "your") pronouns to make it more engaging and personally relevant to readers.
Your revision should preserve the original tone and core message as closely as possible and use one or more first-person and/or second-person pronouns to create a direct, conversational appeal.
For some headlines, this might be easier than for others. Feel free to make changes to the structure or wording of a headline if needed, but make sure the content and tone stay faithful to the original headline.

E. Demographics of Memory Study Participants

Study I participants were between 20 and 74 years old (mean age: 42.41). 29 identified as female and 28 as male and 3 chose not to disclose. 56.14% held a Bachelor's or Master's degree, 21.05% had some college education, 14.04% had only a high school degree and the rest held associate or professional degrees. The majority of participants (64.91%) were employed. When asked about their political views on a five point likert scale ranging from 1 - left to 5 - right, 19.3% indicated political affiliation with the left and 7.02% with the right, whereas the rest fell between. 28.07% of participants indicated they consumed news on news websites more than once a day and 28.07% once a week or less, with the rest falling in-between. 55.36% consumed news on social media more than once a day, and 19.65% once a week or less.

Study II participants were between 22 and 67 years old (mean age: 41.98). 33 identified as female, 21 male, 2 non-binary and 4 chose not to disclose. 60.72% held a Bachelor's or Master's degree, 21.43% had some college education, 8.93% had only a high school degree and the rest held associate or professional degrees. The majority of participants (67.86%) were employed. 36.36% indicated political affiliation with the left and 23.64% with the right, whereas the rest fell between. 34.55% of participants indicated they consumed news on news websites more than once a day and 14.55% once a week or less, with the rest falling

in-between. 66.07% consumed news on social media more than once a day, and 12.5% once a week or less.

Study III participants were between 21 and 68 years old (mean age: 40.23). 54 identified as female, 66 male. 65% held a Bachelor's or Master's degree, 13.33% had some college education, 10% had only a high school degree and the rest held associate or professional degrees. The majority of participants (78.33%) were employed. 25.83% indicated political affiliation with the left and 25% with the right, whereas the rest fell between. 26.89% of participants indicated they consumed news on news websites more than once a day and 25.21% once a week or less, with the rest falling in-between. 65% consumed news on social media more than once a day, and 10% once a week or less.

Aggregated Transformer Attention Measures Predict Reading Times Beyond Surprisal

Lukas Mielczarek, Laura Kallmeyer
Heinrich-Heine-Universität Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf
{lukas.mielczarek, laura.kallmeyer}@hhu.de

Abstract

Transformer language models trained solely for next-token prediction provide strong expectation-based predictors of human reading behaviour via surprisal. A growing literature suggests that additional signals derived from the internal structure of these models – especially from self-attention – may capture memory-related processing difficulty. This paper uses a range of attention-derived structural costs, specifically *attention entropy*, *attention-weighted distance*, *attention-based activation* and *attention-based memory-focus difference*, that are intended to model memory-related processing costs. We define two aggregation schemes for attention patterns across all heads and layers: global metrics computed from an attention distribution aggregated across heads, and head-averaged metrics computed per head and then averaged. For these metrics, we test whether they predict human reading times beyond surprisal: We outline a methodology to evaluate these metrics using linear mixed-effects models on the Provo eye-tracking dataset, controlling for standard predictors such as word length and frequency and including surprisal as a baseline. As a result, we find that unnormalised head-averaged metrics consistently improve prediction beyond surprisal, while this is not the case for globally aggregated and context-scaled variants.

Keywords: transformers, self-attention, reading times, surprisal, attention entropy

1. Introduction

Human reading times (RTs) are strongly influenced by expectation: less predictable words in context incur greater processing difficulty. This relation is formalised as surprisal (Hale, 2001; Levy, 2008), and transformer language models (LMs) provide surprisal estimates (negative log probabilities) that robustly predict human reading behaviour. However, expectation-based accounts do not fully explain processing difficulty (van Schijndel and Linzen, 2021; Arehalli et al., 2023). Memory-based theories attribute additional costs to limitations in encoding and retrieving representations, such as similarity-based interference during retrieval (Lewis, 1993, 1996) and costs associated with long-distance dependencies (Gibson, 2000).

Transformer self-attention offers a potential bridge between these perspectives. Because attention distributes weight over prior tokens, structural properties of attention distributions – such as entropy (Ryu and Lewis, 2021, 2022) or difference in attention distributions between consecutive timesteps (Oh and Schuler, 2022) – have been proposed as proxies for memory-related processing costs. Yet two key issues remain unclear. First, much of the work has focused on attention entropy, omitting formalisations of important theories in the field of sentence processing, questioning the uniqueness of the contribution of entropy. Second, attention in transformers is multi-headed: studies have opted to analyse individual syntactically specialised attention heads or small selections of heads. It is unknown whether all attentional spread in transformer models is associated with processing difficulties (Ryu and Lewis, 2025). We address these issues by asking the following question: **How well do attention-derived structural costs extracted from transformer LMs trained with a pure language-modelling objective predict human reading times beyond surprisal?**

Concretely, we evaluate four memory-motivated attention metrics – entropy, distance, activation, and attention difference – under a unified framework. For each metric, we compare head-averaged and globally aggregated variants, as well as context-length-scaled versions. Using a transformer language model trained solely for next-token prediction, we test whether these metrics predict go-past RTs in the Provo eye-tracking corpus beyond a baseline including frequency, length, surprisal, and spillover effects.

Our main findings are that unnormalised head-averaged metrics consistently improve prediction beyond surprisal, whereas globally aggregated and context-scaled variants show weaker or no effects. Much of the predictive signal overlaps with sentence position, indicating that attention-derived costs are closely tied to expanding contextual memory demands. These findings suggest that transformer attention encodes structural signals aligned with memory-related difficulty, but that their behavioural relevance depends critically on aggregation choices.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 188–202
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Related Work

Recent work has used LMs to connect theories of human sentence processing with measurable RT predictors. Two lines of theory motivate this: expectation-based (Hale, 2001; Levy, 2008) and memory-based accounts (Gibson, 2000).

Expectation-based theories propose that processing difficulty is driven by the unexpectedness of the next word, which can be opertionalised by *surprisal*. Transformer LMs (Vaswani et al., 2017) yield state-of-the-art estimates of surprisal and reproduce robust correlations between surprisal and human reading behaviour (Wilcox et al., 2023).

Memory-based theories instead attribute difficulty to limitations in encoding and retrieving representations during incremental comprehension. These theories vary in their assumed mechanisms and therefore make different predictions about reading patterns. Two influential approaches are particularly relevant here. First, *cue-based retrieval* theories (Van Dyke and Lewis, 2003) assume that incoming material provides cues that probe memory for previously processed items (words, constituents, referents). When multiple candidates match the cues, retrieval suffers from *similarity-based interference*, yielding increased processing time (Gordon et al., 2002, 2006). Second, *Dependency Locality Theory* (DLT) proposes that integrating new material into an evolving dependency structure incurs higher *integration costs* when the relevant dependencies are long (Gibson, 2000).

Beyond interference and locality, memory-based processing theories also motivate *activation/recency* effects, where representations decay unless reactivated, making later retrieval more costly (Lewis and Vasishth, 2005) , and costs associated with *state change* or reallocation of processing resources, linked in modelling work to shifts in attention focus across time (Oh and Schuler, 2022).

A growing set of studies relates transformer self-attention to memory-based mechanisms. Ryu and Lewis (2021) connect attention to cue-based retrieval and use an attention-based entropy measure to model interference effects. Entropy and distance-based metrics on transformer attention predict RTs across both eye-tracking and self-paced reading datasets (Oh and Schuler, 2022; Ryu and Lewis, 2022). Ryu and Lewis (2025) show that attention entropy aggregated over a subset of heads remains predictive, interpreting heads as parallel retrieval operations.

The present work differs from prior studies by (i) evaluating a broader set of attention-derived, psycholinguistically motivated structural cost measures under unified definitions and (ii) systematically comparing head-wise vs. globally aggregated computation (and corresponding normalisations).

3. Psycholinguistically Inspired Processing Metrics

In the following, starting from attention distributions, we devise measures that are intended to model memory-related psycholinguistic processing factors, which in turn are assumed to explain RTs.

3.1. Attention Distribution

Let the matrices $\mathbf{Q}_h$ contain the queries and $\mathbf{K}_h$ the keys of an input sequence at the h-th attention head. Furthermore, let d_k denote the query/key vector dimensionality. At the i-th position in the sequence, we get the following distribution:

$$\mathbf{A}_h^{[i]} = \text{softmax}\left(\frac{\mathbf{Q}_h^{[i]}\mathbf{K}_h^{\mathsf{T}}}{\sqrt{d_k}} + \mathbf{M}\right) \tag{1}$$

where $\mathbf{M}$ is a causal mask with $\mathbf{M}^{[i,j]} = -\infty$ for $j > i$ and 0 otherwise (Vaswani et al., 2017).

Raw attention $\mathbf{A}_h^{[i,j]}$ indicates how much token i attends to token j in attention head h but the quantities combined in multi-head attention are actually weighted sums of $\mathbf{A}_h^{[i]}$ and a matrix of values generated from the input representations $\mathbf{V}_h$. Furthermore, the output transformation $\mathbf{W}^{\mathbf{O}}$ applied to the concatenated head outputs can be decomposed into per-head subcomponents. This means that the contribution to the output of head h which attention from i to j provides crucially also depends on the magnitude of $\mathbf{V}_h$ and $\mathbf{W}^{\mathbf{O}}{}_h$ (Kobayashi et al., 2020). Therefore, following Oh and Schuler (2022), we take the norm of the weighted value vectors from the attention mechanism. Let $\mathbf{b}^{\mathbf{O}}$ be the optional output transform bias term and m the number of attention heads:

$$\mathbf{N}_h^{[i,j]} := \frac{\mathbf{A}_h^{[i,j]} \cdot \left\|\tilde{\mathbf{V}}_h^{[j]}\right\|_2}{\sum_{k=1}^{i} \mathbf{A}_h^{[i,k]} \cdot \left\|\tilde{\mathbf{V}}_h^{[k]}\right\|_2}, \tag{2}$$

$$\text{with} \quad \tilde{\mathbf{V}}_h^{[j]} = \mathbf{V}_h^{[j]}\mathbf{W}^{\mathbf{O}}{}_h + m^{-1}\mathbf{b}^{\mathbf{O}}.$$

Oh and Schuler (2022) also test a variant normalised with respect to residual connections and layer normalisations. However, since they report no improvement for entropy over the simpler version above, we do not adopt it.

A transformer model can consist of a large number of individual attention heads spread over layers and multi-head attention modules. Inspired by Attanasio et al. (2022), we will quantify the contribution of token j to the transformer output at position i by defining a global attention distribution $\bar{\mathbf{N}}^{[i,j]}$ that is simply the average attention token i assigns to token j taken over all heads in the model.

$$\bar{\mathbf{N}}^{[i,j]} := \frac{1}{m}\sum_{h=1}^{m} \mathbf{N}_h^{[i,j]} \tag{3}$$

3.2. Attention Metrics

The attention-based measures we introduce can be computed in conceptually distinct ways (local/global and normalised/not normalised): First, they can be calculated separately for each attention head using the per-head attention distribution (Equation 2) and then averaged across heads (called *local*). We hope that this approach preserves potential functional specialisation across heads, which has been documented in transformer models (e.g. Voita et al., 2019; Clark et al., 2019). Alternatively, the measures can be computed from the globally aggregated attention distribution (see Equation 3). We refer to this as the *global* variant. Secondly, our attention-derived quantities can be expected to increase with the size of the left context. Thus, we also explore normalised measures that control for trivial sentence-length effects.

Attention Entropy and Similarity-Based Interference Cue-based retrieval theories predict similarity-based interference effects: when several memory representations partially match the retrieval cues, retrieval becomes noisy and slower (Van Dyke and Lewis, 2003; Lewis et al., 2006). If attention implements a soft retrieval mechanism, a sharply peaked attention distribution corresponds to selective retrieval of a single dominant candidate, whereas a diffuse distribution reflects competition among multiple candidates.

We operationalise similarity-based interference as the entropy of the attention distribution based on Ryu and Lewis (2021). For the global attention matrix $\bar{\mathbf{N}}$ for input sequence s, attention entropy at position i is defined as

$$\mathbf{H}^{[i]}_{\text{glob}} := -\sum_{j=1}^{i} \bar{\mathbf{N}}^{[i,j]} \log_2 \bar{\mathbf{N}}^{[i,j]}. \tag{4}$$

For the local variant, entropy is computed separately for each head h and then averaged:

$$\begin{aligned} \mathbf{H}^{[i]}_{h} &:= -\textstyle\sum_{j=1}^{i} \mathbf{N}^{[i,j]}_{h} \log_2 \mathbf{N}^{[i,j]}_{h} \\ \mathbf{H}^{[i]}_{\text{loc}} &:= \tfrac{1}{m} \textstyle\sum_{h=1}^{m} \mathbf{H}_h(i,s). \end{aligned} \tag{5}$$

Higher entropy reflects greater competition among partially matching items and is therefore predicted to correlate positively with reading times.

Distance and Dependency Locality Dependency Locality Theory (Gibson, 2000) (DLT) proposes that integration cost increases with the distance between dependent elements measured by the number of intervening referents (usually nouns and verbs). Establishing longer dependencies requires maintaining representations across more intervening material, leading to greater processing cost.

Since attention assigns continuous weights to prior positions, it does not provide a discrete set of retrieved items. Furthermore, transformer LMs do not overtly distinguish between referents and non-referents. Rather than imposing a threshold to identify "retrieved" tokens and classifying dependents, we define a continuous proxy for integration distance via the expected backward distance under the attention distribution:

$$\mathbf{D}^{[i]} := \sum_{j=1}^{i} \mathbf{N}^{[i,j]}(i-j). \tag{6}$$

Larger values indicate longer-distance retrieval and are predicted to correlate with increased processing cost under DLT.

Because expectation is linear, averaging head-wise expected distances yields the same value as computing the expectation from the globally averaged distribution. We do not distinguish two versions for this measure.

Activation and Memory Decay In models of sentence processing based on the cognitive architecture ACT-R (Anderson et al., 2004; Vasishth and Lewis, 2004), activation decreases as a function of time since last retrieval, raising subsequent retrieval cost (Lewis and Vasishth, 2005). Representations decay over time unless they are reactivated.

To approximate this, we define a recency trace over attention matrices: For an matrix $\mathbf{X}$, the degradation function degr is defined recursively as:

$$\begin{aligned} \text{degr}(\mathbf{X})^{[1,j]} &:= 0 \\ \text{degr}(\mathbf{X})^{[i,j]} &:= (1-\mathbf{X}^{[i-1,j]})\ \text{degr}(\mathbf{X})^{[i-1,j]} \\ &\quad +(1-\mathbf{X}^{[i-1,j]}). \end{aligned} \tag{7}$$

Intuitively, if position j is not attended to at $i-1$, its recency trace increases at i; otherwise, the trace is partially reset. The activation-based cost at i is then defined as the expected recency under the current attention distribution (Equation 8). We compute $\mathbf{W}_{\text{glob}}$ using $\bar{\mathbf{N}}$ and $\mathbf{W}_{\text{loc}}$ by computing $\mathbf{W}_h$ per head and averaging across heads.

$$\mathbf{W}^{[i]} := \textstyle\sum_{j=1}^{i} \mathbf{N}^{[i,j]} \text{degr}(\mathbf{N})^{[i,j]} \tag{8}$$

This predictor is large when attention is allocated to items that have not been recently attended, corresponding to retrieval of low-activation representations. Importantly, it captures temporal recency rather than purely structural distance: a representation may be linearly close yet inactive, or distant yet frequently refreshed.

Attention Difference and Reallocation If contextual representations are stored sequentially in memory, reallocating attention across distant positions may incur additional processing cost. In attention-based models, each timestep redistributes retrieval weight over prior tokens, effectively shifting a soft pointer along a linear memory. We quantify this

shift as the expected displacement between consecutive attention distributions.

For positions $i > 1$, we define the reconfiguration cost as the expected absolute displacement between the attention distributions at $i-1$ and i:

$$\mathbf{J}^{[i]} := \sum_{k=1}^{i-1} \sum_{j=1}^{i} \mathbf{N}^{[i-1,k]} \mathbf{N}^{[i,j]} |j-k|. \quad (9)$$

For completeness, we set $\mathbf{J}^{[1]} := 0$.

Intuitively, this quantity is small when attention remains concentrated on similar memory positions across timesteps, and increases as attention shifts toward more distant tokens. Unlike Earth Mover's Distance (EMD) as used by Oh and Schuler, 2022, our formulation captures overall dispersion in retrieval focus rather than the minimal work required to transform one distribution into the other. In particular, it assigns non-zero cost to dispersed but identical distributions, penalising diffuse attention.

As with our other attention-based metrics, we compute both local and global variants.

3.3. Normalisation

Since all proposed metrics increase with larger left contexts assuming uniform attention distributions, we define length-normalised variants that divide each measure at position i by its maximal attainable value. This maps each metric to the interval $[0, 1]$.

For attention entropy, we adopt the method of Oh and Schuler (2022): the maximum occurs under a uniform distribution over $\{1, \ldots, i\}$ and is therefore $\log i$. For expected backward distance, the maximum occurs when all attention is on the leftmost position, yielding a cost of $i-1$. For activation-based cost, the degradation trace of a position that has never been attended grows linearly with its distance from the current position. The maximal activation cost is thus likewise obtained by attending fully to the oldest, maximally degraded position, yielding $i-1$. Finally, for the difference metric, the maximum occurs when attention shifts between opposite extremes of memory in consecutive steps. In this case, the absolute displacement equals $i-1$. Thus, in all three cases, the normalised metric is gained by dividing the measure by $i-1$ (analogous to Oh and Schuler, 2022's normalisation of EMD). We call the normalised measures $\tilde{\mathbf{H}}, \tilde{\mathbf{D}}, \tilde{\mathbf{W}}$ and $\tilde{\mathbf{J}}$.

4. Methodology

4.1. Language Model Training

We trained a transformer LM comparable in architecture and parameter count to `GPT-2 medium`[1] (Radford et al., 2019) and based on the implementation of Mielczarek et al. (2025).

[1] *Generative Pretrained Transformer*

	Value
parameter count	482 590 464
d_{model}	768
vocab size	222 319
f_{FF}	9
dropout	
– embedding	0.14
– feed-forward	0.32
– residual	0.17
learning rate	1.92×10^{-4}

Table 1: Selection of hyperparameters for training the final model.

Split	Train	Validation
loss	2.93	3.45
perplexity	18.70 (70.39)	31.45 (72.96)
final epochs	8.8	

Table 2: Benchmarks on the final model. In parentheses, the corresponding perplexities of `GPT-2 medium` on the sentence-level are provided).

Prior research on attention patterns for RT prediction has focused on the `GPT-2 small` transformer variant (Ryu and Lewis, 2022; Oh and Schuler, 2022; Ryu and Lewis, 2025) and it remains unclear whether the reported effects generalise to transformers with more attention heads. This motives our use of a larger model. The rationale for training a new model was to take into account only intra-sentential contexts, whereas `GPT-2` was trained using a wide context window that crosses sentence boundaries. Since the theories motivating our attention measures predominantly make predictions with respect to sentence processing and are evaluated on the sentence level in prior literature, we aimed to control for extra-sentential confounders.

The same motivation underlies the use of word-level tokenisation instead of Byte Pair Encoding (Gage, 1994) where psycholinguistic theory often employs the simplifying assumption that whole words constitute cognitive units. Punctuation is removed and all words are uncapitalised. Embedding dimensionality is reduced from 1024 to 768 due to the restricted availability of computational resources. Note that our model has a larger overall dimensionality than `GPT-2 medium` (483M vs 355M) due to our large embedding matrix. Furthermore, a hyperparameter search was performed prior to full training. The key hyperparameters that differ from `GPT-2 medium` are listed in Table 1.

All training was conducted on the pre-processed Wikitext-103-v1 dataset (Merity et al., 2016) consisting of over 100 million tokens from Wikipedia. Since we investigate sentence processing, the training contexts are reduced to the sentence-level and a beginning-of-sequence token is included. We en-

force a maximum sentence length of 60. The final model was trained until convergence using early stopping based on validation performance. The selected checkpoint minimised validation perplexity. Table 2 contains information on the final model.

4.2. Psycholinguistic Evaluation

For answering the research question, we will extract the memory-based measures defined in Section 3 from attention scores generated by our language model on a psycholinguistic dataset. Then, we will test whether these measures have an effect on human reading times for that dataset, above and beyond baseline factors that are well-known to influence sentence processing.

Dataset We evaluate predictors on the Provo eye-tracking corpus (Luke and Christianson, 2018). It is based on eye-measurements of 84 native English speakers who were instructed to read 55 paragraphs from diverse sources such as news articles, science magazines and fiction. The paragraphs were presented to the participants in a random order and contained 2.5 sentences each (Luke and Christianson, 2018). From the available metrics in Provo, we use *go-past time* (GPT), which is the sum of all fixation durations including regressive fixations starting with the first fixation on that word up to the point of leaving to the right (Radach et al., 2008). We choose this measure since it is the standard for investigations of high-level meaning integration processes (Radach and Kennedy, 2013) which we assume to be most indicative of any type of memory effects.

Words that were skipped by the participant are not included in the analysis since we assume that they do not cause integrative processes. Where items in Provo correspond to several tokens in our language model, they are tokenised correspondingly and the extracted measures are averaged afterwards.

Linear Mixed-Effects Models We investigate the potential effect of attention-based metrics on GPT with *linear mixed-effects models* (LMEMs), using the `lme4` implementation in `R` (Bates et al., 2015). To assess the significance of a predictor, we compared two nested linear mixed-effects models using a *likelihood ratio test* (LRT). The full model included the predictor as a fixed effect, whereas the reduced model omitted it. Random effects for the predictor to test are included in both models. This allows us to test whether its effect significantly exceeds random group-level variability. Model comparison was conducted via a χ^2 test on the likelihood ratio. The p-value quantifies the probability of obtaining a likelihood ratio statistic at least as large as observed, assuming the reduced model (i.e. no effect of the predictor). We adopt standard practice in rejecting the null hypothesis if p is below the significance threshold of $\alpha = 0.05$.

We also provide the delta in *Akaike Information Criterion* (AIC) and *Bayesian information criterion* (BIC) (see Chapter 2 in Burnham and Anderson, 1998) and the estimated effect sizes.

To maintain generalisability of our models and prevent Type I error inflation (false positives), we aim for a maximal random effects structure, following Barr et al. (2013). However, as a maximal structure is not supported by our data and we encounter issues with the fitting process, we employ the simplification method from Bates et al. (2018).

Baseline Effects We establish a baseline model structure motivated by known factors contributing to RTs. Word frequency and word length are known to be strong predictors (Rayner, 1998). We calculate log frequency using the `wordfreq` Python module Speer (2022). We also include surprisal, i.e. the negative log probability of the respective word given its left context restricted to the sentence-level generated by our LM as a base predictor and we check whether to include features tracking temporal reading progression: sentence and word position inside the paragraph and word position inside the sentence. All predictors are standardised (z-transformed) to allow for faster model fitting.

For the grouping variables, besides the per-participant effects, we consider two hierarchies for nested effects: `lemma` $\rightarrow$ `word` (form) and `paragraph` $\rightarrow$ `sentence` $\rightarrow$ `abs.pos`. Here `abs.pos` corresponds to one specific token.

After finding a maximal model for the three baseline predictors and the given grouping variables, we check for possible *spillover*/*lag*-effects. Processing slowdown is often delayed in reading time data (Ehrlich and Rayner, 1983). We need to account for this in order to ensure that our analysis is not compromised by spurious correlations. Therefore, we consider shifted versions of the predictors, e.g. for the word at position i, we check if we need to take into account word length, frequency and surprisal at $i-1$ (or further). This is done according to the standard procedure: Given spillover window s for a variable, we not only use the value assigned to the current word but also those of the s preceding ones. s is chosen by iteratively performing LRTs for larger spilled-over variants as long as significance is given (cf. Xu et al., 2023). We do this separately for each baseline predictor to prevent overspecification.

5. Results

First, we will outline the results of the LMEM selection process. Then, we will discuss the results of the likelihood ratio tests for the attention measures. Finally, a follow-up analysis will investigate the robustness of our results against spurious effects due

	Estimate	SE
intercept	282.01	5.36
frequency	-7.77	1.95
length	23.11	2.12
surprisal	8.07	1.21
length.1	-7.30	1.23
surprisal.1	4.65	1.22
surprisal.2	2.34	0.85

Table 3: Baseline LMEM effect estimates and standard errors.

to word position.

5.1. Baseline Model

In the following, we describe the baseline effects structure resulting from the model selection process. The complete dataset has 212 436 observations. Conservative cleaning, following standard practices as in Eskenazi (2024); Marsden et al. (2018) (removal of skipped-over words, cut-offs 80–3000 ms, outlier removal by 2.5 SD, removal of sentences with words unknown to the LM tokeniser), leads to 130 151 remaining observations.

Temporal features are not included because the significance threshold was not reached. The `lemma` grouping variable was also discarded due to singularity issues when including random coefficients and non-significance of its random intercept.

Visual inspection confirms that the assumptions that the LRT relies on (linearity of the relationship between predictors and dependent variable, normality of the residuals, homogeneity of variances and normality of the random effects) are met to an acceptable degree. The QQ-plot reveals that the residuals are somewhat right-skewed. We suspect that we could mitigate this by applying a log transformation to the dependent. However, this would reduce interpretability of the model and contradict prior knowledge about the linearity of our base effects, e.g. the linear relationship between surprisal and RTs. Therefore, we rely on the relatively large amount of data to allow reliable estimates.

Additionally, the analysis reveals some irregularities for very low log frequencies. Manual inspection indicates that these extreme values are likely caused by character encoding errors. Therefore, we remove items with values <2 (= occurring less than 100 times per 1 billion words) from the analysis, removing another 550 data points.

The spillover identification process suggests the inclusion of length up to a window of 1 and surprisal with a window of 2. This excludes the first two words per sentence (9240 data points).

Table 3 shows the effect estimates of the baseline model. Correlations are accounted for in the subject grouping variable.

Model Designs Including attention measures required additional simplifications of the random effects structure. In all cases, correlations were not supported by the data. Furthermore, for some candidates, specific random coefficients had to be omitted. The estimated models are given in Table 4 with the first row containing the baseline model without subject-level correlations.

Repeated Testing For each dataset, four candidates are tested with modes weighted/unweighted each and three with global/local features, resulting in 14 statistical tests on the same dataset. This increases the probability of us reporting at least one false positive result and falsely reporting that attention patterns predict RTs. Therefore, we correct the p-values via the *Benjamini–Hochberg* (BH) procedure (Benjamini and Hochberg, 1995).

5.2. Attention-Based Predictions

Overview Table 5 summarises fixed-effect estimates (β; per 1 SD increase), likelihood-ratio tests, raw p-values, BH-corrected q-values, and information-criterion differences (ΔAIC, ΔBIC) for attention-based predictors of go-past time (GPT) relative to their respective baseline models.

Across specifications, three patterns emerge: (i) head-averaged (local), unnormalised metrics consistently improve fit beyond surprisal and lexical controls; (ii) globally aggregated metrics show weaker and less reliable improvements; (iii) context-length scaling largely removes (and sometimes reverses) effects.

Head-Averaged (Local) Metrics All four unnormalised head-averaged metrics – entropy, distance, activation, and difference – significantly improved fit relative to the baseline ($q < 0.05$). Fixed effects were positive and similar in magnitude across predictors ($\beta \approx 3$ ms per SD), indicating that higher attention-derived memory cost is associated with longer go-past times.

Improvements in AIC were modest but consistent across predictors, suggesting small but systematic gains in explanatory power beyond surprisal, word length, frequency, and their spillover terms. Effect sizes are less than half the estimated effect of surprisal (see Table 3). Notably, this contrasts with findings on self-paced reading where attention entropy can exceed surprisal in predictive strength (Oh and Schuler, 2022).

Table 7 provides correlation coefficients for the baseline predictors and the unscaled local attention measures. Correlations for the other predictor candidates can be found in Appendix A. Based on these values, multicollinearity in the LMEMs is not a concern. The attention measures (not included together in a joint LMEM) are strongly correlated,

Candidates	Formula
$\mathbf{H}_{\text{loc}}, \mathbf{D}, \mathbf{W}_{\text{loc}}, \mathbf{J}_{\text{loc}}, \mathbf{W}_{\text{glob}}, \mathbf{J}_{\text{glob}}$	`GPT ~ 1 + freq + len + surp + len.1 + surp.1 + surp.2 + cand + (1 + freq + len + len.1 + surp + cand || subject) + (1 + cand || word) + (1 + surp + surp.1 || paragraph) + (1 + cand || sentence) + (1 + freq + len.1 + cand || abs.pos)`
$\mathbf{H}_{\text{glob}}, \tilde{\mathbf{W}}_{\text{loc}}, \tilde{\mathbf{W}}_{\text{glob}}$	`-(cand | abs.pos)`
$\tilde{\mathbf{H}}_{\text{loc}}$	`-(cand | word)`
$\tilde{\mathbf{D}}, \tilde{\mathbf{H}}_{\text{glob}}$	`-(cand | word) - (cand | sentence)`
$\tilde{\mathbf{J}}_{\text{loc}}$	`-(cand | word) - (surp.1 | paragraph) - (cand | abs.pos)`
$\tilde{\mathbf{J}}_{\text{glob}}$	`-(cand | word) - (cand | abs.pos)`

Table 4: Identified LMEMs for testing the attention-based measures. The second column supplies the LME model descriptions used for individually testing the candidate metrics given in the first column. The term `cand` stands for the respective candidate metric. Rows 2–6 contain the components removed from the maximal model fitted for LRT given in row 1. The notation is explicit – intercepts were not removed. Given in `lme4` syntax. `freq` = frequency, `len` = length, `surp` = surprisal. "`.1`" and "`.2`" signify lagged versions of predictors. "`||`" means that correlations between random effects are not accounted for.

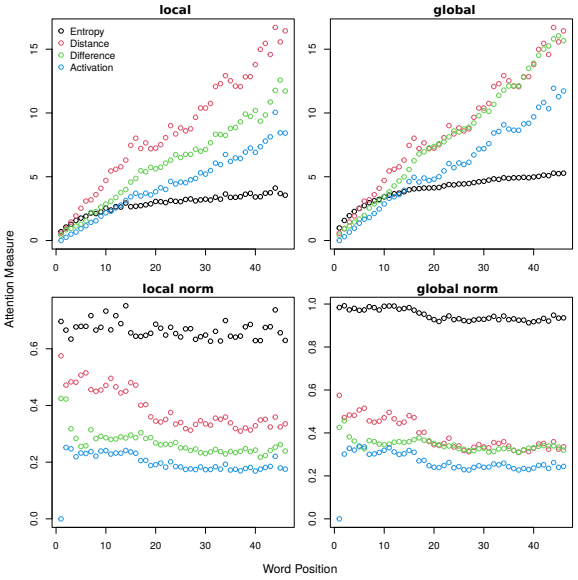


Figure 1: Example attention measures for the sentence *"When a nucleus has all of its proton shells or neutron shells loaded to full capacity, the layers can align so well that each shell can spoon intimately with its attractive neighbors, forming a more compact sphere that fits well within the strong nuclear force's area of influence."*

with the highest correlations of 0.99 occurring between attention distance and activation, and between difference and activation. This suggests that these measures capture closely related aspects of attention structure.

Globally Aggregated Metrics When metrics were computed from the globally averaged attention distribution, results were less consistent. Activation and attention difference remained significant after correction ($q < 0.05$), with positive coefficients comparable to the local variants, whereas global entropy did not reach significance and showed a smaller effect size than local entropy (2.29 vs 2.97).

Inspection of the distributions suggests a mechanistic reason for the weaker global results: after averaging heads, the attention distribution becomes highly diffuse, pushing entropy toward its maximum, thereby reducing variability. Fig. 1 illustrates this for a long sentence: local metrics vary substantially, while global entropy closely tracks its maximum.

These observations indicate that averaging attention *before* metric computation can wash out head-specific structure that is informative for GPT. If attention is interpreted as a retrieval-like mechanism, the contrast between local and global variants suggests that a single pooled "retrieval process" is an oversimplification of how transformers allocate attention. Additionally, our global attention distribution $\bar{\mathbf{N}}$ may be influenced by noisy uninformative attention heads whose contribution to the model is actually low following their output weight $\mathbf{W}^O$. However, our measure takes all attention heads into account equally, without weighing them with respect to their overall magnitude. This may manifest in semi-static noise when computing attention measures on a per-head basis in the local setting (possibly being a contributor to the linear trend observed in Fig. 1), but destroy the global distribution.

At the same time, these findings do not warrant the strong claim that attention *is* cognitive retrieval. A more conservative linking hypothesis is supported: certain attention-derived structural properties correlate with GPT in a direction consistent with memory-based accounts, but the behaviourally relevant signal appears concentrated in specific components as evidenced by the success of Ryu and Lewis (2021, 2022) in using only specific heads, rather than the fully pooled distribution. Architecturally, this suggests that some retrieval-like operations contribute to internal computation without behavioural consequences. It remains a challenge to unify the cognitively plausible view of a single

retrieval mechanism (cf. Timkey and Linzen, 2023) with the multiple parallel and hierarchical attention operations in transformers if we want to explore psycholinguistic hypotheses using these models.

Context-Length Scaling In contrast to the unnormalised predictors, none of the context-length-scaled variants produced significant improvements beyond the baseline after correction. Moreover, several scaled predictors exhibited negative coefficients, indicating shorter predicted GPT with increasing scaled memory cost. The bottom plots in Fig. 1 hint towards a recency bias in attention with the measures declining throughout the sentence (i.e. sublinear growth with word position for distance, activation and difference). The absence of positive scaled effects suggests that GPT aligns more with absolute growth in contextual demands than with proportional attention allocation within an expanding window as suggested by Oh and Schuler (2022).

Magnitude of Improvement Across significant models, ΔAIC values (Table 5) indicate modest but consistent improvements over the baseline. Given the large number of observations, small ΔAIC reductions correspond to systematic likelihood gains; nevertheless, the overall magnitude of improvement remains small relative to the dominant contribution of surprisal. Thus, attention-based metrics explain additional variance in eye-tracking GPT beyond expectation-based and lexical factors, but surprisal remains the dominant predictor, consistent with the findings of Ryu and Lewis (2025).

Robustness to Sentence Position Because the unnormalised attention metrics increase with sentence position, we tested whether their improvements reflect shared variance with word position rather than an independent contribution. Adding position (linear) and log(position) to the maximal baseline model (first row in Table 2) yielded significant positive effects (linear effect: 3.35, p=0.011; log effect: 2.51, p=0.015), confirming a general increase in GPT across sentences not captured by surprisal, length or frequency. Now, controlling for position, the improvements of the attention-based metrics were substantially reduced: only entropy showed a marginal improvement in raw p-values for log(position) resulting in a flipped position effect of -6.85 and an entropy effect of 9.64. However, no effects survived multiple-testing correction (separate correction under the new stricter specification, Table 6).

This suggests that part of the predictive signal in the unnormalised metrics overlaps with generic context growth. Note that this does not imply that attention metrics are spurious. Increased memory costs caused by a larger context might be the cause of the position effect in the first place.

Dependency length Lastly, we correlate $\mathbf{D}$ with dependency lengths extracted from the annotated UD EWT corpus (Silveira et al., 2014). For each word, we computed the sum of distances to all tokens in its left context that are in a dependency relation with it (i.e. its dependents and, where applicable, its head). We find a modest Pearson correlation of 0.22. Manual inspection shows that a large number of short dependency length items were assigned high attention distances (see Appendix B). This could reflect contextualisation needed for next-token prediction in the absence of a recurrent representation.

6. Discussion

In summary, attention-derived structural metrics over all transformer heads provide modest but systematic improvements in predicting reading times beyond surprisal. However, their explanatory contribution depends critically on aggregation choices and overlaps substantially with sentence position.

These results support a cautious view: transformer attention patterns exhibit properties consistent with memory-related processing difficulty, but they should not be equated directly with human retrieval mechanisms. Current results lack evidence for sufficiently distinguishing the precise nature of this effect. Understanding how architectural biases and training objectives give rise to these correlates remains an important direction for future research.

Moreover, additional analyses are necessary to disentangle the contribution of sentence position to RTs from potential memory effects. As the estimated attention effects are already very small and given that our measures seem to naturally correlate strongly with sentence position, reliably detecting a memory effect beyond sentence position might be difficult to achieve with corpus-based methods and require significantly larger amounts of data.

On the other hand, our choice of eye-tracking metric might have contributed to this correlation. GPT includes the duration of regressions to preceding words. If regressions are used to reread words that have been imperfectly detected or forgotten (see Booth and Weger, 2013), a large context would imply the existence of more items of this kind and thus lead to more regressions, increasing the GPT measure.[2] Against this backdrop, reconducting the analysis for lower-level features such as first fixation duration and gaze duration might be advisable. Alternatively, moving-window self-paced reading times would provide a measure that is not confounded by regressive eye movements.

Second, the aggregation choice should be revisited: rather than uniformly averaging heads, weighting heads by their functional contribution (e.g. via output projections) may better preserve

[2]We thank an anonymous reviewer for this hint.

Metric	Variant	Scaling	Effect	SE	χ^2	p	q	ΔAIC	ΔBIC
entropy	local	no	2.97	1.14	6.51	0.011	0.033	-5	5
		yes	1.96	1.13	2.97	0.085	0.099	-1	9
	global	no	2.29	1.13	2.97	0.041	0.072	-2	7
		yes	-1.90	1.04	3.35	0.073	0.093	-1	9
distance		no	3.17	1.26	6.07	0.014	0.033	-4	5
		yes	1.96	0.98	4.39	0.036	0.072	-2	8
activation	local	no	3.15	1.25	6.13	0.013	0.033	-5	5
		yes	-2.04	1.08	3.45	0.063	0.088	-2	9
	global	no	3.16	1.26	6.04	0.014	0.033	-4	6
		yes	-2.09	1.04	3.97	0.046	0.072	-2	8
difference	local	no	3.21	1.24	6.43	0.011	0.033	-4	5
		yes	-1.67	1.06	2.38	0.123	0.133	0	9
	global	no	3.51	1.28	7.20	0.007	0.033	-5	5
		yes	-0.47	0.98	0.00	0.935	0.935	2	11

Table 5: Effects of attention-based predictors on GPT in LMEMs. For each metric, we report the fixed-effect estimate (β; per 1 SD increase), standard error (SE), likelihood-ratio χ^2, raw p-value, BH-corrected q-value, and information-criterion differences relative to the corresponding baseline model (ΔAIC, ΔBIC). Highlighted q-values indicate effects surviving multiple-testing correction (q<0.05).

Metric	Baseline	p	q
entropy	position	0.898	0.898
	log(position)	0.029	0.208
distance	position	0.424	0.484
	log(position)	0.086	0.208
activation	position	0.224	0.358
	log(position)	0.104	0.208
difference	position	0.332	0.443
	log(position)	0.086	0.208

Table 6: p-values and BH-corrected q-values for the effect of local unnormalised predictors controlled for word position.

	GPT	freq	len	surp	$\mathbf{H}_{\text{loc}}$	$\mathbf{D}$	$\mathbf{W}_{\text{loc}}$
freq	-0.16						
len	0.18	-0.76					
surp	0.13	-0.61	0.49				
$\mathbf{H}_{\text{loc}}$	0.02	-0.06	0.09	0.03			
$\mathbf{D}$	0.00	0.03	0.02	-0.07	0.93		
$\mathbf{W}_{\text{loc}}$	0.01	0.00	0.04	-0.03	0.94	0.99	
$\mathbf{J}_{\text{loc}}$	0.01	-0.02	0.06	-0.01	0.93	0.98	0.99

Table 7: Pearson correlations between baseline predictors and local unscaled metrics computed for the observations included in the LMEM analysis.

behaviourally relevant structure. Third, to reduce reliance on post-hoc linking assumptions and maybe to disentangle position effects and various candidates for memory costs, a possible path forward would be to incorporate memory-relevant constraints into modelling itself – either by training objectives that penalise costly attention patterns, by developing process-level models that predict RT as a function of model-internal computation or by developing dynamic models of memory access as discussed by Ryu and Lewis (2025).

Finally, caution is advised when generalising our results to the broad class of transformer-based LMs. Previous work relating attention patterns with RTs has predominantly focused on the `GPT-2 small` variant (Oh and Schuler, 2022; Ryu and Lewis, 2022, 2025). However, given that the predictive power of next-word surprisal decreases for larger transformers (Oh et al., 2022), this raises the question of how model size and architecture influence the predictiveness of other metrics, such as attention entropy. To our knowledge, this question remains unexplored, with the notable exception of Timkey and Linzen (2023), who successfully model interference effects using a single-head architecture. Although we observe modest effects beyond surprisal for a larger model than those used in prior work, these findings do not support strong generalisations. Our differences in results might be due to aggregation scheme or architecture.

However, taken together with the reliable effects reported in previous research, our findings are consistent with the possibility that the predictiveness of aggregate metrics decreases as the number of attention heads increases, potentially due to a growing proportion of heads whose computations do not resemble the retrieval operations posited by psycholinguistic theory. Whether these heads play other cognitively interpretable roles or capture linguistic regularities in alternative ways remains unclear, as does whether increases in the number of heads arising from additional layers and those due to a greater number of parallel heads have comparable effects.

7. Limitations

Several limitations should be acknowledged.

First, the analysis is correlational and conducted on a single eye-tracking dataset. It does not demonstrate a causal role of attention mechanisms in human sentence processing and generalisation to other corpora is not guaranteed.

Second, attention weights as computed here are only indirect proxies for representational contribution and may not fully capture the functional role of information flow within the model due to not accounting for residual connections and layer norms.

Third, the transformer architecture permits parallel retrieval operations via multi-head attention, whereas traditional cue-based retrieval theories assume strong constraints for human memory that allow for only a very limited number of elements to be retrieved when processing a word. The cognitive interpretation of multi-head attention therefore remains speculative.

8. Acknowledgements

We would like to thank Benoit Crabbé for his valuable suggestions and Milica Gašić for sharing computational resources. We also thank the three anonymous reviewers for their helpful feedback. Part of this study was conducted on the Heinrich Heine HILBERT HPC.

9. Code Availability

The `R` notebook with the full linear-mixed effects model selection process and the language model training code are available at `https://github.com/filemon11/MITransformer`.

10. Bibliographical References

John R. Anderson, Daniel Bothell, Michael D. Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An integrated theory of the mind. *Psychological review*, 111 4:1036–60.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2023. Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. ArXiv:2210.12187 [cs].

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based Attention Regularization Frees Unintended Bias Mitigation from Lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119. Association for Computational Linguistics.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Douglas Bates, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2018. Parsimonious mixed models.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Robert W. Booth and Ulrich W. Weger. 2013. The function of regressions in reading: Backward eye movements allow rereading. *MEMORY & COGNITION*, 41(1):82–97.

Kenneth P. Burnham and David R. Anderson. 1998. *Model Selection and Inference. A Practical Information-Theoretic Approach*, first edition. Springer Books. Springer, Newy York, New York, USA.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Kate Ehrlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87.

Michael A. A. Eskenazi. 2024. Best practices for cleaning eye movement data in reading research. *Behavior Research Methods*, 56(3):2083–2093.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 94–126. The MIT Press, Cambridge, Massachusetts, USA.

Peter C. Gordon, Randall Hendrick, Marcus Johnson, and Yoonhyoung Lee. 2006. Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(6):1304–1321.

Peter C. Gordon, Randall Hendrick, and William H. Levine. 2002. Memory-Load Interference in Syntactic Processing. *Psychological Science*, 13(5):425–430.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, New Jersey, USA. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Richard L. Lewis. 1993. *An architecturally-based theory of human sentence comprehension*. Ph.D. thesis, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Richard L. Lewis. 1996. Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research*, 25(1):93–115.

Richard L. Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10(10):447–454.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Emma Marsden, Sophie Thompson, and Luke Plonsky. 2018. A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, 39(5):861–904.

Lukas Mielczarek, Timothée Bernard, Laura Kallmeyer, Katharina Spalek, and Benoit Crabbé. 2025. Modelling expectation-based and memory-based predictors of human reading times with syntax-guided attention. In *Proceedings of the Second Workshop on the Bridges and Gaps between Formal and Computational Linguistics (BriGap-2)*, pages 52–71, Düsseldorf, Germany. Association for Computational Linguistics.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Byung-Doh Oh and William Schuler. 2022. Entropy- and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal. ArXiv:2212.11185 [cs].

Ralph Radach, Lynn Huestegge, and Ronan Reilly. 2008. The role of global top-down factors in local eye-movement control in reading. *PSYCHOLOGICAL RESEARCH-PSYCHOLOGISCHE FORSCHUNG*, 72(6):675–688.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY*, 66(3, SI):429–452.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

K Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *PSYCHOLOGICAL BULLETIN*, 124(3):372–422.

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Soo Hyun Ryu and Richard L. Lewis. 2022. Using transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. *35th Annual Conference on Human Sentence Processing*.

Soo Hyun Ryu and Richard L. Lewis. 2025. Memory for prediction: A Transformer-based theory of sentence processing. *Journal of Memory and Language*, 145:104670.

Robyn Speer. 2022. rspeer/wordfreq: v3.0. [Online; accessed 26-January-2025].

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720, Singapore. Association for Computational Linguistics.

Julie A. Van Dyke and Richard L. Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49(3):285–316.

Marten van Schijndel and Tal Linzen. 2021. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6):e12988.

Shravan Vasishth and Richard L. Lewis. 2004. Modeling sentence processing in ACT-R. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, IncrementParsing '04, pages 82–87, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Ethan G. Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Weijie Xu, Jason Chon, Tianran Liu, and Richard Futrell. 2023. The Linearity of the Effect of Surprisal on Reading Times across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15711–15721, Singapore. Association for Computational Linguistics.

11. Language Resource References

Luke, Steven G. and Christianson, Kiel. 2018. *The Provo Corpus*. Open Science Framework.

Merity, Stephen and Xiong, Caiming and Bradbury, James and Socher, Richard. 2016. *WikiText-103: A Large Language Modeling Corpus*. Salesforce Research.

Natalia Silveira and Timothy Dozat and Marie-Catherine de Marneffe and Samuel Bowman and Miriam Connor and John Bauer and Christopher D. Manning. 2014. *Universal Dependencies English Web Treebank*. Universal Dependencies.

A. Predictor Correlations

length.1	-0.03	0.07	-0.03	-0.07	0.00	0.03	0.04	0.05		
surprisal.1	-0.00	0.06	-0.07	0.00	-0.07	-0.05	-0.04	-0.02	0.55	
surprisal.2	0.01	0.04	-0.04	0.05	-0.08	-0.06	-0.06	-0.06	-0.07	-0.00
	GPT	frequency	length	surprisal	$\mathbf{H}_{\mathrm{loc}}$	$\mathbf{D}$	$\mathbf{W}_{\mathrm{loc}}$	$\mathbf{J}_{\mathrm{loc}}$	length.1	surprisal.1

Table 8: Pearson correlations between spillover versions of baseline predictors and the remaining predictors for the likelihood ratio tests of the local unscaled attention metrics, computed for the observations included in the tests.

$\tilde{\mathbf{H}}_{\mathrm{loc}}$	0.05	-0.24	0.13	0.27	-0.16	-0.13	-0.07			
$\tilde{\mathbf{D}}$	-0.04	0.16	-0.15	-0.13	-0.08	-0.06	-0.01	0.31		
$\tilde{\mathbf{W}}_{\mathrm{loc}}$	-0.02	0.04	-0.08	0.00	-0.02	-0.02	-0.00	0.51	0.86	
$\tilde{\mathbf{J}}_{\mathrm{loc}}$	0.02	-0.19	0.09	0.22	0.02	0.01	-0.03	0.39	0.51	0.68
								$\tilde{\mathbf{H}}_{\mathrm{loc}}$	$\tilde{\mathbf{D}}$	$\tilde{\mathbf{W}}_{\mathrm{loc}}$
$\mathbf{H}_{\mathrm{glob}}$	0.01	-0.01	0.06	-0.04	0.03	-0.05	-0.07			
$\mathbf{D}$	0.00	0.03	0.02	-0.07	0.03	-0.05	-0.06	0.95		
$\mathbf{W}_{\mathrm{glob}}$	0.00	0.03	0.03	-0.06	0.04	-0.04	-0.06	0.95	1.00	
$\mathbf{J}_{\mathrm{glob}}$	0.01	-0.00	0.05	-0.03	0.04	-0.02	-0.05	0.93	0.99	0.99
								$\mathbf{H}_{\mathrm{glob}}$	$\mathbf{D}$	$\mathbf{W}_{\mathrm{glob}}$
$\tilde{\mathbf{H}}_{\mathrm{glob}}$	0.00	-0.08	0.03	0.02	-0.11	-0.12	-0.06			
$\tilde{\mathbf{D}}$	-0.04	0.16	-0.15	-0.13	-0.08	-0.06	-0.01	0.82		
$\tilde{\mathbf{W}}_{\mathrm{glob}}$	-0.04	0.17	-0.16	-0.15	-0.06	-0.04	0.01	0.86	0.97	
$\tilde{\mathbf{J}}_{\mathrm{glob}}$	0.02	-0.13	0.06	0.16	0.00	0.01	-0.06	0.50	0.51	0.42
	GPT	frequency	length	surprisal	length.1	surprisal.1	surprisal.2	$\tilde{\mathbf{H}}_{\mathrm{glob}}$	$\tilde{\mathbf{D}}$	$\tilde{\mathbf{W}}_{\mathrm{glob}}$

Table 9: Pearson correlations between candidate metrics (local scaled, global unscaled and global scaled) and the other predictors for the observations included in the LMEM analyses.

B. Dependency length correlations

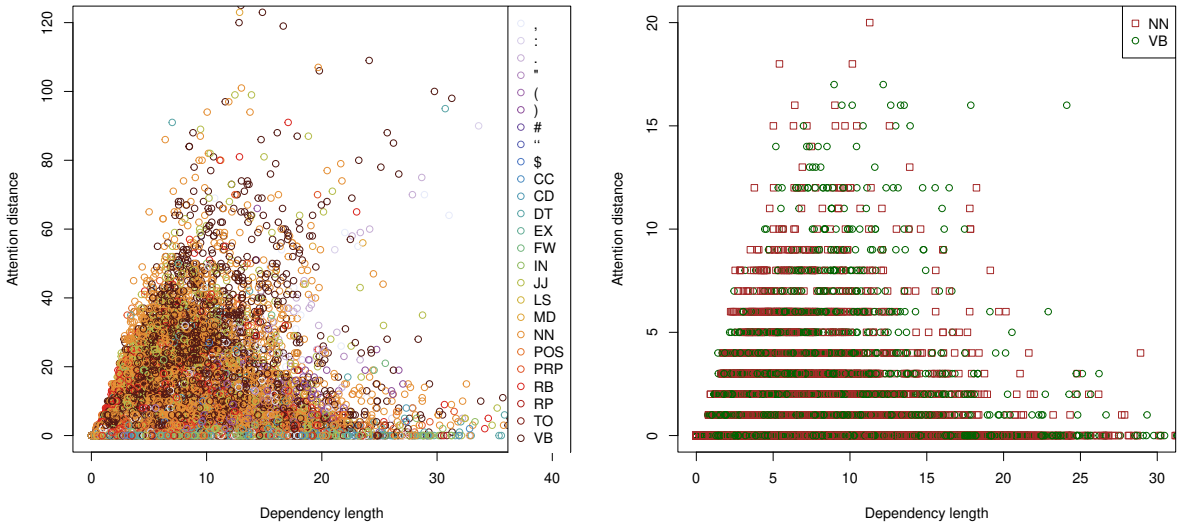


Figure 2: Dependency lengths extracted from the UD English Web Treebank train split plotted against local attention distance generated by our LM. Left plot: dependency length for a token was calculated by summing the surface distances to all items left of it that are connected to it via a relation (both dependents and head). Right plot: distance for a relation was estimated by the number of intervening nouns/verbs and only nouns/verbs were assigned a cost. A few outliers are not shown. Pearson correlation coefficient for the second setting: 0.13.

C. Model Training

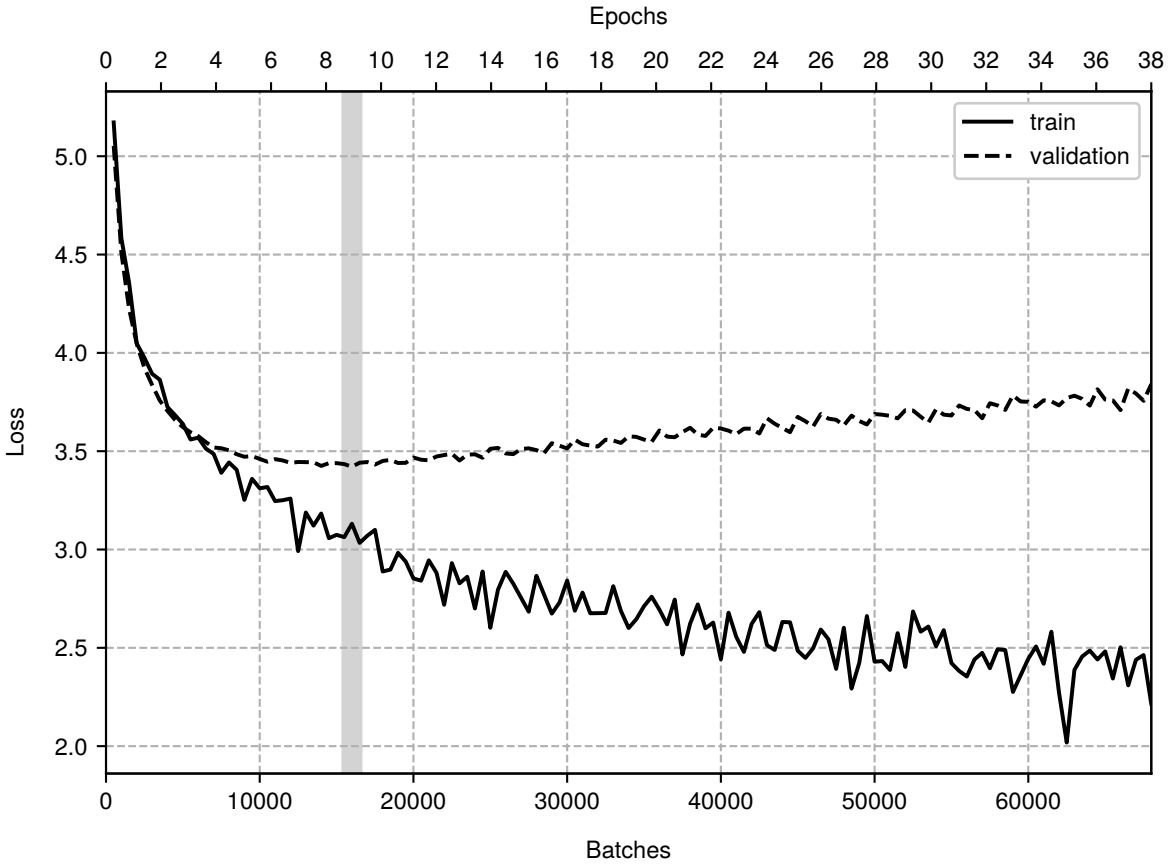


Figure 3: Development of the loss on the training and the development split of the Wikitext corpus during training. The gray vertical bar at 8.8 epochs indicates the region of minimal validation loss which determines the model checkpoint used for psycholinguistic evaluation. One epoch consists of 3 574 204 training sentences. Training was only performed on sentences with a minimum of 3 and a maximum of 60 words. The validation split contains 15 312 sentences.

Resources	Value
batch size	1920 (160 · 12 gradient accumulation)
evaluation interval	every 500 batches
resources	HHU HPC Hilbert
devices	4 Nvidia A100 GPUs
threads (dataloader workers)	8 per process
processor	AMD EPYC 75F3, 2.95 GHz
Peak RAM usage	89.98 Gb DDR4
duration until minimum validation loss reached	10.53 h

Table 10: Model training resources

D. Marginal Effects Plots

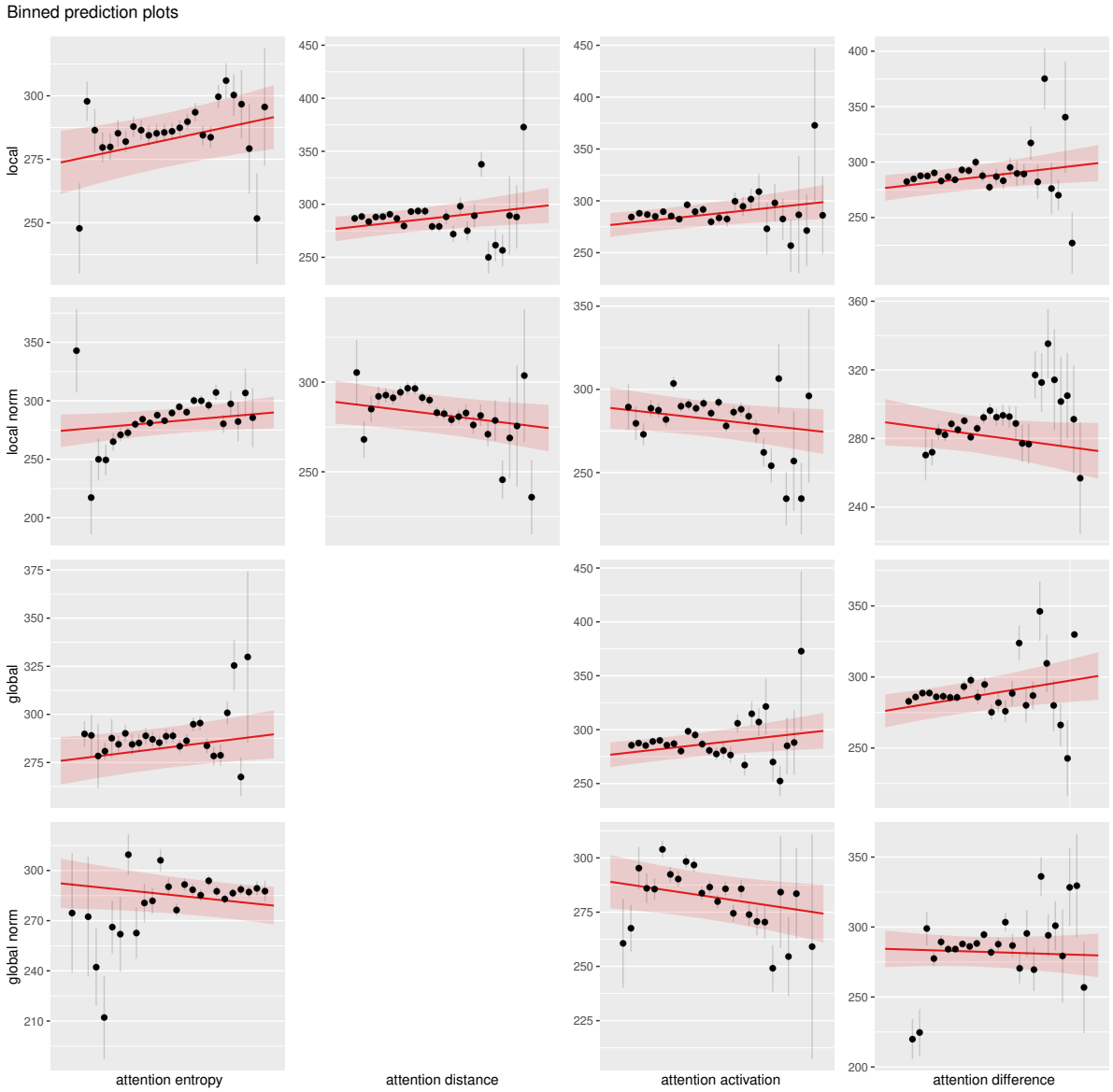


Figure 4: Population-level predicted go-past times as a function of the attention measures (x-axis), based on fixed effects from the mixed-effects models, with other covariates held constant at typical values. Points show mean observed go-past times within evenly spaced bins of the predictor. Shaded areas and error bars indicate 95% confidence intervals.

‘Layer su Layer’: Identifying and Disambiguating the Italian NPN Construction in BERT’s family

Greta Gorzoni, Ludovica Pannitto, Francesca Masini
Alma Mater Studiorum - University of Bologna
greta.gorzoni@studio.unibo.it, ludovica.pannitto@unibo.it, francesca.masini@unibo.it

Abstract

Interpretability research has highlighted the importance of evaluating Pretrained Language Models (PLMs) and in particular contextual embeddings against explicit linguistic theories to determine what linguistic information they encode. This study focuses on the Italian NPN (noun$_i$-preposition-noun$_i$) constructional family, challenging some of the theoretical and methodological assumptions underlying previous experimental designs and extending this type of research to a lesser-investigated language. Contextual vector representations are extracted from BERT and used as input to layer-wise probing classifiers, systematically evaluating information encoded across the model’s internal layers. The results shed light on the extent to which constructional form and meaning are reflected in contextual embeddings, contributing empirical evidence to the dialogue between constructionist theory and neural language modelling.

Keywords: Construction Grammar, Contextual Embeddings, Interpretability in Language Models

1. Introduction

The remarkable empirical performance obtained by Pretrained Language Models (PLMs) across a wide range of tasks has fueled enthusiasm in both computational approaches and theoretical debates about language (Brown et al., 2020). Despite these successes, PLMs remain largely opaque (Rogers et al., 2020). High predictive accuracy does not automatically entail theoretical understanding: it remains unclear what kinds of linguistic information these models encode, how such information is internally structured, and to what extent their representations align with linguistically motivated categories.

Assessing a model’s linguistic competence therefore requires an explicit theoretical characterization of the phenomenon under investigation. Construction Grammar (CxG, Fillmore 1988; Goldberg 1995, 1996) offers a particularly suitable framework in this respect: within CxG, language is conceived as a structured network of conventionalized form–meaning pairings, namely Constructions (Cxns), which may vary in both complexity and schematicity. Since phenomena at the syntax–semantics interface remain comparatively underexplored in interpretability research (Graichen et al., 2026), a constructionist perspective provides a principled way to investigate whether contextual embeddings encode structured form–meaning associations rather than surface-level distributional regularities (Pannitto and Herbelot, 2023; Weissweiler et al., 2022; Rambelli, 2025).

This study focuses on the Italian NPN (noun$_i$-preposition-noun$_i$) constructional family (e.g., *layer su layer* ‘layer upon layer’, Masini 2024a) and builds on previous research on the English NPN Cxn (Scivetti and Schneider, 2025), which showed that BERT encodes information relevant to both Cxn identification and semantic disambiguation. By extending the investigation to Italian, where (as in English) NPN patterns form a network of horizontally related Cxns exhibiting semantic specialization and competition, this work tests whether constructional knowledge is testable across these different languages and across the NPN family of sister Cxns. Our work also addresses a theoretical need within CxG. While constructionist approaches provide a rich descriptive and explanatory framework, they increasingly call for empirical validation. Neural language models, whose representations emerge from large-scale distributional exposure, offer a unique opportunity to test usage-based constructionist hypotheses in a controlled and quantifiable way. By probing contextual embeddings for evidence of constructional organization, this work contributes to the dialogue between linguistic theory and neural language modelling, asking not just whether models perform well, but whether they encode linguistically meaningful structure.

The paper is structured as follows. Section 2 introduces the NPN construction, while Section 3 situates the study within the relevant literature. Section 4 formulates the research questions and details the methodological design. Section 5 introduces the dataset and experimental setup. Sections 6 and 7 present and discuss the results of the identification and semantic disambiguation experiments, respectively.[1]

[1] All code developed for this project is available on GitHub: https://github.com/GretaGorzoni00/NPN

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 203–220
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. The NPN Construction

NPN expressions challenge traditional grammatical categories and motivate a model capable of capturing phenomena along the lexicon–syntax continuum. Formally, the pattern consists of nominal reduplication interrupted by a preposition.

Construction Schema

$Noun_i$ Preposition $Noun_i$

Treating NPN expressions as semi-specified Cxns accounts for both their productivity and their idiosyncratic properties: nominal identity, determiner absence, restricted prepositional choice, as well as characteristic meanings such as succession, distributivity, or contact, which are encoded within the constructional schema itself rather than derived from syntactic operations.

Despite his complex positioning in the constructionist framework (Goldberg, 1996), in his influential analysis of the English NPN pattern, Jackendoff (2008) identifies a limited set of productive prepositions (*by, for, to, after, upon*) and associates them with a restricted inventory of core meanings (e.g., *succession*, *juxtaposition*, *comparison*), alongside numerous idiomatic instantiations. Within this account, NPN is treated as a unified abstract constructional schema pairing a formal template with a family of related meanings. This view has been challenged by Sommerer and Baumann (2021), who argue that the NPN family does not instantiate a single highly abstract Cxn. Instead, they propose a network of horizontally linked, semi-schematic Cxns, emphasizing usage patterns and semantic specialization over maximal abstraction.

The Italian NPN family has been investigated by Masini (2024a), who adopts a similar horizontal-network perspective. Based on corpus evidence from CORIS[2], Masini identifies 8 semi-schematic Cxns, summarized in Table 1. Her dataset amounts to 1,298 types with varying token frequencies (Masini, 2024b). This study focuses only on Cxns with the prepositions *a* 'at/to' and *su* 'on'.

The Italian NPN family exhibits substantial overlap in the noun lexemes licensed by different patterns (Masini, 2024a). Such overlap and semantic competition make the Italian NPN family a particularly informative test case for contextual embeddings, as the coexistence of partially overlapping Cxns with subtly differentiated semantic profiles raises the question of how such distinctions are encoded in distributional representations and whether they are accessible to PLMs.

[2] *Corpus di Riferimento per l'Italiano Scritto*; Favretti et al. 2002.

	Form	Meaning	Types	Tokens
1	$[[x]_{N_i}\ a\ [x]_{N_i}]_{\text{MOD}_j}$	Succession, iteration, distributivity	21	399
2	$[[x]_{N_i}\ a\ [x]_{N_i}]_{\text{MOD}_j}$	Juxtaposition, contact	27	1108
3	$[[x]_{N_i^{sg}}\ su\ [x]_{N_i^{sg}}]_{\text{MOD}_j}$	Succession, iteration, distributivity	71	178
4	$[[x]_{N_i^{pl}}\ su\ [x]_{N_i^{pl}}]_{\text{MOD}_j}$	Greater plurality, accumulation	252	403
5	$[[x]_{N_i}\ per\ [x]_{N_i}]_{\text{MOD}_j}$	Succession, iteration, distributivity	385	3324
6	$[[x]_{N_i}\ per\ [x]_{N_i}]_{\text{SUB.CL}_j}$	Inescapable presupposition	21	21
7	$[[x]_{N_i}\ dopo\ [x]_{N_i}]_{\text{MOD}_j}$	Succession, iteration, distributivity	368	2387
8	$[[x]_{N_i}\ contro\ [x]_{N_i}]_{\text{MOD}_j}$	Juxtaposition, contact	56	170

Table 1: The 8 Italian NPN Cxns postulated by (Masini, 2024a). This work deals with Cxns 1–4.

3. Related work

Recent work has investigated whether LLMs encode constructional knowledge using a variety of experimental designs. One line of research (Tayyar Madabushi et al., 2020; Tayyar Madabushi and Bonial, 2025) examines multiple Cxns organized along a gradient of schematicity, testing whether models generalize across instantiations and whether increasingly schematic patterns remain accessible in contextual embeddings. A complementary approach focuses on individual Cxns and operationalizes constructional knowledge through controlled discrimination tasks, often contrasting target Cxns with carefully designed distractors (Scivetti and Schneider, 2025; Weissweiler et al., 2022). Recently, attention has turned to rare but productive Cxns (Weissweiler et al., 2025), which provide a stringent test for constructional generalization: because they are infrequent, successful modelling cannot rely solely on collostructional frequency effects but must capture construction-specific contributions.

Results converge on a graded pattern: lower-level, lexically anchored Cxns tend to be more accessible to LLMs, while schematic patterns pose greater challenges (Bonial and Tayyar Madabushi, 2024). At the same time, recent work has questioned core assumptions underlying constructional interpretability. Jumelet et al. (2024) argue that model generalizations should not only be measured in terms of performance but also compared to human generalization patterns. Dunn and Eida (2025) further caution against confirmation bias in Cxn probing, emphasizing the importance of detecting false positives and avoiding uncritical expansion of the Constructicon (viz., the network of Cxns) based on model behaviour alone.

This work directly builds upon Scivetti and Schneider (2025), who investigate the English NPN Constructional pattern using a probing framework. In their study, contextual embeddings extracted from BERT are used as input to a linear classifier to assess whether constructional information is encoded in the model's representations. Their dataset includes instances of the NPN Cxn (e.g., 'I was living *moment to moment*') contrasted with superficially similar distractors (e.g., 'In Rome largesse was doled out by *individuals to individuals*'), with lemma-level disjoint train–test splits to prevent lexical memorization. Two experiments target Cxn identification and one addresses semantic disambiguation. Results show that BERT-based probes reliably distinguish Cxns from distractors and capture semantic distinctions, with peak performance in middle-to-late layers, outperforming static embedding baselines.

4. Research Questions and Methodological Design

As Cxns are assumed to be inherently language-specific, probing constructional knowledge requires moving beyond the English-centric focus that characterises much of the existing literature. Moreover, the NPN Cxn occupies an intermediate position on the lexicon–syntax continuum, making it a suitable test case for assessing whether models can capture constructional generalizations that are neither fully schematic nor fully lexically specified.

This study addresses two research questions:

RQ1 Can BERT distinguish instances of the Italian NPN Cxn from distractors?

RQ2 Can BERT distinguish between the construction-specific meanings associated with different NPN instantiations within the Italian NPN family?

Following Scivetti and Schneider (2025), we adopt a probing framework in which contextual embeddings extracted from BERT family's models are used as input to a linear classifier. The preposition is operationally treated as the structural head of the Cxn (Jackendoff, 2008), and its embedding is selected as a primary locus for probing.

To address the research questions, two complementary tasks are implemented: (i) Cxn identification, testing whether the model distinguishes actual NPN instances from formally similar sequences (see Section 6), and (ii) semantic disambiguation, assessing whether the information encoded in the contextual embedings are adequate to distinguish between the construction-specific meanings associated with different NPN instantiations.

Our study extends Scivetti and Schneider (2025) in several important ways. First, the dataset includes different horizontally related Italian NPN semi-schematic Cxns featuring prepositions *a* 'at/to' and *su* 'on' (namely Cxns 1, 2, 3 and 4 in Table 1): this design tests whether embeddings capture constructional generalizations within a network of sister Cxns, with localized competition (Masini, 2024a).

Second, the distractors set is broadened, including distinct Cxns (e.g., PNPN) identified in the literature (see Example 1: similar cases are excluded from Scivetti and Schneider (2025)'s setup) beyond surface-isomorphic patterns: this refines the identification task moving from syntactic discrimination to identification of true form–meaning pairings.

(1) [...] *con una successione da* ***agenzia ad agenzia*** *quasi automatica.*
'[...] with an almost automatic succession from **agency to agency**'.

Third, both the embedding of the `[UNK]` token (used as prepositional substitute) and the embedding of the preposition itself (henceforth, `PREP`) are probed, allowing us to compare abstracted and lexically grounded representations, and to assess how constructional information interacts with prepositional semantics.

5. Methods

5.1. Data

The dataset used in this study (Gorzoni et al., 2026) is derived from the Italian NPN dataset presented in Masini (2024a), extended with full sentential contexts extracted from CORIS[3].

The full dataset contains 3,256 attested instances of the Italian NPN constructional pattern instantiated by the prepositions *a* 'at/to' and *su* 'on'. Following the annotation schema proposed in (Masini, 2024a), each occurrence is manually annotated with one of five semantic labels: *succession/iteration/distributivity*, *greater_plurality/accumulation*, *juxtaposition/contact*, *connection/transition*, and *idiosyncratic*. The dataset further comprises 1,751 distractor instances spanning eight pattern types, for a detailed description see Appendix A. All data were manually cleaned to remove ill-formed sentences. Since the present study focuses exclusively on Cxns (1), (2), (3) and (4) in Table 2, only Cxn instances annotated with the *succession/iteration/distributivity*,

[3]The manually annotated dataset of Italian NPN Constructions and Distractors (Gorzoni et al., 2026), including semantic annotations and inter-annotator agreement data, is archived on Zenodo: https://zenodo.org/records/18268135

greater_plurality/accumulation and *juxtaposition/-contact* label were included in the operative dataset for the analysis. The resulting dataset consists of 1,281 constructional instances and 989 distractors.[4] All instances in the dataset were annotated for their *meaning*, the *lemma* instantiating the Cxn and its *number*. A subset of 100 Cxns (equally balanced across Cxns (1), (2), (3) and (4) in Table 1) were cross-annotated by a group of 5 annotators. Annotation quality of Cxns' meaning was assessed using Cohen's κ and Krippendorff's α (Artstein and Poesio, 2008). Pairwise Cohen's κ shows strong to near-perfect reliability across annotator pairs ($\kappa = 0.79 - 0.91$). Nominal α is high ($\alpha = 0.858$) and further increases when a reduced penalty is assigned to confusions between semantically adjacent labels (*succession/iteration/distributivity* and *greater_plurality/accumulation*, $\alpha = 0.892$). The entire dataset was filtered to only retain longer sentences (> 5 tokens) and at most 30 items with the same lemma per preposition.

For each experiment, data is then split into 5 training and test partitions using an 80/20 ratio. Unlike Scivetti and Schneider (2025), we do not enforce full lemma-level disjointness across splits. Instead, we adopt a modified strategy: lemma–label pairs are never shared between training and test sets, while allowing the same lemma to occur across labels. This design preserves lexical separation at the level of individual labels while retaining cross-label lexical overlap. In particular, it enables evaluation on cases where a lemma appears in training only with the opposite label, providing a stricter test of whether constructional distinctions are recoverable from contextual embeddings beyond lexical identity. Class balancing is applied to both training and test data. Tables B.1–B.4 summarize the composition of the data splits used for both experiments (i) and (ii).

5.2. Models and probing

Probing methods aim to investigate which types of information are encoded in a model's internal representations by training an auxiliary classifier, commonly referred to as a probing classifier. The representations are first extracted from a pre-trained model and then used as input for the probing classifier, together with labels corresponding to a linguistic property that has been explicitly operationalised. The central assumption underlying probing is that the probing classifier's performance reflects the extent to which the target property is accessible in the representations. Crucially, to draw meaningful conclusions about the presence of linguistic information in the embeddings, the probing setup must rely on a deliberately weak classifier paired with a sufficiently complex task (Hewitt and Liang, 2019). This constraint helps ensure that high performance is attributable to information encoded in the representations themselves, rather than to the probe's capacity to learn the task independently.

To evaluate whether constructional information is linearly accessible in contextual embeddings, we train a separate logistic regression probing classifier on embeddings extracted from each layer of the four considered BERT family's models, and track their performance across layers. Scivetti and Schneider (2025), we adopt a BERT-base architecture[5] (Staatsbibliothek and Schweter 2025, 12 layers, 768 hidden units, 12 attention heads) trained on Italian data. We also include multilingual BERT (mBERT[6], Devlin et al. 2019), that enables cross-linguistic comparison and allows partial replication of Scivetti and Schneider (2025) under a unified architecture. To further evaluate the role of monolingual specialization, we test UmBERTo[7] (Parisi et al., 2020), a RoBERTa-based model trained exclusively on Italian corpora: this allows to evaluate the effect of pretraining objectives. Finally, we include XLM-RoBERTa[8] (Conneau et al., 2019), a multilingual model trained on large-scale cross-lingual corpora, as a multilingual companion to UmBERTo. All models are used in inference mode without fine-tuning. Contextual embeddings are extracted layer-wise and evaluated using the same probing protocol across architectures. We consider representations corresponding to the prepositional token and its `[UNK]` substitute.

We also train a control classifier (Hewitt and Liang, 2019) with a random label assigned to each lemma, whose performance should be near chance given the absence of spurious correlations between train and test data. A probing linear classifier is also trained on static embeddings as baseline: differently from Scivetti and Schneider 2025, we adopt FastText (Bojanowski et al., 2017; Joulin et al., 2017) rather than GloVe as a static baseline because its subword-based representations are better suited to morphologically rich languages such as Italian, allowing us to control for lexical and inflectional variation.

5.3. Experimental setup

For the identification task, we perform binary classification (*Construction* vs. *Distractor*). For the disambiguation task, we perform multi-class classification across the three semantic labels. In order

[4] Numerous distractor instances belonging to the same distractor type in the full dataset were excluded in order to ensure heterogeneity.

[5] dbmdz/bert-base-italian-cased

[6] google-bert/bert-base-multilingual-cased

[7] Musixmatch/umberto-commoncrawl-cased-v1

[8] FacebookAI/xlm-roberta-base

to ensure maximum comparability with the previous study on English NPN constructions, we extended our experimental setup to the dataset introduced by (Scivetti and Schneider, 2025). Specifically, their data were converted into our format, subjected to our splitting procedure, and processed through the same embedding extraction and probing classification pipeline adopted for the Italian experiments. In addition, we computed a static baseline using English FastText vectors on the English configuration as well.

6. Identification task

The first experiment evaluates whether contextual embeddings extracted from BERT's models encode sufficient information to distinguish NPN constructions from distractors, and analyzes how the nature of the distractor patterns affect the probing classifier's behaviour.

In Scivetti and Schneider (2025)'s implementation, in fact, the identification task contrasts actual NPN instances with surface-isomorphic patterns. Because the distractor class is syntactically homogeneous and derives from a different structural configuration (e.g., the preposition does not function as the head in the distractor cases), the task may largely rely on syntactic structural cues, which could make it less clear to what extent specific constructional information is being actually tested.

Given the heterogeneous composition of our *distractor* portion of the dataset, three different sampling procedures were implemented[9]:

(i) a fully balanced configuration (henceforth, `SIMPLE`), where both training and test sets contain all distractor types;

(ii) a configuration where training is performed exclusively on structurally distinct constructions (i.e., PNPN, VERBAL and NSUNGIÙ) and testing on the full balanced test set (henceforth, `OTHER` configuration);

(iii) a configuration where, conversely, training is performed exclusively on surface-isomorphic patterns (i.e., *Thematic target*, *N-extended*, *NUM P NUM* and *Proper name inglobation* patterns), and evaluation on the full balanced test set (henceforth, `PSEUDO`).

The aim of these configurations is to investigate whether providing (`OTHER`), as opposed to withholding (`PSEUDO`), structurally informative negative examples that make constructional boundaries explicit has an impact on the performance of the probing classifier.

[9]All the the train test split and in particular the constraint needed for each configuration are menaged through the Python package (Pannitto, 2025)

With regard to configuration (i), as shown in Figure 1, probing classifiers trained on contextual embeddings, regardless of the specific model considered, achieve strong performance in distinguishing NPN Cxns from distractors, both for `[UNK]` and `PREP`.

As in Scivetti and Schneider (2025), the control classifier remains close to chance level, consistent with the selectivity criterion discussed by Hewitt and Liang (2019). The static lexical baseline performs well above chance: FastText probe provides a meaningful lower bound, which is outperformed by contextual embeddings only in middle to late layers, consistently with the interpretation that deeper layers encode more abstract information that is not available to static embeddings alone (Hewitt and Manning, 2019; Tenney et al., 2019; Scivetti and Schneider, 2025).

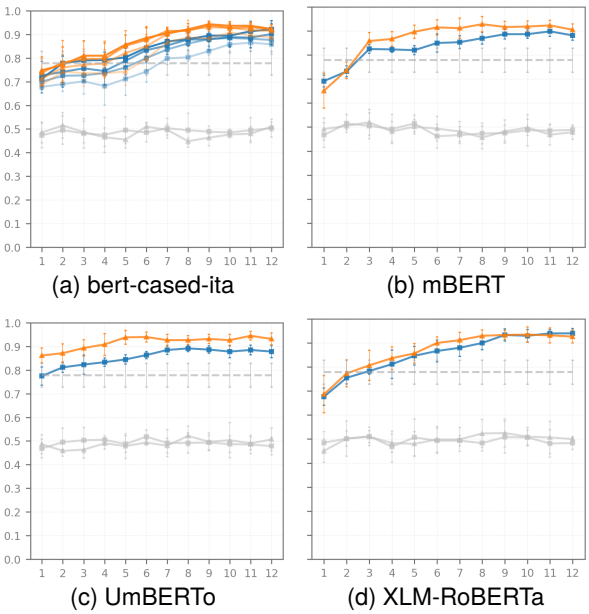


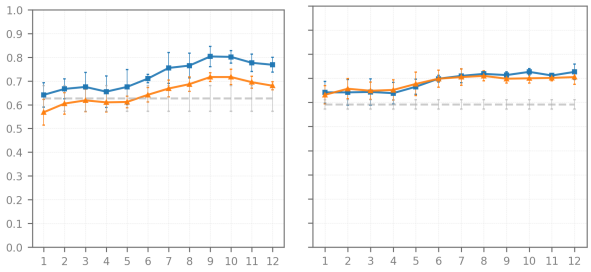
Figure 1: Accuracy of `[UNK]` (red lines, square dots) and `PREP` (orange lines, triangular dots) on Construction identification for the `SIMPLE` configuration. As in the following plots, the accuracy of the five probing classifiers resulting from the five random splits is averaged. Dashed grey line represents FastText baseline. Continuous grey lines refer to control classifiers. Figure (1a) includes decremental training configurations, line shading becomes progressively lighter as the number of training instances decreases (480 → 240 → 120 → 60). No substantial performance differences emerge across configurations. Figure C.1 in Appendix provides a qualitative visualisation of the embedding space across layers.

Results are confirmed on English NPN pattern as well: global performance is consistent both for `[UNK]` and `PREP` embeddings, but GloVe and FastText baselines perform quite differently on the task

(Figure 3).

Finally, the fact that [UNK] and PREP yield similar performance is informative. Replacing the preposition with [UNK] removes direct access to lexical information, yet the probe remains highly accurate. This suggests that successful discrimination does not crucially depend on access to the preposition's lexical semantics, but can be achieved on the basis of information contributed by the surrounding nouns and the broader sentential context. At the same time, the comparable performance of PREP suggests that lexical information about the preposition does not substantially increase accuracy beyond what is already available in the contextual configuration. These results support the conclusion that the probes capture information in BERT representations that reliably distinguishes the NPN Cxn from near-minimal distractor counterparts beyond what is possible through lexical semantic cues alone.

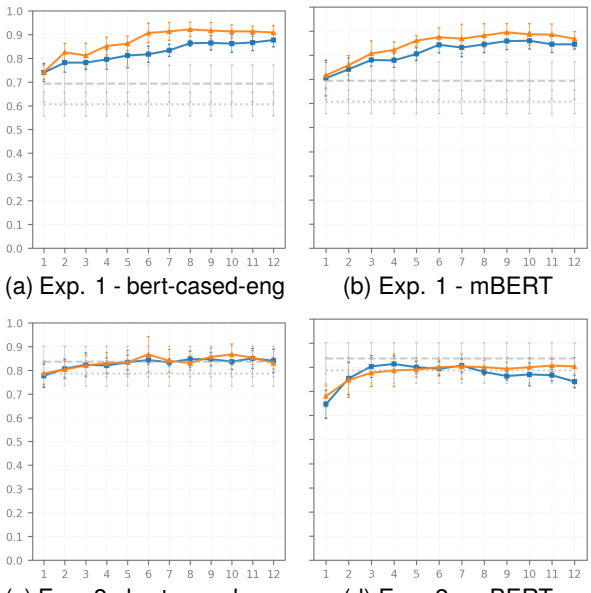
Configurations (ii) and (iii) yield slightly lower overall performance, while still consistently outperforming the baseline (Figure 2).



(a) OTHER (b) PSEUDO

Figure 2: Accuracy of [UNK] (red lines, square dots) and PREP (orange lines, triangular dots) on Construction identification for the OTHER and PSEUDO configurations. Dashed grey line represents FastText baseline.

The distribution of misclassifications (see Appendix D) highlights three main patterns, which are consistently observed across all models in the BERT family. In the SIMPLE configuration, all models exhibit very few classification errors, in line with the high overall accuracies. The vast majority of errors are False Positives, i.e. distractors incorrectly classified as constructions. This pattern suggests that constructional information is robustly encoded in the representations. The PSEUDO configuration illustrates how strongly the nature of the distractor instances defines the task itself. In this setting, the model is exposed to actual NPN constructions alongside distractors that are only superficially isomorphic but lack constructional status. In the test set structurally distinct constructions are more frequently misclassified. This suggests that, given the type of distractors encountered during training, the classifier might effectively learn to distinguish between constructional and non-constructional instances, rather than on deeper structural or construction-specific properties. In the OTHER configuration, PREP performance appears to be influenced by an imbalance in the training data, due to the constraints of this configuration. Since most distractors that instantiate a different Cxn are realised with the preposition *a*, instances with *su* are more frequently misclassified. This suggests that PREP representations are sensitive to prepositional frequency effects. By contrast, [UNK] representations do not seem to exhibit the same bias, indicating a more construction-oriented encoding less dependent on lexical skew. Taken together, these results highlight the importance of carefully defining what counts as a minimal pair for the Cxn under investigation, in order to faithfully translate the underlying theoretical question into a computational task. In probing experiments performance is strongly shaped by the experimental setup, in particular by the definition of distractors and minimal pairs. This suggests that probing results should be interpreted with care, as they reflect not only properties of the model, but also properties of the task design.



(a) Exp. 1 - bert-cased-eng (b) Exp. 1 - mBERT

(c) Exp. 2 - bert-cased-eng (d) Exp. 2 - mBERT

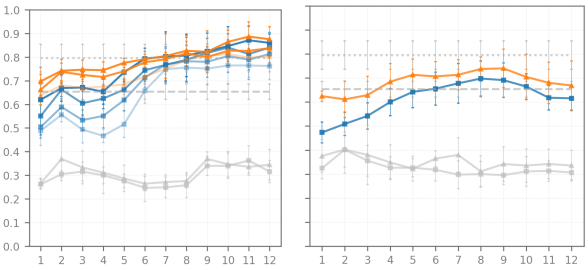
Figure 3: Accuracy of [UNK] (red lines, square dots) and PREP (orange lines, triangular dots) on both experiments on English data from Scivetti and Schneider (2025). Dashed grey line represents FastText baseline. Dotted grey line represents GloVe baseline.

7. Disambiguation task

Given the very high performance achieved in the experiment about the identification of NPN Cxn, ex-

tending the analysis beyond form, we now turn to examining the semantic dimension of the Cxn. Our setup is a multinomial three-class disambiguation problem: we only focus on the Cxn (1), (2), (3) and Cxn (4) in Table 1, which are associated to the three possible meanings of *juxtaposition/contact*, *succession/iteration/distributivity* or *greater_plurality/accumulation*.

Performance for both `[UNK]` and `PREP` embeddings, as shown in Figure 4, are above the baselines. In addition to the static baseline computed on the lemma of the reduplicated noun in the Cxn, this experiment includes a stronger baseline condition. Since one of the defining properties distinguishing Cxn (3) from (4) is the grammatical number of the reduplicated noun, we computed an additional FastText baseline using vectors built from the morphologically inflected noun: this proves substantially stronger than the lemma-based one. BERT models surpass it only in the highest layers, suggesting that lower layers do not encode information beyond what is already available from surface morphological cues, while higher layers capture more abstract constructional distinctions.



(a) bert-cased-ita (b) mBERT

(c) UmBERTo (d) XLM-RoBERTa

Figure 4: Accuracy of `[UNK]` (red lines, square dots) and `PREP` (orange lines, triangular dots) on Construction disambiguation task. Dashed grey line represents FastText baseline. Dotted grey line represents morphological FastText baseline. Continuous grey lines refer to control classifiers. Model (4a) includes decremental training configurations, line shading becomes progressively lighter as the number of training instances decreases (480 → 240 → 120 → 60). No substantial performance differences emerge across configurations. Figures C.2 and C.3, in Appendix, provide a qualitative visualisation of the embedding space across layers.

Error analysis for both `[UNK]` and `PREP` contextual embeddings reveal a highly structured pattern across the BERT family. The probing classifier achieves near-perfect performance in identifying the *greater_plurality/accumulation* and *succession/iteration/distributivity* meanings when instantiated with the preposition *su* 'on'. Although overall performance remains strong, some confusion emerges between *juxtaposition/contact* and *succession/iteration/distributivity* when both are instantiated with the preposition *a* 'at/to'. In other words, the model shows greater difficulty in distinguishing between semi-schematic Cxns that share the same surface preposition. Importantly, *greater_plurality/accumulation* is distinguished not only semantically but also through a salient surface morphological constraint (singular or plural number of the reduplicated noun), which may facilitate classification, as the the additional morphologically informed baseline suggests. Future work could isolate this formal property in a controlled design in order to assess the relative contribution of surface morphology and constructional semantics to probing performance.

A substantial proportion of the errors involves NPN Cxn (1) and (2), for which, respectively, labels *juxtaposition/contact* and *succession/iteration/distributivity* are predicted. These cases frequently involve high-frequency and highly conventionalised types, such as *gomito a gomito* ('elbow to elbow'), *petto a petto* ('chest to chest'), *fronte a fronte* ('forehead to forehead'), and spatial-configuration patterns such as *balcone a balcone* ('balcony to balcony'), *uscio a uscio* ('doorway to doorway'), and *porta a porta* ('door to door').

(2) *Qui lavorano i revisori **gomito a gomito** con i membri di Cosea*.
'Here auditors work **elbow to elbow** with members of Cosea[10]'
ANNOTATED: JUXT./CONTACT;
PREDICTED: SUCC./ITER./DISTRIBUTIVITY

(3) *Faccia a faccia e **petto a petto**, Big Jim serrò le mani sulle braccia di Andy e lo guardò negli occhi.*.
'Face to face and **chest to chest**, Big Jim closed his hands on Andy's arms and looked at him in his eyes.'
ANNOTATED: JUXT./CONTACT;
PREDICTED: SUCC./ITER./DISTRIBUTIVITY

Examples (2) and (3) show that, although the Cxn canonically encodes physical contact or spatial adjacency, the larger event structure evokes dynamic interaction unfolding in time (e.g. working sessions, fighting sequences). The probe may

[10]Cosea is an Italian company, active in the Emilia region.

therefore privilege an interpretation compatible with iterativity or sequential event structure over a purely configurational reading.

A similar oscillation emerges in spatial-distribution expressions.

(4) *Dopo anni passati prima come venditore* ***porta a porta*** *di assicurazioni e di calzature per donne (...) questo è il suo momento..*
'After years spent first as a **door to door** salesman of insurance policies and women's shoes (...) this is his moment.'
ANNOTATED: SUCC./ITER./DISTRIBUTIVITY;
PREDICTED: JUXT./CONTACT

In these cases, such as Example (4), the model appears to oscillate between an interpretation based on contact or adjacency between entities along a spatial path and an interpretation based on distribution over successive sites. The alternation suggests that, for entrenched lexicalised types, the boundary between *juxtaposition/contact* and *succession/iteration/distributivity* could be intrinsically gradient rather than categorical.

These results motivate a more focused examination of the *succession/iteration/distributivity* label, as it is compatible with multiple prepositions and does not rely on a strong surface constraint (unlike *greater_plurality/accumulation*). Its classification therefore provides a more stringent test of whether contextual embeddings capture abstract constructional meaning independently of lexical realisation.

7.1. Semantic generalisation

In order to create an experimental setup that allow us to test the probing classifier on unseen semi-schematic Cxns, we annotated 100 instances on Cxns (5) and (7) from Table 1 with sentential context extracted from CORIS: these have meaning of *succession/iteration/distributivity* and are realised by the prepositions *per* 'by' and *dopo* 'after'.

We trained the probing classifier on the full dataset comprising the three semantic labels together with the distractor instances, and then tested it exclusively on the newly annotated *per* 'by' and *dopo* 'after' instances. The composition of this dataset is shown in Appendix B.5.

Figure 5 reports accuracy across layers for `[UNK]` and `PREP` embeddings, together with static baselines. We can observe that `[UNK]` and `PREP` representations support robust generalisation to unseen prepositions, with performance reaching high accuracy in late layers. The task is intrinsically harder, as it is demonstrated by the drop in performance for both baselines. Nonetheless, results are consistent across different model in BERT family. `[UNK]` starts from moderate performance in the earliest layers and increases steadily, approaching 0.9 in the final layers. `PREP` shows a different

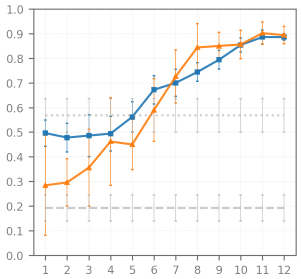


Figure 5: Accuracy of `[UNK]` (red lines, square dots) and `PREP` (orange lines, triangular dots) on Construction disambiguation. Dashed grey line represents FastText baseline. Dotted grey line represents morphological FastText baseline.

trajectory: it is markedly weaker and more variable in the lowest layers, where it performs close to baseline, but in 4–8 layers exhibits a sharp improvement, converging with `[UNK]` in late layers. This pattern suggests that early-layer `PREP` representations are strongly related to lexical information about the preposition, while higher layers increasingly integrate contextual information that can support a semantics-based decision even when the preposition itself is unseen in training.

These findings provide converging evidence that the semantic information exploited by the probing classifier is not merely a learned association between a small set of prepositions and labels. Instead, the late-layer gains for `[UNK]` and especially for `PREP` are consistent with an increasingly abstract encoding of constructional meaning, where the model can map unseen prepositional instantiations (*per* 'by', *dopo* 'after') onto a semantic category encountered in training (*succession/distributivity/iteration*). Under this interpretation, the experiment supports the conclusion that while lower BERT layers, particularly for `PREP`, remain more sensitive to surface lexical form, higher layers encode a representation rich enough to sustain generalisation across competing semi-schematic Cxns (Masini, 2024a) expressing *succession/distribution/iterativity* which can be realised by any of the prepositions *a* 'at/to', *su* 'on', *per* 'by', and *dopo* 'after'.

8. Conclusion

We presented two probing experiments addressing the identification and semantic disambiguation of Italian NPN constructions. To this end, we introduced an extended dataset including both constructional instances and carefully designed distractors, allowing for a controlled evaluation of construction-sensitive encoding.

We extended and enriched the setup of (Scivetti and Schneider, 2025) by comparing `[UNK]` and

PREP contextual embeddings, implementing alternative training configurations grounded in different minimal-pair and distractor definitions, and introducing a generalisation test on unseen prepositions instantiating unseen semi-schematic Cxns.

The aim of this study is to investigate whether the internal representations of models in the BERT family are compatible with a constructionist perspective, with respect to the Italian NPN constructional family. This question is addressed through a set of experiments targeting both form-related and meaning-related properties of the constructions.

The results suggest that distinguishing constructional patterns relies on the joint contribution of multiple factors, including lexical distribution and morphological cues. As indicated by the baseline performance, none of these factors in isolation is sufficient to fully resolve the task; rather, they appear to jointly support the information that is recoverable for classification.

The identification results show that constructional information is detectable in contextual embeddings. However, they also demonstrate that performance is highly sensitive to the composition of the training set: the nature of the distractors and the operationalisation of minimal pairs substantially shape the classification task. In this respect, careful interpretative caution is required, in line with recent observations on task sensitivity in probing studies (see also Dunn and Eida (2025)).

The disambiguation experiment further reveals that static embeddings constitute strong baselines. Lexical information, especially when coupled with morphological cues such as number of the reduplicated noun, proves highly informative and can account for a significant portion of the variance. Crucially, however, when the task involves generalisation to unseen semi-schematic constructions instantiating the same semantic value, contextual embeddings enable the classifier to achieve high accuracy precisely where static vectors show a marked performance drop. This suggests that higher layers encode more abstract constructional regularities that go beyond type-level lexical information.

Taken togheter, the results are consistent with a constructionist perspective, in that successful classification appears to rely on the joint contribution of multiple cues rather than on isolated lexical features.

Multilingual models (mBERT and XLM-R respectively) deserve specific attention: while their performance remain largely comparable to monolingual models in the identification task, the situation appears different for disambiguation. In disambiguation task, in fact, both mBERT and XLM-R underperform their monolingual counterpart. While the difference is negligible for English, the drop in performance is particularly evident for Italian data. This confirms the importance of extending interpretability investigations to languages other than English: taking a language-specific perspective like the one that CxG offers can help evaluating language-specific abilities of multilingual models.

Overall, the findings seem to provide empirically grounded support for an abstraction path in PLMs towards an higher-level construction linking the semi-schematic Cxns associated with the semantics of *succession/distribution/iterativity*, realized by the prepositions *a* 'at/to', *su* 'on', *per* 'by', and *dopo* 'after' within the NPN family.

At the same time, the results remain confined to the representational space within BERT's models and any theoretical implications must therefore be interpreted with caution.

In order to provide empirical grounding for the theoretical debate concerning the schematization of the NPN construction, the probing results are currently being cross-validated through behavioural experiments that directly compare human judgments with model predictions. This step will be essential to determine whether the computational evidence actually reflects cognitively plausible constructional abstractions and to more rigorously assess the theoretical validity of the proposed NPN schematization.

9. Limitations

The present study is subject to several limitations.

First, the analysis is restricted to a single constructional family, namely the Italian NPN Cxs. Although multiple prepositions (*a* 'at/to', *su* 'on', *per* 'by', *dopo* 'after') are included, they instantiate closely related constructions within the same constructional network, differing primarily in degree of productivity and conventionalisation. As such, the findings should be interpreted as evidence about this specific constructional domain rather than as a general account of constructional information in internal representations of the model. Extending the analysis to Cxns with different syntactic profiles and levels of schematicity would be necessary to assess the broader validity of the observed patterns. This is especially relevant given that NPN Cxns predominantly fulfil an adverbial function, which limits the diversity of syntactic configurations examined.

Second, the study focuses on models within the BERT family. While this allows for controlled comparison across variants, these models share core architectural and training properties, including encoder-only structure and a masked language modelling objective. Consequently, the results cannot be straightforwardly generalised to other classes of models. In particular, architectures with different training objectives or representational dynamics may encode constructional information dif-

ferently. Future work should therefore extend the analysis to decoder-only models, in order to assess whether the results observed here are specific to masked language modelling or reflect more general properties of transformer-based representations, including those shaped by next-token prediction.

Third, the methodology is limited to diagnostic probing. This design choice ensures comparability with prior work on the English NPN construction (Scivetti and Schneider, 2025), but it also constrains the scope of the conclusions. In particular, it remains unclear whether the information encoded in the model's internal representations and exploited by the probe is causally involved in the model's behaviour on the task. In other words, while probing suggests that the representational space can be analyzed in terms of constructional information, it does not establish that the model itself relies on this information when making predictions.

10. Ethics Statement

Annotators were recruited within an advanced Master's-level course as part of structured educational activities. Participation was entirely voluntary and had no impact on students' evaluation or academic standing. All participants were informed about the objectives of the study and the intended use of the collected data.

11. Bibliographical References

Ron Artstein and Massimo Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Claire Bonial and Harish Tayyar Madabushi. 2024. A construction grammar corpus of varying schematicity: A dataset for the evaluation of abstractions in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 243–255, Torino, Italia. ELRA and ICCL.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Dunn and Mai Mohamed Eida. 2025. LLMs learn constructions that humans do not know. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 13–23, Düsseldorf, Germany. Association for Computational Linguistics.

R. Favretti, R. Rossini, F. Tamburini, and C. De Santis. 2002. CORIS/CODIS: A corpus of written italian based on a defined and a dynamic model.

Charles J. Fillmore. 1988. The mechanisms of "construction grammar". In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55, Berkeley. Berkeley Linguistics Society.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press, Chicago.

Adele E. Goldberg. 1996. Jackendoff and construction-based grammar. *Cognitive Linguistics*, 7(1):3–20.

Greta Gorzoni, Ludovica Pannitto, and Francesca Masini. 2026. NPN Construction and Distractor dataset.

Nora Graichen, Iria de Dios-Flores, and Gemma Boleda. 2026. The grammar of transformers: A systematic review of interpretability research on syntactic knowledge in language models.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ray Jackendoff. 2008. "Construction after Construction" and Its Theoretical Challenges. *Language*, 84(1):8–28.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Jaap Jumelet, Lisa Bylinina, Willem Zuidema, and Jakub Szymanik. 2024. Black big boxes: Do language models hide a theory of adjective order?

Francesca Masini. 2024a. Costruzioni su costruzioni: idiomaticity and regularity of npn discontinuous reduplications in italian. In *I fraseologismi schematici. Questioni descrittive e teoriche*, pages 51–82. Aracne, Roma.

Francesca Masini. 2024b. Npn discontinuous reduplications in italian: Dataset.

Ludovica Pannitto. 2025. sample-dataset.

Ludovica Pannitto and Aurélie Herbelot. 2023. CALaMo: a constructionist assessment of language models. In *Proceedings of the First International Workshop on Construction Grammars and NLP (CxGs+NLP, GURT/SyntaxFest 2023)*, pages 21–30, Washington, D.C. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

Giulia Rambelli. 2025. *Constructions and Compositionality: Cognitive and Computational Explorations*. Cambridge University Press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Wesley Scivetti and Nathan Schneider. 2025. Construction identification and disambiguation using bert: A case study of npn. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 365–376. Association for Computational Linguistics.

Lotte Sommerer and Andreas Baumann. 2021. Of absent mothers, strong sisters and peculiar daughters: The constructional network of english npn constructions. *Cognitive Linguistics*, 32(1):97–131.

Bayerische Staatsbibliothek and Stefan Schweter. 2025. bert-base-italian-cased (revision 843e404).

Harish Tayyar Madabushi and Claire Bonial. 2025. Construction grammar evidence for how LLMs use context-directed extrapolation to solve tasks. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 190–201, Düsseldorf, Germany. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882. Association for Computational Linguistics.

Leonie Weissweiler, Abdullatif Köksal, and Hinrich Schütze. 2025. Hybrid human-llm corpus construction and llm evaluation for the caused-motion construction. *Northern European Journal of Language Technology*, 11(1):27–57.

Appendix

A. Distractors

PNPN - Construction - (*a*)

Description:
The PNPN Construction $[P\ N_i\ P\ N_i]$ encodes a meaning of *iterativity*, *succession*, and *transition* between elements belonging to the same semantic category. In this Construction, the nominal slots are occupied by the same noun, repeated in a discontinuous form, while the preceding preposition —typically *a* or *da* — specifies the direction or type of relation between the two occurrences.

Example (ID 720): *Il tutto tra l'altro avviene nel giro di pochi mesi con una successione* ***da agenzia ad agenzia*** *quasi automatica*
'All of this, moreover, happens within a few months with an almost automatic succession **from agency to agency**'.

Verbal - Construction - (*a*)

Description:
The verbal Construction $[V\ N_i\ a\ N_i]$ is realised by a restricted set of transitive verbs that allow both a direct object and a prepositional complement introduced by *a*. Among the most frequent are verbs of accumulation or transfer, such as *aggiungere* ('to add'), *unire* ('to join'), *sommare* ('to sum'), and *accostare* ('to bring together'). This Construction exhibits a high degree of semantic conventionalisation: the identity of the nouns generates an intensifying or cumulative meaning that goes beyond the literal meaning of the verb.

Example (ID 1027): *Quel fatto* ***aggiunse irritazione a irritazione****.*
'That fact **added irritation to irritation**'.

NUM P NUM - isomorphic pattern — (*a*)

Description:
This pattern expresses a relationship of **equivalence** or **balance** between two numerical values that are identical or opposed (identical, in the case of superficial identity with NPN)

Example (ID 1498): *Il risultato non si sbloccava e si era sempre sullo* ***zero a zero****.*
'The result did not get unlocked and it was always at **zero to zero**'.

N extended - isomorphic pattern - (*a*, *su*)

Description:
These instances superficially replicate the NPN Construction, but one or both occurrences of the noun are part of a larger complex phrase or multiword expression and do not constitute an autonomous instance of the same syntactic category. Consequently, they do not convey the iterative, distributive, or juxtapositional relations that define actual NPN constructions.

Example (ID 229): *Le due telecamere alternavano* ***primi piani a piani a figura intera*** *da angolazioni diverse.*
'The two cameras alternated **close-ups to full shots** from different angles'.
Example (ID 2193): *La Commissione può ora rimborsare* ***spese generali di base su base forfetaria*** *o versare importi forfetari per piccoli progetti.*
'The Commission can now reimburse basic overhead costs on a flat-rate basis or pay lump-sum amounts for small projects'.

Proper name inglobation - isomorphic pattern — (*su*)

Description:
Accidental lexical repetition in which one or both noun occurrences belong to a proper name. In such cases, the repetition has no constructive or syntactic function, resulting merely from the contiguous presence of the same lemma in two elements of the sentence.

Example (ID 2189): *La posizione del* ***Comune di Arezzo su Arezzo Fiere*** *è chiarissima.*
'The position of the Municipality of Arezzo on Arezzo Fiere is very clear'.

N su N giù - Construction - (*su*)

Description:
Combines nominal repetition with the directional pair *su–giù* ('up, down'), producing a rhythmic and iterative effect. It typically occurs in colloquial or performative contexts, marking repetition or cyclic succession of events or actions, but can also convey an iterative or ironic meaning, evoking repetition and excess in a rhythmically salient way. The su–giù pair contributes to a sense of saturation and cyclical movement.

Example (ID 2174): *Come se ci fosse bisogno di una voce dall' alto che dica* ***"bandiere su bandiere giù"****.*
'As if there were a need for a voice from above saying 'flags up, flags down' '.

Num P Num - isomorphic pattern — (*su*)

Description:
Expresses a proportional or evaluative relation between two numerical values, typically used in quantificational or performative contexts such as voting, scores, or results. The preposition *su* introduces the reference term, establishing a part-whole or ratio relation ("x obtained out of y possible"). When the two numerals are identical, the expression acquires idiomatic and highly conventionaliased value and is especially productive in sports and journalistic language.

Example (ID 2127): *Per nostra fortuna sono stati fecondati **3 su 3**, il 100%, pronti per essere impiantati in utero.*
'Fortunately for us, 3 out of 3, 100'%, were fertilized, ready to be implanted in the uterus'.

Thematic target - isomorphic pattern — (*su*)

Description:
The surface structure N-su-N is preserved, but the function does not correspond to the canonical meanings of the Construction (plurality, iteration, accumulation, distributivity, or connection). These distractors feature a second noun specifying the thematic domain.

Example (ID 2088): *Comunque non spetta a questa introduzione dire delle nostre tre autrici: qui si tratta piuttosto di chiedersi perché, a voler parlare di romanzo popolare in Italia tra i due secoli, si sono scelte tre donne che scrivono per **donne su donne**.*
'Anyway, it is not the role of this introduction to say anything about our three authors: here, it is rather a question of asking why, when one wants to speak of popular novels in Italy between the two centuries, three women who write for women on women were chosen'.

B. Training and Test sets

Label	Prep	Train	Test
CXN	*a*	120	30
CXN	*su*	120	30
CXN (total)	–	240	60
DISTR	*a*	120	30
DISTR	*su*	120	30
DISTR (total)	–	240	60
total	–	**480**	**120**

Table B.1: SIMPLE training test split configuration

Label	Prep	Train	Test
CXN	*a*	70	30
CXN	*su*	70	30
CXN (total)	–	140	60
DISTR	*a*	25	30
DISTR	*su*	115	30
DISTR (total)	–	140	60
total	–	**280**	**120**

Table B.2: PSEUDO training test split configuration

Label	Prep	Train	Test
CXN	*a*	55	30
CXN	*su*	55	30
CXN (total)	–	110	60
DISTR	*a*	105	30
DISTR	*su*	5	30
DISTR (total)	–	110	60
total	–	**220**	**120**

Table B.3: OTHER training test split configuration

Label	Prep	Train	Test
Succession	*a*	60	15
Succession	*su*	60	15
Succession (total)	–	120	30
Accumulation	su	120	30
Juxtaposition	a	120	30
total	–	**360**	**90**

Table B.4: Semantic disambiguation train–test split configuration.

Label	Prep	Train	Test
Succession	*a*	30	–
Succession	*su*	30	–
Succession	*per*	–	50
Succession	*dopo*	–	50
Succession (total)	–	60	100
Accumulation	*su*	60	–
Juxtaposition	*a*	60	–
Distractor	*a*	30	–
Distractor	*su*	30	–
Training set (total)	–	**360**	**100**

Table B.5: *per* and *dopo* train–test split configurations.

C. PCA

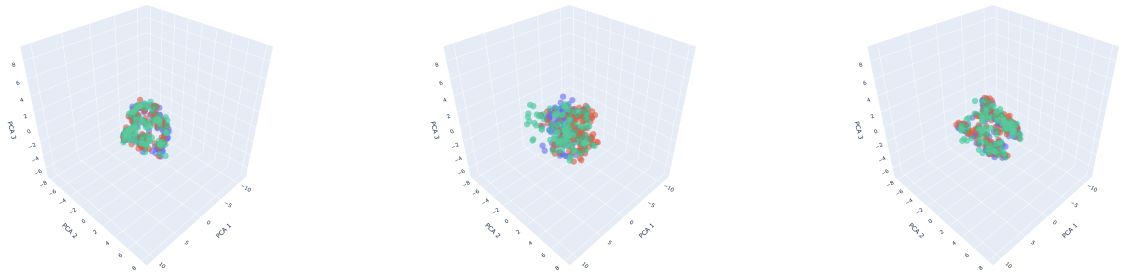
(a) Layer 1 (b) Layer 2 (c) Layer 3

(d) Layer 4 (e) Layer 5 (f) Layer 6

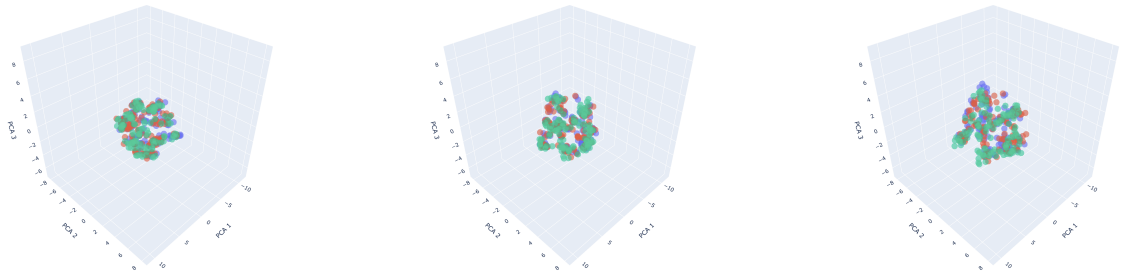
(g) Layer 7 (h) Layer 8 (i) Layer 9

(j) Layer 10 (k) Layer 11 (l) Layer 12

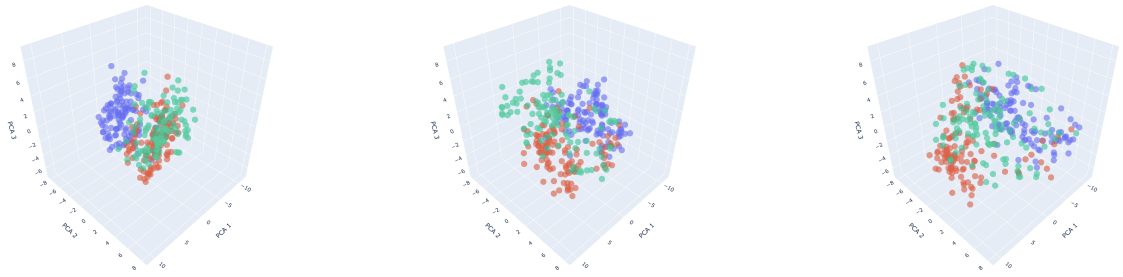
Figure C.1: PCA-based three-dimensional projection of UNK BERT embeddings across layers for NPN constructions (red) and distractors (red). An animated version of the visualisation is available at https://gretagorzoni00.github.io/NPN_contextual_embeddings/.



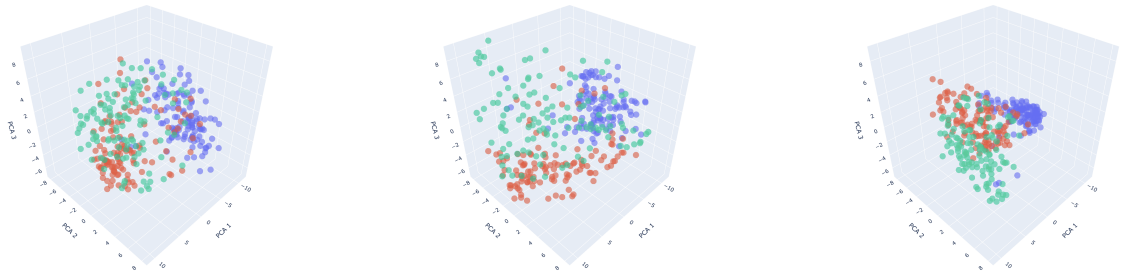
(a) Layer 1 (b) Layer 2 (c) Layer 3



(d) Layer 4 (e) Layer 5 (f) Layer 6

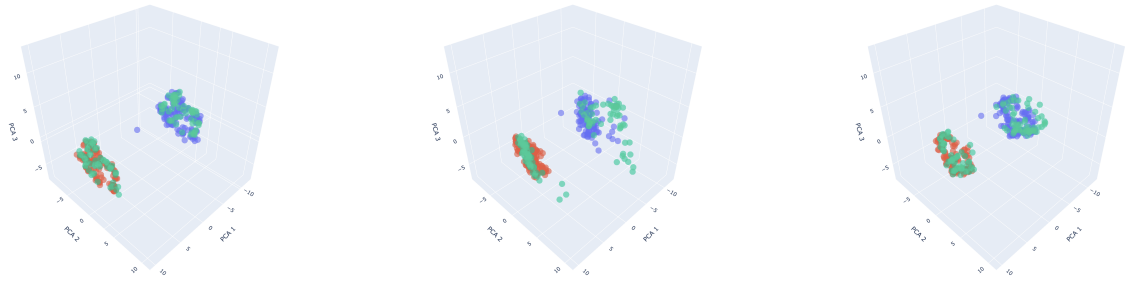


(g) Layer 7 (h) Layer 8 (i) Layer 9

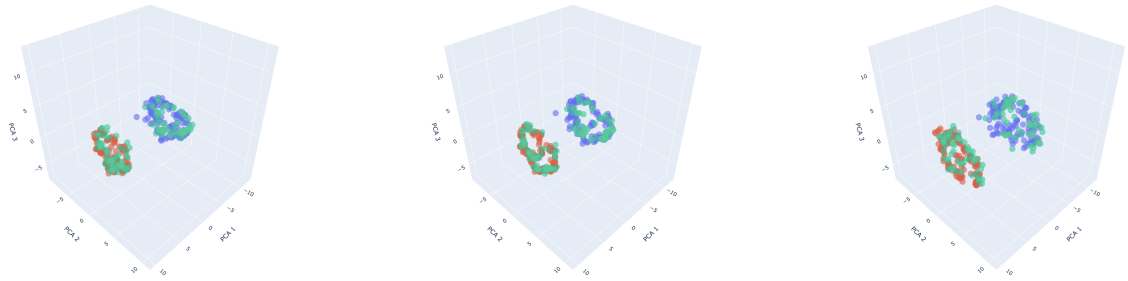


(j) Layer 10 (k) Layer 11 (l) Layer 12

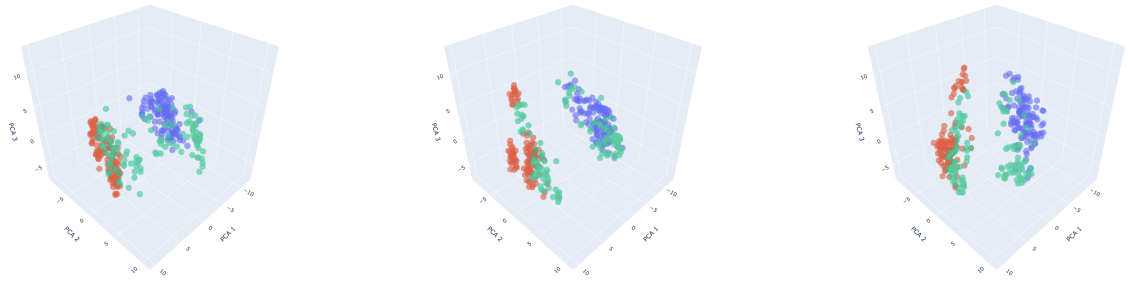
Figure C.2: Three-dimensional PCA projection of UNK-based BERT contextual embeddings across layers for NPN semi-schematic constructions. Data points are colour-coded by semantic value: *juxtaposition/-contact* (red), *greater_plurality/accumulation* (red), and *succession/iteration/distributivity* (green). The visualisation is based on a balanced subset of 360 instances. An animated version of the visualisation is available at https://gretagorzoni00.github.io/NPN_contextual_embeddings/.



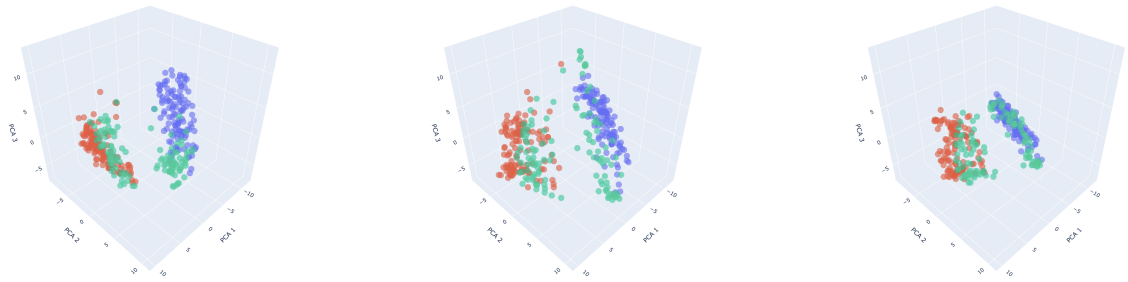
(a) Layer 1 (b) Layer 2 (c) Layer 3



(d) Layer 4 (e) Layer 5 (f) Layer 6



(g) Layer 7 (h) Layer 8 (i) Layer 9



(j) Layer 10 (k) Layer 11 (l) Layer 12

Figure C.3: Three-dimensional PCA projection of PREP-based BERT contextual embeddings across layers for NPN semi-schematic Constructions. Data points are color-coded by semantic value: *juxtaposition/contact* (red), *greater_plurality/accumulation* (red), and *succession/iteration/distributivity* (green). The visualisation is based on a balanced subset of 360 instances. An animated version of the visualisation is available at https://gretagorzoni00.github.io/NPN_contextual_embeddings/.

D. Confusion matrices

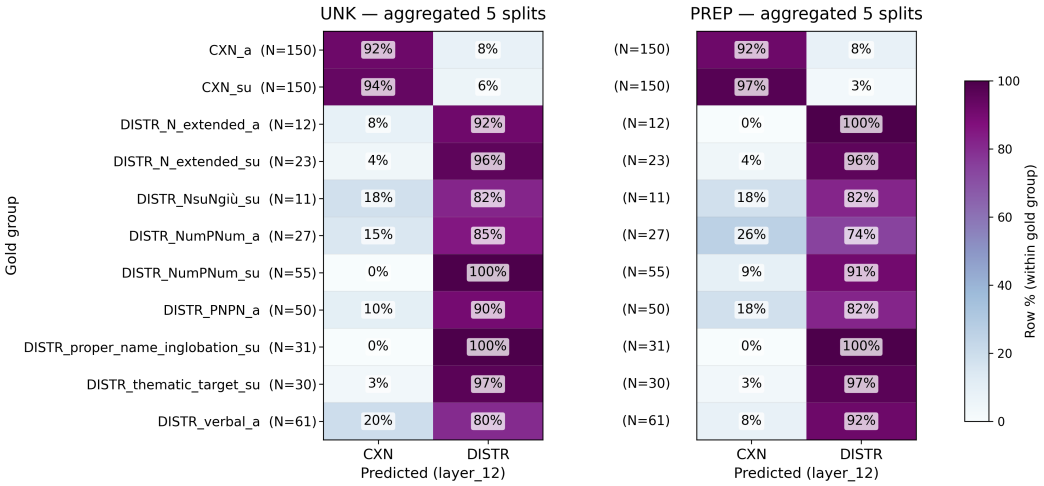


Figure D.1: Confusion matrices for the `SIMPLE` configuration in the Cxn identification experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

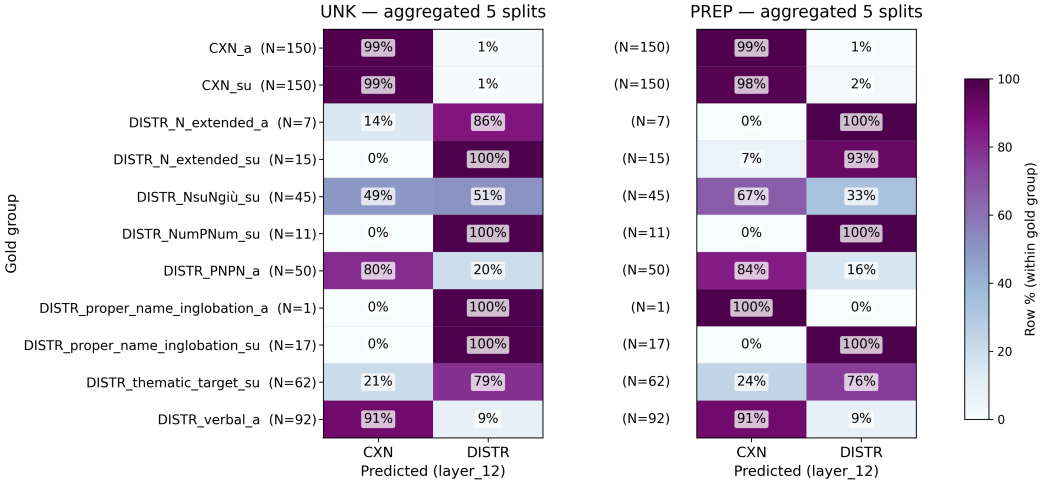


Figure D.2: Confusion matrices for the `PSEUDO` configuration in the Cxn identification experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

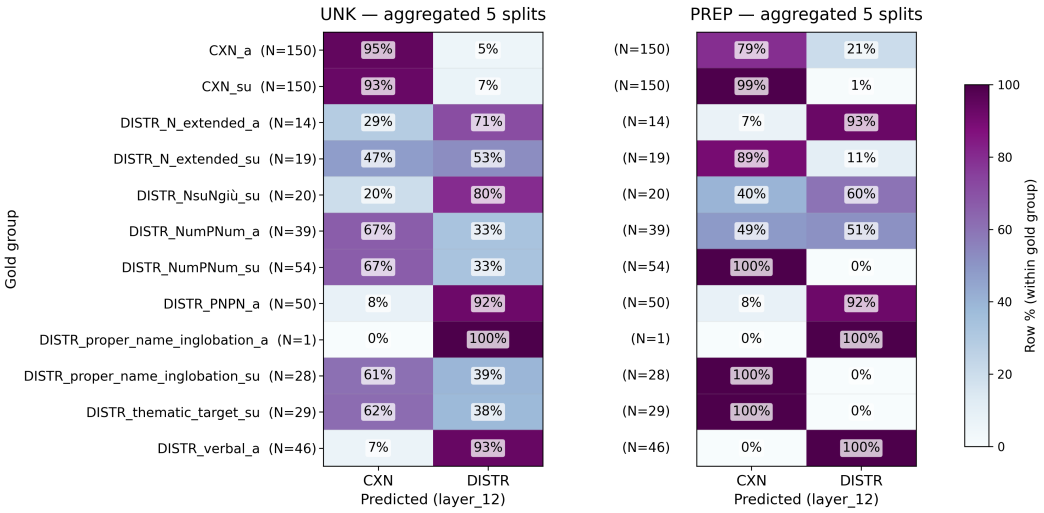


Figure D.3: Confusion matrices for the `OTHER` configuration in the Cxn identification experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

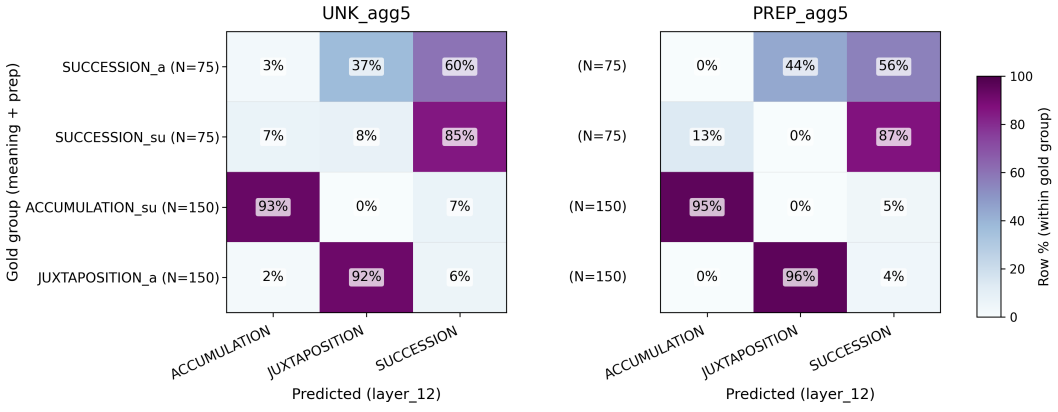


Figure D.4: Confusion matrices for the Cxn disambiguation experiment. Results for the Italian BERT model (`bert-base-italian`) are shown for `[UNK]` embeddings (left) and `PREP` embeddings (right). The results from the five random splits are aggregated by summing the prediction errors across the five corresponding probing classifiers.

Teaching LLMs to unveil tendentious implicit contents of Italian political communication

Walter Paci, Lorenzo Gregori, Alessandro Panunzi
University of Florence

{walter.paci, lorenzo.gregori, alessandro.panunzi}@unifi.it

Abstract

This paper investigates the capacity of Large Language Models (LLMs) to interpret and explain non-bona fide true implicit content within the specific context of Italian political communication. While implicit meaning is a fundamental aspect of pragmatic competence, it is frequently exploited in political discourse through presuppositions and implicatures to convey tendentious messages without overt commitment. Building on previous research that highlighted the limitations of general-purpose models in zero-shot settings, we perform instruction-based fine-tuning on two state-of-the-art open-weight models: Llama3.1 and Qwen2.5. We utilize the IMPAQTS-PID benchmark, a dataset derived from the IMPAQTS corpus, to train these models to generate natural language explanations of manipulative implicit meanings. Our experimental results, validated through manual evaluation by independent annotators, demonstrate that fine-tuning significantly improves model performance compared to previous benchmarks, and that both models perform substantially better at explaining presuppositions than implicatures.

1. Introduction

In natural language use, the meaning conveyed by an utterance rarely coincides entirely with its explicit surface structure. Communication routinely depends on what is left unsaid, making the ability to infer implicit meanings a core aspect of linguistic competence (Grice, 1975; Sperber and Wilson, 1986). Although implicit communication often enhances conciseness and efficiency, it can also function as a rhetorical strategy, enabling speakers to convey messages without overtly committing to them. This is especially apparent in the use of *non-bona fide true* implicit content, a phenomenon that is particularly widespread in political discourse (Lombardi Vallauri, 2017). Political communication frequently exploits presuppositions, implicatures, and other types of implicit contents to deliver information without asserting them (Cominetti et al., 2024). Presuppositions allow speakers to frame contested propositions as already established or mutually accepted background knowledge (Stalnaker, 1978; Beaver, 2001), while implicatures enable the indirect transmission of evaluative or strategically loaded content (Grice, 1975; Levinson, 2000).

With the advent of Large Language Models (LLMs), increasingly sophisticated language processing abilities have emerged, including aspects of pragmatic competence (Ma et al., 2025). Recent research has begun to assess LLMs' capacity to interpret implicit content (Solidjonov, 2025; Cho and mook Kim, 2024). However, most existing studies concentrate on English-language datasets derived from written texts.

In this work, we analyzed the abilities of LLMs to interpret and explain implicit contents in Italian political communication. To this aim, we used a dataset derived from the IMPAQTS corpus as a data source, and performed the fine-tuning of two models, Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct. Generated explanations are manually evaluated by independent annotators; moreover, BERTScore metric is evaluated on this task.

While our experimental setup focuses on a limited number of models, our objective is not to provide an exhaustive comparison across architectures or parameter scales, but to assess whether instruction-based fine-tuning on structured pragmatic data can improve the ability of LLMs to reconstruct implicit content. As such, the models considered in this study are intended as representative instances of current open-weight instruction-tuned systems and not a comprehensive sample of the model landscape.

More broadly, this work relates to the question of how pragmatic reasoning abilities interact with model characteristics such as scale and architecture. While we adopt a controlled fine-tuning setup, our results suggest that improvements in pragmatic interpretation may depend not only on model size, but critically on the availability of structured, task-specific supervision.

2. IMPAQTS corpus

The IMPAQTS (Implicit Manipulation in Politics – Quantitatively Assessing the Tendentiousness of Speeches) project is devoted to the systematic investigation of implicit manipulation in Italian political discourse. Its primary objective is the creation of a large-scale, multimodal corpus of political

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 221–229
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Speech Type	Speeches	Words
Parliamentary speech	561 (39.99%)	889,769 (43.11%)
Rally	283 (20.17%)	480,983 (23.30%)
Party assembly	137 (9.76%)	229,379 (11.11%)
Statement in person	231 (16.46%)	299,404 (14.51%)
Broadcast statement	164 (11.69%)	126,427 (6.13%)
New media statement	27 (1.92%)	37,971 (1.84%)
Total	1403	2,063,933

Table 1: IMPAQTS corpus size

speeches annotated for instances of tendentious or questionable content conveyed implicitly. By focusing on the manipulative potential of linguistic implicitness, the project addresses a crucial dimension of political communication: the strategic use of presuppositions, implicatures, and other non-explicit meanings to present debatable content as taken for granted.

The IMPAQTS corpus comprises approximately 1,400 monologic speeches delivered by 150 Italian politicians between 1946 and 2023, totaling around 2.5 million tokens (see Table 1). The corpus includes monologues produced directly by politicians belonging to six types: parliamentary speeches, rally speeches, party assembly speeches, statements in presence, broadcast statements, and new media statements. The corpus is diachronically structured to cover nearly the entire history of the Italian Republic, subdivided into three major political periods (1946–1972; 1972–1994; 1994–2023). These divisions correspond to significant historical and institutional turning points and allow for the investigation of long-term changes in political rhetoric and implicit strategies.

The annotation framework of IMPAQTS is grounded in the distinction between *bona fide true* and *non-bona fide true* implicit contents. Only the last ones, that are potentially manipulative, are annotated and classified in the corpus. This criterion enables a principled differentiation between implicitness used for rhetorical manipulation and implicitness serving purely pragmatic or discourse-economy functions. The IMPAQTS annotation has been manually performed by expert linguists and contains, for each*non-bona fide true* implicit content, its classification and its explanation.

3. Previous Work

This work builds on a line of research investigating if LLMs can interpret and explain implicit content in Italian political communication, using data derived from the IMPAQTS corpus.

Paci (2025) evaluated pretrained LLMs on two tasks: a binary detection task on implicit content and a binary classification task where model had to classify a content as an implicature or a presupposition. Across nine multilingual models and multiple prompting strategies, models achieved unsatisfactory results in detection and remained close to chance in classification, suggesting that models did not exhibit any relevant kind of pragmatic reasoning capabilities when confronted with naturally occurring and strategically employed pragmatic language.

Subsequently, Paci et al. (2025) introduced the IMPAQTS-PID benchmark, containing over 30k implicit passages paired with expert explanations, and tested models in both multiple-choice and open-ended explanation settings. All tested models struggled to reconstruct manipulative implicit meanings: even the best-performing system remained far below the estimated performance ceiling, and fully correct explanations were produced only in a minority of cases. Chain-of-Thought prompting yielded slightly better results but they remain far from satisfactory. When models are required to freely generate an explanation of the implicit meaning rather than selecting among alternatives, performance drops sharply and fully correct explanations are produced only in a minority of cases. This suggests that the limitation does not merely concern classification or retrieval abilities, but the reconstruction of the intended communicative act itself.

Building on these findings, the IMPOLS shared task (Gregori et al., 2026) formalized the problem as a benchmark including detection, implicature/presupposition classification, and implicature type classification (models had to choose between conventional and (conversational) particularized or generalized implicatures), highlighting the difficulty of modeling context-dependent pragmatic inference in political speech.

Beyond political discourse, a growing body of work has investigated the ability of LLMs to perform pragmatic inference, particularly in relation to conversational implicatures and related phenomena. Recent studies have shown that, while models can capture certain scalar implicatures under controlled conditions, their performance remains unstable and highly sensitive to prompting strategies and contextual framing (Hu et al., 2023; Cho and mook Kim, 2024). Earlier benchmarks and analyses have sim-

ilarly highlighted that pragmatic inference in LLMs is often shallow, with models relying on surface-level heuristics rather than robust reasoning about speaker intentions (Min et al., 2022).

The present work extends these research lines by moving from evaluating general-purpose LLMs to explicitly training models to generate explanations of manipulative implicit content: we fine-tune Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct on explanation pairs derived from IMPAQTS-PID and evaluate their outputs through human annotation and automatic similarity metrics.

4. Dataset

IMPAQTS-PID[1] (Paci et al., 2025) is a benchmark derived from the IMPAQTS corpus, that focuses specifically on implicatures and presuppositions, and includes more than 30,000 annotated instances accompanied by wide contextual windows to provide sufficient information for context interpretation. The benchmark is designed as a multiple choice task: each occurrence contains 4 possible answers about the content that is implicitly conveyed by the sentence. Only one answer is correct.

For the current experiment, we used IMPAQTS-PID as a dataset, discarding the multiple choices, and focusing on implicit content explanation. The dataset has been split into train and test set, obtaining 25,000 pairs of implicit contents and their explanation to exploit for training purposes (see Table 2).

	Train	Test	Total
Implicatures	11,802	3,123	14,925
Presuppositions	13,198	3,699	16,897
Total	25,000	6,822	31,822

Table 2: Dataset numbers

4.1. Dataset example

Below we report an example of a dataset instance, containing a transcription excerpt with an implicit content (within <s> and </s> tags), the implicit type, and the implicit explanation.

> **Transcription with context**: Ognuno di noi si è tagliato lo stipendio e c'è qualcuno che non ci crede. C'è qualcuno che non ci crede, ancora. Gli porti i bonifici e ti dice: "Io non ci credo". Se sono così inetti non ci parliamo. C'è un 44% invece di <s>persone che non votano, che sono disperate e che sono disilluse</s>.
>
> *Each of us has cut our salary, and there are still people who don't believe it. There are people who still don't believe it. You show them the bank transfers and they say, "I don't believe it." If they are that inept, then we have nothing to say to each other. Meanwhile, there is a 44% of <s>people who do not vote, who are desperate and disillusioned</s>.*
>
> **Implicit type**: Implicature
>
> **Implicit content explanation**: Chi non vota è disperato e disilluso.
>
> *Those who do not vote are desperate and disillusioned.*

[1] https://github.com/WalterPaci/IMPAQTS-PID

5. Experiment

5.1. Fine-tuning

In this work, we exploited the training dataset to specialize two LLMs, Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct, through fine-tuning to perform implicit content explanation.

To contextualize the performance of the specialized models, we compare them against a strong general-purpose baseline. As reference, we adopt the zero-shot explanation setting introduced in Paci et al. (2025), where GPT-4o-mini is prompted to explicitly formulate the implicit content without task-specific training. This baseline represents the behavior of a general LLM relying solely on its pre-trained pragmatic competence, allowing us to assess whether improvements derive from instruction-based specialization rather than from model scale alone.

The selection of models reflects a controlled comparison between strong open-weight instruction-tuned systems (Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct) and a general-purpose closed model baseline (GPT-4o-mini). While these models differ in architecture and training regimes, the goal of this comparison is to evaluate the impact of task-specific supervision relative to general-purpose capabilities.

5.2. Evaluation method

The evaluation was conducted through the manual annotation of 50 explanations generated by each language model: Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct fine-tuned on IMPAQTS-PID dataset. The annotation was carried out independently by three annotators that were asked to evaluate the correctness of the generated explanations on a 0-4 scale, considering the following score meaning:

0. missing output;
1. wrong explanation;

2. the explanation contains correct elements, but it's wrong;

3. the output contains the correct explanation plus other elements;

4. the explanation is fully correct.

5.3. Results

Table 3 reports the number of explanations generated by the two fine-tuned models, manually evaluated and grouped by score, alongside an approximate distribution for the baseline model. While the results for the fine-tuned models are computed on our annotated sample with full access to instance-level labels, the baseline distribution is reconstructed from statistics reported in prior work (Paci et al., 2025).

However, the baseline results are not directly comparable at the same level of granularity: we cannot recover instance-level annotations, annotator agreement, or fully aligned evaluation conditions. For this reason, the baseline should be interpreted as an approximate but informative reference point, rather than as a strictly matched experimental condition.

The results for the fine-tuned models are based on gold labels obtained by aggregating the scores assigned by three independent annotators. In all cases, at least two annotators agreed on the same score, allowing for straightforward label selection through majority voting.[2] A comparison of score distributions reveals a clear difference in output behavior between general-purpose and fine-tuned models. While the baseline exhibits a broad dispersion across all score categories, including intermediate scores (2–3), the fine-tuned models show a more concentrated distribution, with no instances of score 3 and a higher proportion of fully correct explanations. This is expected, given the fine-tuning strategy reported in Appendix 9.

The agreement among annotators is moderate, with a Krippendorff's α between 0.76 and 0.78 in the two models (Table 4).

The overall accuracy on the manually annotated test set is 0.54 for Qwen and 0.46 for Llama (see Table 5). For both the fine-tuned models and the baseline, these values are computed by considering only the explanations that received the highest score (4). Notably, the baseline achieves an overall accuracy of 0.22, confirming that fine-tuning substantially improves performance on this task.

Surprisingly, the generated explanations are much more valid when the implicit type is a presupposition, while a lower accuracy is observed for implicatures. This results can be interpreted in light of the different cognitive and linguistic properties of presuppositions and implicatures: presuppositions are typically associated with lexical or syntactic triggers (e.g., definite descriptions, factive verbs), which constrain the space of possible inferences and make the implicit content more directly recoverable from the surface form (Stalnaker, 1978; Beaver, 2001). In contrast, implicatures are highly context-dependent and require reconstructing the speaker's communicative intention, often relying on pragmatic reasoning about relevance, contrast, and discourse goals (Grice, 1975; Sperber and Wilson, 1986; Levinson, 2000).[3]

From a modeling perspective, this suggests that fine-tuned LLMs can internalize mappings between surface cues and implicit content when these mappings are structurally constrained, but still struggle when inference requires integrating broader discourse context and speaker intentions. This asymmetry mirrors the distinction between local, form-driven inference and global, discourse-level reasoning, highlighting a key limitation of current LLMs in pragmatic interpretation. In other words, fine-tuning appears to primarily enhance trigger-based reconstruction, while leaving intention-driven inference comparatively underdeveloped.

We also observe that the fine-tuned models consistently produce concise answers, as evidenced by the absence of outputs evaluated with score 0 (missing answer) and score 3 (correct but containing additional information), in contrast to the baseline distribution. This difference can be interpreted in light of known generation biases in general-purpose LLMs. Prior work has shown that such models tend to exhibit a verbosity bias, where longer responses are systematically preferred or produced independently of their actual informativeness (Zheng et al., 2023). Similarly, LLMs have been shown to favor additive strategies, introducing additional or loosely related content beyond what is strictly required by the task (Ji et al., 2023). In this perspective, the presence of intermediate outputs (score 3) in the baseline can be seen as a byproduct of over-generation, where partially correct explanations are embedded within broader, less focused responses. By contrast, instruction-based fine-tuning has been shown to improve alignment with task objectives and to constrain generation toward more targeted outputs (Ouyang et al., 2022), which is consistent with the more concise and focused behavior observed in our models.

To provide a more fine-grained comparison with

[2] Direct fine-grained level comparability of the baseline with our model is limited due to differences in annotation protocols: Paci et al. (2025) used a single expert evaluator to evaluate outputs.

[3] This interpretation is supported by our qualitative analysis conducted in Section 5.4.

Score	Llama3.1-8B			Qwen2.5-7B			Baseline (GPT-4o-mini)		
	Impl.	Pres.	Tot.	Impl.	Pres.	Tot.	Impl.	Pres.	Tot.
4	4	19	23	7	20	27	19	13	32
3	0	0	0	0	0	0	7	6	13
2	9	5	14	8	6	14	18	27	45
1	11	2	13	9	0	9	30	28	58
0	0	0	0	0	0	0	1	1	2

Table 3: Distribution of explanation quality scores for fine-tuned models and baseline. Baseline counts are reconstructed from Paci et al. (2025) using zero-shot results over 150 samples (75 implicatures, 75 presuppositions).

	Krippendorff's α	Kendall's τ
Llama	0.78	0.40
Qwen	0.76	0.40

Table 4: Inter-annotator agreement (Krippendorff's α) and annotators' score correlation with BERTScore (Kendall's τ)

	Llama	Qwen	Baseline
Impl.	0.17	0.29	0.25
Presupp.	0.73	0.77	0.17
Overall	0.46	0.54	0.22

Table 5: Accuracy of fine-tuned models and the baseline on explanation generation task.

the baseline, we leverage the disaggregated statistics distinguishing between implicatures and presuppositions: the baseline shows slightly higher performance on implicatures than on presuppositions, while fine-tuned models exhibit the opposite pattern, achieving substantially higher accuracy on presuppositions. This contrast suggests that task-specific supervision enables models to better exploit structural cues associated with presuppositions, while implicature interpretation remains more challenging due to its reliance on context-dependent pragmatic reasoning.

Finally, we measured BERTScore (Zhang et al., 2019) on our test set to assess whether this metric can reliably estimate explanation quality. BERTScore is a semantic similarity metric based on contextual embeddings that evaluates the alignment between generated and reference texts, and is commonly used in natural language generation as a proxy for semantic adequacy. To this aim, we computed the Kendall's τ correlation between the BERTScore and the gold labels, resulting in 0.40 for both fine-tuned models (see Table 4). This value indicates a moderate correlation with human judgments, suggesting that BERTScore alone is not sufficient to reliably select correct explanations.

5.4. Qualitative analysis

A manual inspection of the errors reveals that most failures do not originate from lexical or compositional misunderstanding, but from an incorrect reconstruction of the communicative act performed by the speaker. In particular, models tend to interpret utterances propositionally, while political discourse frequently encodes stance through presupposition, rhetorical structure, and indirect contrast. As a consequence, models systematically mis-handle (i) metadiscursive negation, (ii) generic polemical reference, and (iii) rhetorical accusations expressed through interrogative or self-praise constructions. The problem concerns pragmatic orientation rather than semantic content: models paraphrase the literal proposition instead of recovering the speaker's intended evaluative target.

One phenomenon that often occurs in both models' errors is the application of the grammatical negation to the embedded proposition instead of recognizing that the negation targets an attitude or stance toward the whole context.

> **Original**[4]**:** La considerazione non vale più per nessuno, né, per quanto ci riguarda, la nostalgia per il passato, né per quanto riguarda gli alti o i finti innovatori che assisterebbero il nostro paese.
> **IMPAQTS:** Vi sono finti innovatori che assistono il nostro paese.
> **Qwen:** In realtà gli alti e finti innovatori non assistono il Paese.
> **Llama:** In realtà gli alti e finti innovatori non assistono il nostro paese.
> '**Original:** *Consideration no longer applies to anyone, nor nostalgia for the past, nor to the so-called innovators assisting our country.*
> **IMPAQTS:** *There are fake innovators assisting the country.*
> **Qwen:** *In reality the so-called innovators do not assist the country.*

[4]We only report the relevant part of the following examples and not the whole context given to the model for space reasons.

Llama: *In reality the so-called innovators do not assist our country.*'

When the target of criticism is intentionally generic (e.g., an unspecified governing coalition), models replace it with a concrete entity, introducing hallucinated referents. In the following example, the model also states something that is not present nor relevant in the excerpt.

Original: mettiamo in guardia dal rischio di un esecutivo frutto di una nuova maggioranza tra diversi, una maggioranza improvvisata, che esiste solo in Parlamento e non nel Paese, una maggioranza che non rispecchia assolutamente la maggioranza degli elettori italiani.
IMPAQTS: Quella attuale è una maggioranza tra diversi, improvvisata, che esiste solo in Parlamento e non nel Paese, che non rispecchia assolutamente la maggioranza degli elettori italiani.
Llama: La maggioranza tra FI e i suoi alleati è improduttiva.
'**Original:** *We warn against a government based on a heterogeneous and improvised parliamentary majority not representative of voters.*
IMPAQTS: *The current majority is heterogeneous and unrepresentative.*
Llama: *The majority between FI and its allies is unproductive.*'

Models also tend to interpret self-praise constructions literally instead of contrastively; political self-attribution often presupposes a negative comparison with previous governments, like in the example that follows. Here, the speaker implies having done something previous governments failed to do, but models simply restate an explicit proposition.

Original: Rivendico con orgoglio di aver fatto quello che gli italiani avevano dimenticato si potesse fare, ovverosia mantenere una promessa elettorale.
IMPAQTS: I governi precedenti non mantengono alcuna promessa elettorale, mentre Salvini sì.
Qwen: Il M5S ha mantenuto la promessa elettorale.
Llama: Il PD ha mantenuto la promessa elettorale di abolire i decreti sicurezza.
'**Original:** *I proudly claim to have done what Italians had forgotten was possible: keeping an electoral promise.*
IMPAQTS: *Previous governments do not keep promises, while the speaker does.*
Qwen: *M5S kept the promise.*
Llama: *PD kept the promise of abolishing the security decrees.*'

Another recurrent phenomenon is the strengthening of rhetorical questions into intentional claims: in the example that follows, an accusatory rhetorical question encoding criticism through presupposition is interpreted as an explicit intentional statement.

Original: si è domandato pochi giorni fa in presenza del ministro Napolitano perché, dopo due anni di governo della sinistra, non è giunta una sola voce, un invito forte da parte del governo a denunciare a chi subisce estorsioni.
IMPAQTS: Non è giunto un solo invito da parte del Governo a denunciare chi subisce estorsioni.
Llama: Il governo non è interessato a denunciare le estorsioni.
'**Original:** *He asked why, after two years of a left-wing government, no strong invitation to report extortion has been issued.*
IMPAQTS: *No invitation to report extortion was issued.*
Llama: *The government is not interested in reporting extortion.*'

These errors consistently arise when interpretation requires identifying the illocutionary function of the utterance before deriving its propositional content. Models instead derive meaning locally and only subsequently attempt inference, which leads to distortions in argumentative discourse. This suggests that, while fine-tuning improves the mapping between linguistic form and implicit content, current models still struggle with forms of inference that require global discourse interpretation rather than local trigger-based reconstruction, in line with recent findings on the limitations of LLMs in pragmatic reasoning (Hu et al., 2023; Cho and mook Kim, 2024; Ma et al., 2025). Importantly, this limitation cannot be attributed solely to model scale: previous work has shown that larger general-purpose models in zero-shot settings still fail to reliably reconstruct implicit content in political discourse (Paci et al., 2025), and that performance in complex reasoning tasks depends critically on the structure and relevance of the input signals rather than on scale alone (Min et al., 2022). This indicates that pragmatic competence does not robustly emerge just from scaling, but depends crucially on the availability of task-relevant training signals. In this perspective, our results suggest that even relatively simple instruction-based fine-tuning on structured pragmatic data can lead to substantial improvements, highlighting the importance of data construction and task formulation alongside model size.

6. Conclusions

In this study, we investigated the ability of Large Language Models to unveil and explain manipulative implicit communication in Italian political discourse. Our experiments demonstrate that while general-purpose models previously struggled with these tasks, specialized fine-tuning provides a significant performance boost.

Fine-tuning Qwen2.5 and Llama3.1 led to accuracy scores of 0.54 and 0.46, respectively, for fully correct explanations. Moreover, there is a big difference in how models handle different pragmatic phenomena; both models were highly effective at explaining presuppositions (reaching up to 0.77 accuracy) but found implicatures considerably more difficult to reconstruct.

We found that BERTScore has only a moderate correlation with human expert judgment and cannot be reliably used on its own to identify correct explanations.

These findings, together with previous evidence on the limitations of general-purpose models in pragmatic tasks, suggest that improvements in this domain depend not only on model scale but also on the availability of structured and task-specific training data, which enables models to better capture discourse-level and intention-driven aspects of meaning.

In future work, we will extend this line of research by exploring how task-specific supervision interacts with broader model characteristics, including scale and architectural differences, in order to better understand the conditions under which pragmatic reasoning abilities can emerge in LLMs.

7. Limitations

This study focuses on a limited number of models and does not aim to provide a comprehensive comparison across architectures, parameter scales, or training paradigms. In particular, we consider two open-weight instruction-tuned models (Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct) and a single closed-model baseline (GPT-4o-mini), which are intended as representative instances rather than as an exhaustive sample of current LLMs.

As a consequence, we do not systematically investigate how factors such as model architecture, parameter size, or language specialization (e.g., Italian-specific or multilingual models) influence the ability to reconstruct implicit content. Similarly, while the inclusion of a strong closed-model baseline provides a useful reference point, our analysis does not extend to a broader range of proprietary systems.

These limitations imply that our findings should be interpreted primarily as evidence of the effectiveness of instruction-based fine-tuning on structured pragmatic data. A more extensive evaluation across a wider variety of models, including larger-scale and closed-source systems, is necessary to fully assess the role of scale, architecture, and training data in pragmatic reasoning.

8. Bibliographical References

David I Beaver. 2001. *Presupposition and assertion in dynamic semantics*, volume 29. CSLI publications Stanford.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, and Alessandro Panunzi. 2024. Impaqts: a multimodal corpus of parliamentary and other political speeches in italy (1946-2023), annotated with implicit strategies. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024*, pages 101–109.

Lorenzo Gregori, Walter Paci, and Saccone Valentina. 2026. IMPOLS at EVALITA 2026: Overview of the IMPOLS Task. In *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*, Bari, Italy. CEUR.org.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Edoardo Lombardi Vallauri. 2017. Implicits as evolved persuaders. In *Pragmemes and theories of language use*, pages 725–748. Springer.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Walter Paci. 2025. "it's a further exercise in futility": implicit content detection and classification in italian political discourse. a pilot study. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 2(2).

Walter Paci, Alessandro Panunzi, and Sandro Pezzelle. 2025. They want to pretend not to understand: The limits of current LLMs in interpreting implicit content of political discourse. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15569–15593, Vienna, Austria. Association for Computational Linguistics.

Dilyorjon Solidjonov. 2025. Pragmatic competence without embodiment? evaluating llm performance on implicature, presupposition, and speech acts.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

9. Appenidix

Models fine-tuning parameters

The following prompt has been used to perform instruction tuning of Llama and Qwen models.

> Il seguente estratto di un discorso politico contiene un contenuto implicito non bona fide vero tra i tag <s> e </s>. Esplicita questo contenuto implicito.
>
> *The following excerpt from a political speech contains a non-bona fide true implicit content marked between the tags <s> and </s>. Make this implicit content explicit.*

The following parameters have been used for Llama and Qwen models fine-tuning.

> num_train_epochs=3
>
> per_device_train_batch_size=2
>
> gradient_accumulation_steps=8
>
> learning_rate=2e-5
>
> lr_scheduler_type="cosine"
>
> warmup_ratio=0.1

Manual evaluation results

The following tables reports the judgments of each annotator per score. Both partial values on presuppositions and implicatures, and the overall values are reported.

Score	4	3	2	1	0
Total items					
annot 1	22	0	15	13	0
annot 2	19	0	15	16	0
annot 3	24	0	13	13	0
Implicatures					
annot 1	4	0	10	10	0
annot 2	3	0	7	14	0
annot 3	6	0	8	10	0
Presuppositions					
annot 1	18	0	5	3	0
annot 2	16	0	8	2	0
annot 3	18	0	5	3	0

Table 6: Number of explanation generated by fine-tuned Llama3.1-8B-Instruct evaluated by each annotator per score.

Score	4	3	2	1	0
Total items					
annot 1	26	0	13	11	0
annot 2	24	0	12	14	0
annot 3	26	0	15	9	0
Implicatures					
annot 1	6	0	8	10	0
annot 2	6	0	5	13	0
annot 3	7	0	8	9	0
Presuppositions					
annot 1	20	0	5	1	0
annot 2	18	0	7	1	0
annot 3	19	0	7	0	0

Table 7: Number of explanation generated by fine-tuned Qwen2.5-7B-Instruct evaluated by each annotator per score.

Examining Algebraic Recombination for Compositional Generalisation

Joaquin Cardona Ruiz, Antske Fokkens, Lucia Donatelli
Vrije Universiteit Amsterdam
j.cardonaruiz@student.vu.nl, {antske.fokkens, l.e.donatelli}@vu.nl

Abstract

We introduce DEAR, a compositional semantic parser that extends the LEAR model (Liu et al., 2021) to handle non-projective syntactic derivations. We evaluate our parser on SLOG (Li et al., 2023), a compositional generalisation benchmark focused on structural generalisation. We find that DEAR achieves stronger performance on syntactically complex examples than projective baselines, even as it incurs a modest decrease in overall accuracy due to a larger search space. Analysis of model behaviour confirms a fundamental trade-off: while projective architectures yield efficient, benchmark-aligned learning, non-projective mechanisms provide the expressive capacity needed to capture richer linguistic dependencies. These findings contribute to current discussions on how more linguistically faithful semantic parsers can be developed beyond benchmark-specific biases.

Keywords: Compositional generalisation, Semantic parsing, Non-projectivity

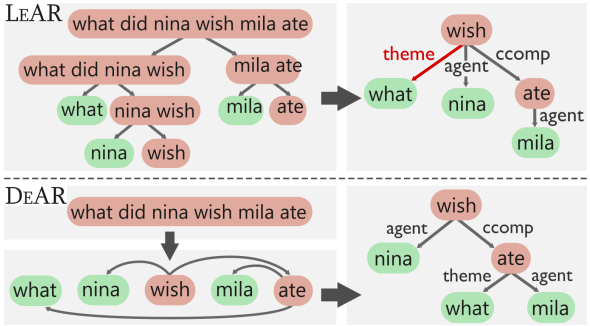


Figure 1: DEAR extends LEAR's compositional parsing architecture by replacing its constituency structure with a dependency structure. This allows DEAR to handle non-projective arguments that LEAR fails to capture, as seen in red.

1. Introduction

Compositional generalisation, the ability to recombine known linguistic components to derive the meanings of novel constructions (Partee, 2004), is an essential component of human language acquisition and understanding. Despite their broad linguistic coverage, sequence-to-sequence (seq2seq) models often fail to generalise compositionally in semantic parsing tasks (Lake and Baroni, 2018; Dehghani et al., 2019a; Herzig and Berant, 2021). When presented with structures requiring the systematic recombination of known elements, such models tend to overfit to surface-level correlations rather than capturing underlying syntactic–semantic regularities that make human generalisation effective.

Models designed to be explicitly compositional, such as LEAR (Liu et al., 2021) and the AM-Parser (Groschwitz et al., 2018), address this limitation by deriving an intermediate syntactic representation from the input that can be algebraically recombined into a target semantic graph. These architectures can achieve strong compositional generalisation while being very data efficient. Nevertheless, syntactic assumptions built into these models often impose projectivity constraints that limit their overall expressivity. As a result, they struggle to capture the flexible, non-local dependencies that characterise natural language *in the wild* beyond more predictable tendencies of structures found in benchmark datasets (Shaw et al., 2021) (Sec. 2).

To examine the tension between *expressivity* and *efficiency* in compositional language models, we introduce DEAR (**De**pendency **A**lgebraic **R**ecombination), a compositional semantic parser that extends LEAR's architecture to model a wider range of structures (Sec. 3). DEAR's most notable introduction is the use of an unconstrained dependency intermediate representation, rather than a constituency one, which allows for the modelling of non-projective structures (Fig. 1). This design draws inspiration from empirical evidence of non-projective structures across languages (Havelka, 2007) and theoretical work demonstrating that non-projective mechanisms possess greater semantic expressive capacity than projective ones (Venant and Koller, 2019). All code developed for this study is publicly available.[1]

We explore the performance and architectural implications of both models through the SLOG benchmark (Li et al., 2023) (Sec. 4, 5). SLOG maps English sentences to logical forms representing their meaning, and features a generalisation split that requires models to recombine seen structures

[1] https://github.com/JoaquinCardonaRuiz/thesis-compositionality

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 230–249
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

to generalise to new ones. This focus on *structural generalisation* is crucial, given that recent work has demonstrated that lexical generalisation, the generalisation of individual lexical units to unseen positions, is easier than structural generalisation (Weißenhorn, 2022). Notably, SLOG includes instances of non-projectivity, the presence of crossing edges in an ordered dependency tree (analogous to the non-contiguity of constituents in a constituency tree, see Hays, 1964 and Nivre, 2008). The dataset is thus an ideal setting to probe whether more expressive parsing mechanisms confer empirical benefits.

This study makes the following contributions:

1. A novel compositional parser, DEAR, which extends LEAR to support non-projective syntactic structures. DEAR outperforms projective parsers on specific structurally complex phenomena.
2. An empirical study of expressivity–efficiency trade-offs in compositional parsing, showing how controlled non-projectivity impacts both search complexity and generalisation.

Our results affirm the potential of principled linguistic design in semantic parsers to generalise beyond benchmark-specific biases to more versatile natural language structures, and highlight the value of interpretability in the development and evaluation of NLP tools (Sec. 6).

2. Compositional Generalisation

The Principle of Compositionality states that the meaning of a complex expression is determined by the meanings of its parts and the rules combining them.[2] This principle underlies humans' ability to generate infinite meanings from finite elements (Partee, 2004; Chomsky, 1965; Fodor and Pylyshyn, 1998).

Recent work has examined whether language models can replicate this ability in systematic semantic parsing tasks (Lake and Baroni, 2018; Dehghani et al., 2019a), raising the question of why compositional generalisation matters for NLU systems. From a pragmatic perspective, compositionality in a language model serves as a domain-specific form of parameter-sharing, reducing the number of parameters needed by sharing components to process similar structures in arbitrary contexts. Learning to identify common patterns independently of position, rather than learning each structure anew, removes a hard barrier to a model's ability to generalise, and is a major factor for sample efficiency (Lake et al., 2015): a model which can leverage compositionality to correctly assess infinite recombinations of seen components needs significantly fewer training examples to generalise.

More conceptually, the in-distribution approach often employed in supervised learning is not well suited to evaluate the linguistic capabilities of modern NLU systems, as generalisation becomes hard to distinguish from memorisation of their vast training datasets (Kim et al., 2022). To examine a model's ability to generalise out-of-distribution, and therefore to learn, evaluation can be grounded on the principles of human understanding of language. The compositionality of language means humans innately display a capacity for *compositional generalisation*, recombining known components to derive the meaning of unseen constructions (Partee, 2004). Semantic parsing benchmarks can target this capability by presenting primitives and basic structures during training, but withholding examples for evaluation which require recombining these elements into unseen positions and structures.

Importantly, compositional behaviour depends on how models represent and constrain syntax. Most relevant for this study is the constraint of projectivity, which in dependency syntax requires that for every arc from a head to a dependent, all words linearly intervening between them are dominated by that head (Jurafsky and Martin, 2026). Most sentences in a language are projective (Melčuk, 1988), yet non-projective dependencies arise naturally in valid constructions, most notably in languages with flexible word order (Havelka, 2007), but also in English, for instance in cases of long-distance *wh*-movement. In projective systems, composition therefore occurs only between adjacent constituents (Nivre, 2008). While this constraint simplifies inference, it limits expressivity: natural language frequently exhibits non-projective dependencies that require non-local composition (Howell and Zamaraeva, 2018; Venant and Koller, 2019). This section surveys datasets and modelling approaches that probe this tension between structural simplicity and expressive capacity in compositional generalisation.

2.1. Datasets

Semantic parsing benchmarks designed to specifically target compositional generalisation have overwhelmingly been synthetic in nature, and thus not representative of language *in the wild* (Shaw et al., 2021). The most notable early efforts are SCAN (Lake and Baroni, 2018) and CFQ (Dehghani et al., 2019a). Both map simple sentences to computer commands, and feature **generalisation sets**, which present sentences with known atoms in previously unseen positions. The dataset examined in this work, SLOG (Li et al., 2023), and its predecessor COGS (Kim and Linzen, 2020), introduce sentences closer to natural language in diversity

[2] Usually attributed to Gottlob Frege; see Janssen (2012).

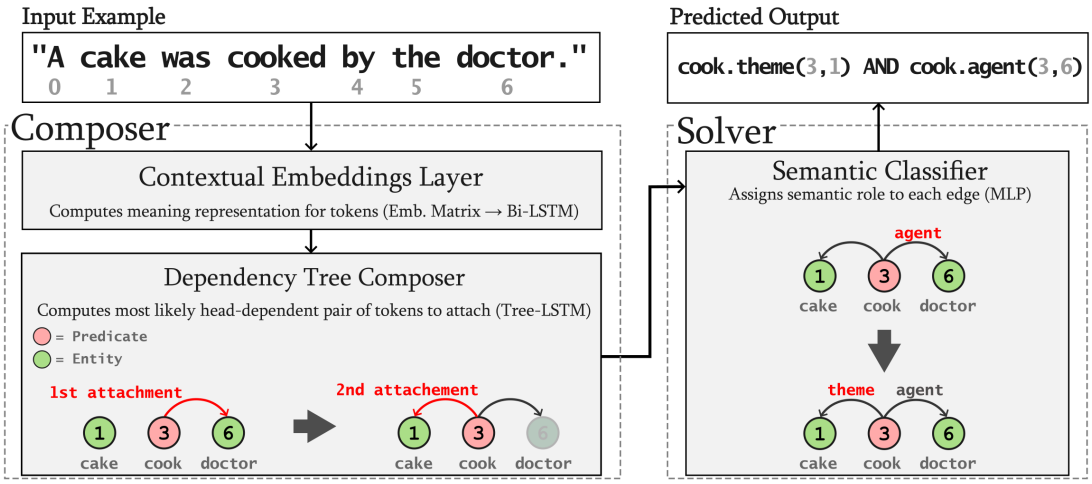


Figure 2: Overview of DEAR's architecture, showcasing an example from COGS. The input sentence (top-left) gets fed into the *Composer* module (left), which creates a dependency tree. Then the *Solver* (right) labels it, and produces an output (top-right).

and complexity, mapped to meaning representations based on the Lambda Calculus. This frames semantic parsing as a sequence-to-sequence problem, as shown in Example (1).

(1) **Sentence**: *Liam liked to run.*
Target Form: `like.agent(x_1, x_0) AND like.xcomp(x_1, x_3) AND run.agent(x_3, x_0)`

SLOG extends COGS to more structurally complex generalisation cases, most notably non-projective dependencies. Test cases are grouped into categories of 1000 cases each, in which a specific systematic gap is evaluated. Where most categories in COGS are focused on *lexical generalisation*, SLOG targets all its categories towards structural recombination, which requires recombining entire syntactic structures to arrive at previously unseen ones.

2.2. Approaches

Despite recent seq2seq models displaying seemingly impressive linguistic capabilities in both comprehension and production, they perform inconsistently on compositional generalisation benchmarks when presented with linguistic structures withheld during training (Lake and Baroni, 2018; Loula et al., 2018; Andreas, 2020; Oren et al., 2020; Dehghani et al., 2019a; Herzig and Berant, 2021). This suggests a lack of the same systematic generalisation capabilities humans possess, especially when it comes to structural—rather than lexical—recombination (Zhu et al., 2021; Yao and Koller, 2022; Zheng and Lapata, 2022; Li et al., 2023).

Compositional parsers architecturally model a latent structure of the input expression and recombine its components to derive the correct meaning representation. Many efforts have focused on the SCAN dataset, making use of either quasi-synchronous context-free grammars (Kim, 2021; Qiu et al., 2022) or constituency trees (Liu et al., 2020, 2021; Herzig and Berant, 2021) as latent representations. LEAR extends constituency-based approaches on simpler benchmarks to COGS, achieving near-perfect performance (97.7% accuracy). It learns a latent constituency tree through a Tree-LSTM (Tai et al., 2015) and assigns semantic operations to each node through several MLPs. Another compositional parser, the AM parser (Groschwitz et al., 2018), instead implements a dependency parser based on stacked Bidirectional LSTMs and an MLP supertagging component. Its application to the SLOG benchmark in Li et al. (2023) yields outstanding results (70.8% accuracy).

3. DEAR

We introduce DEAR, a compositional semantic parsing model that builds on LEAR (Liu et al., 2021). As in LEAR, the model consists of two components: a *Composer*, which induces a latent syntactic structure, and a *Solver*, which assigns semantic relations to structural edges. The two stages are optimised through reinforcement learning using separate reward signals. An overview of DEAR's components, alongside an example input and output is shown in Figure 2.

The primary departure from LEAR lies in the choice of structural representation. Whereas LEAR constructs binary constituency trees, DEAR induces a *dependency* tree over abstracted lexical items. Dependency-based composition allows the model to directly represent predicate–argument relations and, crucially, does not impose projectivity constraints during tree construction. As a result, DEAR can produce crossing dependencies when

they are semantically licensed, making the architecture more suitable for naturalistic text and languages exhibiting freer word order. An overview of LEAR's architecture, similar to the one presented for DEAR in Figure 2, is shown in Appendix B.

3.1. Input Token Classification and Representation

Following LEAR, DEAR assumes a predefined lexicon of *content words* in the training corpus, each typed as either an *entity* or a *predicate* according to SLOG's context-free grammar. Input tokens that appear in this lexicon are labelled with their respective type, and those that do not are labelled as *non-content words*. The model therefore does not identify predicate–argument structure from raw text; instead, it focuses on determining the correct relations between a known closed set of primitives. This design isolates the model's ability to generalise to new structures, without confounding effects from lexical ambiguity. Details on the construction of this lexicon are provided in Appendix A.2.

All tokens are mapped to 256-dimensional embeddings and fed into a bidirectional LSTM encoder. The encoder consists of independent forward and backward LSTMs with hidden size 256. Their outputs are concatenated (yielding 512 dimensions) and passed through a learned linear projection to 512 dimensions, followed by a $\tanh$ activation. The resulting vector is then split evenly into a hidden state and a cell state (h_i, c_i) for each token, which serve as its initial representation for downstream composition. Tokens also retain their label as either *entity*, *predicate*, or *non-content word*.

3.2. *Composer*: Structure Induction Via Dependency Composition

The *Composer* iteratively constructs a dependency tree over *content words*. Its architecture, described below, is additionally illustrated in Figure 3, alongside that of the Solver. As in LEAR, *non-content words* are masked out from the set of candidate nodes, although information about them is preserved in the contextualised BiLSTM representations of *content words*.

At each iteration, the Composer evaluates every ordered pair (i, j) of remaining content-word nodes by first computing a candidate composed parent state via the binary Tree-LSTM:

$$(\tilde{h}_{i\leftarrow j}, \tilde{c}_{i\leftarrow j}) \;=\; \mathrm{TLSTM}\big((h_i, c_i), (h_j, c_j)\big), \qquad (1)$$

where the Tree-LSTM computes gate activations from a linear projection of the concatenated child hidden states and combines child cell states (see Appendix A.1 for notes on our implementation of Tai et al. (2015)'s Tree-LSTM). Each candidate composed hidden state is then scored by taking its dot product with a learned parameter vector $\mathbf{q}$, producing a single real number used as the unnormalised logit score for the arc:

$$s(i, j) \;=\; \tilde{h}_{i\leftarrow j}^{\top}\mathbf{q}. \qquad (2)$$

A categorical distribution over valid arcs is formed from these scores after masking out invalid attachments (self-loops and previously attached dependents). During training, an arc is sampled from this distribution; at test time, the highest-scoring arc is selected deterministically. The chosen dependent is composed into the head (the head's state is replaced by the composed state) and removed from further consideration. This iterative process continues until a single root remains. Because attachment decisions depend only on pairwise token representations and no span-based constraints are imposed, the resulting trees may be non-projective.

This design contrasts with LEAR, which iteratively scores all valid adjacent spans of content words and merges the best-scoring span into a single constituent. LEAR 's design enforces binary-branching constituency structure and therefore cannot produce crossing edges.

Furthermore, because the representations of chosen dependent tokens are composed into that of their corresponding heads, the head's state becomes a summary of the entire subtree rooted at that node, enabling the model to perform genuine recursive composition.

3.3. *Solver*: Semantic Labelling

Given the induced tree, the *Solver* assigns a semantic relation to each directed edge. We train separate feed-forward classifiers (one-hidden-layer MLPs) for different head–dependent type configurations: *predicate–entity*, *predicate–predicate*, and *entity–entity*. Edges of type *entity–predicate* are deterministically assigned a *relative clause* label.

Each MLP takes as input the concatenation of the hidden states of the head and dependent:

$$\lambda_{h,d} = \mathrm{MLP}_{\tau(h),\tau(d)}([h_h; h_d]). \qquad (3)$$

Concretely, each classifier consists of a linear projection from 2×256 to 256, followed by a ReLU activation and a final linear layer mapping to the label set associated with that type pair (e.g., *agent*, *theme*, *recipient* for *predicate–entity* edges).

Post-processing expands relative clauses and introduces the reentrancies required by the target logical form, converting the dependency arborescence into a directed graph that no longer satisfies strict tree constraints. Further details on post-processing are provided in Appendix A.4.

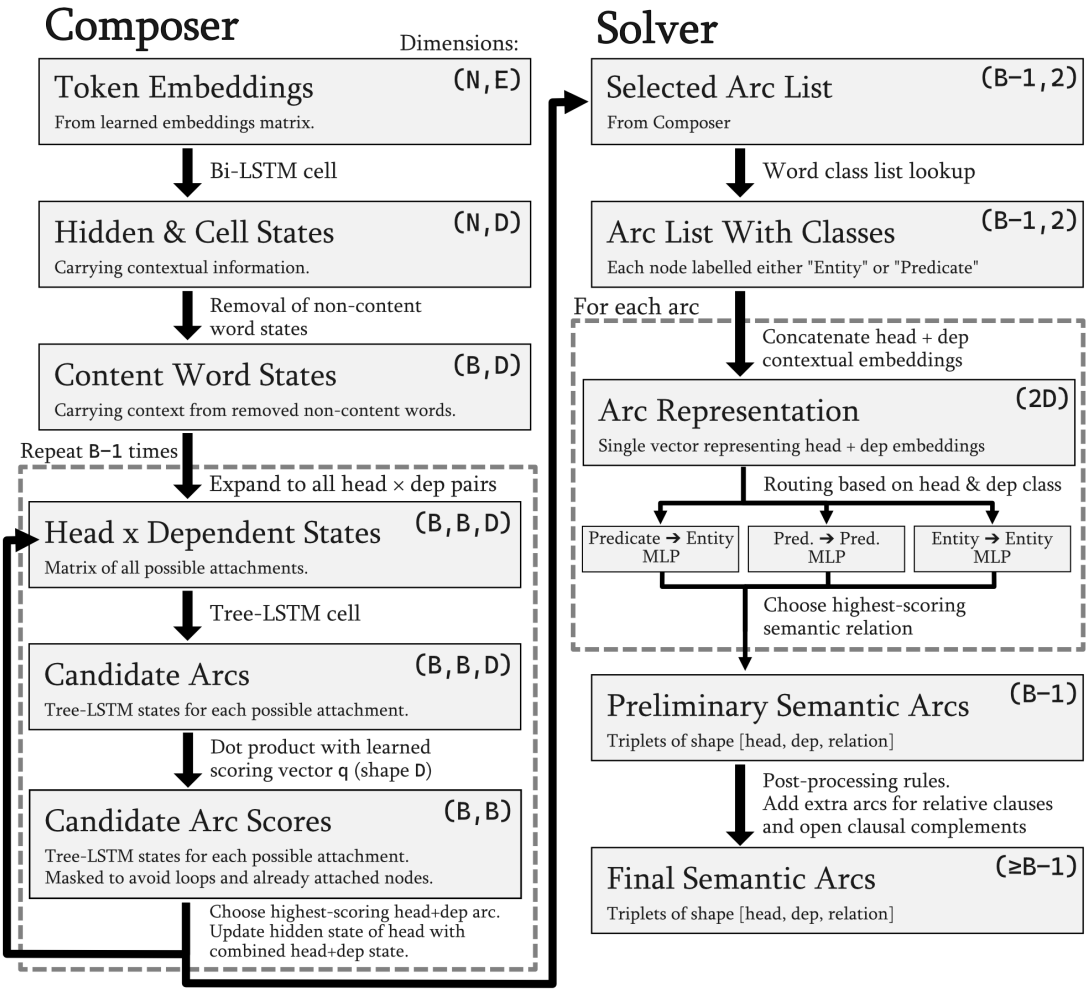


Figure 3: Detailed diagram of DEAR's architecture. The Composer (left column) takes the input sentence and predicts dependency edges, and the Solver (right column) labels those edges with semantic relations, producing a final labelled graph. At each step, the dimensions of the vector representations are shown to the right of the box, where N is the number of input tokens, B is the number of content words ($B \leq N$), E is the word embedding dimension, and D is the hidden state dimension, used for the Bi-LSTM, Tree-LSTM, and MLPs.

3.4. Reinforcement Learning Objective

Training follows the hierarchical reinforcement-learning setup of Liu et al. (2021). For each input, the model samples K complete derivations (we use K=15), each consisting of a compositional structure proposed by the *Composer* and a semantic labelling of all induced edges produced by the *Solver*. For a predicted derivation we denote by $P = \{(h, d, \lambda)\}$ its labelled edges, and by $\bar{P}$ the same edges without labels; G and $\bar{G}$ denote the corresponding gold sets.

The structural and semantic components receive separate rewards. The *Composer* receives a reward for correctly recovering unlabelled dependency arcs, while the *Solver* receives a reward for correctly labelling those edges that were structurally correct:

$$R_{\text{comp}} = H\Big(\tfrac{|\bar{P}\cap\bar{G}|}{|\bar{P}|},\ \tfrac{|\bar{P}\cap\bar{G}|}{|\bar{G}|}\Big)$$

$$R_{\text{solv}} = \begin{cases} \dfrac{|P \cap G|}{|\bar{P} \cap \bar{G}|}, & |\bar{P} \cap \bar{G}| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

For each sampled derivation we record the summed log-probability of the Composer's discrete attachment decisions ($\log p_{\text{comp}}$) and the summed log-probability of the Solver's label decisions ($\log p_{\text{solv}}$). Each module is then updated *independently* using REINFORCE with a simple mean-baseline. Here $E[R_{\text{comp}}]$ and $E[R_{\text{solv}}]$ denote the empirical mean reward over the K sampled derivations for the corresponding module, and p_{comp} and p_{solv} are the model's own probabilities assigned to the specific sequence of sampled decisions:

$$\begin{aligned} \mathcal{L}_{\text{comp}} &= -\big(R_{\text{comp}} - E[R_{\text{comp}}]\big)\ \log p_{\text{comp}} \\ \mathcal{L}_{\text{solv}} &= -\big(R_{\text{solv}} - E[R_{\text{solv}}]\big)\ \log p_{\text{solv}} \end{aligned} \tag{5}$$

We optimise the model using two separate Adadelta optimisers. The *high-level* optimiser updates all parameters associated with structural decisions—namely the *Composer*, the Tree-LSTM, and the input encoder—while the *low-level* optimiser updates only the Solver's MLP classifiers. Gradients for the two modules are kept disjoint: after computing the relevant loss, we apply gradient clipping to the corresponding parameter subset and perform an optimiser step only for that subset. This ensures that the *Composer* receives no gradient signal from semantic mislabelling, and that the *Solver* is trained solely from label rewards on structurally correct arcs. In this respect our model differs from LEAR, which trains both modules using a single shared reward.

4. Experimental Setup

We evaluate the DEAR model introduced in this work on the SLOG benchmark, and report per-category accuracies. Additionally, we evaluate the LEAR model (Liu et al., 2021) and two pre-trained large language models (Llama 3 8b, Mistral 7b) to provide context for DEAR's performance and examine the impact of the different model architectures. We re-report Li et al. (2023)'s results for the AM-Parser and a vanilla Transformer. Details on computational resources and hardware setup can be found in Appendix C.2, and the hyperparameters chosen for each model can be found in Appendix C.1. We evaluate the models based on accuracy: the percentage of samples in which the model produces a semantically equivalent output to the ground truth. Reported results are measured on SLOG's generalisation set.

4.1. Adapting LEAR to SLOG

LEAR's functionality is heavily reliant on heuristics and pre-computed alignment tables specific to the COGS dataset it was developed for. Adapting the model to process the SLOG dataset was not trivial. Implemented changes are outlined below.

Content questions: The SLOG dataset incorporates syntactic structures not present in COGS, amongst them content questions. In their target logical form, *wh*-words are abstracted with a generic *interrogated entity* symbol, rather than being treated as indexed tokens, as shown in Example (2). In order to keep LEAR's architecture unchanged, sentences were pre-processed to treat *wh*-words as *entity* words.

(2) **Sentence**: *Who did Henry adore?*
Target Form: `adore.agent(x_3,x_2) AND adore.theme(x_3,?)`
Adjusted Form: `adore.agent(x_3,x_2) AND adore.theme(x_3,x_0)`

Standalone NP: SLOG introduces sentences without a root predicate, absent from COGS, as shown in Example (3). To account for LEAR's architectural constraints, which demand a root predicate to assign arguments to, a *null predicate* is introduced without an index value, and assigned the root NP as a theme.

(3) **Sentence**: *The cake on a tree.*
Target Form: `cake.nmod.on(x_1, x_4)`
Adjusted Form: `null.theme(x, x_1) AND cake.nmod.on(x_1, x_4)`

Relative Clauses: Unlike COGS, SLOG features relative clauses, as shown in Example (4). LEAR's output representation was updated to allow relative clause tags to be present, though the model remains unable to predict such relations due to being architecturally restricted from creating *entity→predicate* edges.

(4) **Sentence**: *Liam was lent the rose that was liked.*
Target Form: `lend.recipient(x_2, x_0) AND lend. theme(x_2, x_4) AND rose.relcl(x_4, x_7) AND like.theme(x_7, x_4)`

These changes enable the evaluation of LEAR on the SLOG dataset, which served as a crucial step towards the development of the DEAR model, informing the architectural changes to be made, and providing a baseline for comparison.

5. Results & Error Analysis

Per-category results are presented in Table 1 and examined in detail in the following sections. While the most relevant categories are briefly described below, Appendix D contains a complete description of all categories, alongside example sentences, for ease of reference. Overall, compositional models consistently outperform seq2seq ones. DEAR is the only such model to achieve non-zero performance in SLOG's non-projective category, *Wh Q. long movement*. Pre-training aids performance of seq2seq models, but obfuscates the systematic gap between training and evaluation data, leading to potential overestimation of compositional capabilities (Kim et al., 2022).

5.1. LEAR's Performance on SLOG

LEAR achieves an accuracy of 83.2% on SLOG's in-distribution validation set and 54.2% on its generalisation set. Several of SLOG's generalisation categories present a significant challenge to LEAR's architecture. Below is an overview of those that prove architecturally impossible for the model to accurately predict, and therefore have an in-category accuracy of 0%.

RC in subj. NPs, RC in ind. object NPs & Ind. object-extracted RC: These categories introduce systematic gaps by presenting models with relative

Category	LEAR$^{\kappa}$	DEAR$^{\kappa}$	Transf.†	AM-Parser$^{\dagger\kappa}$	LLaMA 3$^{\pi}$	Mistral$^{\pi}$
Shallower PP rec.	100.0	99.4	98.7	100.0	96.0	94.9
Deeper PP rec.	100.0	38.0	13.1	100.0	21.5	29.3
Shallower tail CP rec.	100.0	99.0	32.6	100.0	80.1	66.8
Deeper tail CP rec.	100.0	31.9	0.2	100.0	8.5	7.9
Shallower center emb.	**0.0**	**95.9**	0.0	100.0	17.3	0.0
Deeper center emb.	**0.0**	**23.4**	0.0	99.5	0.1	0.3
PP in subj. NPs	66.9	**16.9**	0.0	57.6	46.0	16.2
PP in ind. object NPs	67.4	**0.6**	42.5	90.4	74.9	61.3
RC in subj. NPs	**0.0**	**3.6**	0.0	55.8	22.6	28.0
RC in ind. object NPs	**0.0**	**1.8**	34.4	74.4	62.4	47.5
Ind. obj-extracted RC	**0.0**	**99.9**	4.7	0.0	6.1	0.0
Dir. object wh Q.	48.7	70.1	2.8	29.4	5.5	40.6
Ind. object wh Q.	100.0	94.1	35.9	41.4	11.1	17.7
Active subj. wh Q.	100.0	99.0	96.7	99.8	96.5	97.7
Passive subj. wh Q.	100.0	100.0	27.4	100.0	18.7	40.1
Wh Q. modified NPs	38.0	14.9	17.6	55.6	27.3	31.5
Wh Q. long movement	**0.0**	**7.8**	4.0	0.0	10.9	33.6
Overall	54.2	53.1	24.2	70.8	35.6	36.1

Table 1: Main Results. Mean Accuracy (%) for models evaluated on each of SLOG's categories. Label κ indicates a compositional model, and π indicates a pre-trained model. Results labelled † are re-reported from Li et al. (2023) for comparison.

clauses in previously unseen grammatical positions. This does not present a challenge to LEAR's capability for generalisation, but to its inductive bias. Relative clauses introduce *semantic reentrancy*, wherein a single entity fills a role in more than one predicate. LEAR's semantic dependency tree does not allow for reentrancy, as all entities get assigned a role in a single predicate. Importantly, this is a failure of LEAR's *Solver* component, not of its *Composer*. The *Composer* module is capable of producing the correct syntactic structure. However, because of the *Solver*'s limitations outlined above, it is never rewarded for doing so.

This pattern holds for all sentences featuring relative clauses in both the training and generalisation sets. 34.5% of predictions for the *RC in subject NP* category, 53.6% for *RC in indirect object NP*, and 25.3% for *Ind. object-extracted RC* contain only errors caused by this architectural limitation. The remaining predictions contain at least one *unforced* error, which consists of entities being assigned to erroneous roles by the *Solver*. **Shallower center embedding & Deeper center embedding:** The accuracy of 0% on these categories is due to the same reasons as the categories above: LEAR is incapable of modelling relative clauses.

Wh Q. long movement. This category presents sentences that exhibit *long wh*-movement, wherein the *wh*-word is separated from its corresponding predicate by a subordinate clause. This makes the structures in this category non-projective. LEAR's *Composer* creates phrases by selecting one pair of adjacent tokens (or previously composed phrases) in each iteration to merge. This adjacency restriction means the model is incapable of producing such non-projective trees, and thus the accuracy is 0%.

(5) **Sentence**: *What did Liam wish that the father ate?*
Latent syntax:
`[[what [liam wish]] [father ate]]`
Resulting edges:
`agent(wish, liam), theme(wish, what), ccomp(wish, eat), agent(eat, father)`

In Example (5), the *wh*-word *what*, which should be the *theme* of predicate *ate*, is separated from its predicate by several tokens. LEAR's prediction erroneously creates an edge attaching the *wh*-word to its nearest predicate, highlighted in red. This limitation is inherent to LEAR's Composer, which can only attach tokens to adjacent phrases, and is a major motivation for the development of DEAR in the present study.

5.2. DEAR's Performance on SLOG

DEAR achieves an accuracy of 97.8% on SLOG's in-distribution validation set and 53.1% on its generalisation set. It presents non-zero accuracies on all categories in SLOG, unlike comparable models, as can be seen in Table 1. Nevertheless, in-category performance is generally lower than other compositional models, which may stem from the relation between model complexity and the limited dataset size. This is further discussed in Section 6. Details on time and computational expense can be found in Appendix C.2. Below is an overview of the categories in which the results achieved by DEAR differ substantially from those achieved by LEAR.

Deeper PP recursion & Deeper tail CP recursion: DEAR achieves near perfect accuracy on the categories presenting *shallower* recursions than those seen in training. In the categories presenting

Category	5	6	7	8	9	10	11	12
PP recursion	93.6	91.2	68.8	40.0	20.8	1.6	1.6	0.8
Comp. Reward	99.0	97.7	90.9	82.4	75.5	66.0	62.7	58.1
Solv. Reward	100.0	100.0	99.9	99.8	99.7	99.9	99.7	99.9
Tail CP recursion	70.4	59.2	30.4	11.2	6.4	2.4	0.8	0.8
Comp. Reward	92.5	88.0	80.8	71.6	66.6	69.6	64.7	60.9
Solv. Reward	98.6	99.3	99.1	99.1	99.5	99.1	99.6	99.6
Center embedding	93.6	61.6	20.8	3.2	0.8	0.8	0.0	0.0
Comp. Reward	98.3	89.9	76.4	63.4	58.2	55.4	48.0	45.5
Solv. Reward	99.4	95.4	93.1	94.0	95.2	93.5	93.0	90.7

Table 2: Mean Accuracy (%), Mean Composer Reward (%) and Mean Solver Reward (%) achieved by DEAR at each level of Recursion Depth (columns) for the categories Deeper PP recursion, Deeper tail CP recursion, & Deeper center embedding (rows). Deeper recursion levels trigger a fall in *Composer* reward and overall accuracy, but not on *Solver* reward.

deeper recursion, however, performance is mixed. A closer inspection of the results shows that the model struggles with the deeper levels of recursion, with performance dropping as the number of nested clauses in the sentence grows, as can be seen in Table 2.

This indicates DEAR is able to generalise up to recursion depths unseen during training, but is limited in its capacity to accurately model long or deeply nested sentences. The *Solver* rewards being mostly unaffected by sentence depth show that this is mostly a limitation of DEAR's *Composer*.

Shallower center embedding & Deeper center embedding: Thanks to the inclusion of *entity* to *predicate* edges in DEAR, and the resulting ability for the *Solver* to tag dependencies as relative clauses, these categories are able to be captured with non-zero accuracy. However, the same pattern of decreasing accuracy with increasing depth can be seen in Table 2.

RC in subj. NPs, RC in ind. object NPs & Ind. object-extracted RC: Similarly to the categories involving center embedding, the integration of relative clauses to the model allows for these categories to be predicted for with non-zero accuracy. Nevertheless, accuracies on *RC in subj. NPs* and *RC in ind. object NPs* remain low. Alongside the similarly low scores in *PP in subj. NPs* and *PP in ind. object NPs*, this reflects the model's difficulties transferring multi-token sequences to positions unseen during training, which these four categories require. This is further confirmed by the near-perfect performance on *Ind. obj-extracted RC*.

Notably, *RC in subj. NPs*, which requires NPs seen in direct object positions to be generalised to subject positions, sees mostly failures in the *Composer*. These involve either correctly parsing the transferred NP, but misplacing it as the root of the entire sentence, or fully failing to capture it. These syntactic failures occur in 51% of sentences in this category, with only 14% displaying any *Solver* errors. *RC in ind. object NPs*, which requires direct object NPs to be generalised to indirect object positions, presents mostly *Solver* errors, which appear in 56% of sentences, as opposed to 20% for the *Composer*. *Solver* errors in both categories almost entirely consist of mislabelled roles within relative clauses.

Wh Q. long movement: 92.2% of sentences in this category are classified incorrectly, exhibiting the same fault present in LEAR's predictions. The *wh*-words introducing the questions are attached to their nearest predicate, rather than to the correct distant predicate. DEAR is architecturally capable of modelling such sentences correctly, as indicated by the non-zero accuracy, but in practice falls short of consistently identifying them. This limitation of the *Composer* seems to apply uniformly to most sentences in this category, without an identifiable pattern in the few sentences that are predicted correctly. This points to the relationship between the increased model complexity of DEAR and the limited training data as the culprit, which will be discussed in Section 6.

6. Discussion

Our results show that the choice of latent syntactic representation substantially affects both a model's expressivity and its computational efficiency. LEAR 's *Composer* encodes a strong inductive bias aligned with simple English syntax. This projective design limits expressivity for non-projective constructions but sharply reduces the syntactic search space compared to DEAR 's dependency-based approach. Concretely, LEAR 's space grows exponentially with sentence length, following the Catalan sequence, whereas DEAR 's grows super-exponentially according to Cayley's formula (See Appendix A.5). As a result, DEAR 's search space becomes quickly intractable for long or recursive constructions, such as PP and CP embeddings.

Sentence length also mediates generalisation. On SLOG's validation set—dominated by short sentences—DEAR outperforms LEAR (97.8% to 83.2%), but performance equalises on the gen-

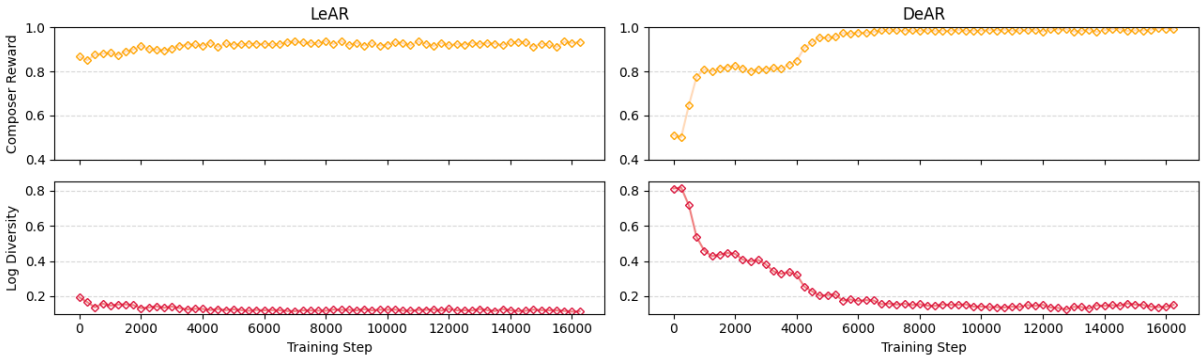


Figure 4: Mean *Composer* Reward (Top) and Diversity (Bottom) during training for LEAR (Left) and DEAR (Right). Each data-point aggregates 500 sentences.

eralisation set (53.1% to 54.2%), which includes longer and more complex sentences. This inversion reflects the combinatorial growth of DEAR 's hypothesis space with increasing length.

Training dynamics further reveal this trade-off (Figure 4). Visualisation of model predictions shows that DEAR initially explores a much wider range of syntactic hypotheses (diversity ≈ 0.8 to LEAR 's ≈ 0.2), leading to slower convergence and less consistent structural reuse. Although DEAR ultimately achieves higher *Composer* rewards, these do not translate to improved generalisation—likely because LEAR 's constrained search promotes more transferable compositional patterns.

These findings are in line with the well-established inverse relationship between the size of the search-space in a reinforcement learning algorithm and its sample-efficiency (Yu, 2018), as well as computational efficiency in dependency parsers in general (Gómez-Rodríguez, 2016). A larger training set would likely be the most impactful factor in allowing DEAR to converge to a stable and transferable solution. Still, there are other considerations beyond the quantity of data available. The models' Lambda-style compositional architectures, which enforce a strict separation between syntactic and semantic representations, mean that semantic placeholders cannot be used to limit the search space (e.g. a predicate having filled all its expected semantic roles, thus not allowing more to be attached, see Venant and Koller (2019)). Still, other linguistically motivated constraints on DEAR's dependency-based *Composer* are feasible. Much work has been dedicated to explore potential constraints of non-projective dependency parsers that allow for the expressivity necessary to model natural language without the combinatorial explosion of the search space that leads to poor efficiency (Kahane et al., 1998; Nivre, 2006; Tratz and Hovy, 2011; Kuhlmann, 2013). The interpretability of DEAR may prove valuable in exploring the impacts of such constraints in the future. We hope this study serves to highlight the necessity of such theoretically-informed methods to incorporate non-projectivity while constraining expressivity in ways that are minimally detrimental to performance on natural, multilingual samples.

7. Conclusion

We introduce DEAR, a dependency-based compositional semantic parser that extends the expressivity of LEAR by handling non-projective syntax. Our experiments on the SLOG benchmark demonstrate that relaxing projectivity constraints enables DEAR to capture a broader range of syntactic–semantic dependencies. However, the additional expressivity comes at the cost of overall performance, particularly in deeply nested sentences, confirming a known trade-off between achieving the expressivity necessary to capture natural language and maintaining efficiency by reducing a model's search space. While systems that rely heavily on dataset-specific assumptions like LEAR are beneficial in demonstrating the value of compositional architectures, designs with increased expressivity are inevitably necessary as benchmarks evolve to better approximate the structural diversity of natural language.

Future work can explore integrating linguistically motivated constraints to temper DEAR 's combinatorial explosion while retaining its expressive advantages. Additionally, larger and more structurally diverse datasets could reveal whether the observed trade-offs persist in more naturalistic settings.

8. Limitations

Limitations of this study include the synthetic nature of the datasets utilised, which limits the applicability of any results to real-world natural language data (see Shaw et al., 2021). A focus on English as an input language provides another, more general limitation, as different languages present unique challenges to the assumptions made by syntactic models, particularly regarding the frequency of non-projective structures. Lastly, the framing of the task assumed by the chosen datasets is an inherent simplification of all that could potentially be captured within semantic parsing. Semantic factors like quantifiers, events and their place in time, conditionals and truth conditions, and co-reference are left out of scope (see Flickinger et al., 2014).

From a methodological standpoint, the results reported here are based on a single experimental run per model due to computational constraints, and no systematic hyper-parameter search was conducted. While this limits the robustness of the quantitative comparisons, we verified that the observed trends were stable across preliminary exploratory runs. Future work should aim to replicate these results under multiple random seeds and systematically tuned hyper-parameters to ensure statistical reliability.

A natural progression of this work and of related studies may be to expand existing compositional generalisation benchmarks to more closely reflect the diversity and complexity of natural language, to cover languages beyond English, and to adopt more comprehensive meaning representations. Furthermore, the potential of recent architectural innovations such as Universal Transformers (Dehghani et al., 2019b; Tan et al., 2023) in explicitly compositional models would be a fruitful area for further work.

9. Ethical considerations

This work presents a methodological contribution to compositional semantic parsing in a benchmark-specific setting, with no immediate real-world deployment. As such, it is not expected to pose direct risks of harm or misuse. We note, however, that the study involves fine-tuning two large language models, which entails a non-negligible environmental cost due to energy consumption associated with GPU usage. All other experiments, namely the training of LEAR and DEAR, were performed on CPU hardware, with comparatively minimal resource demands. Finally, all datasets used in this study are synthetic and publicly available, containing no personally identifiable information or sensitive content.

10. References

Jacob Andreas. 2020. Good-Enough Compositional Data Augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics (ACL).

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50th edition. The MIT Press, Cambridge, MA.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019a. Measuring Compositional Generalization: A Comprehensive Method on Realistic Data. In *International Conference on Learning Representations*.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019b. Universal transformers. In *International Conference on Learning Representations*.

Dan Flickinger, Emily M. Bender, and Stephan Oepen. 2014. Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 875–881, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jerry A. Fodor and Zenon W. Pylyshyn. 1998. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Jonas Groschwitz, Matthias Lindemann, Meaghan Fowlie, Mark Johnson, and Alexander Koller. 2018. AMR dependency parsing with a typed semantic algebra. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1831–1841, Melbourne, Australia. Association for Computational Linguistics (ACL).

Carlos Gómez-Rodríguez. 2016. Restricted non-projectivity: Coverage vs. efficiency. *Computational Linguistics*, 42(4):809–817.

Jiří Havelka. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 608–615, Prague, Czech Republic. Association for Computational Linguistics (ACL).

David G. Hays. 1964. Dependency theory: A formalism and some observations. *Language*, 40(4):511–525.

Jonathan Herzig and Jonathan Berant. 2021. Span-based semantic parsing for compositional generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 908–921, Online. Association for Computational Linguistics (ACL).

Kristen Howell and Olga Zamaraeva. 2018. Clausal modifiers in the grammar matrix. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2939–2952, Santa Fe, NM, USA. Association for Computational Linguistics (ACL).

Theo Janssen. 2012. Compositionality: Its Historic Context. In *The Oxford Handbook of Compositionality*. Oxford University Press, Oxford.

Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing*, 3rd edition. Draft of January 6, 2026. https://web.stanford.edu/~jurafsky/slp3/.

Sylvain Kahane, Alexis Nasr, and Owen Rambow. 1998. Pseudo-projectivity: A polynomially parsable non-projective dependency grammar. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Najoung Kim and Tal Linzen. 2020. COGS: A Compositional Generalization Challenge Based on Semantic Interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics (ACL).

Najoung Kim, Tal Linzen, and Paul Smolensky. 2022. Uncontrolled Lexical Exposure Leads to Overestimation of Compositional Generalization in Pretrained Models. *arXiv:2212.10769*.

Yoon Kim. 2021. Sequence-to-sequence learning with latent neural grammars. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 26302–26317, Red Hook, NY, USA. Curran Associates Inc.

Marco Kuhlmann. 2013. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 4487–4499, Stockholm, Sweden. Proceedings of Machine Learning Research (PMLR).

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A Structural Generalization Benchmark for Semantic Parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics (ACL).

Chenyao Liu, Shengnan An, Zeqi Lin, Qian Liu, Bei Chen, Jian-Guang Lou, Lijie Wen, Nanning Zheng, and Dongmei Zhang. 2021. Learning Algebraic Recombination for Compositional Generalization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1129–1144, Online. Association for Computational Linguistics (ACL).

Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, and Dongmei Zhang. 2020. Compositional generalization by learning analytical expressions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 11416–11427, Red Hook. Curran Associates Inc.

João Loula, Marco Baroni, and Brenden Lake. 2018. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 108–114, Brussels, Belgium. Association for Computational Linguistics (ACL).

Igor A. Melčuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany, NY.

Joakim Nivre. 2006. Constraints on non-projective dependency parsing. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–80, Trento, Italy. Association for Computational Linguistics (ACL).

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving

Compositional Generalization in Semantic Parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2482–2495, Online. Association for Computational Linguistics (ACL).

Barbara Partee. 2004. Compositionality. In *Compositionality in Formal Semantics: Selected Papers by Barbara H. Partee*. Blackwell Publishing, Malden.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Pawel Nowak, Tal Linzen, Fei Sha, and Kristina Toutanova. 2022. Improving Compositional Generalization with Latent Structure and Data Augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4341–4362, Seattle, WA, USA. Association for Computational Linguistics (ACL).

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional Generalization and Natural Language Variation: Can a Semantic Parsing Approach Handle Both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics (ACL).

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, Beijing, China. Association for Computational Linguistics (ACL).

Shawn Tan, Yikang Shen, Zhenfang Chen, Aaron Courville, and Chuang Gan. 2023. Sparse universal transformer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 169–179, Singapore. Association for Computational Linguistics (ACL).

Stephen Tratz and Eduard Hovy. 2011. A fast, accurate, non-projective, semantically-enriched parser. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1268, Edinburgh, Scotland, UK. Association for Computational Linguistics (ACL).

Antoine Venant and Alexander Koller. 2019. Semantic expressive capacity with bounded memory. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 65–79, Florence, Italy. Association for Computational Linguistics (ACL).

Pia Weißenhorn. 2022. Investigations on what makes compositional generalization challenging. Master's thesis, Saarland University.

Yuekun Yao and Alexander Koller. 2022. Structural generalization is hard for sequence-to-sequence models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5048–5062, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).

Yang Yu. 2018. Towards Sample Efficient Reinforcement Learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence Early Career*, pages 5739–5743. International Joint Conferences on Artificial Intelligence Organization (IJCAI).

Hao Zheng and Mirella Lapata. 2022. Disentangled Sequence to Sequence Learning for Compositional Generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4256–4268, Dublin, Ireland. Association for Computational Linguistics (ACL).

Wang Zhu, Peter Shaw, Tal Linzen, and Fei Sha. 2021. Learning to generalize compositionally by transferring across semantic parsing tasks. *arXiv:2111.05013*.

A. Appendix: Implementation Notes

This appendix contains notes on the implementation of the LEAR model as presented by Liu et al. (2021) and as modified in this study for the SLOG dataset, as well as on the implementation of the DEAR model.

A.1. Tai et al. (2015)'s Tree-LSTM in LEAR and DEAR

LEAR's *Composer* merges the hidden and cell states of selected nodes into those of a single parent node using Tai et al. (2015)'s N-ary Tree-LSTM. Nevertheless, their implementation of the Tree-LSTM cell differs from that of the original authors in a significant way.

Tai et al. define their cell transition equations for the kth child of node j as follows:

$$i_j = \sigma\Big(\sum_{\ell=1}^{N} U_{\ell}^{(i)}\, h_{j\ell} + b^{(i)}\Big), \tag{6}$$

$$f_{jk} = \sigma\Big(\sum_{\ell=1}^{N} U_{k\ell}^{(f)}\, h_{j\ell} + b^{(f)}\Big), \tag{7}$$

$$o_j = \sigma\Big(\sum_{\ell=1}^{N} U_{\ell}^{(o)}\, h_{j\ell} + b^{(o)}\Big), \tag{8}$$

$$u_j = \tanh\Big(\sum_{\ell=1}^{N} U_{\ell}^{(u)}\, h_{j\ell} + b^{(u)}\Big), \tag{9}$$

$$c_j = i_j \odot u_j \; + \sum_{\ell=1}^{N} f_{j\ell} \odot c_{j\ell}, \tag{10}$$

$$h_j = o_j \odot \tanh(c_j). \tag{11}$$

Liu et al. mostly maintain this architecture for LEAR's *Composer*, but crucially replace Equation (10) with the following:

$$c_j = \tanh\Big(i_j \odot u_j \; + \sum_{\ell=1}^{N} f_{j\ell} \odot c_{j\ell}\Big) \tag{12}$$

In developing the DEAR model, it was chosen to revert to the original formulation of c_j expressed in Equation (10). The reason for this is twofold.

Firstly, it was assumed to be an error on the part of LEAR's authors, given that the hyperbolic tangent is applied to c_j by invoking PyTorch's `tanh_` method rather than `tanh` when computing Equation (11). Because `tanh_` is an in-place operation, this results in the hyperbolic tangent being both used in the calculation of h_j *and* applied to the final value of c_j before being returned.

```
h = o * c.tanh_()
```
(13)

Secondly, this additional hyperbolic tangent was found to cause numerical instability in certain hyper-parameter configurations which particularly affected DEAR. Thus, it was chosen to modify this part of the cell to restore the behaviour described by Tai et al., by making use of the out-of-place `tanh`. This results in an identical value for h_j, so Equation (11) remains unaffected, but the value of c_j is no longer being computed using a hyperbolic tangent.

```
h = o * c.tanh()
```
(14)

A.2. Pre-Computed Tables

LEAR and DEAR make use of several pre-computed alignment tables that provide the models with dataset-specific information about tokens. Below is a categorised description of the tables in use, with statistics as of Liu et al. (2021)'s original implementation for the COGS dataset:

A.2.1. Lemmatisation:

The file `enct2dect` maps all *content words* found in the dataset to their corresponding lemmas. Given the simple nature of the dataset's vocabulary, the only content words whose lemma differs from their original form are predicates conjugated in the past tense, which are mapped to their unconjugated counterparts. There are a total of 731 tokens in this table, 591 of which have lemmas identical to their original form. There are 643 resulting unique lemmas, due to predicates in both their present and past forms being present.

A.2.2. Token Classification:

There are three tables that serve to classify tokens. These are `entity`, `unac_predicate`, and `caus_predicate`. Tokens can be classified as either an *entity* if they are on the `entity` table, or a *predicate*, if they are in either `unac_predicate` or `caus_predicate`. If a token is in none of the tables, it is deemed not to be a content word, and thus does not become a node during parsing. The `entity` table has 509 entries, `unac_predicate` has 21, and `caus_predicate` 113, for a total of 643 content words.

A.2.3. Indexing:

The file `encode_tokens` maps all tokens in the dataset to a unique, incremental ID based on their index in the table. It contains 741 entries, which account for all 731 content words plus 10 non-content words, which are: "*on*", "*by*", "*was*", "*beside*", "*the*", "*in*", "*that*", "*to*", "*a*", and ".".

In summary, there are 741 distinct tokens present in the dataset. 731 of them are considered content

words. These content words are lemmatised to 643 distinct values, 509 of which are classified as *entities* and 134 as *predicates*.

A.3. Pre–Computed Tables for SLOG

The alignment tables utilised by LEAR and DEAR, detailed in Appendix A.2, were expanded to encompass the additional vocabulary present in SLOG. An overview of the changes is shown below:

A.3.1. Entities:

4 new tokens were added to the entities list due to being present in SLOG but not in COGS. These consist of 2 regular entities *"gardener"* and *"monastery"*, alongside the 2 words introducing *wh*-questions, *who* and *what*. Note that the word *"gardener"* is present in the publicly available version of SLOG (Li et al., 2023) in the form *"gardner"*. This was manually changed to *"gardener"* in the data used for this study to avoid confusion.

A.3.2. Predicates:

All predicates present in SLOG are also present in COGS, so there are no additions to the `caus_predicate` and `unac_predicate` tables. However, 33 predicates that are only present in their past conjugation in COGS appear in their present conjugation in SLOG. These forms were correspondingly added to the lemmatisation table `enct2dect`. The list of such predicates can be seen below:

`dream, declare, pass, loan, say, award, mail, promise, prove, post, think, hope, sell, feed, confess, bring, lend, expect, believe, forward, give, send, slip, serve, wish, wire, rent, return, offer, hand, mean, support, imagine`

A.4. DEAR Post-processing

Once DEAR's *Composer* has produced a valid dependency tree, and its *Solver* has labelled all edges, the resulting labelled dependency tree lacks the edges corresponding to re-entrant entities, which can be deterministically derived from the existing labels. This is performed in two steps, *Relative Clause Reentrancy* and *Open Complement Reentrancy*, both of which are illustrated in Figure 5.

A.4.1. Relative Clause Reentrancy

Relative clauses are represented as an edge from the root *entity* to the *predicate* that modifies it. These relationships trigger an additional edge in the opposite direction, expressing the semantic role the *entity* plays in its *predicate*. While this edge is created manually, based on the existence of the original edge labelled as a relative clause (`relcl`), it is labelled by applying the same MLP mechanism as any other *predicate→entity* edge.

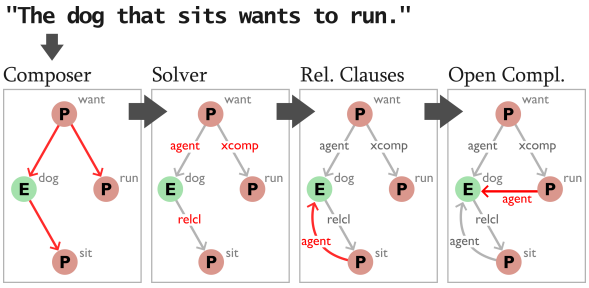


Figure 5: Diagram illustrating DEAR's two post-processing steps. From left to right, the *Composer* outputs a dependency tree, the *Solver* labels edges with semantic relations, and the post-processing steps (*Rel. Clauses* and *Open Compl.*) deterministically add the extra edges required to match the target representation. In this example sentence, the entity *"dog"* is the agent of all three predicates.

A.4.2. Open Complement Reentrancy

Predicates positioned as arguments of other predicates can be labelled as *open clausal complements* (`xcomp`), in which case both share an agent. Given the existence of such a relation in the *Composer*'s output, an additional edge is inserted to express this agent reentrancy.

A.5. Search Space Growth by Sentence Length

As discussed in Section 6, LEAR and DEAR's syntactic search spaces, the number of possible trees they could construct, grow with the sentence length (the number of content words), in different ways. LEAR's *Composer* follows the Catalan sequence, detailed below

$$P_{\text{LeAR}}(n) = \frac{(2n)!}{(n+1)!n!} \sim \frac{4^n}{n^{\frac{3}{2}}\sqrt{\pi}} \tag{15}$$

Whereas DEAR's follows Cayley's formula, as below:

$$P_{\text{DeAR}}(n) = n^{n-2} \tag{16}$$

Table 3 compares the corresponding amount of possible trees for sentences of lengths commonly found in the SLOG dataset.

B. Appendix: LEAR's Architecture

An overview of LEAR's architecture, and an example input and output, is shown in Figure 6.

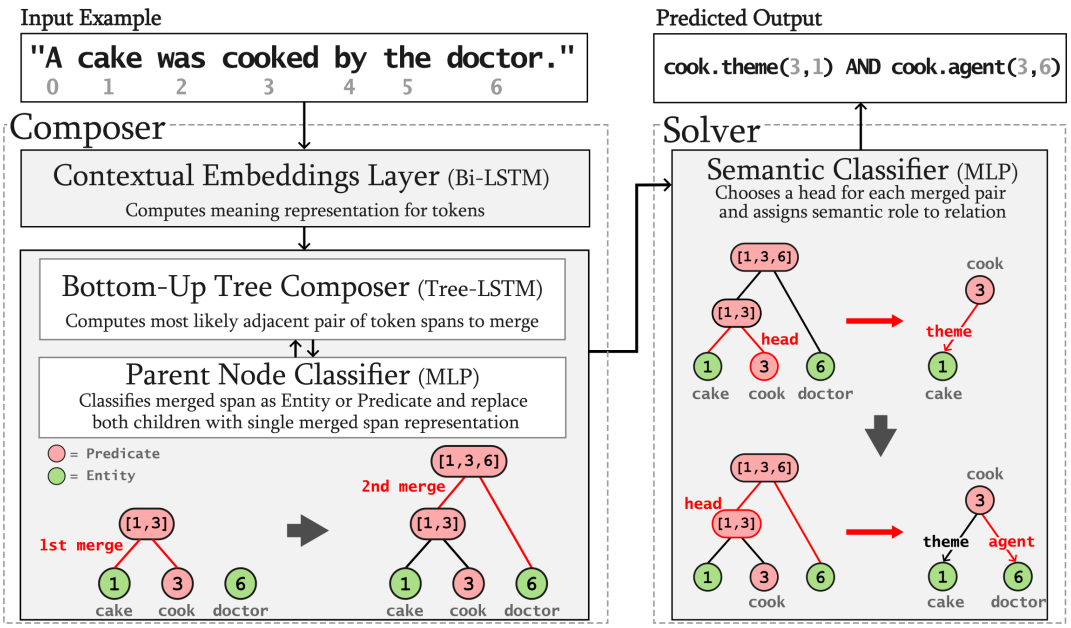


Figure 6: Overview of LEAR's architecture, showcasing an example from COGS. The input sentence (top-left) gets fed into the *Composer* module (left), which creates a constituency tree. Then the *Solver* (right) predicts a dependency structure from each constituent, and produces an output (top-right).

Length	LEAR	DEAR
2	1	1
3	2	3
4	5	16
5	14	125
6	42	1,296
7	132	16,807
8	429	262,144
9	1,430	4,782,969
10	4,862	100,000,000
11	16,796	2,357,947,691
12	58,786	61,917,364,224
13	208,012	1,792,160,394,037
14	742,900	56,693,912,375,296

Table 3: Comparison of LEAR and DEAR's search spaces across sentence lengths.

C. Appendix: Training Notes

C.1. Hyper-parameters

Due to constrained resources, a systematic hyper-parameter search was not performed in this study. Additionally, reported results were produced by a single training run, so a seed-sensitivity analysis remains for future work. The architectural differences between the LEAR and DEAR models, and DEAR's increased complexity in particular, make it unsuitable to compare them using identical hyper-parameters. Specifically, LEAR's learning rate results in numerical instability when applied to DEAR's larger architecture, so a lower one was chosen. Additionally, DEAR's increased search-space benefits from a higher number of samples in its REINFORCE loop. Table 4 below shows a list of the models' principal hyper-parameters, alongside a description for each.

The hyper-parameters used to fine-tune the LLMs in this study, Llama3 8b and Mistral 7b, are shown below.

- **LoRA configuration:** rank $r = 16$, $\alpha = 32$, dropout $= 0.05$, bias = none
- **Target modules:** `[q_proj, k_proj, v_proj, o_proj, up_proj, down_proj, gate_proj]`
- **Quantization:** 4-bit NF4 with double quantization; compute dtype = `bfloat16`/`float16` depending on GPU support
- **Optimiser:** `paged_adamw_8bit`
- **Learning rate:** 5×10^{-5} with cosine scheduler and warmup ratio $= 0.03$
- **Batch size:** 1 (train), 2 (evaluation)
- **Gradient accumulation steps:** 16
- **Epochs:** 1
- **Precision:** `fp16` enabled, `bf16` disabled
- **Gradient checkpointing:** enabled
- **Sequence length:** 512 tokens
- **Evaluation interval:** every 500 steps
- **Logging interval:** every 10 steps
- **Scheduler warmup ratio:** 0.03
- **Inference parameters:** batch size = 128, maximum generation length = 1024 tokens

Hyper-parameter	Description	LEAR	DEAR
Hidden Layer Size	Vector size of token embeddings, LSTM states, and MLP inputs.	128	256
High Learning Rate	Learning rate for the *Composer* parameters, including embeddings.	1.0	0.1
Low Learning Rate	Learning rate for the *Solver* parameters.	0.5	0.05
Batch Size	Batch Size for gradient accumulation.	1	8
Number of Samples	Number of Monte-Carlo samples explored for each sentence.	10	15
Epochs	Number of training epochs during training.	1	3
Optimisation Method	Stochastic Gradient Descent algorithm employed during training.	AdaDelta	AdaDelta

Table 4: Hyper-parameters chosen for the LEAR and DEAR models.

C.2. Resource and Time Requirements

All training for both LEAR and DEAR was performed on an `AMD Ryzen 7 5825U` CPU, at 2GHz, running as a single-threaded process. Training time was found to scale mostly linearly with the number of parameters, determined by the *Hidden Layer Size* hyper-parameter, the number of Monte–Carlo samples performed, and the number of epochs needed, which was affected by learning rates. Whereas LEAR achieved its maximum performance on epoch 1, over-fitting afterwards, DEAR required at least 3 epochs to adequately train. Table 5 below shows a summary of training times for LEAR and DEAR.

While the increased training time exhibited by DEAR as compared to LEAR is inherent to its architecture and the chosen hyper-parameters (see Appendix C.1), it is likely not reflective of parallelised GPU performance. This is because the increased Monte-Carlo samples, as well as the larger hidden vector size, would not scale the computational expense to the same degree.

Fine-tuning of Llama3 8b and Mistral 7b was done on an A100 GPU, with a training duration of approximately 5 hours each.

D. Appendix: Dataset Categories

Below are all the categories in the generalisation set for both COGS and SLOG, alongside its identifier within the dataset (for reference when working with the code published alongside this paper), a description of the category, and an example taken from the original authors' publications (Kim and Linzen, 2020; Li et al., 2023).

D.1. COGS: Lexical Recombination Categories

D.1.1. Subject → Object (common noun):

(ID: `subj_to_obj_common`) Entities consisting of common nouns only seen in subject positions during training, presented in object positions in the generalisation set.
Train.: *A **hedgehog** ate the cake.*
Gen.: *The baby liked the **hedgehog**.*

D.1.2. Subject → Object (proper noun):

(ID: `subj_to_obj_proper`) Entities consisting of proper nouns only seen in subject positions during training, presented in object positions in the generalisation set.
Train.: ***Lina** gave the cake to Olivia.*
Gen.: *A hero shortened **Lina**.*

D.1.3. Object → Subject (common noun):

(ID: `obj_to_subj_common`) Entities consisting of common nouns only seen in object positions during training, presented in subject positions in the generalisation set.
Train.: *Henry liked a **cockroach**.*
Gen.: *The **cockroach** ate the bat.*

Model/Dataset	Mean Epoch Time (minutes)	# Epochs	Total Time (minutes)
LEAR (SLOG)	73.4	1	73.4
DEAR (SLOG)	394.7	3	1184.2

Table 5: Training times, in minutes, for the LEAR and DEAR models on the SLOG dataset.

D.1.4. Object → Subject (proper noun):

(ID: `obj_to_subj_proper`) Entities consisting of proper nouns only seen in object positions during training, presented in subject positions in the generalisation set.
Train.: *The creature grew **Charlie**.*
Gen.: ***Charlie** worshipped the cake.*

D.1.5. Primitive noun → Subject (common noun):

(ID: `prim_to_subj_common`) Common noun entity that only appears as a stand-alone primitive in training, and is presented as the subject in the generalisation set.
Train.: ***shark***
Gen.: *A **shark** examined the child.*

D.1.6. Primitive noun → Subject (proper noun):

(ID: `prim_to_subj_proper`) Proper noun entity that only appears as a stand-alone primitive in training, and is presented as the subject in the generalisation set.
Train.: ***Paula***
Gen.: ***Paula** sketched William.*

D.1.7. Primitive noun → Object (common noun):

(ID: `primn_to_obj_common`) Common noun entity that only appears as a standalone primitive in training, and is presented as the object in the generalisation set.
Train.: ***shark***
Gen.: *A chief heard the **shark**.*

D.1.8. Primitive noun → Object (proper noun):

(ID: `primn_to_obj_proper`) Proper noun entity that only appears as a standalone primitive in training, and is presented as the object in the generalisation set.
Train.: ***Paula***
Gen.: *The child helped **Paula**.*

D.1.9. Primitive verb → Infinitival argument:

(ID: `prim_to_inf_arg`) Verb that only appears as a stand-alone primitive in training, and is presented in an infinitival open clausal complement in the generalisation set.
Train.: ***crawl***
Gen.: *A baby planned to **crawl**.*

D.1.10. Active → Passive:

(ID: `active_to_passive`) Verbs used in active constructions during training appear in passive constructions in the generalisation set.
Train.: *The crocodile **blessed** William.*
Gen.: *A muffin **was blessed**.*

D.1.11. Passive → Active:

(ID: `passive_to_active`) Verbs used in passive constructions during training appear in active constructions in the generalisation set.
Train.: *The book **was squeezed**.*
Gen.: *The girl **squeezed** the strawberry.*

D.1.12. Object-omitted transitive → Transitive:

(ID: `obj_omitted_transitive _to_transitive`) Verbs seen with omitted objects during training appear with explicit objects in the generalisation set.
Train.: *Emily **baked**.*
Gen.: *The giraffe **baked a cake**.*

D.1.13. Unaccusative → Transitive:

(ID: `unacc_to_transitive`) Verbs displaying causative alternation are used in their intransitive form during training and in their transitive form in the generalisation set.
Train.: *The glass **shattered**.*
Gen.: *Liam **shattered** the jigsaw.*

D.1.14. Double object dative → PP dative:

(ID: `do_dative_to_pp_dative`) Verbs displaying dative alternation are shown in their double object dative form during training and in prepositional dative form in the generalisation set.
Train.: *The girl **teleported Liam the cookie**.*
Gen.: *Benjamin **teleported the cake to Isabella**.*

D.1.15. PP dative → Double Object Dative:

(ID: pp_dative_to_do_dative) Verbs displaying dative alternation are shown in their prepositional dative form during training and in double object dative form in the generalisation set.
Train.: *Jane **shipped the cake to John**.*
Gen.: *Jane **shipped John the cake**.*

D.1.16. Agent NP → Unaccusative subject:

(ID: only_seen_as_transitive_subj_as_unacc_subj) Entities seen as agentive subjects in training are presented as theme subjects in non-causative predicates in the generalisation set.
Train.: *The **cobra** helped a dog.*
Gen.: *The **cobra** froze.*

D.1.17. Theme NP → Object-omitted transitive subject:

(ID: only_seen_as_unacc_subj_as_obj_omitted_transitive_subj) Entities seen as theme subjects in non-causative predicates in training are presented as agentive subjects of transitive predicates without an explicit theme in the generalisation set.
Train.: *The **hippo** decomposed.*
Gen.: *The **hippo** painted.*

D.1.18. Theme NP → Unergative subject:

(ID: only_seen_as_unacc_subj_as_ unerg_subj) Entities seen as theme subjects in non-causative predicates in training are presented as agentive subjects of intransitive predicates in the generalisation set.
Train.: *The **hippo** decomposed.*
Gen.: *The **hippo** giggled.*

D.2. COGS: Structural Recombination Categories

D.2.1. Object modification → Subject modification:

(ID: obj_pp_to_subj_pp) Object noun phrases containing prepositional phrases in training appear instead as subject noun phrases with similar PPs in the generalisation set.
Train.: *Noah ate the **cake on the plate**.*
Gen.: *The **cake on the table** burned.*

D.2.2. Depth generalization: Sentential complements:

(ID: cp_recursion) Sentences displaying deeper nested sentential complements in the generalisation set than in the training set.
Train.: *Emma said that Noah knew that the cat danced.*
Gen.: *Emma said that Noah knew that Lucas saw that the cat danced.*

D.2.3. Depth generalization: PP modifiers:

(ID: pp_recursion) Sentences displaying deeper nested prepositional phrases in the generalisation set than in the training set.
Train.: *Ava saw the ball **in the bottle on the table**.*
Gen.: *Ava saw the ball **in the bottle on the table on the floor**.*

D.3. SLOG

D.3.1. Prepositional phrases (PP) max depth 4 → depth 5–12

(ID: PP_5-12) Prepositional phrase nesting seen to a depth of 4 in training is extended to depths 5–12 in the generalisation set.
Train.: *Ava saw the ball **in the bottle on the table**.*
Gen.: *Ava saw the **cat in the box on the mat on the bed on the floor in the room**.*

D.3.2. Tail CP recursion max depth 4 → depth 5–12

(ID: CP_5-12) Recursive sentential complements (CPs) nested to a depth of 4 in training are extended to 5–12 in the generalisation set.
Train.: *Ava believed that Emma said that a fish froze.*
Gen.: *Ava said that Emma liked that Max believed that Noah found that Liam saw that the cat slept.*

D.3.3. Center embedding max depth 4 → depth 5–12

(ID: center_embed_5-12) Center-embedded relative clauses nested to a depth of 4 in training are extended to 5–12 in the generalisation set.
Train.: *Eva saw the **cat that the horse that the dog liked** chased.*
Gen.: *Ava held the dress that a store that a girl that a boy that a cat that a man drew saw loved knew sold.*

D.3.4. PP max depth 4 → depth 3

(ID: PP_3) Prepositional phrase nesting seen to a depth of 4 in training is reduced to depth 3 in the generalisation set.
Train.: *Emma saw the **ball in the bottle on the table on the floor in the office**.*
Gen.: *Ava saw the **cat on the mat on the floor in the office**.*

D.3.5. Tail CP recursion max depth 4 → depth 3

(ID: `CP_3`) Tail-recursive CPs trained at depth 4, tested at shallower depth 3.
Train.: *Ava believed that Emma said that Max found that a cat saw that a fish froze.*
Gen.: *Ava said that Emma liked that Max believed that the cat slept.*

D.3.6. Center embedding max depth 4 → depth 3

(ID: `center_embed_3`) Center embedded RCs trained at a depth of 4 are presented at a depth of 3 in the generalisation set.
Train.: *Eva saw the cat that the horse that the dog that the man that the girl loved found* liked chased.
Gen.: *Emma bought the dress that the store that the woman that Mike knew liked sold.*

D.3.7. PP in direct object NPs → PP in subject NPs

(ID: `PP_modif_subj`) Prepositional phrase modifiers that appear in direct object noun phrases during training are presented in subject noun phrases in the generalisation set.
Train.: *Noah ate the **cake on the plate**.*
Gen.: *The **cake on the table** burned.*

D.3.8. PP in direct object NPs → PP in indirect object NPs

(ID: `PP_modif_iobj`) Prepositional phrase modifiers that appear in direct objects during training are presented in indirect objects in the generalisation set.
Train.: *Noah ate the **cake on the plate**.*
Gen.: *Max gave a fish to a **cat on a table**.*

D.3.9. RC in direct object NPs → RC in subject NPs

(ID: `RC_modif_subj`) Relative clauses modifying direct objects during training are presented modifying subject noun phrases.
Train.: *Noah saw the **cat that froze**.*
Gen.: *The **cat that froze** smiled.*

D.3.10. RC in direct object NPs → RC in indirect object NPs

(ID: `RC_modif_iobj`) Relative clauses modifying direct objects during training are presented modifying indirect object noun phrases.
Train.: *Noah saw the **cat that froze**.*
Gen.: *Max gave a fish to a **cat that ran**.*

D.3.11. Subject, direct object-extracted RC → Indirect object-extracted RC

(ID: `RC_iobj_extracted`) Relative clauses introduce subjects and direct objects in training, the generalisation set presents relative clauses introducing indirect objects.
Train.: *Noah saw the cat that gave a fish to Liam*. | Noah saw the cat that Liam liked _.
Gen.: *Noah saw the cat that Emma gave a cake to _*.

D.3.12. Subject, direct object *wh*-questions → Indirect object *wh*-questions

(ID: `Q_iobj_ditransV`) Generalisation from subject *wh*-questions to ones interrogating indirect object arguments.
Train.: *Who saw the cat?* | *What did Emma see _?*
Gen.: *Who did Noah give the cake to _?*

D.3.13. Subject, object *wh*-q of simple transitives → Active subject *wh*-questions

(ID: `Q_subj_active`) Transitive predicates seen in training in non-interrogative contexts are presented in the generalisation set as part of *wh*-questions interrogating the active subject.
Train.: ***Who saw** the cat?* | *Emma **wanted** to sleep.*
Gen.: ***Who wanted** to sleep?*

D.3.14. Subject, object *wh*-q of simple transitives → Passive subject *wh*-questions

(ID: `Q_subj_passive`) Transitive predicates seen in training in non-interrogative contexts are presented in the generalisation set as part of *wh*-questions interrogating the passive subject.
Train.: ***Who** did Emma see _?* | *The boy **was found** by Emma*
Gen.: ***Who was helped** by Emma?*

D.3.15. Subject, object *wh*-q of simple transitives → Direct object *wh*-questions with ditransitive verbs

(ID: `Q_dobj_ditransV`) Generalisation from transitive *wh*-questions in training to ditransitive ones in the generalisation set.
Train.: ***What** did Emma **see** _?* | *Emma **gave** a fish **to** the cat*
Gen.: ***What** did Emma **give** _ to the cat?*

D.3.16. Subject, object *wh*-q of simple transitives → *Wh*-questions with modified NPs

(ID: `Q_modified_NPs`) Generalisation from *wh*-questions with simple nouns in training to ones with modifying PPs in the generalisation set.
Train.: ***What*** *did* ***the cat*** *see _?* | *the cat* ***on the mat***
Gen.: *What did* ***the cat on a table*** *see _?*

D.3.17. Subject, object *wh*-q of simple transitives → *Wh*-questions long movement

(ID: `Q_long_mv`) *Wh*-dependencies span longer distances than in training, testing generalisation to long-distance movement.
Train.: ***What*** *did the cat* ***see*** *_?* | *Emma* ***said that*** *the cat saw a fish.*
Gen.: ***What*** *did Emma* ***say that*** *the cat* ***found*** *_?*

Can LLMs Simulate Human Behavioral Variability? A Case Study in the Phonemic Fluency Task

Mengyang Qiu[1] **Zoe Brisebois**[2] **Siena Sun**[1]
[1]Department of Speech-Language Pathology, Saint Elizabeth University, United States
[2]Department of Psychology, Trent University, Canada
{mqiu, ssun}@steu.edu zoebrisebois@trentu.ca

Abstract

Large language models (LLMs) are increasingly explored as substitutes for human participants in cognitive tasks, but their ability to simulate human behavioral variability remains unclear. This study examines whether LLMs can approximate individual differences in the phonemic fluency task, where participants generate words beginning with a target letter. We evaluated 34 distinct models across 45 configurations from major closed-source and open-source providers, and compared outputs to responses from 106 human participants. While some models, especially Claude 3.7 Sonnet, approximated human averages and lexical preferences, none reproduced the scope of human variability. LLM outputs were consistently less diverse, with newer models and thinking-enabled modes often reducing rather than increasing variability. Network analysis further revealed fundamental differences in retrieval structure between humans and the most human-like model. Ensemble simulations combining outputs from diverse models also failed to recover human-level diversity, likely due to high vocabulary overlap across models. These results, together with converging evidence from different cognitive and linguistic tasks, highlight key limitations in using LLMs to simulate human behavior.

Keywords: large language models, phonemic fluency, behavioral variability, cognitive simulation

1. Introduction

Large language models (LLMs) have rapidly advanced in recent years, achieving impressive performance across a wide range of natural language tasks. As a result, researchers have become increasingly interested in using LLMs as experimental tools in cognitive and behavioral science. Some even propose that LLMs could replace human participants in certain studies, offering scalable and efficient alternatives for simulating human behavior (Dillion et al., 2023).

This idea has intuitive appeal: LLMs can generate fluent responses on demand, and arguably encode the "wisdom of the crowd" from their massive training data (Trott, 2024). For instance, Hansen and Hebart (2022) found that GPT-3 could generate semantic features for concepts that not only mirrored the distribution of human-generated features but also matched human norms in their ability to predict similarity, relatedness, and category membership.[1]

On the other hand, a growing body of work has pointed out that LLMs may lack a crucial feature of human language and cognition: variability. Cuskley et al. (2024) argue that LLMs are limited by their training on written language, which captures only a narrow slice of human communicative behavior. Zanotto and Aroyehun (2024) offer empirical support, showing that LLM-generated texts exhibit substantially lower linguistic diversity than human-written texts across a range of features, including syntax, vocabulary, and style.

This lack of variability is also evident when examining how LLMs perform in tasks designed to probe semantic or associative structures. In a semantic fluency task where participants named as many animals as possible, Wang et al. (2025b) found that LLM-generated semantic networks were structurally different from those of humans. Compared to human networks, those of LLMs exhibited weaker local associations, poorer global integration, and greater rigidity in semantic organization. Similarly, Haim et al. (2025) examined word association networks in a STEM-related mindset task and found that GPT-3.5 produced networks that were notably sparser and less interconnected than those generated by human participants.

It is worth noting that Wang et al. (2025b) recognized this limitation and attempted to address it by prompting LLMs to role-play 30 different occupations, effectively simulating multiple distinct participants. However, the resulting semantic networks still failed to match the flexibility and associative structure of human data. Crucially, real human participants, even those drawn from relatively ho-

[1]However, recent work has also revealed distributional misalignments between human and language model productions. For example, Jacobs et al. (2024) found that language models systematically misranked human responses and produced poorly aligned response distributions in a cloze completion task, where participants produce a single word to complete a given sentence context.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 250–263
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

mogeneous groups such as college students, consistently exhibit meaningful individual differences. These differences may reflect not only variation in the amount, type, and organization of semantic knowledge, but also differences in executive function, and strategy use. Of course, because LLMs are optimized to encode and reproduce stable semantic relationships, consistency across prompts may reflect a design feature. From this perspective, the internal coherence and semantic regularity that make LLMs so effective at language modeling may also constrain their ability to simulate the variability that characterizes human cognition.

If semantic-based tasks limit LLMs' ability to simulate human behavioral variability, it raises the question: how would they perform on a task that relies less on meaning? One such task is phonemic (or letter) fluency, where individuals are asked to produce as many words as possible that begin with a particular letter, such as *F*, within a fixed time limit. Unlike semantic fluency, this task requires an effortful search through the mental lexicon based on orthographic or phonological form—an unnatural retrieval strategy in everyday language use. Prior analyses have shown that phonemic fluency performance is influenced by both lexical properties, such as word frequency, age of acquisition, and phonetic similarity, and cognitive factors like switching patterns (Cho et al., 2021).

The goal of the present study is to explore whether LLMs can simulate inter-participant variability in phonemic fluency, or whether their outputs remain rigid, perhaps driven by frequency-based or alphabetic heuristics. Unlike previous work that introduced arbitrary personas such as fictional occupations (Wang et al., 2025b), we ground our simulations in real participant metadata. Extensive research in psychology has consistently shown that age and education are among the most reliable predictors of verbal fluency performance (Tombaugh et al., 1999; Gladsjo et al., 1999; Villalobos et al., 2023). By providing LLMs with demographic and performance information drawn directly from real human participants, the simulation can be more meaningfully compared against actual human data.

Furthermore, previous studies comparing LLM and human cognitive and language performance have typically relied on one or a small number of models. The present study offers a substantially more comprehensive evaluation, testing 34 distinct models across 45 configurations from major closed-source and open-source providers. This broad coverage allows us to identify patterns that hold across model families and architectures, rather than drawing conclusions from a single system.

It also enables us to address a fundamental concern about LLM-based simulation: using the same model to simulate multiple participants means those simulated individuals inherently share the same foundational knowledge, making it difficult to generate authentically varied outputs regardless of prompt manipulation (Wang et al., 2025a). To overcome this limitation, we investigate whether combining outputs from diverse models through ensemble sampling—where each simulated participant's response is drawn from a randomly selected model—can better approximate human-level diversity, leveraging architectural and training differences across models to collectively recover variability that no single model captures alone.

In preview, our findings suggest a clear answer: LLMs are highly capable of producing fluent and correct responses in the phonemic fluency task, but none of them—individually or in ensemble—match the full scope of human behavioral variability. The remainder of the paper is organized as follows. Section 2 describes the human *F* fluency dataset and the LLM simulation procedure. Section 3 evaluates participant-level response counts and lexical diversity. Section 4 presents item-level analyses of production frequency distributions and linguistic predictors. Section 5 compares word co-occurrence network structures. Section 6 examines ensemble-based diversity. Section 7 concludes with a discussion of implications and future directions.

2. Method

2.1. Human data source

We used phonemic fluency data collected by Qiu and Johns (2021)[2], who investigated noun- and verb-based semantic fluency across two experiments (see also Qiu et al., 2021). In both experiments, participants first completed a phonemic fluency task using the letter *F* as a familiarization trial before proceeding to semantic tasks. A total of 106 native English speakers (age: $M = 35.59, SD = 10.04$; education: $M = 14.92$ years, $SD = 2.01$) were recruited via Amazon Mechanical Turk and completed the tasks online. Participants were instructed to produce as many words as possible beginning with the letter *F* within one minute, excluding proper nouns, numbers, and same-lemma variants. Audio recordings were collected for all participants.

Following the original study's procedure, we used the Google Speech-to-Text API[3] to transcribe the recordings, followed by manual review to exclude errors such as repetitions, out-of-category responses, same-lemma variants (e.g., *fish*, *fishing*,

[2]The verbal fluency data are publicly available on the Open Science Framework under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

[3]https://cloud.google.com/speech-to-text

fished), and counting behavior. After corrections, participants produced an average of 16.89 words ($SD = 4.84$).

2.2. LLM simulation procedure

We evaluated a diverse set of closed-source and open-source LLMs. For closed-source models, we used the official APIs from OpenAI[4], Anthropic[5], Google[6], and xAI[7]. For open-source models, we accessed models from Meta, Moonshot, Alibaba, Zhipu AI, DeepSeek, and Mistral via the Together.ai API[8].

All models received the same full prompt, which included the participant's age, education level, and number of correct responses, along with task instructions identical to those given to human participants (see Appendix A for the complete prompt). The prompt instructed models to role-play as the corresponding human participant and generate words beginning with *F* under a one-minute constraint, outputting one word per line with no additional text. Each model was prompted independently 106 times, once per human participant, with each prompt constituting a separate API call with no memory of previous calls.

In total, we evaluated 34 distinct models. Given that many recent models support hybrid thinking or reasoning modes, we additionally tested these models with and without extended thinking enabled, and at different reasoning effort levels where applicable, resulting in 45 configurations.

3. Participant-Level Analysis

3.1. Number of responses

Human participants produced an average of 16.89 correct responses within the one-minute time constraint ($SD = 4.84$). To evaluate whether LLMs adhered to the performance constraints specified in the prompt, we computed the Mean Absolute Error (MAE) between LLM-generated and human response counts for each configuration. Configurations with MAE ≤ 1.69 (i.e., within 10% of the human mean) were considered to have successfully simulated human-like production rates.

Of the 45 configurations tested, 33 met this criterion, indicating that the majority of models generally adhered to the instructed constraints. The remaining 12 configurations significantly overproduced relative to human participants. Among these, OpenAI's GPT-4 Turbo (MAE $= 57.17$) and o3 (MAE $= 35.39$) showed the most extreme overproduction, generating outputs several times longer than human responses. Surprisingly, two of Anthropic's latest models, Claude 4.6 Sonnet (Thinking) (MAE $= 3.19$) and Claude 4.6 Opus (MAE $= 3.87$), also failed to adhere to the performance constraints despite the explicit inclusion of the number of correct responses in the prompt.

Among the 33 configurations with MAE ≤ 1.69, we further examined whether any individual simulations deviated substantially from the corresponding human participant's response count. Configurations containing any simulation where $|\text{LLM} - \text{Human}| > 5$ were flagged as having outliers. Of the 33 configurations, 21 produced no outliers across all 106 simulations and were retained for subsequent analyses. Details of included and excluded configurations are provided in Appendix B.

3.2. Variability of responses

To evaluate the variability of responses across human participants and LLM-simulated participants, we computed the Type-to-Token Ratio (TTR) and the Idiosyncratic Type-to-Total Type Ratio (ITTTR). TTR captures lexical diversity by dividing the number of unique words (types) by the total number of responses (tokens). A higher TTR indicates a wider range of vocabulary produced. ITTTR measures how many of those types were idiosyncratic (i.e., produced by only one participant) relative to the total number of types. This provides insight into how consistent or individualized the generated words are across participants within the group (Castro et al., 2021).

As shown in Table 1, human participants produced the highest number of unique types (476) and idiosyncratic types (201), resulting in a TTR of 0.27 and an ITTTR of 0.42. No LLM approached this level of variability. The most diverse LLM output came from Claude 3.7 Sonnet (226 types, TTR $= 0.13$, ITTTR $= 0.32$), followed by o3-mini (193 types, TTR $= 0.11$, ITTTR $= 0.30$). Even these best-performing models produced fewer than half the unique types observed in human data.

Several patterns emerged across model families. For Anthropic models, non-thinking configurations consistently produced more diverse outputs than their thinking-enabled counterparts (e.g., Claude 3.7 Sonnet: 226 vs. 137 types; Claude 4.5 Sonnet: 110 vs. 77 types; Claude 4 Sonnet: 74 vs. 73 types). Additionally, the earlier Claude 3.7 Sonnet substantially outperformed more recent Anthropic models in lexical diversity, suggesting that newer models may prioritize consistency over variability.

For OpenAI models, reasoning capacity appeared to benefit diversity. GPT-5.2 showed a clear gradient, with higher reasoning effort producing more unique types (High: 158, Medium: 149, Low:

[4] https://openai.com/api/
[5] https://www.anthropic.com/api
[6] https://ai.google.dev/gemini-api/docs
[7] https://x.ai/api
[8] https://www.together.ai

139). The most diverse OpenAI model, o3-mini (193 types), is itself a reasoning-focused model. Despite these differences across models, all LLM configurations remained well below human-level variability in both TTR and ITTTR, highlighting a persistent tendency toward output rigidity across architectures and providers.

Table 1: Type-to-token ratios (TTR) and idiosyncratic type-to-total type ratios (ITTTR) across human participants and LLM-simulated participants

Model	Types	Idio. Types	Tokens	TTR	ITTTR
Human	**476**	**201**	**1790**	**0.27**	**0.42**
OpenAI					
o3-mini	193	58	1790	0.11	0.30
GPT-5.2 (High Reasoning)	158	40	1790	0.09	0.25
GPT-5.2 (Medium Reasoning)	149	38	1790	0.08	0.26
GPT-3.5 Turbo	148	48	1783	0.08	0.32
GPT-5.2 (Low Reasoning)	139	35	1790	0.08	0.25
GPT-5	130	33	1790	0.07	0.25
GPT-4o	128	38	1773	0.07	0.30
GPT-5 mini	95	18	1790	0.05	0.19
GPT-4.1	73	20	1790	0.04	0.27
Anthropic					
Claude 3.7 Sonnet	226	73	1781	0.13	0.32
Claude 3.7 Sonnet (Thinking)	137	43	1788	0.08	0.31
Claude 4.5 Sonnet	110	32	1784	0.06	0.29
Claude 4.5 Sonnet (Thinking)	77	18	1790	0.04	0.23
Claude 4 Sonnet	74	12	1795	0.04	0.16
Claude 4 Sonnet (Thinking)	73	18	1789	0.04	0.25
Google					
Gemini 2.5 Flash	149	51	1794	0.08	0.34
Gemini 2.5 Pro	83	20	1790	0.05	0.24
Gemini 3 Pro	73	20	1790	0.04	0.27
xAI					
Grok-4.1-fast (Reasoning)	62	8	1787	0.03	0.13
Open-Source Models					
Kimi K2	96	25	1787	0.05	0.26
Qwen3-2507	93	22	1781	0.05	0.24

4. Item-Level Analysis

4.1. Distribution of production frequency

Previous studies have observed that word production in verbal fluency tasks typically follows Zipf's law, a type of power-law relationship where word frequency is inversely proportional to its rank, expressed as $f(r) \propto 1/r^{\alpha}$, where $f(r)$ denotes the frequency of a word at rank r and α is the scaling exponent (e.g., Taler et al., 2020). This distribution is characterized by a small number of words that are produced frequently across participants, followed by a "long tail" of words that are produced infrequently, often appearing only once or twice in the entire dataset.

To estimate α, we performed linear regression on the log-transformed rank and frequency values; the slope of the fitted line provides the α estimate, and R^2 indicates goodness of fit. A larger α reflects a steeper decline, indicating that top-ranked words dominate more heavily with less diversity in the long tail. Only model configurations that produced at least 100 unique types were included in this analysis, resulting in 11 configurations alongside the human data (Figure 1).

Human participants' word production closely followed the expected Zipfian pattern, with $\alpha = 0.89$ and $R^2 = 0.92$. The most frequently produced word was "fun", mentioned 34 times, followed by "fire" (30 times) and "fan" (28 times). In contrast, LLMs demonstrated a notably different pattern. While all models showed adequate fit to a power-law distribution (R^2 ranging from 0.86 to 0.92), their consistently higher α coefficients (1.19–1.53) reflect steeper declines in frequency, with top-ranked words dominating the distribution more heavily than in human data. Among the LLMs, Claude 3.7 Sonnet exhibited the closest α to the human distribution ($\alpha = 1.19$, $R^2 = 0.89$), though still notably steeper than the human pattern.

4.2. Linguistic variables and production frequency

Previous research has shown that verbal fluency performance is influenced by multiple linguistic variables, including word frequency, familiarity, word duration, and age of acquisition (Cho et al., 2021; Taler et al., 2020). In this study, we examined the relationship between production frequency and core linguistic variables obtained from the English Lexicon Project (Balota et al., 2007), including word length (number of letters), word frequency (log frequency per million words from SUBTLEX corpus), orthographic neighborhood size (number of words that differ by one letter), phonological neighborhood size (number of words that differ by one phoneme), semantic neighborhood size (number of semantically related words), and age of acquisition (estimated age when a word is typically learned). As with the Zipf analysis, only model configurations with at least 100 unique types were included, resulting in 11 configurations alongside human data. We first conducted simple correlation analyses between these variables and production frequency.

As shown in Table 2, all linguistic variables except for semantic neighborhood size were significantly correlated with production frequency in human data. Word frequency showed the strongest positive correlation ($r = 0.55$), followed by age of acquisition ($r = -0.52$) and orthographic neighborhood size ($r = 0.43$). Among the LLMs, Claude 3.7 Sonnet most closely approximated the human pattern, with highly similar correlation coefficients across all variables. o3-mini followed a similar trend but with generally weaker correlations, particularly for orthographic neighborhood size. The remaining models showed progressively weaker alignment, with many losing significance for neighborhood variables while retaining word frequency

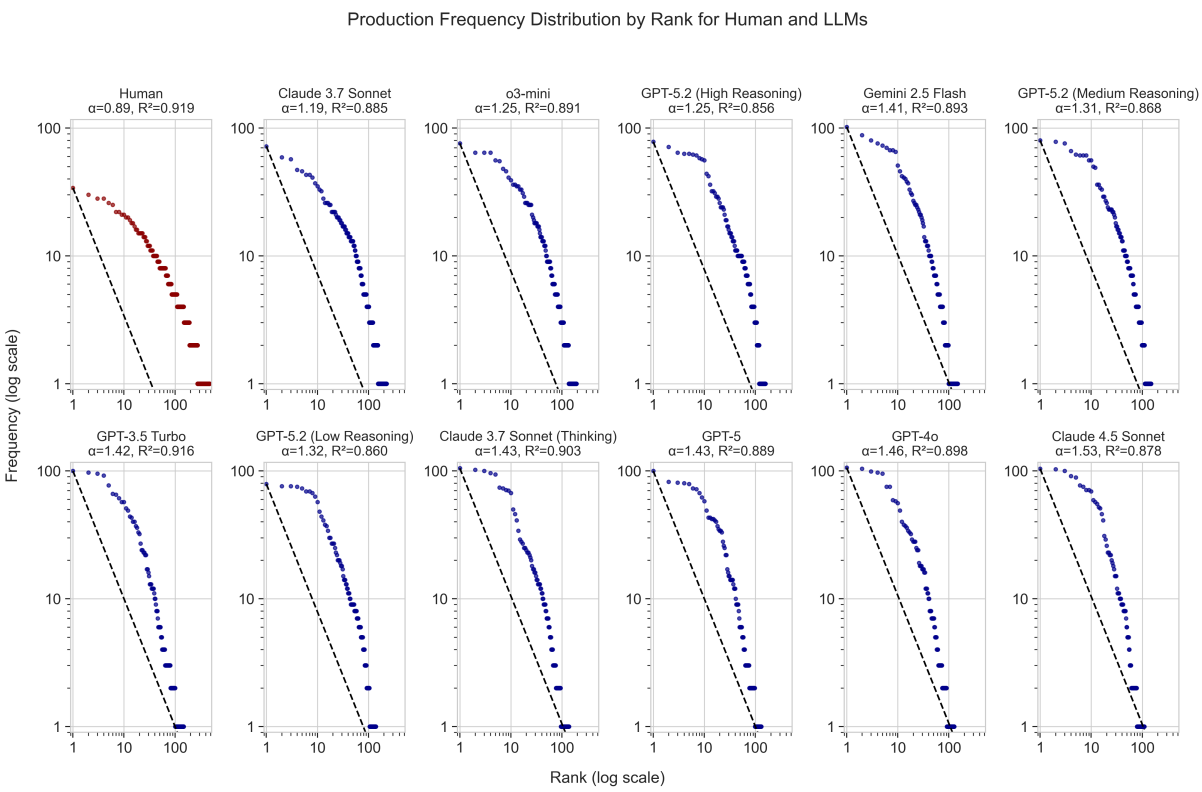


Figure 1: Production frequency distribution by rank for human participants (red) and 11 LLM configurations (blue) with at least 100 unique types. Both axes are log-scaled. Dashed lines represent fitted power-law models, with α and R^2 shown for each panel.

and age of acquisition as reliable predictors. Semantic neighborhood size remained non-significant across all models, consistent with the phonemic nature of the task.

To formally establish the factors that influence production frequency, we conducted stepwise regression analyses with all linguistic variables. Full statistical details are provided in Appendix C, but a consistent pattern emerged: linguistic variables explained the most variance in production frequency for human responses (adjusted $R^2 = .396$), followed closely by Claude 3.7 Sonnet (adjusted $R^2 = .356$), with substantially less variance explained for o3-mini (adjusted $R^2 = .179$). Claude 3.7 Sonnet most closely matched the human pattern in both variance explained and predictors retained, with word frequency and age of acquisition emerging as significant in both. Orthographic neighborhood size was uniquely important for humans, while word length played a stronger role in Claude 3.7 Sonnet and o3-mini.

5. Network Analysis

Building on our earlier findings, we focused the network analysis exclusively on comparing human and Claude 3.7 Sonnet outputs. Claude consistently exhibited the most human-like patterns–response count, variability, production frequency distribution, and linguistic predictors of word choice.

5.1. Network construction approach

To better understand the organizational structure of word production in letter *F* fluency, we constructed correlation-based networks—a common approach in cognitive science for examining the relationships between concepts (Borodkin et al., 2016; Li and Qiu, 2025; Siew and Guru, 2023). This approach is based on the premise that word co-occurrence patterns in fluency tasks reflect underlying associative relationships: the tendency to produce word b given that word a was also produced indicates a shared associative link between those items. By aggregating these co-occurrence patterns across participants, it is possible to estimate a network structure where nodes represent words and edges capture the strength of association (Borodkin et al., 2016).

Given the different number of nodes in the full networks (476 words for humans, 226 for Claude), direct comparison of network metrics may be confounded by size differences (Borodkin et al., 2016). Therefore, we selected 197 words that were produced by both humans and Claude to ensure a fairer comparison.

Table 2: Correlations between linguistic variables and production frequency across models

Model	Length	WF	Ortho_N	Phono_N	Sem_N	AoA
Human	-0.39^{***}	0.55^{***}	0.43^{***}	0.38^{**}	0.01	-0.52^{***}
Claude 3.7 Sonnet	-0.34^{***}	0.54^{***}	0.33^{***}	0.31^{***}	-0.02	-0.49^{***}
o3-mini	-0.29^{***}	0.40^{***}	0.24^{**}	0.36^{***}	-0.004	-0.36^{***}
GPT-5.2 (High Reasoning)	-0.05	0.31^{***}	-0.06	-0.01	-0.03	-0.37^{***}
Gemini 2.5 Flash	-0.18^{*}	0.43^{***}	0.17^{*}	0.16^{*}	-0.03	-0.51^{***}
GPT-5.2 (Medium Reasoning)	-0.07	0.31^{***}	-0.05	-0.04	-0.04	-0.39^{***}
GPT-3.5 Turbo	-0.16	0.30^{***}	0.16	0.14	0.08	-0.44^{***}
GPT-5.2 (Low Reasoning)	0.003	0.25^{***}	-0.10	-0.12	-0.04	-0.32^{***}
Claude 3.7 Sonnet (Thinking)	-0.23^{**}	0.43^{***}	0.23^{**}	0.20^{*}	-0.10	-0.43^{***}
GPT-5	-0.09	0.30^{***}	-0.01	-0.07	0.04	-0.48^{***}
GPT-4o	-0.04	0.03	-0.05	-0.05	0.10	-0.27^{**}
Claude 4.5 Sonnet	-0.15	0.35^{***}	0.12	0.09	-0.05	-0.50^{***}

Note. Length = word length; WF = log word frequency (SUBTLEX); Ortho_N = orthographic neighborhood size; Phono_N = phonological neighborhood size; Sem_N = semantic neighborhood size; AoA = age of acquisition.
$^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$

The network construction involved three steps: (1) creating binary participant-by-word matrices indicating which words each participant produced, (2) computing word-word cosine similarity matrices based on co-occurrence patterns across participants, and (3) applying the Triangulated Maximally Filtered Graph (TMFG; Massara et al., 2017) algorithm to retain only the strongest associations while preserving network connectivity.

5.2. Structural comparison of human and Claude networks

We compared the structural properties of networks constructed from human and Claude 3.7 Sonnet outputs using two standard global metrics: clustering coefficient (CC) and average shortest path length (ASPL). CC quantifies the degree to which nodes in a network tend to cluster together (i.e., the likelihood that two neighbors of a node are also connected), while ASPL reflects the average number of steps required to connect any two nodes, indexing the global efficiency of the network (Siew et al., 2019).

Both networks were constructed from the same set of 197 shared words to control for network size. The human network showed a higher CC (0.42) and longer ASPL (5.32) compared to Claude (CC $= 0.37$, ASPL $= 4.40$). To assess whether the observed structures differed from random expectations, we simulated 1,000 Erdős-Rényi random networks (Erdős and Rényi, 1960) matched in size and density. One-sample z-tests confirmed that both human and Claude networks significantly deviated from randomness, exhibiting substantially higher CCs and slightly higher ASPLs than random networks (human: $z = -2460.1$ for CC, $z = -3765.1$ for ASPL; Claude: $z = -2105.2$ for CC, $z = -2410.6$ for ASPL; all $p < .001$).

To statistically compare the two networks, we conducted a subnetwork bootstrap procedure: 1,000 partial networks were generated for each system by randomly sampling 50% of nodes and recalculating network metrics (Borodkin et al., 2016; Qiu et al., 2021). As shown in Figure 2, human subnetworks had significantly higher CCs ($t = 19.71$, $p < .001$) and longer ASPLs ($t = 11.48$, $p < .001$) than Claude subnetworks. These differences indicate that human responses are organized into tighter local clusters with weaker global integration, whereas Claude responses form a more evenly connected network with greater global efficiency.

To complement the network analysis, we conducted a representational similarity analysis (RSA; Nili et al., 2014) comparing the original pairwise similarity matrices that underlie network construction. The Spearman correlation between the lower triangular portions of the human and Claude matrices was modest ($\rho = .22$, $p < .001$), providing additional evidence that Claude organizes or retrieves words differently than humans. Taken together with the network metric results, this suggests that while Claude can produce human-like words and respond to similar lexical variables, the global structure and retrieval dynamics it exhibits remain meaningfully distinct from human performance.

6. LLM Ensemble Analysis

The preceding analyses demonstrate that no single LLM configuration captures the full scope of human behavioral variability. A natural question is whether combining outputs from multiple models—each potentially drawing on different architectures, training data, and alignment strategies—could better approximate the distributional diversity observed in human responses. While recent ensemble methods in NLP have primarily focused on improving task accuracy through techniques such as majority

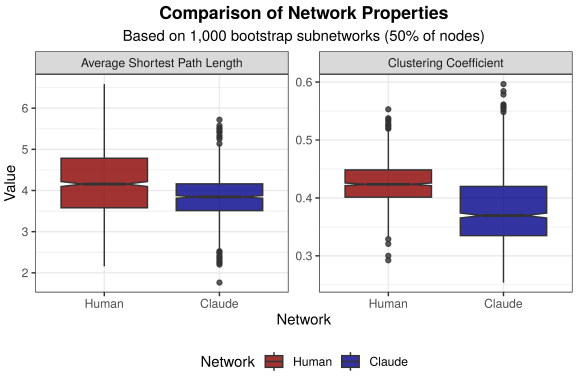


Figure 2: Clustering coefficient and average shortest path length for human and Claude networks based on 1,000 bootstrap subnetworks (50% nodes)

voting (Li et al., 2024), our goal is the opposite: to explore whether ensembling can simulate the distributional variability characteristic of human behavior.

6.1. Ensemble construction

To construct ensemble outputs, we used a participant-level random sampling procedure. For each of the 106 human participants, we assembled a pool of eligible LLM outputs from all 33 configurations with MAE ≤ 1.69, excluding any outlier simulations for that participant (see Appendix B). This resulted in 29 to 33 available LLM outputs per participant, depending on how many configurations produced outliers for that individual.

For each ensemble simulation, we independently and randomly selected one LLM output per participant from their available pool, producing 106 simulated participants from a mixture of models. This procedure was repeated $1{,}000$ times, yielding $1{,}000$ ensemble samples.

6.2. Ensemble results

Figure 3 shows the distributions of variability metrics across the $1{,}000$ ensemble simulations, with human values indicated by the red vertical lines. Across all metrics, the ensemble distributions fell far short of human-level variability. The mean number of unique types across simulations was 179.13 ($SD = 9.82$, range: 143–214), compared to 476 for human participants. Similarly, mean TTR was 0.100 ($SD = 0.005$, range: 0.080–0.119), well below the human TTR of 0.27. The same pattern held for idiosyncratic types and ITTTR.

Notably, the ensemble simulations did not substantially exceed the diversity of the best-performing individual model. Claude 3.7 Sonnet alone produced 226 unique types with a TTR of 0.13—values that fall within or near the upper tail of the ensemble distribution for types but above the ensemble range for TTR.

The Zipf analysis of ensemble outputs reinforced this conclusion. Across $1{,}000$ simulations, the mean α coefficient was 1.29 ($SD = 0.03$, range: 1.18–1.37), substantially steeper than the human value of $\alpha = 0.89$ and comparable to individual model values (e.g., Claude 3.7 Sonnet: $\alpha = 1.19$; o3-mini: $\alpha = 1.25$).

6.3. Vocabulary overlap across models

To understand why ensembling failed to increase diversity, we computed pairwise vocabulary overlap coefficients between all 21 models retained for participant-level analyses (Figure 4). The overlap coefficient between two vocabularies A and B is defined as:

$$\text{Overlap}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{1}$$

A value of 1.0 indicates that the smaller vocabulary is entirely contained within the larger one, while lower values indicate greater lexical divergence.

The overlap coefficient was high across the board (mean $= 0.74$, $SD = 0.11$). Within Anthropic models, overlap was especially pronounced (mean $= 0.90$), with some pairs reaching 0.99, indicating that switching between model versions barely changes the word pool. OpenAI models showed somewhat more internal diversity (mean $= 0.72$), though this was largely driven by GPT-3.5 Turbo, an older architecture with the lowest overlap with other OpenAI models.

Even across providers, roughly three-quarters of any model's vocabulary appeared in every other model. This pervasive lexical overlap explains why combining outputs from diverse models yields minimum gains in diversity: the models are drawing from a largely shared word pool, and ensembling amplifies high-frequency core vocabulary rather than expanding into the long tail of rare or idiosyncratic responses.

7. Discussion and Conclusion

This study evaluated whether LLMs can simulate the behavioral variability observed in human phonemic fluency performance. Across 34 models and 45 configurations, no model—individually or in ensemble—captured the full scope of human variability, despite many producing plausible average response counts and showing sensitivity to similar linguistic features.

A key contribution of this study is its comprehensive coverage of models. Our results reveal that LLMs do not perform equally well on this task, and

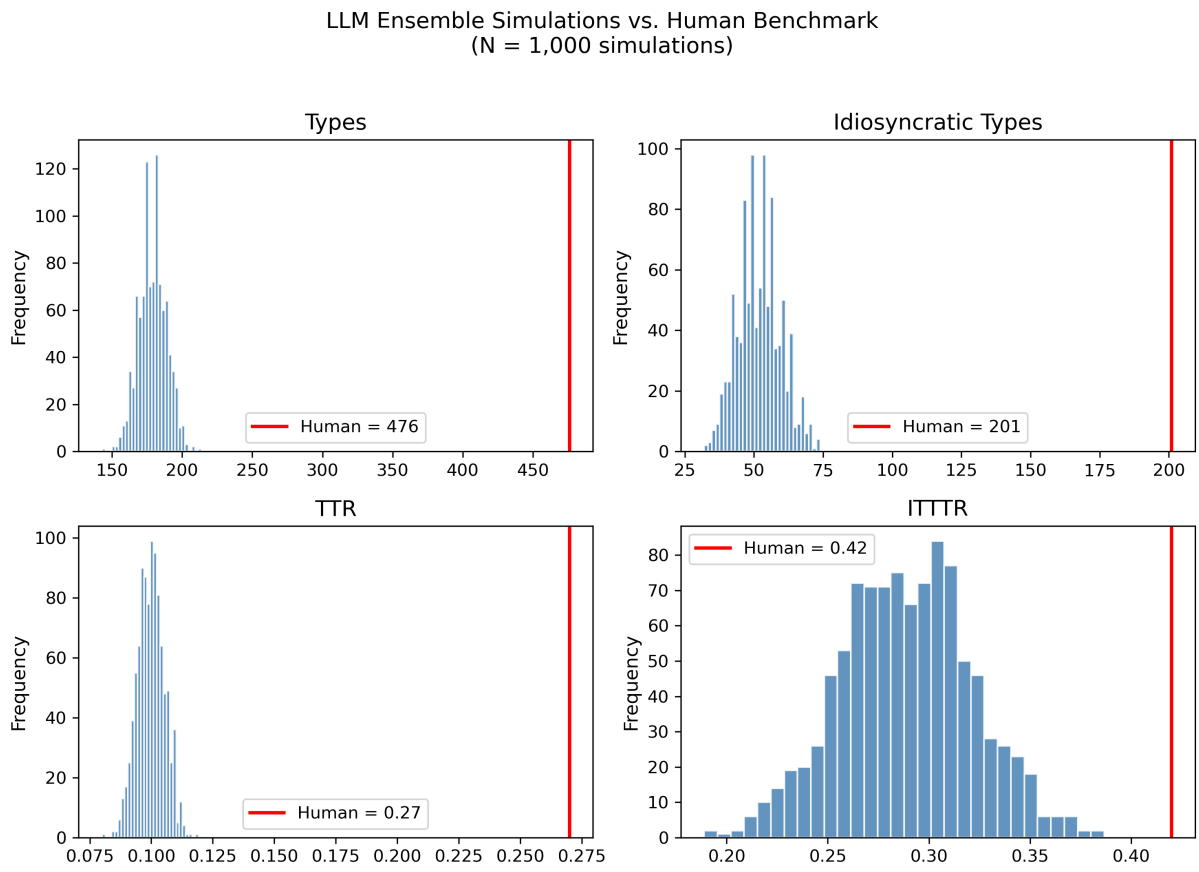


Figure 3: Distributions of variability metrics across $1{,}000$ LLM ensemble simulations. Red vertical lines indicate human values. All ensemble distributions fall substantially below human benchmarks.

importantly, newer models are not necessarily better. Within Anthropic, the earlier Claude 3.7 Sonnet substantially outperformed more recent models in lexical diversity, and enabling thinking mode consistently reduced rather than increased variability. These findings suggest that researchers conducting LLM-based simulations of human behavior should evaluate multiple models across different providers rather than defaulting to the latest release, as model selection can substantially influence the conclusions drawn.

Beyond differences in variability, our analyses also revealed divergent retrieval mechanisms. At the item level, orthographic neighborhood size played a unique role for human participants, whereas LLMs showed greater sensitivity to word length, suggesting that humans rely more on form-based associations during phonemic retrieval while LLMs prioritize more surface-level features. The network analyses reinforced this divergence: human responses exhibited stronger local clustering but weaker global integration, whereas Claude 3.7 Sonnet optimized for network-wide efficiency. Together, these structural mismatches highlight a core limitation—LLMs can produce plausible words but fail to approximate the associative dynamics of human memory search.

A fundamental concern motivating our ensemble analysis was that using the same model to simulate multiple participants means those simulated individuals share the same foundational knowledge, making it difficult to generate authentically varied outputs (Wang et al., 2025a). Our ensemble approach addressed this by drawing each simulated participant's response from a randomly selected model. Yet even this strategy failed: the pairwise vocabulary overlap analysis revealed that models across providers draw from a largely shared word pool, with within-family overlap reaching 0.90 for Anthropic models. This convergence likely reflects the fact that current LLMs are fundamentally trained on similar internet-scale data using transformer-based architectures, leading to shared biases that amplify high-frequency, prototypical responses at the expense of the rare and idiosyncratic words that characterize human individuality. This pattern raises broader concerns about representational bias: if LLMs systematically suppress low-frequency responses, they risk reinforcing dominant linguistic patterns while marginalizing the diversity inherent in real human populations (Guo et al., 2024).

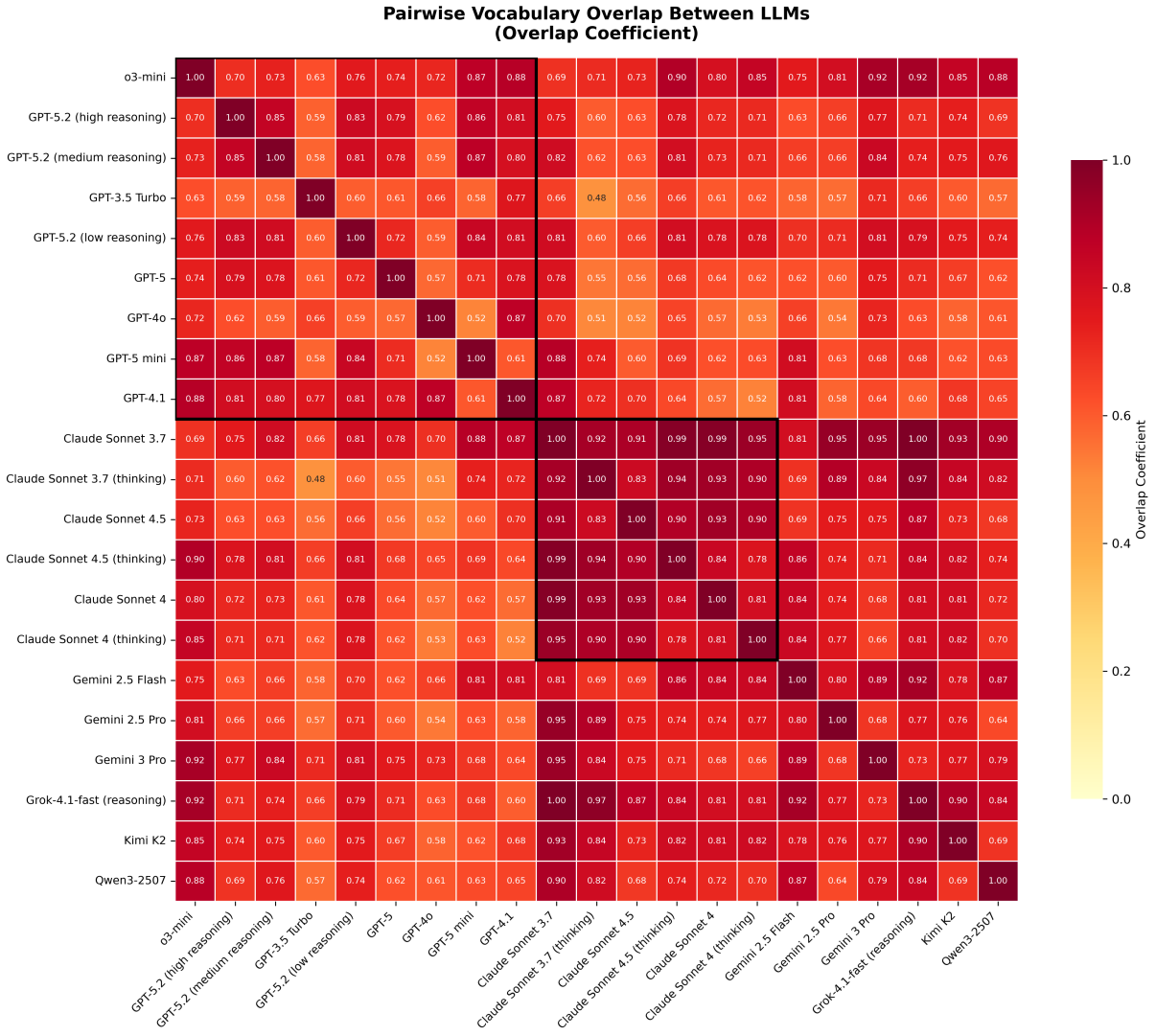


Figure 4: Pairwise vocabulary overlap coefficients between the 21 LLM configurations. Higher values (darker) indicate greater shared vocabulary.

Among all models tested, Claude 3.7 Sonnet consistently emerged as the most human-like, and performed far better than other configurations. Why this particular model outperforms others, including more recent models from the same provider, remains an open question. Future work could focus on extending this model's capabilities through more targeted persona modeling, potentially incorporating linguistic and cognitive profiles (e.g., selective reading exposure; Qiu et al., 2024) to better simulate participant-level differences. More broadly, rather than substituting LLM-simulated individuals for real human participants, a more productive approach may be to treat LLMs as a baseline system–one that captures central tendencies in linguistic knowledge–against which the uniqueness and flexibility of real human behavior can be measured.

While researchers would not typically use LLMs to elicit responses to a phonemic fluency task, the patterns observed in the current study reinforce a broader caution against using LLMs as substitutes for human participants in cognitive and behavioral tasks. Whether the task draws on semantic knowledge, as in animal fluency (Wang et al., 2025b), orthographic and phonological form, as in the present letter *F* fluency task, or contextual prediction integrating semantic and syntactic constraints, as in cloze production (Jacobs et al., 2024), LLM outputs consistently show reduced variability, more rigid response distributions, and divergent organizational structures compared to human data. These converging findings offer a strong case for developing new strategies if LLMs are to be used meaningfully in the simulation of human behavior.

Limitations

This study focuses on a single verbal fluency task (letter *F*) in English, limiting generalizability across

task types, letters, and languages. While phonemic fluency provides a stringent test of form-based retrieval, different letters may elicit distinct lexical or structural patterns. Similarly, results may differ in languages with different orthographic or phonological systems. Our analysis uses a single prompt design across model generations and thinking mode configurations; other prompting techniques (e.g., few-shot examples) or fine-tuned models may yield different patterns. Additionally, while we grounded our simulations in real participant metadata, age and education are not the only sources of individual variation in verbal fluency. Factors such as profession, domain expertise, and accumulated life experience also contribute to behavioral differences, and incorporating these richer individual-level traits may be necessary to drive more human-like variability. Future work should expand to multilingual and cross-task comparisons and evaluate whether alternative sampling or persona modeling strategies better capture human behavioral variability.

Ethical Considerations

The human data used in this study are drawn from Qiu and Johns (2021), a publicly available dataset under a CC-BY 4.0 license. The dataset includes anonymized verbal fluency responses alongside participant age and education information, with all personal identifiers removed. According to the original publication, all data collection procedures received institutional ethics approval, and participants provided informed consent. In our study, only de-identified demographic and performance information was used to construct LLM prompts.

Our findings also raise ethical concerns regarding the use of LLMs to simulate human behavioral variability. LLMs exhibit output rigidity, favoring high-frequency, prototypical responses and failing to capture natural human diversity. If used as substitutes for human participants without acknowledging this limitation, LLM-based simulations risk systematically underrepresenting the variability present in real populations, particularly marginalizing rare or idiosyncratic responses.

Bibliographical References

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods*, 39(3):445–459.

Katy Borodkin, Yoed N Kenett, Miriam Faust, and Nira Mashal. 2016. When pumpkin is closer to onion than to squash: The structure of the second language lexicon. *Cognition*, 156:60–70.

Nichol Castro, Taylor Curley, and Christopher Hertzog. 2021. Category norms with a cross-sectional sample of adults in the united states: Consideration of cohort, age, and historical effects on semantic categories. *Behavior Research Methods*, 53:898–917.

Sunghye Cho, Naomi Nevler, Natalia Parjane, Christopher Cieri, Mark Liberman, Murray Grossman, and Katheryn AQ Cousins. 2021. Automated analysis of digitized letter fluency data. *Frontiers in Psychology*, 12:654214.

Christine Cuskley, Rebecca Woods, and Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Paul Erdős and Alfréd Rényi. 1960. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Julie Akiko Gladsjo, Catherine C Schuman, Jovier D Evans, Guerry M Peavy, S Walden Miller, and Robert K Heaton. 1999. Norms for letter and category fluency: demographic corrections for age, education, and ethnicity. *Assessment*, 6(2):147–178.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *arXiv preprint arXiv:2411.10915*.

Edith Haim, Lars van den Bergh, Cynthia SQ Siew, Yoed N Kenett, Daniele Marinazzo, and Massimo Stella. 2025. Cognitive networks highlight differences and similarities in the stem mindsets of human and llm-simulated trainees, experts and academics. *arXiv*.

Hannes Hansen and Martin N Hebart. 2022. Semantic features of object concepts generated with gpt-3. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Cassandra L Jacobs, Loïc Grobol, and Alvin Tsang. 2024. Large-scale cloze evaluation reveals that token prediction tasks are neither lexically nor semantically aligned. *arXiv preprint arXiv:2410.12057*.

Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need. *arXiv*.

Yu Li and Mengyang Qiu. 2025. A network analysis of the semantic evolution of 'fruit' and 'stone' in Tibeto-Burman languages. *Poznan Studies in Contemporary Linguistics*, 61(2):121–172.

Guido Previde Massara, Tiziana Di Matteo, and Tomaso Aste. 2017. Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, 5(2):161–178.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4):e1003553.

Mengyang Qiu, Nichol Castro, and Brendan T. Johns. 2021. Structural comparisons of noun and verb networks in the mental lexicon. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, pages 1649–1655.

Mengyang Qiu, Nichol Castro, and Brendan T. Johns. 2024. Estimating type of print exposure across aging through author production. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 1175–1181.

Mengyang Qiu and Brendan T. Johns. 2021. A distributional and sensorimotor analysis of noun and verb fluency. *PsyArXiv*.

Cynthia SQ Siew and Anutra Guru. 2023. Investigating the network structure of domain-specific knowledge using the semantic fluency task. *Memory & Cognition*, 51(3):623–646.

Cynthia SQ Siew, Dirk U Wulff, Nicole M Beckage, and Yoed N Kenett. 2019. Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2019.

Vanessa Taler, Brendan T Johns, and Michael N Jones. 2020. A large-scale semantic analysis of verbal fluency across the aging spectrum: Data from the Canadian Longitudinal Study on Aging. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 75(9):e221–e230.

Tom N Tombaugh, Jean Kozak, and Laura Rees. 1999. Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology*, 14(2):167–177.

Sean Trott. 2024. Large language models and the wisdom of small crowds. *Open Mind*, 8:723–738.

Dolores Villalobos, Lucía Torres-Simón, Javier Pacios, Nuria Paul, and David Del Río. 2023. A systematic review of normative data for verbal fluency test in different languages. *Neuropsychology Review*, 33(4):733–764.

Qian Wang, Zhenheng Tang, and Bingsheng He. 2025a. Can llm simulations truly reflect humanity? a deep dive. In *ICLR Blogposts 2025*.

Ye Wang, Yaling Deng, Ge Wang, Tong Li, Hongjiang Xiao, and Yuan Zhang. 2025b. The fluency-based semantic network of LLMs differs from humans. *Computers in Human Behavior: Artificial Humans*, 3:100103.

Sergio E Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv*.

A. Prompt Design

Figure A1 shows the complete prompt used to simulate the *F* fluency task. The prompt included three pieces of participant-level information: age, education level, and number of correct responses.

Age and education were included because they are among the most reliable predictors of verbal fluency performance, consistently accounting for significant variance in both phonemic and semantic fluency across languages and populations (Tombaugh et al., 1999; Gladsjo et al., 1999; Villalobos et al., 2023). These variables were drawn directly from the human dataset rather than fabricated, ensuring ecological validity in the simulation.

In terms of the inclusion of the number of correct responses, Wang et al. (2025b) found that LLMs produced comparable numbers of animal names to human participants when prompted with only a one-minute time constraint and no performance information, suggesting that LLMs can approximate typical human production rates for semantically driven tasks. However, in our pilot testing of the phonemic fluency task, we found that prompting models with only age and education information led to significant overproduction. This likely reflects the greater difficulty of phonemic fluency for humans, who must perform an effortful, non-semantic search through the mental lexicon—a constraint that does not apply to LLMs. To ensure realistic response counts and enable meaningful comparisons at the item and network levels, we therefore also provided the number of correct responses as an additional constraint in the prompt.

```
prompt = (
    "In a verbal fluency task, you will be asked to say as many words as you can
        think of that conform to a specified criterion within 1 minute. "
    "Please DO NOT say words that are proper nouns (like Bob or Boston), numbers
        , or the same words with different endings (for example, love -> loves,
        lover, loving). Please DO NOT say phrases or sentences. Please DO NOT
        use a dictionary, internet, or other external help.\n\n"
    "Criterion: Words that begin with the letter F\n\n"
    "A human participant completed this task on Amazon Mechanical Turk.\n\n"
    "Here is their demographic and performance information:\n"
    f"Age: {age}\n"
    f"Highest degree: {education}\n"
    f"Number of correct responses: {num_correct}\n\n"
    "Now, please imagine that you are this participant. Your task is to generate
        as many words as possible that begin with the letter F, as if you were
        speaking aloud, and you have exactly one minute to do so. Respond in a
        way that reflects how this human participant might perform under this
        timed condition. "
    "Output only the words, one per line. Do not include any introductions,
        explanations, or extra text. If you add anything other than the words,
        you will be disqualified from the task."
)
```

Figure A1: Full prompt used to simulate the *F* fluency task, including participant age, education, and number of correct responses, with instructions identical to those given to human participants

B. Model Selection Criteria

Tables A1 and A2 summarize the performance of all 45 configurations in terms of adherence to the instructed number of responses. Mean Absolute Error (MAE) was used to evaluate each configuration across all 106 simulations, capturing the average magnitude of deviation between LLM-generated and human response counts regardless of direction:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i^{\text{LLM}} - y_i^{\text{Human}}| \tag{2}$$

where $N = 106$ is the number of participants and y_i is the number of correct responses for participant i. Configurations with MAE ≤ 1.69 (i.e., 10% of the human mean of 16.89) were considered to have successfully adhered to the performance constraints.

We additionally report Mean Bias Error (MBE), which captures the direction of deviation:

$$\text{MBE} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{\text{LLM}} - y_i^{\text{Human}}) \tag{3}$$

Positive MBE values indicate systematic overproduction relative to human participants, while negative values indicate underproduction. Because LLM-generated responses were subjected to the same error-checking criteria applied to human data—with repetitions being the most common error type when errors occurred, though overall error rates were very low—many models that closely followed the instructed number of responses tended to show slightly negative MBE values after error removal.

Table A1 also reports the number of outlier simulations per configuration, defined as cases where $|y_i^{\text{LLM}} - y_i^{\text{Human}}| > 5$. Of the 33 configurations meeting the MAE criterion, 21 produced no outliers and were retained for participant-level and item-level analyses. The remaining 12 configurations with outliers were excluded from these analyses due to unreliable individual simulations. However, for the LLM ensemble analysis, these 12 configurations were included after removing their outlier simulations, in order to maximize the diversity of LLMs contributing to the ensemble.

C. Stepwise Regression Analyses of Linguistic Predictors of Production Frequency

For human data, the stepwise regression yielded a significant model [$F(3, 390) = 86.85, p < .001$] that accounted for 40.0% of the variance (adjusted $R^2 = .396$). Word frequency emerged as the strongest predictor ($\beta = .34, t = 6.04, p < .001$), followed by orthographic neighborhood size ($\beta = .254, t = 5.91, p < .001$), and age of acquisition ($\beta = -.18, t = -3.17, p = .002$). This indicates that more frequent, orthographically dense, and earlier-acquired words were produced more often in the task.

For Claude 3.7 Sonnet, stepwise regression likewise produced a significant model [$F(3, 215) =$

Table A1: Model configurations with MAE ≤ 1.69. Configurations with zero outliers were retained for participant-level and item-level analyses; those with outliers were included in ensemble analyses only (after removing outlier simulations).

Provider	Model	MAE	MBE	Outliers
OpenAI	GPT-5	0.00	0.00	0
	GPT-5 mini	0.00	0.00	0
	GPT-5.2 (High Reasoning)	0.00	0.00	0
	GPT-5.2 (Medium Reasoning)	0.00	0.00	0
	GPT-5.2 (Low Reasoning)	0.00	0.00	0
	o3-mini	0.00	0.00	0
	GPT-4.1	0.06	−0.06	0
	GPT-3.5 Turbo	0.07	−0.07	0
	GPT-4o	0.16	−0.16	0
	GPT-5.2	1.50	1.46	2
Anthropic	Claude 4.5 Sonnet (Thinking)	0.00	0.00	0
	Claude 4 Sonnet (Thinking)	0.01	−0.01	0
	Claude 3.7 Sonnet (Thinking)	0.02	−0.02	0
	Claude 4.5 Sonnet	0.06	−0.06	0
	Claude 3.7 Sonnet	0.12	−0.08	0
	Claude 4 Sonnet	0.14	0.05	0
	Claude 4.6 Sonnet	0.75	0.71	3
	Claude 4 Opus	0.84	−0.07	4
	Claude 4.1 Opus	1.06	0.62	6
	Claude 4.5 Opus	1.43	1.30	10
	Claude 4.6 Opus (Thinking)	1.48	1.46	9
Google	Gemini 2.5 Pro	0.00	0.00	0
	Gemini 3 Pro (High Thinking)	0.00	0.00	0
	Gemini 2.5 Flash	0.04	0.04	0
	Gemini 3 Pro (Low Thinking)	0.85	0.72	2
	Gemini 3 Flash	0.86	−0.86	8
xAI	Grok-4.1-fast (Reasoning)	0.05	−0.03	0
	Grok-4-fast (Reasoning)	0.06	0.06	1
Open-Source	Kimi K2	0.03	−0.03	0
	Qwen3-2507	0.08	−0.08	0
	GLM-4.6	0.49	−0.47	3
	DeepSeek-V3.1	0.97	0.93	2
	Llama 4 Maverick	1.29	1.27	9

MAE = Mean Absolute Error; **MBE** = Mean Bias Error (positive = overproduction, negative = underproduction); **Outliers** = number of participants with |LLM−Human| > 5.

$41.13, p < .001$] explaining 36.5% of the variance in production frequency (adjusted $R^2 = .356$). As with humans, word frequency was the strongest predictor ($\beta = .37, t = 5.16, p < .001$). However, unlike humans, orthographic neighborhood size was not retained in the model; instead, word length emerged as a significant predictor ($\beta = -.20, t = -3.56, p < .001$). Age of acquisition remained a significant predictor ($\beta = -.19, t = -2.59, p = .010$), consistent with the human pattern. o3-mini's regression model explained substantially less variance, only 18.8% (adjusted $R^2 = .179$)

Table A2: Model configurations with MAE > 1.69, excluded from all analyses.

Model	MAE	MBE
OpenAI		
o4-mini	3.16	3.16
GPT-5.1	3.91	3.85
o3	35.39	34.67
GPT-4 Turbo	57.17	57.17
Anthropic		
Claude 4.5 Haiku	3.12	2.92
Claude 4.6 Sonnet (Thinking)	3.19	3.19
Claude 4.6 Opus	3.87	3.81
xAI		
Grok-4-fast	2.64	2.43
Grok-4.1-fast	3.67	3.31
Open-Source Models		
Cogito v2.1	3.20	2.16
Mistral-Small-24B	4.05	1.48
Ministral-3-14B	18.62	18.62

MAE = Mean Absolute Error; **MBE** = Mean Bias Error.

[$F(2, 181) = 20.94, p < .001$], retaining just two predictors: word frequency ($\beta = .33, t = 4.59, p < .001$) and word length ($\beta = -.19, t = -2.60, p = .010$).

Translation from the Information Bottleneck Perspective: an Efficiency Analysis of Spatial Prepositions in Bitexts

Antoine Taroni, Ludovic Moncla, Frédérique Laforest
INSA Lyon, CNRS, Lyon 1 Université, LIRIS, UMR5205,
69621 Villeurbanne, France
first.last@insa-lyon.fr

Abstract

Efficient communication requires balancing informativity and simplicity when encoding meanings. The Information Bottleneck (IB) framework captures this trade-off formally, predicting that natural language systems cluster near an optimal accuracy-complexity frontier. While supported in visual domains such as colour and motion, linguistic stimuli such as words in sentential context remain unexplored. We address this gap by framing translation as an IB optimisation problem, treating source sentences as stimuli and target sentences as compressed meanings. This allows IB analyses to be performed directly on bitexts rather than controlled naming experiments. We applied this to spatial prepositions across English, German and Serbian translations of a French novel. To estimate informativity, we conducted a pile-sorting pilot-study ($N = 35$) and obtained similarity judgements of pairs of prepositions. We trained a low-rank projection model ($D = 5$) that predicts these judgements (Spearman $\rho = 0.78$). Attested translations of prepositions lie closer to the IB optimal frontier than counterfactual alternatives, offering preliminary evidence that human translators exhibit communicative efficiency pressure in the spatial domain. More broadly, this work suggests that translation can serve as a window into the cognitive efficiency pressures shaping cross-linguistic semantic systems.

Keywords: Information Theory, Spatial Language, Spatial Cognition, Information Bottleneck, Linguistic Efficiency, Translation

1. Introduction

Languages partition semantic domains in remarkably diverse ways. For example, Levinson, 2003 and Levinson et al., 2003 intensively document how spatial relations are named differently by speakers of typologically diverse languages. Yet beneath this diversity, systematic regularities emerge when naming spatial relations, but also within the semantic domains of colour and kinship *inter alia* (Kemp et al., 2018). One influential standpoint invokes the principle of efficient communication to explain these regularities (Gibson et al., 2019; Kemp et al., 2018; Zaslavsky et al., 2018). According to this view, when speakers of the same language balance the informational and cognitive load in their utterances, the mapping from meanings to words is optimised with respect to that trade-off between informativity and simplicity. Languages around the world achieve different realisations of this trade-off through evolution (Kirby et al., 2015).

The Information Bottleneck (IB) framework (Tishby et al., 2000) was put forward by Zaslavsky et al., 2018 to operationalise this trade-off for natural languages. It states that languages achieve near-optimal compression of meanings from an information-theoretic perspective, and has recently gained substantial empirical support. Usually, an IB analysis focuses on a specific semantic domain, and relies on naming data (e.g. natural descriptions of stimuli such as pictures or videos). The underlying assumption is that a stimulus evokes a mental representation, which is efficiently compressed into a description. The communicative cost of the description (complexity) is then compared to the relevant information it conveys about the stimuli (informativity), across natural and hypothetical languages. Prior IB applications relied on various stimuli like colour chips (Zaslavsky et al., 2018), static visual referents from images (Gualdoni et al., 2023), and kinematic events in videos of human locomotion (Langlois et al., 2025). However, to our knowledge, no IB analysis has considered linguistic stimuli.

We propose to regard a word in context as a stimulus in its own right and to perform an IB analysis in the semantic domain of spatial relations. To do so, we use a novel type of data: bitexts, e.g. translations of source sentences into target ones (illustrated in Table 1). We frame the act of translation as a constrained optimisation problem in the IB sense, and assume that human translators aim for exact meaning preservation. A word of interest from a source sentence is then considered a stimulus, which triggers its own mental representation, which is then linguistically encoded in the target language.

We tested this new methodology on a cornerstone of spatial language: spatial prepositions. These terms overtly mark for spatial relations between entities, and are highly prone to engage spatial representations (Landau and Jackendoff, 1993). We first extracted cross-linguistic variations of these spatial terms in four languages, namely detecting

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 264–275
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

id	French	English	German	Serbian
n1909	le pilote [...] prit place **sur** la passerelle	The pilot [...] took his place on the bridge	der Lootse [...] und nahm seinen Platz auf dem Steg	pilot [...] zauze mesto na komandnom mostu
n3688	Mudge [...] se maintenait **dans** la ligne droite	Mudge [...] kept in a straight line	Mudge [...] hielt sich in gerader Linie	Mody se drzxao prave linije
n3110	[...] il reprit la direction de l'est **à travers** le massif accidenté des monts Wahsatch.	[...] it took an easterly direction towards the jagged Wahsatch Mountains.	[...] nahm er wieder östliche Richtung durch den unregelmäßigen Hauptstock des Wahsatchgebirges.	[...] krenuo je prema istoku preko masiva planine Vesecy.
n2481	Quelques matelots [...] descendirent le pauvre garçon **dans** une cabine des secondes	Several sailors [...] carried the poor Frenchman down into the second cabin	Einige Matrosen [...] trugen den armen Jungen in eine Cabine zweiten Ranges	Nekoliko mornara [...] odnesosxe jadnog momka u jednu kabinu druge klase
n1243	le guide dirigerait l'éléphant **vers** la pagode de Pillaji	the guide should direct the elephant towards the pagoda of Pillaji	der Führer solle den Elephanten nach der Pagode Pillaji leiten	vodicy uputi slona ka pagodi u Piladyi

Table 1: Excerpt of (LIFAT et al., 2024). Spatial prepositions in the French source sentences are emphasised. `id` refers to the original identifier as released in the corpus.

occurrences in source sentences from LIFAT et al., 2024, and aligning them with their target counterparts. This allowed us to assess the complexity of spatial prepositions translations. Furthermore, to estimate the informativity of these translations, we also modelled the psychological space of spatial relations by leveraging pile-sorting results from a pilot study ($N = 35$). We show that a low-dimensional space suffices to predict similarity judgements. Ultimately, we provide preliminary evidence for greater communicative efficiency of natural translations compared to random and counterfactual systems, in the semantic domain of spatial relations.

2. Theoretical framework

This section outlines how the IB theory has been applied to language. Tracing back to the Rate-Distortion theory (Cover, 1999), communication between a speaker and a listener is formalised as a joint search for an encoder and a decoder that enable lossy data compression. In this schematic communicative scenario, the speaker seeks a linguistic encoder q that maps meanings to forms, which is minimally complex, and simultaneously minimises the *distortion* between the intended meaning M, and the listener's reconstructed meaning $\hat{M}$. An operationalisation of this trade-off was first proposed by Tishby et al., 2000, and later applied to colour-naming (Zaslavsky et al., 2018), and other semantic domains (Mollica et al., 2021; Chen et al., 2023; Langlois et al., 2025).

In this line of work, the Complexity of the stochastic encoder q is quantified by the mutual information[1] $I_q(M;W)$ between two random variables: M and W[2]. The mental representations the speaker wishes to communicate are denoted by M, and W represents the discrete linguistic forms they are mapped onto, according to the policy $p_q(w|m)$ of the encoder q. Conversely, the *distortion* between the speaker's intended meaning M, and the listener's reconstructed meaning $\hat{M}$, is measured using the expected Kullback-Leibler distance, $\mathbb{E}_q[D(M||\hat{M}]$. Minimising this distortion is mathematically equivalent to maximising the quantity $I_q(W;U)$, which is defined as the Accuracy of the encoder (Zaslavsky, 2019). In this formulation, U represents the possible states of the world that the speaker perceives and describes, and about which the listener forms probabilistic inferences.

The core trade-off between the compression of mental representations (measured by Complexity) and the preservation of meaningful information (measured by Accuracy) translates directly to optimisation. For any fixed amount of Accuracy, the encoder and decoder are optimised to ensure that Complexity is minimised. Introducing a Lagrange multiplier β, this trade-off can be framed as a con-

[1] The mutual information is a general measure of how related two random variables are. It measures the reduction in uncertainty in one variable when observing the other: $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$, with $H(X)$ being the entropy of X. Note that the joint distribution $p(X,Y)$ is all we need to compute the mutual information: $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\cdot p(y)}$. It is a symmetric measure: $I(X;Y) = I(Y;X)$.

[2] We refer to random variables using upper-case letters, and realisations with lower-case letters.

strained optimisation problem with the following objective function:

$$\mathcal{F}_\beta[q] = I_q(M;W) - \beta I_q(W;U) \quad (1)$$

where β controls the trade-off. Any incremental 1-bit growth in Accuracy of a given optimal encoder incurs a cost of β bits in Complexity. All natural or artificial encoders can be located in the Accuracy-Complexity information plane. From the Data Processing Inequality it follows that Complexity serves as an upper bound for Accuracy. Consequently, any encoders lying above the optimal frontier defined by the set $\{\mathcal{F}^*_\beta \mid 0 \leq \beta < +\infty\}$ are unachievable under the assumptions of the IB framework.

3. General methodology

Within a specific semantic domain, an IB analysis relies on two types of data to compute the Complexity and Accuracy of an encoder q: naming data and human similarity judgements. Examples of naming data include colour names (when describing colour chips in Zaslavsky et al., 2018), verbs of human gait (when describing a person moving in short videos, in Langlois et al., 2025), object names (when describing real-world images in Gualdoni et al., 2023), or spatial relations (when describing drawings of spatial scenes in Khetarpal et al., 2013). In this work, we apply the IB framework to the translation task. In this section, we detail how naming data and similarity judgements can be recovered from translation data, and how Complexity and Accuracy measures are derived.

Using a bitext, we treat each occurrence of a term of interest within its sentential context U as a stimulus that evokes a meaning representation M. Consider, for example, $u = \text{the apple } \underline{\text{in}} \text{ the bowl}$. Here, the spatial relation between the apple and the bowl is not only determined by the lexical item $\underline{\text{in}}$ alone, but emerges from the full sentential context in which it is embedded. Crucially, two occurrences of the same lexical item evoke the same meaning M if and only if they appear in identical sentential contexts, that is, when the same phrase is repeated in the bitext corpus. Under this assumption, M is contextually grounded rather than lexically fixed. The evoked representation M is then encoded using a lexical item W in the target language. The schema $U \rightarrow M \rightarrow W$ summarises how translations are generated under these assumptions.

3.1. Naming data

Given a bitext corpus, our goal is to align terms of interest from source sentences with their counterparts from target sentences. For example, the French spatial prepositions from Table 1 ($u_1 = \text{le pilote prit place } \underline{\text{sur}} \text{ la passerelle}$ and $u_2 = \text{Mudge se maintenait } \underline{\text{dans}} \text{ la ligne droite}$) map to their English equivalents $w_1 = \text{on}$ and $w_2 = \text{in}$. This process yields alignment tables like the one presented in Table 2. From these alignments, we can derive the conditional probability distribution $p_q(w|m)$ where m ranges over each distinct meaning and w ranges over the target lexicon. Assuming a uniform prior distribution $p(m)$ over all meanings in the source text [3], we compute the joint distribution: $p_q(w,m) = p_q(w|m)p(m)$. From this, we derive the Complexity $I_q(M;W)$ (see footnote 1 for details).

id	French	Automatic alignment with English				
		above	in	into	on	towards
n1909	et prit place **sur** la passerelle	0	0	0	**1**	0
n3688	se maintenait **dans** la ligne droite	0	**1**	0	0	0
n3110	la direction de l'est **à travers** le massif	0	0	0	0	**1**
n2481	descendirent le pauvre garçon **dans** une cabine	0	0	**1**	0	0
n1243	le guide dirigerait l'éléphant **vers** la pagode de Pillaji	0	0	0	0	**1**

Table 2: French to English excerpt of computed alignments of spatial prepositions from LIFAT et al., 2024. Each row is a one-hot encoding of the source preposition over the target lexicon. The source preposition is emphasised in bold.

3.2. Similarity judgements

Collecting contextual similarity judgements of source terms grows quadratically in the size of the bitext. Being able to estimate the pairwise similarity between any pair of source terms is therefore crucial. As part of our proposed methodology, we first select a subset of terms in their sentential context $\{u_i \mid 1 \leq i \leq K\}$ of reasonable size K, and collect human pairwise similarity judgements $\{sim(u_i,u_j) \mid 1 \leq i,j \leq K\}$. For ease of exposition, we denote these empirical similarities as $sim_{i,j}$. Based on these pairwise similarity judgements, we then train a model to predict the similarity $\widehat{sim}_{i,j}$ for any new unseen pair of terms. Following Zaslavsky et al., 2018 and Chen et al., 2023, we derive a subjective belief distribution on world states conditional on meanings $p(u|m)$, which is estimated as

$$p(u_j|m_i) \propto \exp\left(\gamma \cdot \widehat{sim}_{i,j}\right) \quad (2)$$

[3] Zaslavsky et al., 2018 show that replacing an empirically derived prior reflecting actual communicative need with a simple uniform prior has little qualitative influence on the core finding. We use therefore a uniform prior distribution in this first step

Here, γ is a temperature parameter that controls the penalty for mistaking two distinct stimuli[4]. Note that when IB models a subjective belief as a distribution on world states U, it captures the ambiguity a term of interest can have in its context. Using the distributions $p(u|m)$, $p_q(w|m)$, and $p(m)$, we can now compute the joint distribution: $p_q(w,u) = p_q(w|m)p(m)p(u|m)$. From this, we derive the Accuracy $I_q(W;U)$.

4. Experiments

For the remainder of this work, we focus on the specific semantic domain of spatial relations. These relations are typically marked by prepositions as discussed extensively in the literature (for French linguistic aspects we refer to Vandeloise, 1991; Borillo, 1998; Le Pesant, 2012, for a broad discussion on spatial cognition we refer to Landau and Jackendoff, 1993). We also consider "absolute use" of prepositions (Le Pesant, 2012), which refers to instances where a preposition appears without its standard complement because the referent is understood from context (e.g., Je me gare devant, I park in front, Ich parke davor and Parkiram ispred). More importantly, we consider a preposition to have a spatial sense when its surrounding context triggers a concrete spatial mental imagery, involving entities that can be physically touched or pointed at. This operational definition is theoretically grounded in the notions of *cline of concreteness* (Brysbaert et al., 2014) and *cline of collocation* (Pragglejaz, 2007).

Throughout our experiments[5], we utilised English, German, and Serbian translations of Jules Verne's French adventure novel, *Le tour du monde en 80 jours*. The sentence-level alignment across these four texts was previously processed by LIFAT et al., 2024.

4.1. Naming data: detection and alignment of spatial prepositions

We first collected occurrences of prepositions with potential spatial sense, namely searching for instances from the lexical inventory of spatial prepositions established by Le Pesant, 2012. An initial disambiguation of each occurrence was then carried out jointly by three annotators until consensus was reached. The lexical and semantic disambiguation steps yielded 1,312 prepositions with a spatial sense. Using the Large Language Model `mistral-large-2512` (Mistral AI, 2025) in a few-shot setting, these French prepositions were subsequently aligned with their English, German and Serbian counterparts. The alignment table for Table 1 is illustrated in Table 2. The few-shot prompt for French-to-English alignment is available in Appendix A. To evaluate the quality of these automatic alignments, we sampled 100 preposition pairs for each target language. Errors of alignment were defined as the alignments of source prepositions to either inadequate target prepositions, or to any distinct other parts of speech, such as nouns or verbs. This evaluation yielded precision scores of 81% for French–English, 91% for French–German, and 94% for French–Serbian. Out of the 1,312 source prepositions, 580 were successfully aligned across all three target languages. The remaining instances were discarded, primarily because the source preposition was not relexified with a preposition in at least one of the translations[6]. The cross-linguistic discrepancy between automatic alignments is discussed in section 5.3.

Furthermore, we computed contextual embeddings for each source preposition, using `xlm-roberta-large` (Conneau et al., 2020). We averaged and stacked the last four hidden layers, yielding embeddings of size $d = 4,096$. We then applied K-means clustering ($K = 30$) to obtain a representative subset of 30 French sentences, each with a spatial preposition, for downstream analysis. We found that $K = 30$ reflected both the distribution of French spatial terms in the corpus, and the embedding space (see Appendix D for a dimension reduction). The clustering and dimension reduction were implemented using `scikit-learn` (Pedregosa et al., 2011).

4.2. Models of human similarity judgements

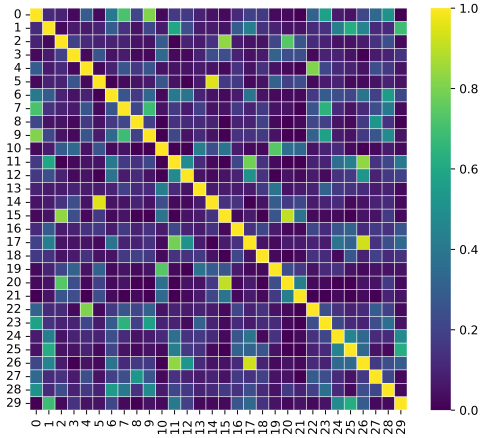
Our objective is to predict the human-perceived similarity between two distinct spatial relations expressed by spatial prepositions from any source sentence.

For this purpose, we designed a pile-sorting task based on the $K = 30$ samples obtained from our clustering step. To assess the feasibility of these similarity judgements, we conducted an anonymised pilot study (see section 7) with a cohort of colleagues ($N = 35$), all of whom were native French speakers. The reported patterns should be interpreted as preliminary evidence. The 30 samples were printed on individual cardboard cards. The participants were orally instructed to group the

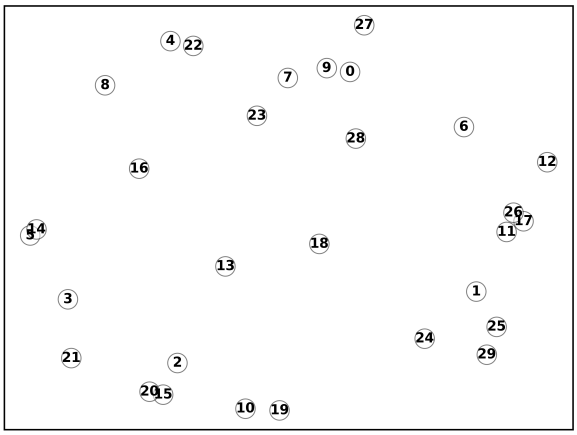
[4]We set γ so that $p(u|m)$ is defined by a softmax over similarity scores. Alternative methods for grounding the distribution $p(u|m)$ should be explored in future work.

[5]Data and code are available at https://github.com/antoineOYO/spatial-terms-translations

[6]This is also known as *zero-marking*: consider the use of the preposition in Je suis à la maison versus I'm home. We refer to Beekhuizen, 2025 for related analysis.



(a) Proportion of participants (N=35) who grouped 2 spatial relations into the same pile.



(b) MDS visualisation of the pairwise similarities.

0: et prit place **sur** la passerelle
1: car **dans** cet immense et admirable port de New York
2: le guide dirigerait l'éléphant **vers** la pagode
3: ce voyage s'accomplissait régulièrement **autour du** monde
4: **sous** ses goélettes, l'Henrietta marcha
5: **entre** ces deux importantes cités
6: il avait pu s'embarquer **à bord du** Rangoon
7: **sur** le bûcher enflammé se détachait alors le corps
8: Phileas Fogg prit une chaise et s'assit **près de** la cheminée
9: je ne le vois jamais **sur** le pont
10: et **par** Rock-Island, il entrait dans l'Illinois
11: Je suis resté couché **dans** ma cabine
12: Étant entré **chez** un barbier chinois pour se faire raser
13: la vapeur s'échappait **par** les soupapes
14: qui s'étendent **entre** le Japon et le Nouveau Monde
15: un paquebot qui ne partait que le lendemain [...] **pour** Hong-Kong.
16: se trouver [...] **au milieu** de voyageurs
17: descendirent le pauvre garçon **dans** une cabine des secondes
18: un bruit retentit **au – dehors**
19: il reprit la direction de l'est **à travers** le massif
20: dût-il pousser **jusqu'à** Macao
21: un parcours de deux cent vingt milles **depuis** Hong-Kong
22: la tête de la procession apparut **sous** les arbres
23: couraient **sur** des tuyaux de pipe
24: se maintenait **dans** la ligne droite
25: La goélette [...] semblait voler **dans** l'air
26: descendre **dans** la cabine
27: il s'arrêtait **devant** une vaste case
28: Au lieu d'arriver **sur** l'Henrietta
29: encore en faveur **dans** le Céleste Empire

(c) Prepositions (emphasised) and context of the 30 spatial relations sorted by the participants

Figure 1: Empirical similarity data from the pile-sorting task.

samples into as many piles as they deemed necessary based on their perceived similarity; there was no time limit and participants could revise their groupings until satisfied. The full instructions (see Appendix B) were printed and available to each participants throughout the whole task. The distribution of number of piles produced by the participants is displayed in Appendix C.

Following Khetarpal et al., 2013, we define the pairwise similarity $sim_{i,j}$ as the proportion of all participants who sorted u_i and u_j into the same pile. A visualisation of the proportions is shown in Figure 1a. From the list of pairwise similarities, a Multidimensional Scaling (MDS)[7] was then computed using `scikit-learn` implementation (Figure 1b). Both visualisations suggest that the similarity structure of spatial relations is complex and graded rather than discrete or reducible to surface form (discussed in section 5.2).

We modelled these similarity judgements using three approaches based on the contextual embeddings F_i of size d we computed in the previous step. As a baseline, we computed the Cosine Similarity between z-scored embeddings, following Timkey and van Schijndel, 2021. We compared this baseline against two learned models: a Ridge regression and a low-rank projection model (replicating Jha et al., 2023). The Ridge regression model is a simple linear regression, where the size of the coefficients is penalised with a parameter α. It is trained on the vertical stack of three features: the Hadamard product $F_i \odot F_j$, their L1 distance, and their cosine similarity. Conversely, the low-rank projection model learns a mapping matrix $P \in \mathbb{R}^{D \times d}$ to project embeddings into a D-dimensional space, predicting similarity as $\widehat{sim}_{i,j} = (PF_i)^\top (PF_j)$. This model minimises the mean squared error loss with L2 regularisation:

$$\min_{\lambda,P} \frac{1}{2} \sum_{i,j} ||sim_{i,j} - \widehat{sim}_{i,j}||^2 + \lambda ||P||^2 \quad (3)$$

Hyperparameters (including the Ridge α, the projection rank D, and the regularisation penalty λ) were optimised within 6-fold inner validation splits, and the generalisation error (measured via Spearman's ρ) was computed across 6 outer test folds using a nested cross-validation procedure. Model

[7] Note that this MDS is for visualisation purpose only.

performance (see Table 3) is reported using Spearman's rank correlation coefficient ρ, as it is better suited for evaluating monotonic relationships than Pearson's r_P, which evaluates strictly linear ones (De Winter et al., 2016). Results presented in Table 3 indicate that a simple cosine similarity does not strongly correlate with the sortings we collected, whereas the low-rank projection model achieves a high Spearman correlation of $\rho = 0.7800 \pm 0.0622$, with dimension $D = 5$.

Model	**Spearman** ρ
Cosine Similarity (Baseline)	0.079
Ridge Regression	0.647 ± 0.056
Low-rank Projection	
$D = 1$	0.263 ± 0.140
$D = 5$	$\mathbf{0.780 \pm 0.062}$
$D = 10$	0.762 ± 0.053
$D = 15$	0.751 ± 0.053
$D = 20$	0.749 ± 0.034
$D = 50$	0.648 ± 0.070
$D = 100$	0.578 ± 0.101

Table 3: Spearman correlation between model-predicted pairwise similarities and human pile-sorting judgements. Models all take pairs of contextual embeddings as input. Learned models performance is computed from 6-fold nested cross validation.

4.3. Computation of encoders locations in the information plane

To benchmark the attested translation encoders, we generated 3x10,000 counterfactual "perturbed encoders" by randomly permuting 1%, 5% and 10% of the rows within the alignment tables. We also generated 300,000 uniform-random translation encoders. We then computed the Accuracy and Complexity for attested, perturbed and fully random translation encoders using the implementation provided by Zaslavsky et al., 2018. The optimal frontier was computed using the reversed deterministic annealing algorithm implementation from Chen et al., 2023, and was traced by iterating over 100 values of β log-linearly spaced between 1 and 2^{20}. Attested, perturbed and random translations, along with the optimal frontier, are displayed in the information plane from Figure 2a. We observe that attested translation encoders lie far from the steepest part of the theoretical bound, which is discussed in section 5.3.

Within the information plane, the distance of any given encoder q to the optimal frontier defines its deviation from optimality, as established by Zaslavsky et al., 2018:

$$\epsilon_q = \min_{\beta} \frac{1}{\beta} \left(\mathcal{F}_\beta[q] - \mathcal{F}^*_\beta \right), \tag{4}$$

The deviations of translation encoders is shown in Figure 2b. Results indicate that as the degree of perturbation increases, the efficiency of translation encoders decreases.

5. Discussion

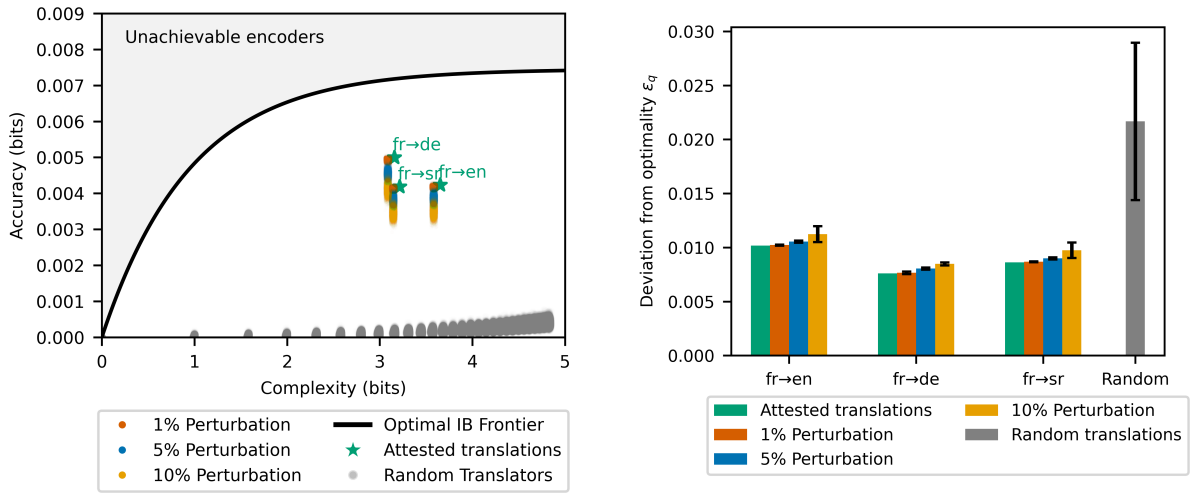
5.1. Characterising cross-linguistic use of spatial prepositions

The automatic detection and alignment of spatial prepositions in bitexts has received limited attention until recent studies such as those by Viechnicki et al., 2024 and Beekhuizen, 2025. For detection, we handled the bitext differently by choosing not to apply a syntactic filter to the source sentences. This choice is based on the observation that spatial terms appear in a variety of syntactic contexts, and range from single terms (sur, dans), to nominal locutions (à bord du, au milieu de). With respect to alignment, Viechnicki et al., 2024 rely on the GIZA++ word alignment tool (Och and Ney, 2003). In contrast, the RUIMTE model proposed by Beekhuizen, 2025 avoids traditional word alignment, instead identifying spatial relation markers by calculating statistical associations. Direct performance comparisons between these methods are difficult, as these previous works evaluated highly morphologically complex languages. While our pipeline introduces a dependency on few-shot extraction tools, it generalises to any bitext without requiring morphologically complex language-specific alignment models.

5.2. Estimation of similarity between spatial relations evoked by prepositions

The correlation between judgements of word sense relatedness and computational representations relatedness has been widely investigated using both static embeddings (Hill et al., 2015), contextual embeddings and Large Language Models (Nair et al., 2020; Marjieh et al., 2022). Specifically, how context modulates the perceived meaning similarity between word pairs in written sentences has been graded by human subjects in many experiments (Huang et al., 2012; Pilehvar and Camacho-Collados, 2019; Armendariz et al., 2020; Vulić et al., 2020). However, most of these studies focus on nouns, adjectives or verbs. Our pilot study, conversely, was concerned with relations between entities, and the prepositions that mark them.

Results presented in Table 3 indicate that a simple cosine similarity does not strongly correlate with



(a) Information plane of attested, perturbed and random encoders of spatial prepositions. Dots are jittered slightly along the Complexity axis for visualisation purposes.

(b) Deviation from optimality for attested, perturbed and random encoders. Black error bars indicate standard deviations.

Figure 2: How much do attested and random systems deviate from optimality?

our collected human sortings, whereas the low-rank projection model achieves a high Spearman correlation of $\rho = 0.7800 \pm 0.0622$. Interestingly, low-rank projection models achieve the highest correlation scores with $D = 5$. This suggests that higher-dimensional projections likely lead to overfitting. These findings indicate that low-rank projections may be more effective than standard cosine similarity for predicting contextual similarities in spatial relations.

Aside from modelling, we acknowledge that our subset of 30 spatial prepositions in context does not guarantee cognitive representativeness of the domain of spatial relations. However, results from Figure 1a and 1b suggest that colexification does not explain alone the similarity structure of spatial relations. We assume that modelling the semantic similarity of spatial relation in a continuous D-dimensional space accounts for the graded similarity space we observed, and that a sufficiently large bitext should provide broad coverage of the domain.

However, a lexical bias may arise during the pile sorting task, where shared word forms inflate human judgements of semantic similarity. As a result, similarity ratings may partly reflect surface-form overlap rather than purely relational semantics. This introduces a potential language-specific skew, whereby the lexical structure of French influences perceived relatedness. To mitigate this effect, future work should control for form overlap in the stimuli and compare across source languages, in order to quantify the contribution of lexical structure to similarity judgements.

We also acknowledge that this subset is not fully representative of all linguistic means available for encoding spatial relations. While we focused on prepositions as a first step, alternative encoding mechanisms must be accounted for in future research, such as interactions between the preposition and the verb (see for example the verb-framing typologies from Talmy, 1983 and Aurnague, 2012).

5.3. Efficiency of attested systems

In Figure 2a we observe that attested translations of spatial prepositions lie closer to optimal systems than random ones. To further investigate whether this is a feature of "reasonable" translations, we generated counterfactual translations with varying degrees of perturbation. The results from Figure 2b indicate that as the degree of perturbation increases, the deviation to optimality increases. We believe that this comparison would benefit from including another baseline, such as a surjective mapping from meanings to terms, as used by Strickland and Chemla, 2018. This is equivalent to removing the non-deterministic assumption on the encoder q (Strouse and Schwab, 2017). Future work should also test how altering the similarity geometry affects Accuracy. Nevertheless, these results suggest that attested systems exhibit a pressure for efficiency under the IB framework.

We conclude by addressing why attested encoders deviate from optimality, and the confounds inherent to translation data. This question lies at the crossroads of research on language efficiency and translation process (Teich et al., 2020). Translating involves genre conventions and notably stylistic incentives. This is illustrated by observations from

the corpus: while German and Serbian translations of *Le Tour du monde en 80 jours* remain stylistically conservative, the English version acts as a "free" translation with added details. This divergence is clear in the original corpus, where LIFAT et al., 2024 found and explicitly annotated changes in meaning, but were found only in the English version. From an information-theoretic perspective, these changes in meaning represent a deviation from strict communicative efficiency, as the translator prioritised semantic elaboration or expressivity over the compact encoding predicted by the IB optimum. Yet, a conservative translator might also imitate some of the lexical choices of the French original despite what would be natural in their language (Volansky et al., 2015; Lim et al., 2024). The alignment of preposition-to-preposition we performed (see section 4.1) is by design too coarse to capture the aforementioned phenomena. Therefore we argue that IB analysis of translation data must narrow the scope to relexification of terms of interest, such as prepositions, in order to capture only one aspect of the translation process.

6. Conclusion

This work proposes a novel methodology for analysing communicative efficiency via translation data. We performed an IB analysis on spatial prepositions across three translations of a French novel. Our preliminary results suggest that the translation of spatial prepositions exhibits pressure toward optimal trade-offs between informativity and complexity, consistent with the IB literature, and more broadly, recent evidence from language efficiency. An important next step is to apply this approach to typologically diverse, genealogically distant languages, allowing us to better separate general efficiency pressures from patterns arising through shared linguistic ancestry. We believe we presented a plausible way to extend to other semantic domains, other linguistic forms than prepositions, and scale beyond fully elicited naming datasets while staying grounded in natural speech.

7. Ethics Statement

Written informed consent was obtained from all participants. Pile-sorting data consisted only of lists of sentences. These data were fully anonymised at the point of collection, such that no personal data were retained; the data from the pile-sorting experiment therefore fall outside the scope of the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679). Given the minimal-risk nature of this feasibility study (involving no sensitive information and no identifiable data) institutional ethics committee approval was not sought.

8. Acknowledgements

We sincerely thank our colleagues at Laboratoire d'InfoRmatique en Image et Systèmes d'information for participating in the pile-sorting experiment. We further thank the participants of the SIGTYP workshop for their stimulating discussions and valuable feedback during the presentation of this work. We are also grateful to the anonymous reviewers for their constructive suggestions, which helped us improve this paper.

9. Bibliographical References

Carlos Santos Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France. European Language Resources Association.

Michel Aurnague. 2012. De l'espace à l'aspect: les bases ontologiques des procès de déplacement. *Corela. Cognition, représentation, langage*, (HS-12).

Barend Beekhuizen. 2025. Spatial relation marking across languages: extraction, evaluation, analysis. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 571–585, Vienna, Austria. Association for Computational Linguistics.

Andrée Borillo. 1998. *L'espace et son expression en français*. Editions Ophrys.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Sihan Chen, Richard Futrell, and Kyle Mahowald. 2023. An information-theoretic approach to the typology of spatial demonstratives. *Cognition*, 240:105505.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Joost CF De Winter, Samuel D Gosling, and Jeff Potter. 2016. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological methods*, 21(3):273.

Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. 2019. How efficiency shapes human language. *Trends in cognitive sciences*, 23(5):389–407.

Eleonora Gualdoni, Charles Kemp, Yang Xu, and Gemma Boleda. 2023. Quantifying informativeness of names in visual space. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–882, Jeju Island, Korea. Association for Computational Linguistics.

Aditi Jha, Joshua C. Peterson, and Thomas L. Griffiths. 2023. Extracting Low-Dimensional Psychological Representations from Convolutional Neural Networks. *Cognitive Science*, 47(1):e13226.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic Typology and Efficient Communication. *Annual Review of Linguistics*, 4(1):109–128.

Naveen Khetarpal, Grace Neveu, Asifa Majid, Lev Michael, and Terry Regier. 2013. Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35. Issue: 35.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.

Barbara Landau and Ray Jackendoff. 1993. "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238.

Thomas A. Langlois, Roger Levy, Nidhi Seethapathi, and Noga Zaslavsky. 2025. Efficient compression in locomotion verbs across languages. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Denis Le Pesant. 2012. Essai de classification des prépositions de localisation. In *3ème Congrès Mondial de Linguistique Française (CMLF 2012)*, pages 921–936.

Stephen C Levinson. 2003. *Space in language and cognition: Explorations in cognitive diversity*, volume 5. Cambridge University Press.

Stephen C Levinson, Sérgio Meira, et al. 2003. 'natural concepts' in the spatial topologial domain–adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79(3):485–516.

LIFAT, LLL, and Université de Belgrade. 2024. En80jours. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. Simpson's Paradox and the Accuracy-Fluency Tradeoff in Translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–103, Bangkok, Thailand. Association for Computational Linguistics.

Raja Marjieh, Pol Van Rijn, Ilia Sucholutsky, Theodore R Sumers, Harin Lee, Thomas L Griffiths, and Nori Jacoby. 2022. Words are all you need? language as an approximation for human similarity judgments. *arXiv preprint arXiv:2206.04105*.

Mistral AI. 2025. Introducing mistral 3. Mistral Documentation.

Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. 2021. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49):e2025993118.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Group Pragglejaz. 2007. MIP: A Method for Identifying Metaphorically Used Words in Discourse. *Metaphor and Symbol*, 22(1):1–39.

Brent Strickland and Emmanuel Chemla. 2018. Cross-linguistic regularities and learner biases reflect "core" mechanics. *PLOS ONE*, 13(1):e0184132. Publisher: Public Library of Science (PLoS).

Daniel J Strouse and David J Schwab. 2017. The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630.

Leonard Talmy. 1983. *How Language Structures Space*, pages 225–282. Springer US, Boston, MA.

Elke Teich, José Martínez Martínez, and Alina Karakanta. 2020. Translation, information theory and cognition. In *The Routledge Handbook of Translation and Cognition*. Routledge. Num Pages: 16.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Claude Vandeloise. 1991. *Spatial prepositions: A case study from French*. University of Chicago Press.

Peter Viechnicki, Kevin Duh, Anthony Kostacos, and Barbara Landau. 2024. Large-Scale Bitext Corpora Provide New Evidence for Cognitive Representations of Spatial Terms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1089–1099, St. Julian's, Malta. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, et al. 2020. Multi-simlex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics*, 46(4):847–897.

Noga Zaslavsky. 2019. *Information-theoretic principles in the evolution of semantic systems*. Ph.D. thesis, The Hebrew University of Jerusalem Jerusalem, Israel.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

Appendix

A. Prompt for French to English few-shot alignment

ALIGNMENT OF SPATIAL TERMS
You are an expert linguist and translator specialised in French and English. French is always the source language, and the target language is English. You will be provided with a source sentence in French, and the target sentence in English. In the source sentence, a spatial term is wrapped with <loc> and </loc>. This spatial term marks a concrete spatial relationship. Your task: align the spatial term in the source sentence with its translation in the target sentence, if any.
How: wrap with <loc> and </loc> the corresponding spatial term in the target sentence, if any. Also return the normalized form of the target spatial term (e.g., "on the right of" to "on the right").
Constraints:
1. Spatial terms only include prepositions, adverbs, and locutions.
2. Zero-alignment: if you cannot find an adequate spatial term in the target sentence corresponding to the translation of the source spatial term, don't

wrap anything in <loc> from the target sentence and return "None" for the normalized form.
3. If the meaning of the source spatial term is translated only by a verb or a noun in the target sentence, don't wrap anything in <loc> from the target sentence and return "None" for the normalized form.
Example 1: basic alignment
INPUT = { "source_sentence" : "Passepartout demeura seul <loc>dans</loc> la maison de Saville-row", "target_sentence": "Passepartout remained alone in the house of Saville-row", }
OUTPUT = { "target_sentence": "Passepartout remained alone <loc>in</loc> the house of Saville-row", "target_normalized_term": "in" }
Example 2: zero-alignment
INPUT = { "source_sentence" : "Il rentra <loc>chez</loc> lui à dix heures.", "target_sentence": "He returned home at ten o'clock.", }
OUTPUT = { "target_sentence": "He returned home at ten o'clock.", "target_normalized_term": "None" }
Example 3: complex locution
INPUT = { "source_sentence" : "Au moment où l'Henrietta appareillait,tous quatre étaient <loc>à bord</loc>.", "target_sentence": "all four were aboard at the moment when the Henrietta was setting sail.", }
OUTPUT = { "target_sentence": "all four were <loc>aboard</loc> at the moment when the Henrietta was setting sail.", "target_normalized_term": "aboard" }
Example 4: zero-alignment
INPUT = { "source_sentence" : "Il a marché <loc>juqu'à</loc> la ville.", "target_sentence": "He reached the city by foot.", }
OUTPUT = { "target_sentence": "He reached the city by foot.", "target_normalized_term": "None" }
# Now is your turn. Provide the output in JSON format as in the examples.

B. Instructions for pile sorting

French instructions (as printed for the participants): "Sur chaque billet est exprimée une relation spatiale entre 2 entités. Une préposition en gras vous signale cette relation. Formez des piles regroupant les relations spatiales qui expriment des configurations spatiales similaires. Vous pouvez faire autant de piles que vous le souhaitez (y compris des piles d'un seul billet si une relation vous semble unique). Le numéro sur les billets est aléatoire."
A possible English translation: "On each card, a spatial relationship between 2 entities is expressed. A bolded preposition indicates this relationship. Form piles grouping the spatial relationships that express similar spatial configurations. You can create as many piles as you like (including stacks with a single card if a relationship seems unique). The numbers on the cards are random."

C. Pairwise similarity modelling

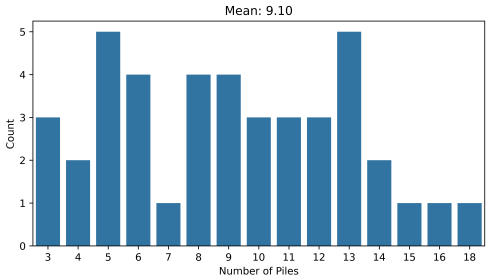


Figure C-1: Distribution of the number of piles across participants (N=35).

D. Dimension reduction of the embedding space

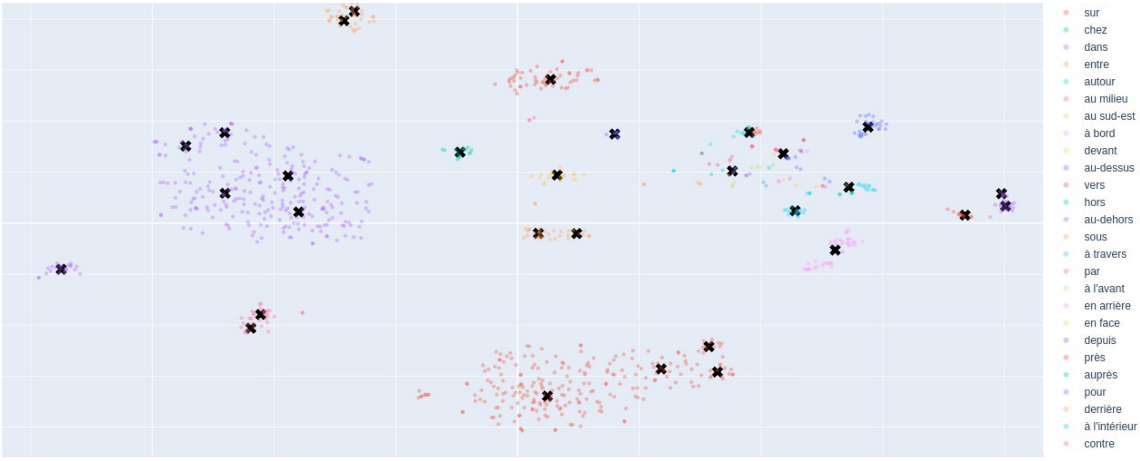


Figure D-1: t-SNE visualisation of the 30 K-means centroids (computed from stacked final-layer representations of `xlm-roberta-large`) and embeddings of the most frequent French prepositions

Causal Inferences Are Driven by Noun Concept Specificity: Evidence from Self-Paced Reading and Large Language Model Surprisal

Fabian Schlotterbeck*, **Raphael Barth***
University of Tübingen, Independent Researcher
fabian.schlotterbeck@uni-tuebingen.de, raphael.barthe@web.de
*co-first authors

Abstract

Inferring causal coherence relations has been argued to affect incremental language comprehension, particularly the integration of subsequent material. Building on prior psycholinguistic work, we investigated whether noun phrase (NP) specificity modulates the integration of explicit causal explanations. In a German self-paced reading study, we manipulated NP specificity in sentences containing implicit causality verbs, causally interpretable relative clauses, and explicitly causal subordinate clauses introduced by *because*. NP specificity systematically affected reading times in the *because*-clause, supporting the view that discourse-level causal expectations are derived incrementally. Within a frame-semantic account, more specific NPs activate more tightly constrained scenario structures, thereby strengthening causal expectations. Surprisal values from GPT-2 models replicated the critical interaction, though discrepancies with human reading times emerged. Overall, the findings support interactive, multi-component models of language processing and also speak to the relation between abstract psycholinguistic models and surprisal-based models of predictive processing affecting reading behavior.

Keywords: Causal Inferences, Coherence Relations, Elicitures, Noun Concept Specificity, Incremental Processing, Self-Paced Reading, Surprisal, Large Language Models

1. Introduction

Natural language involves drawing inferences at multiple levels. On the level of discourse comprehension one of the more elusive kinds of inferences are so called conversational elicitures (Cohen and Kehler, 2021). These inferences have been distinguished from entailments, presuppositions and implicatures. Two of their defining features are that they cannot be linked to a specific trigger and that they involve general cognitive biases, such as a tendency to draw causal connections between events. One classical case are causal inferences in discourse. In (1-a), for example, the relative clause tends to be understood as a causal explanation of the event described in the matrix clause although no unequivocal triggers for this inference can be identified. This is assumed to arise from the cognitive bias to construct causal relations between events, particularly when establishing coherence relations between discourse contents.

(1) a. The boss fired the employee who was always late for work.
b. The boss fired the employee because he was always late for work.

Relative clauses in this type of structure are often used restrictively to identify a referent. It is, however, assumed that the specific event described in the sentence in (1-a) in combination with the causality bias favor a causal coherence relation between the contents of the matrix and the relative clause. The resulting interpretation is similar to the interpretation of (1-b) where the relative pronoun *who* is replaced with *because he*.

Cases like these have not only attracted attention in theoretical linguistics but have also featured in psycholinguistic studies (Hoek et al., 2021; Rohde et al., 2011) where the interpretation of constructions like in (1-a) have been tested from the perspective of modularity (Fodor, 1983). In modular, as compared to interactive, processing models the processing of the phonological, morphological, syntactic and sematic features of phrases and sentences strictly precedes the assignment of coherence relations between them. In contrast to this hypothesis and in support of interactive architectures, empirical results discussed below showed that causal interpretation affects the incremental processing of the relative clause immediately.

Here we add to this line of work by taking a combined psycholinguistic and computational modeling perspective. Our contribution is twofold. Firstly, we identify another factor affecting the incremental processing of causal elicitures, besides the ones that have already been identified in previous research. In particular, we show that the specificity of the noun concepts that are involved affects incremental understanding of relative clauses as in (1-a), at least in German where verb information often follows rather than precedes the relative clause in comparable constructions. We discuss these results in relation to frame semantics (Fillmore, 1982) and predictive processing of coherence relations.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 276–285
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

Secondly, we show that critical effects related to the incremental assignment of coherence relations can be captured using surprisal calculated from the large language model (LLM) GPT-2, both in the German item set manipulating noun specificity and in an original item set manipulating the type of the relative clause, as well as its German translation. The second finding is partly inline with Surprisal Theory (Levy, 2008; Hale, 2001), which assumes that processing difficulty is determined by surprisal, but we also observe substantial deviation between surprisal and reading times. These findings are relevant regarding the questions how surprisal from LLMs is related to discourse processing architectures assumed in psycholinguistics and whether surprisal from LLMs suffices to capture sentence processing difficulty.

2. Previous studies on incremental processing of elicitures

In an ingenious design, Rohde et al. 2011 put interactive processing models to a tough test. They made use of known attachment preferences (formulated, e.g., in the Garden Path model of Frazier, 1987) of relative clauses in constructions such as (2) and tested whether these are modulated during online processing by causal expectations, induced, e.g, by certain types of verbs.

(2) a. John babysits the children of the musician who is generally arrogant and rude.
b. John babysits the children of the musician who are generally arrogant and rude.
c. John detests the children of the musician who is generally arrogant and rude.
d. John detests the children of the musician who are generally arrogant and rude.

According to the Garden Path model and supported by substantial empirical data, the relative clauses in (2) are preferably attached to the second, hierarchically lower NP, *the musician*, because this results in the simpler structure, as compared to high attachement to *the children*. The Garden Path model assumes modularity and, in particular, syntax-first processing (cf. Fodor, 1983; Chomsky, 1965) where initial syntactic parsing preferences are blind to other aspects of the sentence especially discourse-related effects. For these reasons, the plural disambiguation towards the dispreferred high attachment reading should cause processing difficulty. In contrast, to this hypothesis Rohde et al. 2011 adopted an interactive perspective (e.g. McRae et al., 1998; Tyler and Marslen-Wilson, 1977) and predicted that expectations for causal explanations, as induced, e.g., by the verb *detest*, can modulate parsing preferences incrementally. They contrasted implicit causality (IC) verbs like *detest*, which trigger strong discourse expectations for causal explanations involving the head NP, e.g. *the children*, with non-IC verbs like *babysit* that do not by themselves induce such explanations. The key idea is that the low-attachment preference is weakened to make room for a causal reading if an explanation is expected and high attachment can provide one. Rohde et al. 2011 thus predicted a smaller Garden Path effect on the number disambiguation in (2-c/d) than (2-a/b). This prediction was confirmed in a self-paced reading experiment providing evidence that causal coherence relations are inferred incrementally and may affect parsing preferences during online processing. This supports interactive processing models and challenges modular approaches like the Garden Path model.

Building on this design, Hoek et al. 2021 tested in an eyetracking during reading study whether causal inferences also modulate incremental processing of subsequent explicit causal relations as would be expected if they are drawn immediately, ruling out an explanation of the results of Rohde et al. 2011 based simply on the next-mention bias associated with IC-verbs (Bott and Solstad, 2023). They added a causal subordinate clause introduced by *because*, as in (3-a), in addition to the relative clause and examined whether causal expectations affect processing of the *because*-clause. In their study, causal expectations where modulated by the contents of the relative clause, which either functioned well as explanations as in (3-a/c) or not as in (3-b/d). If the relative clause is interpreted causally, integrating a second explanation introduced by *because* should be difficult because there arises a conflict between two potential causes in the discourse structure. They manipulated relative clause type (causal vs. neutral) and embedding connector (*because* vs. *so that*). While *because* introduces a causal explanation, *so that* introduces a consequential relation and should therefore not be sensitive to prior explanatory saturation.

(3) a. Diane fired the guy from the London office who was embezzling money because astoundingly he hired a stripper for the Christmas party.
b. Diane fired the guy from the London office who was here last month because astoundingly he hired a stripper for the Christmas party.
c. Diane fired the guy from the London office who was embezzling money and so astoundingly he hired a lawyer to sue the company.
d. Diane fired the guy from the London

office who was here last month and so astoundingly he hired a lawyer to sue the company.

The results revealed the critical interaction: with causal, but not consequential, subordinate clauses, causal relative clauses increased processing difficulty in the second subordinate clause. This effect was interpreted as evidence that an initial causal inference saturates the expectation for an explanation (in line with the Empty Slot Theory, Bott and Solstad, 2023), making a second explanation harder to integrate into the discourse model. Overall, these findings indicate that causal inferences, and elicitures more generally, are derived incrementally during online processing.

Hoek et al. 2021 found an early effect already on the adverb, e.g. *astoundingly*, directly following the connective. This effect was not particularly large in the measures directly related to first pass reading (they report $\beta = 59.40, SE = 29.52, t = 2.01, p < .05$ for regression path duration and a non-significant effect in first pass durations). The largest effect was found in total fixation durations and this could be driven by regressions initiated in the relative clause, which they did not analyze. That material within the subordinate clause is driving the effect is plausible since, when the connector is encountered, there are still ways to integrate another explanation (e.g. one that targets another level in the causal model). For example, the sentence in (4) provides a secondary explanation at another level without causing any conflict and this only becomes evident towards the end of the second subordinate clause. We will approach the question where in the sentence effects of causal inferences manifest in our surprisal-based simulations of the Hoek et al. 2021 data.

(4) Diane fired the guy from the London office who was embezzling money because obviously she hates bad money-management.

3. Previous work on LLM surprisal as a model of processing difficulty during reading

Surprisal theory (see Hale, 2001; Levy, 2008, for two different conceptions) equates processing difficulty with the contextual probability of an incoming linguistic unit, formally defined as the negative logarithm of its conditional probability, $-\log P(w_i|context)$. Lower-probability words incur higher processing costs, and difficulty is thought to be driven by to the reallocation of probability mass over competing structural analyses upon encountering new input. Empirically, surprisal correlates robustly and approximately linearly with reading times (e.g. Smith and Levy, 2013), capturing predictability effects and many aspects of syntactic ambiguity resolution (e.g. Arehalli et al., 2022).

Recent work using transformer-based language models (Vaswani et al., 2017) further refines this picture. While larger models typically achieve lower perplexity (i.e. better next-word prediction), their surprisal estimates may fit human processing measures worse, leaving more variance unexplained (Oh and Schuler, 2023). One explanation that has been proposed is that larger models utilize an unrealistic amount of knowledge in their next-word predictions which surpasses human capacities employed during real time processing. In fact, medium-sized models such as GPT-2 have been argued to occupy a sweet spot: they capture broad distributional regularities relevant to human expectations but not to an unrealistic degree, thus trading-off complexity and predictive accuracy. As a result, GPT-2–based surprisal seems to align more closely with human reading times than surprisal derived from substantially larger models. Human-like predictive behavior may depend not only on overall predictive accuracy but also on aspects like capacity constraints. For this reason we use GPT-2 in our simulations below.

At the same time, surprisal-based models sometimes underestimate the magnitude of processing disruption, especially in cases that have gained substantial attention in psycholinguistic studies. An important example are garden-path effects (e.g. Arehalli et al., 2022) where comprehenders are assumed to strongly commit to an initially preferred parse (e.g. the low attachment preferenmce discussed above). The reason may simply be that predicted difficulty at disambiguation points is attenuated because expectation-based models distribute probability mass across a larger set of analyses in parallel, relative to humans. Earlier commitments to an interpretation in humans vs. LLMs may be incentivized by the fact that, during conversation, language is not only parsed but also used in subsequent cognitive processes, involving, e.g., inference drawing or decision making. Another reason why processing disruption may sometimes be underestimated by surprisal estimates maybe that aspects related to memory-based costs (involved, e.g., in retrieval or reanalysis operations) contribute to the relatively large disambiguation penalties in humans (see, e.g., Oh and Schuler, 2022, for a hybrid approach incorporating such aspects).

Against this background, it is interesting to see whether the subtle coherence-related effects described in the previous section and also the data from our own self-paced reading study are captured well by LLM surprisal. This is for several reasons. Firstly, causal inferences may be part of the just-mentioned extra cognitive processes

that may not be reflected in surprisal and may encourage early processing commitments in humans. Causal elicitures are, moreover, thought to depend on general cognitive biases related to causal world and discourse models that need not be reflected in LLM surprisal. Finally, memory processes may be involved in coherence-related effects because at least two propositional contents need to be retrieved from or held in memory in order to assign a coherence relation. If this happens incrementally while a sentence is parsed, the memory burden may be especially large. Therefore, causal elicitures pose an interesting challenge to Surprisal Theory.

4. NP specificity and the design of the current study

In the present study, we carried out a self-paced reading experiment combining central ingredients from previous research with an additional crucial manipulation, namely the specificity of the NP concepts involved. The underlying idea is that more specific NP concepts foster stronger activation of scenarios linked to the causal expectations triggered by IC verbs (in the broadest sense). Verbs and their arguments are assumed to be associated with prototypical scenarios, called frames (Fillmore, 1982), and causal relations often arise within such frames. More specific NPs are linked to a smaller set of frames and thus causal expectations arising form within one of these frames may exert a larger effect as compared to less specific NPs, which are linked to a larger set of frames. Thus, just as verb type, relative clause type and connector type modulate effects related to causal inferences (see previous research section), the specificity of NP concepts may exert a comparable influence.

To test this, we adopted the items from Hoek et al. 2021 that contained *because*-clauses and added a manipulation of the NP specificity, as in (5-a/c) vs. (5-b/d) (all followed by (5-e)). The experiment was conducted in German, which resulted in translations with a different word order: the verb followed the NPs and the relative clause. Consequently, at the point of processing the relative clause, it is primarily the NP information that determines whether a causal interpretation is established. Following the verb, and parallel to the design of Hoek et al. 2021, a subordinate clause introduced by *weil* (because) was presented, and processing difficulty was measured in this clause. Across all experimental items, three regions of interest for self-paced reading were defined after the connector. These were: a pronoun, e.g. *er* (he), an adverbial, e.g. *plötzlich* (suddenly) and the rest of the clause.

(5) a. Die Vermieterin hat den Mieter, der
the landlady has the tenant who
letztes Jahr ihre Wohnung
last. year her apartment
überschwemmt hat, verklagt
flooded has sued
'The landlady sued the tenant who flooded her apartment last year'

b. Die Frau hat den Mann, der
the woman has the man who
letztes Jahr ihre Wohnung
last year her apartment
überschwemmt hat, verklagt
flooded has sued
'The woman sued the man who flooded her apartment last year '

c. Die Vermieterin hat den Mieter, der
the landlady has the tenant who
letztes Jahr eine Wohnung gesucht
last year a apartment sought
hat, verklagt
has sued
'The landlady sued the tenant who was looking for an apartment last year'

d. Die Frau hat den Mann, der
the woman has the man who
letztes Jahr eine Wohnung gesucht
last year a apartment sought
hat, verklagt
has sued
'The woman sued the man who was looking for an apartment last year'

e. ...weil er plötzlich anfing im
...because he suddenly started in_the
Gebäude zu rauchen.
building to smoke.
'...because he suddenly started smoking in the building.'

In addition to the self-paced reading study, we computed surprisal values from LLMs, in particular a German version of GPT-2, for the connector and the ROIs inside the *because*-clause. We conducted such a surprisal-based analysis of the English Hoek et al. 2021 items, their German translation and our German items with the NP specificity manipulations. This served three purposes. First, we asked whether the coherence-related effects under investigation are reflected in surprisal, which constitutes an interesting question in its own right (see discussion in the previous section). Second, we aimed to assess whether the effects observed in the original items from Hoek et al. 2021 would also emerge in our translated German materials, potentially reflecting the manipulation of NP specificity. Third, we sought to further localize the source of the effects by examining where within the relative clauses the

relevant differences arise. Since effects in the Hoek et al. 2021 study were mainly driven by total fixation duration, it is possible that effects in surprisal, which have been linked to first pass reading (Smith and Levy, 2013; Shain et al., 2024), only show up relatively late within the relative clause. This would be in line with the idea that integration is only possible and difficulty of the integration process is only determined after sufficient information about the content of the *because* clause has been processed, as in example (4). Following the the chronological order in the study, we report the surprisal-based analyses first and then present the reading time experiment.

5. Methods

5.1. Surprisal Simulation

Surprisal was computed using the *llm-surprisal*[1] tool with the pretrained models *gpt2-large*[2] (774M paramteres) and *german-gpt2-larger*[3] (137M parameters) with the default temperature setting. For each item, token-level surprisal values (negative log-probabilities) were extracted and summed within each predefined region of interest (ROI). Thus, the reported values reflect cumulative surprisal per region rather than mean per-token surprisal. Under a surprisal-based account of incremental processing, cumulative surprisal within an ROI is predicted to be proportional to processing difficulty.

5.2. Self-paced reading experiment

5.2.1. Participants and procedure

100 participants were recruited over prolific.co and redirected to `pavlovia.org`, where the experiment started automatically in full-screen mode. Participants were distributed randomly to one of four lists. They read sentences in a self-paced reading paradigm using the moving-window technique (Just and Carpenter, 1976). The latency between key presses served as the reading time measure. After the final word of a sentence, a comprehension task followed. Responses were given using the left arrow key ("yes") or the right arrow key ("no"). The next trial then commenced automatically. Participants gave informed consent.

The experiment began with a brief introduction and detailed instructions. Participants were explicitly informed that the comprehension questions following each sentence had a single unambiguous correct answer and that no ambiguity was intended. They were further advised that previously presented words could not be revisited. At the end of the experiment, participants were debriefed and provided with a completion link to confirm participation. The total duration of the experiment was approximately 15–20 minutes. Participant received 5 € compensation.

5.2.2. Materials

We constructed 28 items by translating the conditions with *because*-clauses from Hoek et al. 2021 into German. In doing so, several adjustments were necessary. First, we modified the structure of the embedded clause, as the succession of elements differs between English and German due to word order differences. We ensured that, following the connective, each item contains (i) a pronoun (e.g., *er*), (ii) an adverbial (e.g., *offensichtlich*), and (iii) a longer phrase, which constitutes the third region of interest (ROI). Second, the word order in the matrix clause differs from the original English examples. In our German materials, the structure is as follows: first NP, second NP, relative clause, and finally the verb. Third, we systematically adjusted the specificity of the noun phrases. Since specificity is not fully consistent in the original items, we established a constant level of specificity in our base versions (role nouns as in (5)). Based on these translations, we constructed an additional set of conditions in which the noun phrases are less specific (either generic nouns like *man* and *woman* in (5) or proper names). If the specificity manipulation could not be carried out in a satisfactory way in a given item, it was removed from the list and replaced by an item that was generated from scratch (7 in total). An example item is provided above in (5). Overall, this yielded a 2 × 2 factorial design (Relative Clause Type × NP Specificity) with both factors manipulated within items and within participants. The 28 experimental items were distributed over four lists according to an Latin square. They were accompanied in each list by 27 filler items. In addition, 32 items from a related experiment were added which contained also causal inferences but of a different kind, namely the causal interpretation of a adjectival modification (*The drunk/young driver had an accident*).

5.3. Statistical analyses

Surprisal values and reading times were both analyzed using linear mixed-effects models.[4] These

[1] https://github.com/tmalsburg/llm_surprisal

[2] https://huggingface.co/openai-community/gpt2-large

[3] https://huggingface.co/stefan-it/german-gpt2-larger

[4] Data and analysis scripts will be made available at https://osf.io/rckpt/overview?view_only=d99e4d5aedea40698790000efc52494f

models were fit using the lme4 package (Bates et al., 2015) for the statistical computing software R. Besides fixed effects for the experimental manipulations and their interaction, they included random intercepts for items and (if applicable) participants as well as random slopes for the experimental manipulations (without interaction) provided this allowed for convergence and did not lead to overfitting. We analyzed the four ROIs in the causal (*because*) or consequential (*and so*/*woraufhin*) clause (i.e. the connector, pronoun, adverb and rest) and for the simulation study we also analyzed the cumulative surprisal in all four ROIs if no effects in the individual ROIs were found. We computed one-tailed tests (with $\alpha = .05$), as we tested for a directed hypothesis, i.e. a specific type of interaction consisting in reduced processing disruption in *because*-clauses after conditions with restrictive relative clauses assumed to weaken causal inferences. In addition to the model specifications, all summaries of statistical models in all ROIs with significant effects and corresponding plots are provided in the accompanying repositories.

For the self-paced reading experiment, the data were cleaned in three steps. First, the entire distribution of reading times for each participant were inspected visually to spot unusual behavior. This led to the exclusion of one data set which seemed to be of non-human origin because it showed a multimodal RT distribution with five modes almost constantly across the experiment. Second, participants were removed that answered more than 5 out of 20 designated control questions in filler trials incorrectly. This affected not a single participant. Second, trials with RT shorter than 200 ms or more than 2.5 standard deviations above the mean in each ROI were removed. This affected 9.2% of the data.

6. Results

6.1. Surprisal simulation

Analyses of the surprisal values revealed the critical interaction in each of the three item sets. In the original item set from Hoek et al. 2021 (cf. example (3)), the critical interaction was found immediately on the connector *because* / *so that* ($t = -1.905, p = .033$). In that region, restrictive relative clauses led to higher surprisal than causal ones, but this difference was smaller for causal vs. consequential continuations (see Figure 1). The critical interaction was, furthermore, marginal in the last ROI containing all the words after the pronoun ($t = -1.487, p = .071$). In that region, we observed a cross-over pattern where causal relative clauses led to numerically higher surprisal in causal continuations than restrictive ones, but the opposite pattern was observed for consequential continuations.[5]

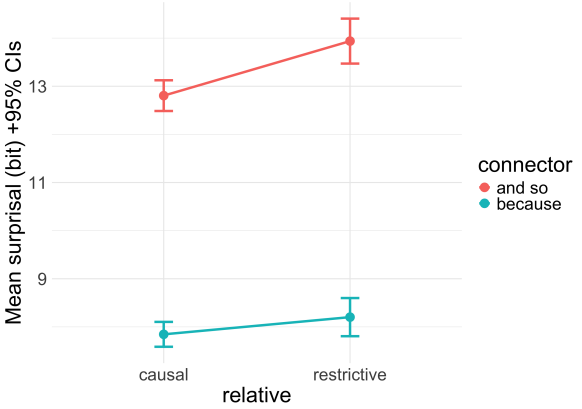


Figure 1: Surprisal in the connector ROI of the Hoek et al. 2021 items.

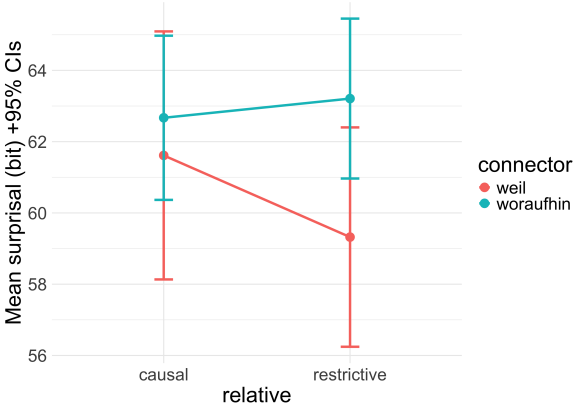


Figure 2: Cumulative surprisal over the entire subordinate continuation clause in the German translation of the Hoek et al. 2021 items. *weil*: 'because'; *woraufhin*: 'and so'.

In the German translation of the item set from Hoek et al. 2021, the critical interaction was not significant in any of the individual ROIs. However, when the connector and the three following regions were analyzed together and cumulative surprisal was computed across all four regions, the predicted interaction emerged ($t = -1.89, p = .034$; see Figure 2). The interaction was due to a crossover pattern where causal relative clauses led to lower surprisal in consequential continuations but higher values in causal ones.

The simulation of the NP Specificity manipulation

[5] A first surprisal analysis of these items was presented in the M.A.-thesis of the second author (Barth, 2025). This analysis was based on a smaller version of gpt-2[6] (with 124M parameters) and it only revealed the critical interaction in cumulative surprisal values over the entire subordinate clauses. In that analysis causal relative clauses led to numerically lower surprisal in consequential continuations but higher values in causal ones.

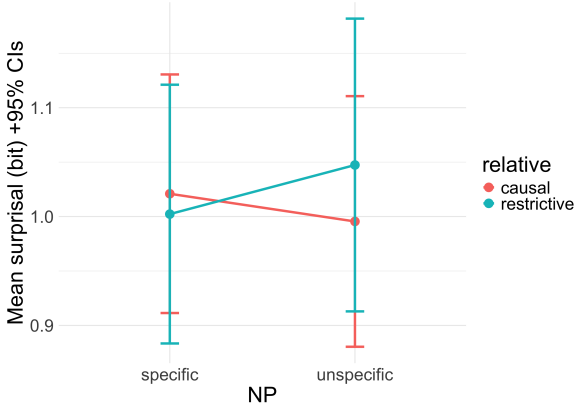


Figure 3: Surprisal in the ROI containing the pronoun for the NP-specificity effect in German.

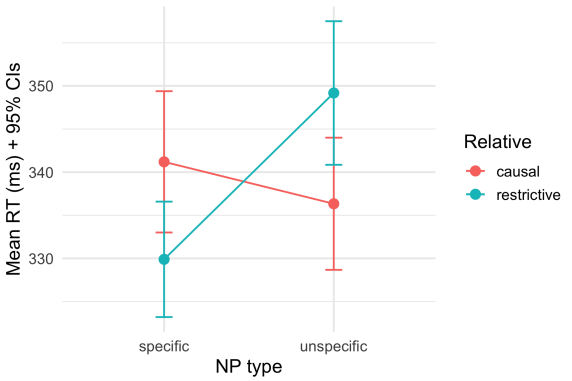


Figure 4: The NP specificity effect in German. Mean reading times and 95% CIs for the adverbial region in the *because* clause.

revealed the critical interaction in the ROI containing the pronoun (see Fig 3; $t = 2.603, p = .007$). In this region, specific NPs led to higher surprisal in combination with causal as compared to restrictive relative clauses whereas the opposite pattern was observed for unspecific NPs. The same numerical pattern was already present on the connector (*because*) preceding the pronoun, but it did not lead to a significant interaction effect in that ROI ($t = 1.583, p = .063$).

6.2. Self-paced reading

Significant effects in the reading time analysis were restricted on the second spillover region, i.e., the adverbial, for example, *plötzlich* ('suddenly'). In this region, the predicted interaction between NP Specificity and the Relative Clause Type, restrictive vs. causal, was significant ($t = 1.702, p = 0.045$). For the unspecific NPs, restrictive relative clauses led to longer reading times than causal ones, whereas the opposite was observed for the specific NPs (see Figure 4). No other effects were significant (all $|t| < 1$).

7. Discussion

In the present study, we built on previous research on causal inferences (Rohde et al., 2011; Hoek et al., 2021), which have been termed conversational elicitures in theoretical linguistics (Cohen and Kehler, 2021) and identified as a central type of inference in natural language. Psycholinguistic evidence has shown that such causal inferences are derived incrementally during online processing and affect the interpretation of upcoming material. In particular, once a causal explanation has been inferred, additional explicit explanations introduced by *because*-clauses become more difficult to integrate, as reflected in increased reading times. Extending these findings, we manipulated the specificity of the noun concepts involved. We observed in our self-paced reading study that NP specificity modulated the difficulty of integrating an upcoming *because*-clause.

We interpret this effect within a frame-semantic perspective (Fillmore, 1982). On this view, verbs and also the noun phrases filling their thematic roles are associated with prototypical scenarios, or frames. More specific noun phrases (e.g., *landlord*) evoke a narrower and more strongly constrained set of such scenarios than less specific noun phrases (e.g., *woman*). Since causal inferences crucially rely on these scenarios as an interface between linguistic and world knowledge, stronger scenario activation should amplify causal expectations. Accordingly, when a specific NP strongly activates a limited set of frames, additional causal material may lead to increased integration costs due to expectation saturation.

Our results show that NP specificity modulates the difficulty of integrating further causal explanations in a systematic way. What was, however, unexpected in the reading-time data was the observation that, in combination with unspecific noun phrases, the causal relative clause lead to faster integration of the *because*-clause than did the restrictive relative clause. One explanation for this observation could be that the cue provided by unspecific noun phrases may be too weak to trigger a strong causal inference. Given the general bias toward establishing causal coherence relations, the explicit *because*-clause may therefore enhance discourse felicity and be processed relatively easily, as it satisfies an independently strong expectation for causal coherence.

Turning to the surprisal-based modeling, we would like to highlight and discuss two central observations. First, across all simulations we conducted, i.e. the original dataset from Hoek et al. 2021, the German translations of these items, and our newly

constructed materials manipulating NP specificity, the critical interaction that has been taken as evidence for causal inferences affecting incremental processing was found. In this respect, LLM-based surprisal provides a useful model of reading behavior in a theoretically important domain. Causal inferences are particularly relevant here because, unlike phenomena like core syntactic parsing, semantic processing or certain pragmatic inferences, they operate at the interface between linguistic and world knowledge. The fact that LLM surprisal captures the relevant interaction suggests that predictive language models are sensitive to discourse-level regularities of this kind.

Second, despite the observed convergence, systematic discrepancies between human reading times and surprisal-based estimates were observed. In particular, there was a cross-over pattern in the reading time effects reported by (Hoek et al., 2021) that was absent from our surprisal estimates. In their data, the conditions with causal continuations led to a slow down when they contained causal as compared to restrictive relative clauses whereas the opposite effect was observed in conditions with consequential continuations. In addition, the critical effects were observed one region earlier in our suprisal estimates than in the reading time data sets. This is not particularly surprising given that surprisal has been shown empirically to affect reading behavior in spillover regions (e.g. Smith and Levy, 2013). One possible explanation for discrepancies in the pattern of effects could thus be that effects on reading behavior are determined by the surprisal of the previous word as well as processes directly related to the current region. In addition, the phenomena under investigation may recruit cognitive resources that go beyond predictive processing. In particular, memory-related processes such as maintaining previously processed sentence material and linking it to scenarios retrieved from long-term memory may contribute to integration costs in ways that are not directly reflected in surprisal (e.g. Oh and Schuler, 2022). Moreover, human reading behavior may also be shaped by adaptation effects arising within an experimental setting. Participants are repeatedly exposed to structurally and semantically similar sentences, which may lead to adaptation over the course of the experiment, as shown in previous research (Fine et al., 2010). By contrast, surprisal is computed independently for each item, without any possibility of adaptation.

Moreover, our surprisal estimates suggest that some of the effects in the study of Hoek et al. 2021 may not be strictly local. While they reported relatively early effects, the largest effects they found were in total fixation duration. Surprisal estimates provided some indication that part of this effect may be driven by surprisal values of the final region of the sentence. This is also corroborated by the results in our German translation of their items, where no reliable effects emerged in the individual ROIs but the critical interaction only emerged when surprisal was summed across the entire subordinate continuation clauses. In this case, the system appears to accumulate information across the clause before integration difficulty becomes fully manifest. From a theoretical perspective, this gradual accumulation is plausible, as the incompatibility between competing cues may only become evident once sufficient propositional content has been processed (cf. the discussion of example (4) above).

Our data also provide some information regarding discussions about the type of LLM-surprisal that is best suited for capturing reading time effects. In analyses based on reading time corpora, Oh and Schuler 2023 found that smaller models, e.g. smaller GPT-2 models, provide the best fit for reading time data. In our study, we used smaller models for the German than for the English items. In the second simulation, i.e. in the simulation of our German translation of the items from Hoek et al. 2021, the smaller models only revealed the critical interaction in cumulative surprisal across the entire subordinate continuation clause. The same general pattern was also found in previous analyses conducted in the context of the M.A. thesis of the second author (Barth, 2025), who used a similar-sized model for the original English items and also found the critical interaction only in cumulative surprisal values (cf. the analyses provided in the accompanying repository). In contrast, the larger GPT-2 model revealed early effects already on the connector, in line with the study of Hoek et al. 2021. One possibility is that with regard to the type of discourse-related effects we studied here an increase in model size from the smaller GPT-2 variants may yield a closer match to reading time data.

Besides model size, another dimension that has been discussed in the literature is the layer from which surprisal is extracted. All our analyses as well as the systematic comparison by Oh and Schuler 2023 were based on surprisal from the final layer of LLMs. In a recent study, Kuribayashi et al. 2025 showed, however, that surprisal extracted from internal layers of larger models captures data from reading time corpora as well as surprisal from the final layer of smaller variants or even better.[7] An interesting question for future research would be whether this also holds for the type of discourse-related effects studied here.

With regard to the debate surrounding modular vs. interactive psycholinguistic processing models, our data are somewhat ambiguous. While they

[7] We would like to thank an anonymous reviewer for pointing this out to us.

show that effects related to discourse coherence affect incremental language processing, in line with interactive models, we also found some indication of effects that emerge relatively late in the relevant *because*-clauses, as would be expected from a modular perspective. We would like to propose though that our data are more compatible with a nuanced interactive perspective where effects due to the integration of causal relations into the discourse model show up immediately but are still modulated by information later on, reflecting the vast space of potential unfolding discourse structures and resulting in no immediate effects already on the connective.

8. Acknowledgements

We would like to thank three anonymous reviewers for helpful comments on an earlier version of this paper. We also thank Irene Rapp, Stefan Engelberg, and Robin Hörnig for helpful comments and discussion on the M.A. project of Raphael Barth on which the paper is based. FS received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 538184825.

9. Bibliographical References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL 2022)*. Association for Computational Linguistics.

Raphael Barth. 2025. Psycholinguistische Untersuchung des Einflusses von Nominalphrasenkonzepten auf die kausale Inferenzbildung. Master's thesis, University of Tübingen, Tübingen, Germany.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Oliver Bott and Torgrim Solstad. 2023. The production of referring expressions is influenced by the likelihood of next mention. *Quarterly Journal of Experimental Psychology*, 76(10):2256–2284.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Jonathan Cohen and Andrew Kehler. 2021. Conversational eliciture. *Philosophers Imprint*, 21(12):1–26.

Charles Fillmore. 1982. Frame semantics. *Linguistics in the Morning Calm. Seoul*, pages 111–137.

Alex Fine, Ting Qian, T. Florian Jaeger, and Robert Jacobs. 2010. Syntactic adaptation in language comprehension. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics*, pages 18–26, Uppsala, Sweden. Association for Computational Linguistics.

Jerry A Fodor. 1983. *The modularity of mind*. MIT press, Cambridge, MA.

Lyn Frazier. 1987. Sentence processing: A tutorial review. In Max Coltheart, editor, *Attention and Performance 12: The Psychology of Reading*, pages 559–586. Lawrence Erlbaum Associates.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American chapter of the Association for Computational Linguistics*.

Jet Hoek, Hannah Rohde, Jacqueline Evers-Vermeul, and Ted J.M. Sanders. 2021. Expectations from relative clauses: Real-time coherence updates in discourse processing. *Cognition*, 210:104581.

Marcel Adam Just and Patricia A Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 8(4):441–480.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13:1743–1766.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in online sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Hannah Rohde, Roger Levy, and Andrew Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Nathaniel J. Smith and R. Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–19.

Lorraine K. Tyler and William D. Marslen-Wilson. 1977. The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior*, 16(6):683–692.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Topic-context Dependency on Continuous Semantic Reconstruction of Language from fMRI Signals

Fermin Travi[*,†], **Agustín Delmagro**[*], **Diego Fernández Slezak**[*,†], **Bruno Bianchi**[*,†], **Juan E Kamienkowski**[*,†,‡]

[*]Instituto en Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

[†]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

[‡]Maestría de Explotación de Datos y Descubrimiento del Conocimiento, Universidad de Buenos Aires

ftravi, bbianchi, juank@dc.uba.ar

Abstract

Recent advances in neural language decoding have enabled reconstruction of perceived speech from fMRI signals using large language models and brain encoders. While these systems achieve impressive semantic fidelity, the extent to which training data biases signal encoding and the resulting decoded output remains unclear. Here, we investigate how topic-specific training constrains language decoding by partitioning 84 auditory stories into two semantic clusters and training brain encoders on topic-specific subsets across three participants. We find that encoders trained exclusively on one topic systematically bias decoded outputs toward their training domain: family-related stories are reclassified as such when reconstructed with family-trained encoders, but not otherwise. Moreover, topic-constrained encoders perform significantly above chance only within their training topic and fall below random baseline when decoding out-of-topic stimuli. These findings underscore the need for more diverse and semantically rich training data, while also raising questions about the breadth of semantic variability that current brain encoders can effectively capture.

Keywords: language, fMRI, voxelwise encoding

1. Introduction

The convergence of increasingly powerful large language models (LLMs) with growing fMRI data availability has enabled the development of neural language decoding systems that reconstruct perceived or imagined speech from brain activity (Tang et al., 2023; Antonello and Huth, 2024; Antonello et al., 2023). Broadly, the architecture of these decoders typically involves training language-to-BOLD signal encoders as voxelwise linear regressions on extensive, subject-specific datasets. During decoding, an LLM generates plausible text continuations, which are then encoded to predicted voxel activity; the continuation whose predicted activity best matches the observed brain signal is selected as the decoded output. Remarkably, decoded texts often capture the semantic content of stimuli with high fidelity (Tang et al., 2023). These systems have advanced both language neuroscience (by enabling more detailed accounts of how language is represented in the brain (Antonello et al., 2021; Chen et al., 2024, 2025; Gwilliams et al., 2025)) and natural language processing (by improving and deepening our understanding of how LLMs process language (Toneva and Wehbe, 2019; Zhou et al., 2024)), exemplifying the productive interplay between AI and neuroscience.

Despite this success, an important question remains: to what extent does the semantic content represented in the stimuli systematically bias signal encoding and its downstream decoding? While recent work by Chen et al. (2025) has examined topic-related functional specialization in the brain (demonstrating that different brain regions show preferential activation for distinct topics), this focuses on neural organization rather than on how training constraints shape the encoder's interpretation of brain signals. Here, we directly address how semantic biases in training data propagate through to decoded outputs. We partition stimuli into two distinct topics using a largely unsupervised approach, then train brain encoders on topic-specific subsets and analyze the resulting biases in decoded text. Our findings reveal that topic-constrained training produces decoders that systematically reconstruct stimuli as belonging to their training topic, while performing below chance on stimuli from other semantic domains. These results demonstrate that language encoders can become over-specialized to narrow semantic domains during training, systematically biasing decoded outputs toward those domains. This finding has important implications for both the reliability of brain-computer interface applications and the interpretations of language processing results.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 286–292
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

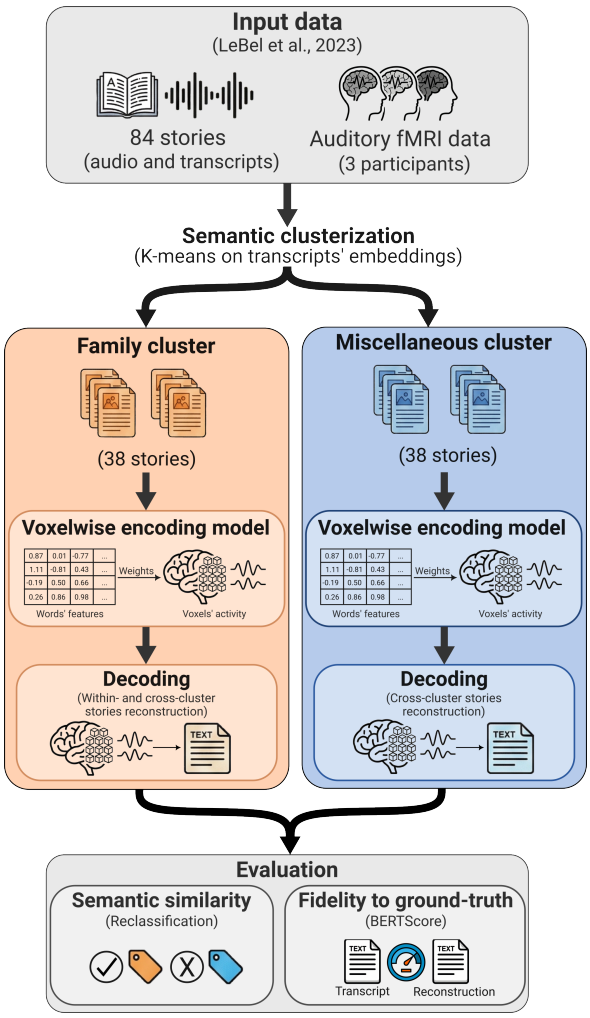


Figure 1: Methodological pipeline for topic-specific language decoding. Narrative stories are partitioned via semantic clusterization into Family (orange) and Miscellaneous (blue) topics. For each cluster, voxelwise encoding models (GPT-2 based) are trained and used to reconstruct stories from brain activity. Reconstructions are evaluated for thematic consistency and linguistic fidelity against ground-truth transcripts using BERTScore.

2. Methods

To evaluate how semantic topics influence auditory fMRI signal encoding and downstream language decoding, we conducted a semantic classification on the 84 stories from the LeBel et al. (2023) dataset, focusing on the data from three participants who listened to all the stories. First, we divided these stories into two semantic clusters (*Family* and *Miscellaneous*). Then, we trained topic-specific voxelwise encoding models to map the text stimuli to the observed brain activity. The decoding process then used these same encoding models to reconstruct the held-out stories: candidate text was generated, mapped to predicted voxel activity via the encoding model, and compared to the observed fMRI signal to select the most likely continuation (Tang et al., 2023). We evaluated the resulting reconstructions based on their semantic similarity to their original cluster and their fidelity to the ground-truth transcript (Fig. 1).

2.1. Dataset

The data analysed pertain to BOLD fMRI responses from the dataset released by LeBel et al. (2023), where participants listened to natural, narrative stories extracted from the podcasts The Moth Radio Hour (77 stories) and Modern Love (7 stories) (LeBel et al., 2023). We focused our analysis on the three participants (S1, S2, and S3) that listened to all 84 stories (over 16 hours). These stories possess up to five different tags that describe their content, with the exception of five stories that contain no tags. These were tagged by providing their transcript and the set of available tags to Gemini 2.5 Pro. The most frequent tag is 'Family' (30 stories), followed by 'Relationships' (21 stories), 'Health/Medicine' (17 stories) and 'Jobs/Employment' (16 stories). We kept only the most frequent tag to describe the main topic of each story (Fig 2A), combining 'Health/Medicine' and 'Jobs/Employment' into 'Health/Jobs' and assigning 'Other' to those stories that did not possess any frequent tag.

2.2. Semantic Clusterization

To classify the stories according to their semantic content in a largely unsupervised manner, we made use of OpenAI's `text-embedding-3-small` embedding model to transform the stories' transcripts into 1536 dimensional text embeddings. These were then used as input to a K-Means algorithm (K=2, scikit-learn v1.7.0).

2.3. Voxelwise encoding models

We made use of the code base released by LeBel et al. (2023) to train the voxelwise encoding models, albeit with some changes. Instead of employing GPT-1 (Radford et al., 2018) as the base LLM of the encoding model, we made use of GPT-2 Small (Radford et al., 2019) (approximately 124 million parameters) fine-tuned to use word level tokenization[1], as it was shown to perform more similarly to humans (Vaidya et al., 2023). The encoding model hyperparameters remained unchanged (layer nine was used for extracting text features, with a context window of five words). For within-cluster evaluation, we performed leave-one-out cross-validation

[1] https://github.com/HuthLab/lm-repeating-text

to maximize the amount of training data in the encoding model. Training time for each fold averaged four hours, while decoding time of each story averaged two hours. This was performed on an Intel Core i7-11700 CPU, NVIDIA GeForce RTX 3060 12GB vRAM GPU, and 64GB DDR4 memory.

2.4. Decoding stories from fMRI signals

Decoding stories from fMRI consists of two stages (Tang et al., 2023). First, a voxelwise encoding model is trained to map a text stimulus to predicted brain activity, as described above. Second, this same encoding model is used to decode the held-out stories: at each time step, the language model generates up to 200 possible text continuations (beginning with a fixed set of start words). Repetitions are filtered out of these extension words, to which we added the filtering of the unknown token to prevent it from derailing text generation. Each candidate continuation is then encoded into predicted voxel activity using the trained encoding model, and the candidate whose predicted activity best matches the observed fMRI signal is selected. This process is repeated iteratively to reconstruct the full story. A random model is defined by replacing the encoder model with a random weighing of the possible continuations.

The resulting reconstructions are then evaluated by recategorizing them into the previously defined semantic clusters and by comparing them against the ground-truth transcription. The latter is done in each time window by means of BERTScore (Zhang et al., 2020) using `microsoft/deberta-xlarge-mnli`, as it tops the benchmarks in human evaluation. BERTScore is an automatic evaluation metric that measures the similarity between generated text and reference text using pre-trained contextual embeddings. Instead of relying on exact string matching, it computes the sum of maximum pairwise cosine similarities between the token embeddings of the candidate and reference sentences. By employing greedy matching, each token is aligned to its most similar counterpart in the other sentence to compute precision, recall, and an F1 measure. An inverse document frequency (idf) file is provided in the dataset for evaluation, containing 1000 sentences for each text. These values are incorporated to apply importance weighting to token matches. However, two of the stories are missing ("From Boyhood to Parenthood" and "Where there's smoke"), so they were left out of the evaluation.

3. Results

The stories' classification into topics was primarily unsupervised. First, text embeddings were computed from the stories' transcripts to generate lower-dimensional semantic descriptions. Subsequently, unsupervised clustering (K-means, K=2; silhouette score 0.37) was applied to these embeddings, dividing the stories into two topics. One semantic cluster, designated *Family,* predominantly contains family-related narratives, while the other, *Miscellaneous,* comprises a mixed set of stories, as indicated by their primary tags (Table 1 and Fig 2A). To ensure equal cluster size, eight randomly selected stories from the Miscellaneous cluster were excluded. Word clouds (Fig 2B) of the main content words for each cluster further support this distinction, clearly showing a high frequency of kinship terms (e.g., "father," "mother," "dad," "daughter") in the Family cluster and mixed terms in the other.

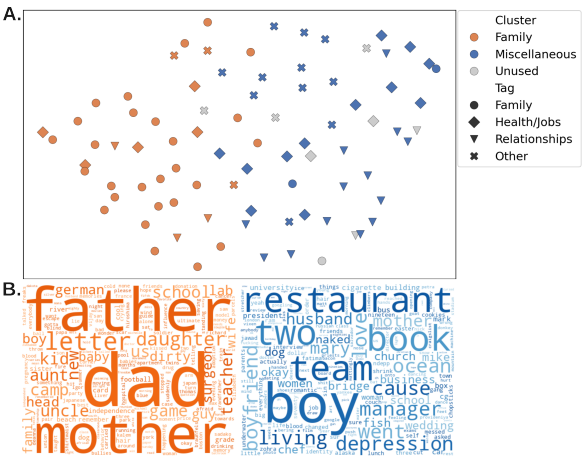


Figure 2: **A.** UMAP components of the stories' embeddings, clusterized with K-means (K=2; 0.37 silhouette score). The most frequent tag was selected for each story, as they usually possess more than one. **B.** Word cloud of the Family stories cluster (left) and Miscellaneous stories cluster (right).

Using this division, we trained voxelwise encoders (Tang et al., 2023) for all three participants on the Family stories using leave-one-out cross-validation (37 stories per training fold). This yielded 114 individual encoding models (38 per participant), plus six models trained on entire clusters (two per participant). These encoders were then used to decode either the held-out story (within-topic decoding) or the full held-out cluster (cross-topic decoding). We evaluated the resulting transcriptions in two ways: their semantic similarity to the original cluster and their fidelity to the ground-truth story transcription (see Section 2.4).

To assess semantic similarity, we computed text embeddings of the decoded transcriptions and clustered them using the previously fitted K-means algorithm. When decoded with the Family-trained encoder, 36 of 38 Family stories remained in the same cluster for at least one participant. In contrast, when using an encoder trained exclusively on

	Family cluster	Miscellaneous cluster
Number of stories	38	38
Total words (average)	68,293 (1,797)	74,531 (1,961)
Total unique words	6,081	6,644
Most common tags	Family (27), Health (10)	Relationships (13), Jobs (11)
Shared unique words	51% (3,108)	47% (3,108)

Table 1: Descriptive statistics of the stories in each semantic cluster.

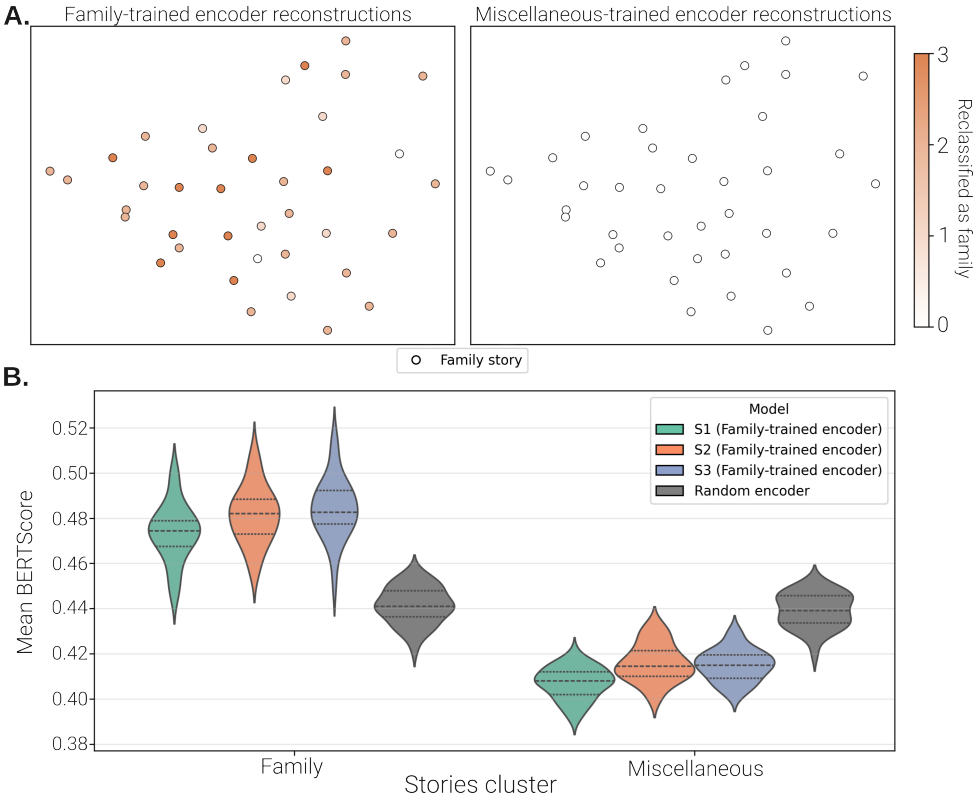


Figure 3: **A.** Clustering of the reconstructed Family stories using an encoder trained only on Family stories (left) and one trained only on Miscellaneous stories (right). Color intensity corresponds to how many reconstructions (of all three participants) have been clusterized as Family. **B.** Reconstruction performance as measured with BERTScore on both the Family and Miscellaneous stories using the Family encoder of all three participants. In grey, the performance of a random encoder is depicted.

Miscellaneous stories, all reconstructions shifted from the Family cluster to the Miscellaneous cluster (Table 2 and Fig 3A). This reveals a strong dependency on training topics: encoders are biased to interpret fMRI signals as referring to their specific semantic training context.

Encoder	Ground-truth stories cluster	Reconstructed Family	Reconstructed Misc.
Family-trained	Family	76	38
	Miscellaneous	0	114
Misc.-trained	Family	0	114

Table 2: Stories categorization after being reconstructed with Family-trained and Miscellaneous-trained encoders.

To assess reconstruction fidelity, we computed BERTScore (Zhang et al., 2020) between ground-truth transcriptions and decoded text in each time window, using a random encoder as baseline (Tang et al., 2023). Figure 3B shows average BERTScore for each story across participants when using the Family-trained encoder to reconstruct both within-cluster (Family) and cross-cluster (Miscellaneous) stories. The Family-trained encoder significantly outperformed the random encoder for Family stories (mean BERTScore: 0.47, 0.48, 0.48 ± 0.13 for S1, S2, S3, respectively, vs. 0.44 ± 0.01 for random; $p < 0.001$, Wilcoxon rank-sum test), but performed significantly worse on Miscellaneous cluster stories (mean: 0.41, 0.42, 0.42 ± 0.01 vs. 0.44 ± 0.01; $p < 0.001$, Wilcoxon rank-sum test). These results demonstrate limited generalization of brain encoders trained on specific topics, and

caution against overinterpreting their predictions when applied to out-of-domain fMRI signals.

Finally, we examined the most prominent brain regions ranked by Pearson correlation between Family-trained encoder predictions and ground-truth fMRI signals from Family stories, averaged across stories for each participant (Fig. 4). To achieve this, we kept the resulting top 10% of voxels and parcelled them into ROIs using Schaefer's atlas (2mm³, 400 ROIs) (Schaefer et al., 2018). A strong dominance from the left hemisphere can be observed in S3 (who yielded the highest correlation), particularly in brain regions associated with the language and speech perception networks (Fedorenko et al., 2024; Damera et al., 2023), encompassed in the atlas by the Somatomor and temporal Default mode networks (Table 3). This is consistent with prior work on language decoding from fMRI signals (Tang et al., 2023), and also provides a more nuanced understanding of the functional aspects involved. The average 90th percentile correlation was 0.16 (std 0.02; 8,113 voxels) for S1, 0.18 (std 0.03; 9,426 voxels) for S2, and 0.20 (std 0.05; 9,556 voxels) for S3.

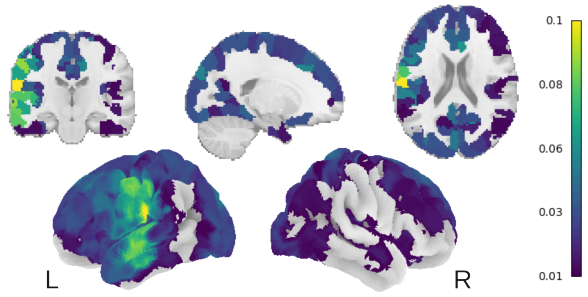


Figure 4: Mapping to Schaefer's atlas (Schaefer et al., 2018) with 400 ROIs of the 10% voxels with the highest correlation between its activity and the estimation of the Family encoder on Family stories in S3.

Network	Hem.	N°	Corr.
Somatomotor	L	11	0.11
Somatomotor	L	16	0.09
Default Mode Temporal	L	8	0.09
Somatomotor	L	15	0.09
Somatomotor	L	13	0.08
Default Mode Temporal	L	4	0.08
Default Mode Temporal	L	2	0.08
Somatomotor	L	2	0.08
Somatomotor	L	12	0.07
Somatomotor	L	1	0.07

Table 3: ROIs with the highest correlation in S3. Hem. refers to hemisphere, and N° refers to the ROI number.

4. Discussion

In this work, we explored how specific semantic domains influence the decoding of language from fMRI signals. Voxelwise encoding models are becoming increasingly more prominent in the field (Dupré la Tour et al., 2025), helping to unveil localized semantic coding in the brain (Chen et al., 2025) and to leverage data-driven models for testing of language neuroscience hypotheses (Antonello et al., 2023). Conversely, alignment of LLMs to predictive models of brain signals is proving to be a productive endeavour, as language processing performance of the former improves in accordance to its alignment with the latter (Toneva and Wehbe, 2019; Zhou et al., 2024). As such, understanding the potential biases, limitations, and mitigation strategies of these tools becomes essential. Our results reveal a strong dependence on the topics included in their training, where narrowing them to specific contexts (e.g., "Family") forces the interpretation of brain activity as consistently relating to those domains, resulting in degraded performance on out-of-domain content. Even though the case studied here is taken to an extreme, it serves to highlight an aspect that has generally been overlooked in the field: the potential biases that may incur if the semantic domains employed are not diverse enough.

It is a well-established principle in machine learning that divergence between training and test data yields degraded performance. In the standard case, a model trained on a narrow distribution generalizes poorly to unseen domains, producing outputs that are simply less accurate or less confident (Quiñonero-Candela et al., 2022). In the results presented here, however, the effect is qualitatively different: topic-constrained encoding models do not merely fail to generalize, but actively reconstruct out-of-domain stimuli in the direction of their training domain, with out-of-topic performance falling below the random baseline. This points to a form of domain bias that is stronger than a simple loss of generalization: the model does not become uninformative in the face of unfamiliar input, but rather imposes a systematic distortion, pulling reconstructions toward a familiar but incorrect semantic space. This distinction matters because it implies that deploying narrowly trained encoding models on ecologically diverse stimuli could introduce structured errors that are harder to detect than mere noise and, potentially, harder to correct for.

What constitutes a diverse enough semantic domain, however, remains an open question. A related consideration raised by this work is how generalizable these tools can effectively be. Given that BOLD signals are inherently slow and spatially coarse, voxelwise encoding models may face in-

herent constraints on the breadth of semantic diversity they can capture. Moving forward, a stronger emphasis should be made both to the variety and lexical richness of stimuli used during data collection and to systematic evaluation of cross-domain generalization.

Limitations

Some of the limitations on this work stem from the usage of a previously published dataset, which was not specifically designed to test the hypothesis inquired here. In particular, topics are not clearly divided, nor balanced. The semantic cluster Miscellaneous comprises stories with diverse topics, instead of a single one, which would have helped to contrast the effects of training in one specific semantic domain versus another. Another limitation pertains to the amount of training data: dividing stories in two independent sets implies halving the number of samples, yielding a substantial loss in performance. Even though we have tried to maximize the amount of training data by performing leave-one-out cross validation, it involved a great cost in compute time (nearly 2,000 hours or 82 days).

Ethics statement

This study utilizes a previously published, de-identified fMRI dataset. As the data was collected under protocols approved by the original host institution's Institutional Review Board (IRB) and is publicly available for research use, this work did not require additional ethics committee approval.

5. Acknowledgements

J.E.K. received research grants from CONICET (PIP 11220220100240CO). F.T., B.B., D.F.S., and J.E.K. were funded by the CONICET and UBA.

6. Bibliographical References

Richard Antonello and Alexander Huth. 2024. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language (Cambridge, Mass.)*, 5(1):64–79.

Richard Antonello, Javier Turek, Vy Vo, and Alexander Huth. 2021. Low-dimensional structure in the space of language representations is reflected in brain responses. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pages 8332–8344, Red Hook, NY, USA. Curran Associates Inc.

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2023. Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, 36:21895–21907.

Catherine Chen, Tom Dupré la Tour, Jack L. Gallant, Daniel Klein, and Fatma Deniz. 2024. The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, 7(1):284. Publisher: Nature Publishing Group UK London.

Jiaqi Chen, Richard Antonello, Kaavya Chaparala, Coen Arrow, and Nima Mesgarani. 2025. Quantifying Semantic Functional Specialization in the Brain Using Encoding Models of Natural Language. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–90, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Srikanth R. Damera, Lillian Chang, Plamen P. Nikolov, James A. Mattei, Suneel Banerjee, Laurie S. Glezer, Patrick H. Cox, Xiong Jiang, Josef P. Rauschecker, and Maximilian Riesenhuber. 2023. Evidence for a Spoken Word Lexicon in the Auditory Ventral Stream. *Neurobiology of Language*, 4(3):420–434.

Tom Dupré la Tour, Matteo Visconti di Oleggio Castello, and Jack L. Gallant. 2025. The Voxelwise Encoding Model framework: A tutorial introduction to fitting encoding models to fMRI data. *Imaging Neuroscience*, 3:imag_a_00575.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. 2024. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Rémi King. 2025. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, 122(42):e2422097122.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. 2023. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1):555. Publisher: Nature Publishing Group.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, Neil D. Lawrence, Michael I. Jordan, and Thomas Dietterich, editors. 2022.

Dataset Shift in Machine Learning. Neural Information Processing series. MIT Press, Cambridge, MA, USA.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, 28(9):3095–3114.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866. Publisher: Nature Publishing Group.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Aditya Vaidya, Javier Turek, and Alexander Huth. 2023. Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 58–69, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

Yuchen Zhou, Emmy Liu, Graham Neubig, Michael J. Tarr, and Leila Wehbe. 2024. Divergences between Language Models and Human Brains. *Advances in Neural Information Processing Systems*, 37:137999–138031.

Disentanglement and Compositionality of Letter Identity and Letter Position in Variational Auto-Encoder Vision Models

Bruno Bianchi[1,@], Aakash Agrawal[2], Stanislas Dehaene[2,3,4], Emmanuel Chemla[4,5], Yair Lakretz[4,5,@]

1. CONICET-UBA, FCEN, ICC-Departamento de Computación. Buenos Aires, Argentina;
2. Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin, Gif/Yvette, France;
3. Collège de France, Paris, France; **4.** PSL University, Paris, France;
5. LSCP, Ecole Normale Supérieure, CNRS, Paris, France,
bbianchi@dc.uba.ar; yair.lakretz@gmail.com

Abstract

Human readers can accurately count how many letters are in a word (e.g., 7 in "buffalo"), remove a letter from a given position (e.g., "bufflo") or add a new one. The human brain of readers must have therefore learned to disentangle information related to the position of a letter and its identity. Such disentanglement is necessary for the compositional, unbounded ability of humans to create and parse new strings, with any combination of letters appearing in any positions. Do modern deep neural models also possess this crucial compositional ability? Here, we tested whether neural models that achieve state-of-the-art on disentanglement of features in visual input can also disentangle letter position and letter identity when trained on images of written words. Specifically, we trained beta variational autoencoder (β-VAE) to reconstruct images of letter strings and evaluated their disentanglement performance using CompOrth - a new benchmark that we created for studying compositional learning and zero-shot generalization in visual models for orthography. The benchmark suggests a set of tests, of increasing complexity, to evaluate the degree of disentanglement between orthographic features of written words in deep neural models. Using CompOrth, we conducted a set of experiments to analyze the generalization ability of these models, in particular, to unseen word length and to unseen combinations of letter identities and letter positions. We found that while models effectively disentangle surface features, such as horizontal and vertical 'retinal' locations of words within an image, they dramatically fail to disentangle letter position and letter identity and lack any notion of word length. Together, this study demonstrates the shortcomings of state-of-the-art β-VAE models compared to humans and proposes a new challenge and a corresponding benchmark to evaluate neural models.

Keywords: Reading, Deep Learning, Compositionality, Auto-Encoders.

1. Introduction

Reading is an invention of human modern culture. Unlike other domains of visual processing such as faces, reading skills are not innate and require extensive practice (Dehaene et al., 2015). Reading acquisition therefore must rely on the development of a new neural mechanism in the human brain. Given that letters are the building blocks of words, to process new words, the neural mechanism needs to identify single letters in the input, their positions in a word, and compose this information to process entire words. While brain imaging has localized where in the brain visual processing of words occurs (Cohen et al., 2002), what is the precise neural mechanism that enables us to recognize words is largely unknown.

Recent advances in deep neural models have drastically improved the accuracy on Optical Character Recognition (OCR) tasks (Li et al., 2023). Deep neural models can now achieve similar-to-human performance on a variety of tasks, including vision and language. Although neural models are often considered black boxes, full access to their neural computations during processing is possible. Analyzing the properties of these networks provides new opportunities to study neural mechanisms underlying orthographic processing, and in particular, into how letter-identity and letter-position information are extracted from raw images and then composed together to encode whole words.

To study this question in neural models, we developed *CompOrth* - a battery of tests, which evaluate compositionality in models and their generalization performance. CompOrth provides several tests, which can be used to both evaluate the 'behavioral' performance of the models, as well as study *neural* mechanisms in the model. The tests are designed in a way that directly probes the question of whether a neural model extracts and disentangles letter-identity and letter-position information from raw images, and whether it can compose them together to encode entire words.

The main hypothesis of this work is that vision models can achieve such *functional* disentanglement of letter identity and position, required to succeed on CompOrth, by *neurally* disentangling this information, representing it in separate units of the model. This is since neural disentanglement could facilitate composition of letter identity and letter po-

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 293–308
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

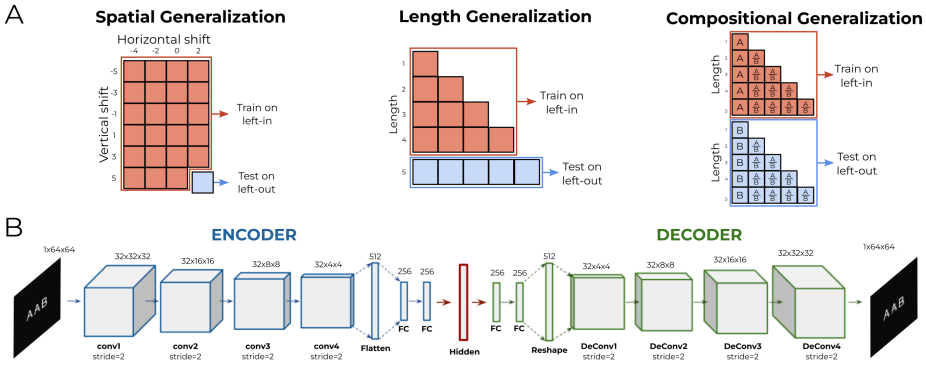


Figure 1: **(A)** The CompOrth Benchmark: schemes of the three types of generalization tests. Each test consists of several splits into training and test sets, ensuring that in each split, the test set contains images generated from different combinations of factors than those in the training set. **(B)** Model architecture: an illustration of the architecture of an auto-encoder for processing images of written words. The size of the hidden layer (bottleneck) was varied as part of the experimentation.

sition in downstream computations, thus improving generalizations to unseen combinations of letters and positions. Beta Variational Auto-Encoders (β-VAEs) are a leading model for neural disentanglement (Higgins et al., 2017), and, interestingly, they have been shown to also align with neural activity recorded from the primate cortex. Specifically, Higgins et al. (2021) showed that disentangled representations in neural models enhance prediction of Macaque neural activity, aligning artificial and biological processing. This suggested that the face region in the Macaque's brain disentangles such factors of variations. Here, we thus study β-VAEs using CompOrth, testing whether strong neural disentanglement in the models could lead to improved performance on CompOrth, given its requirement for both disentanglement and compositional abilities as of humans.

Our results show that β-VAEs learn to disentangle surface properties of written words, such as 'retinal' horizontal and vertical position. However, β-VAEs dramatically fail to disentangle identity and positional information of letters, and to combine them together. We show that such disentanglement of letter identity and position is lacking both at the behavioral and neural levels of the model. Furthermore, we show that β-VAEs lack a robust notion of word length, failing to generalize to unseen word lengths in CompOrth. Finally, we provide arguments for why other types of neural models might suffer from the same limitation, and therefore suggest CompOrth as a challenging benchmark for future models.

2. Related Literature

Literature on cognitive sciences contains competing theories about how words are neurally encoded in the human brain. Based on experiments in humans, early studies theorized the presence of letter combination detectors (Dehaene et al., 2005), i.e., letters combine to form bigrams, which are further combined to form larger n-grams and finally words. Other theories suggested neural encoding based on open bigrams - neurons that are tuned for higher-order combinations of letters, which are not necessarily adjacent (Grainger and Whitney, 2004). Compositional theories suggest a neural encoding scheme, whereby letter identities and letter positions are directly combined to form words (Agrawal et al., 2020), in line with early suggestions in the literature (Davis and Bowers, 2004). More recently, the lexical categorization model (LCM) was suggested, positing that linguistic processing is optimized by enabling swift retrieval of meaning for known words while effectively discarding meaningless letter combinations (Gagl et al., 2022). Predictions from such theories about the computational operations underlying orthographic processing were often contrasted using neural data recorded from brain data (e.g., Taylor et al., 2013).

Modern neural models now provide new opportunities to study how words can be neurally encoded, possibly informing the above debate. One of the computational tasks in the field, most closely related to reading, is OCR. Various neural models were suggested for OCR, including Recurrent Neural Networks (Breuel et al., 2013), Convolutional Neural Networks (CNNs) (Zhang et al., 2017), Transformers (Li et al., 2023; Azadbakht et al., 2022), and hybrid architectures (Naseer and Zafar, 2019; Jain et al., 2017). Some studies have also tried to identify neural mechanisms in models trained on OCR, or similar tasks. For example, Hannagan et al. (2021) studied neural activations of units of CNNs that were trained to recognize

images of objects and words. They found that a compositional encoding scheme emerged also in the models. During training, the models developed special units that encode either letter identity or ordinal letter position in a word (but not n-grams). This evidence from neural models thus provides support in favor of a compositional code based on single letters and their positions (Agrawal et al., 2020).

OCR focuses on developing models capable of extracting text from images, a task complicated by various sources of noise. This noise can stem from inherent variations in handwriting styles and scripts, as explored in the survey by Baldominos et al. (2019), which examines handwritten character recognition using the MNIST and EMNIST datasets (Cohen et al., 2017). Another significant source of noise is the degradation of physical documents due to age or environmental factors, as discussed by Fontanella et al. (2020), who investigate pattern recognition and AI techniques for cultural heritage preservation, including dealing with deteriorated materials. These challenges have driven the development of numerous benchmarks and approaches in scene text recognition, aiming to address specific issues. Shi et al. (2016) pioneered end-to-end trainable neural networks for image-based sequence recognition, a foundational work in the field. More recently, Lyu et al. (2022) introduced MaskOCR, leveraging masked encoder-decoder pretraining for improved text recognition. Du et al. (2022) proposed SVTR, a scene text recognition model based on a single visual model, simplifying the architecture. Wang et al. (2021) presented a novel scene text recognizer with a visual language modeling network, bridging the gap between visual and linguistic understanding.

The complexities of scene text recognition extend beyond just noise. Researchers have tackled the challenge of capturing global semantic context, as demonstrated by Yu et al. (2020) with their semantic reasoning networks, Wan et al. (2020) with TextScanner's ordered character reading, Cui et al. (2021) with their representation-enhanced encoder-decoder framework, and Bhunia et al. (2021b) who explore joint visual semantic reasoning with a multi-stage decoder. Handling text in diverse orientations is another key area, addressed by Zhang et al. (2020) with AutoSTR's efficient backbone search and Yan et al. (2021) who utilize primitive representation learning. The interplay between visual processing and language models is crucial, as investigated by Fang et al. (2021) who propose a human-like bidirectional and iterative language model, and Bautista and Atienza (2022) who explore scene text recognition with permuted autoregressive sequence models. Furthermore, researchers have focused on specific challenges like degraded image quality, as in Mou et al. (2020) PlugNet which uses a pluggable super-resolution unit, the misrecognition of contextless text images studied by Yue et al. (2020) with their RobustScanner that enhances positional clues, and unseen character sequences addressed by Bhunia et al. (2021a) through iterative text recognition by distilling from errors. Finally, the issue of limited labeled data has been tackled by Baek et al. (2021) who explore scene text recognition with fewer labels by leveraging real datasets.

However, in contrast to these other challenges in OCR, the primary objective of CompOrth is to identify models that achieve *compositionality*. That is, models that can recognize words with new combinations of letter identities and letter positions, unseen during training, similarly to how humans process new words (Barton et al., 2014). To achieve compositionality, a model would need to learn to identify single letters and then efficiently compose them with their positions, recursively, when encoding entire words. To focus on compositionality, CompOrth thus simplifies much of the problem by eliminating several sources of noise present in OCR corpora, as described below.

3. General Setup

3.1. The CompOrth Benchmark

3.1.1. Stimuli:

The stimuli for CompOrth were designed to probe the encoding of single letters and their positions, minimizing other information. Strings ('words') in each test contain only two uppercase letters (e.g., 'A' and 'B') in Arial font, size 12. We generated all 62 possible words of 1 to 5 letters (e.g., 'A', 'B', 'AA', ..., 'BBBBB'). Images were created by varying word location ('retinal' location) and character spacing, resulting in 11,904 images. All featured white letters on a black background (Figure A.1).

Retinal location was adjusted by shifting the string vertically and horizontally; zero displacement centers the string, while a displacement of 1 moves it one pixel up and right. Strings were shifted up to 4 pixels from the center in all directions (Figure A.1). Spacing variation decoupled a letter's retinal (absolute) location from its (relative) word position by adjusting inter-character spacing from -2 to +1 pixels, avoiding overlap with 0 as the default.

3.1.2. Three Generalization Tests:

CompOrth contains three generalization tests: (1) *Spatial Generalization* (Figure 1A-Left), across 'retinal' positions; (2) *Length Generalization* (middle), to unseen word lengths; and (3) *Compositional Generalization* (right), to unseen letter identity and position combinations. The latter evaluates if mod-

els develop abstract notions of position and identity. Each test used multiple training-test splits where the test set included factor combinations excluded from training. Specifically, one level of a generative factor (e.g., a specific retinal position) was reserved for testing (Figure 1A-Left, blue square).

Spatial Generalization: For each possible combination of x-y-shift, all words with that combination are used as the left-out while the rest of the words are used as the left-in set (Figure 1A, left).

Length Generalization: For each length, all words with that length are used as the left-out set, while the rest of the words are used as the left-in set (Figure 1A, middle).

Compositional Generalization: For each relative position of each letter (e.g., 'A' in 2^{nd} position), all words with that combination (e.g., "AA", "BA", "AAA", "BAB", etc.) are used as the left-out set, while the rest of the words are used as the left-in set (Figure 1A, right).

3.2. Model Architecture

We benchmarked CompOrth using Variational Auto-Encoders (VAEs; Kingma and Welling, 2013), including β-VAEs (Higgins et al., 2017). Following Higgins et al. (2017), models used 4 convolutional and 2 fully connected layers for the encoder, with a mirrored decoder architecture (Figure 1B). Auto-encoders can be tested on unseen words, whereas standard feed-forward classifiers have finite output units (Hannagan et al., 2021), making evaluation on unseen words difficult without retraining. Moreover, β-VAEs can be optimized for neural disentanglement, encouraging individual latent units to encode generative factors like letter identity and position. For training, we used a batch size of 64 samples. Training and Evaluation were conducted with Nvidia Quadro RTX 8000 48 GB GPUs. The whole experiment, including grid search, took about 72 hours. For the grid search, we optimized for the following hyper-parameters: Beta ($2^i, i \in \{0, ..., 7\}$), size of the latent layer ($2^i, i \in \{3, ..., 7\}$) and initial learning rate ($10^i, i \in \{-4, ..., -2\}$). The optimal learning rate across all combinations in the grid search was consistently 10^{-4}. We set the maximal number of epochs to 1000, which we verified to be large enough to reach full convergence in all cases. We also verified that the hyper-parameters that led to the best performance lie within the range of explored values, and not at its edges (Figure A.1).

3.3. Model Evaluation

Reconstruction Loss: For model selection (Figure 3), we used the standard reconstruction loss for visual AEs, which is calculated based on the pixel-by-pixel difference between the original and reconstructed images Suganuma et al. (2018).

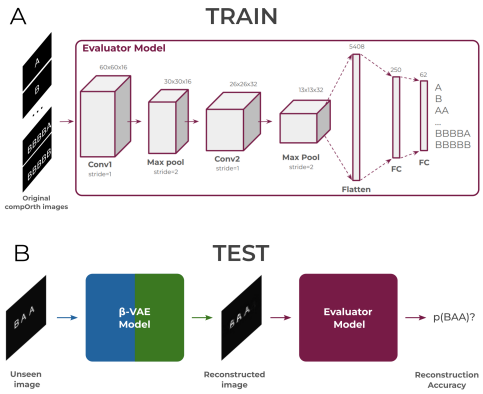


Figure 2: **Evaluator Model: (A)** Pixel-by-pixel loss poorly quantifies reconstruction quality; blurry images can yield low loss despite unrecognizable letters. To ensure a reliable measure of word recognizability, we trained an 'Evaluator Model'—a CNN-based classifier—on all CompOrth words. Its output probability serves as a proxy for the recognizability of β-VAE reconstructions. **(B)** Reconstruction Accuracy: This metric is defined as the classification accuracy of the Evaluator model on all reconstructed images.

Reconstruction Accuracy: However, to test compositionality, relying solely on reconstruction loss might not be sufficient. This is since a large number of pixels in the images could simply be black, leading to a low reconstruction loss even when the reconstructed word is poorly recognizable. Furthermore, blurry or unidentifiable letters in reconstructed images can reduce reconstruction loss without genuinely improving word recognizability. (see examples in Figure 3B). To address this, we defined a new evaluation metric, Reconstruction Accuracy, which quantifies word *recognizability*.

To compute Reconstruction Accuracy, we trained an *Evaluator*—a feed-forward CNN classifier—on all CompOrth words (Figure 2A). It was trained on original images, achieving perfect accuracy via 5-fold cross-validation. The Evaluator comprises two convolutional and two linear layers, with a softmax output across 62 words. This methodology is akin to a GAN 'discriminator' (Goodfellow et al., 2014), but used only for testing.

Reconstruction Accuracy was then defined as the probability assigned by the Evaluator to the target image (Figure 2B). For example, given an image from the test set, such as "ABABA", the image is presented first to the beta-VAE model. The output from the beta-VAE, the reconstructed image, is then given as input to the Evaluator. The Reconstruction Accuracy for this image is therefore the probability

that the Evaluator assigns, at its output layer, to the word "ABABA". In sum, the Evaluator produces a value in the range [0, 1] for how well the word image was reconstructed by the beta-VAE model. How recognizable it is.

Metrics for Disentanglement (MIG and MIR): To quantify neural disentanglement, Chen et al. (2018) used the Mutual Information Gap (MIG), whereas Whittington et al. (2022) suggested the Mutual Information Ratio (MIR). MIG scores high if each factor is encoded in a single neuron. In contrast, MIR scores high when each neuron responds to a single factor. MIR thus allows high scores when multiple neurons respond to the same unique factor. Since compositionality can be achieved even if several neurons extract the same information, we used MIR to test if neural disentanglement facilitates behavioral disentanglement of letter-identity and position in CompOrth.

4. Results

4.1. Model selection

We first optimized for the hyperparameters of the β-VAE models, using nested cross-validation. Figure 3A illustrates model selection, by showing both the reconstruction loss and the MIR for all models in the grid-search. Since both good reconstruction and disentanglement are desired properties of a model, the optimal models lie on the Pareto front (purple circles) of the problem (Goodarzi et al., 2014). No other models outperform them in both criteria simultaneously. In what follows, we therefore report results based on average performance across all optimal models on the Pareto front. We later analyze particular cases from these models.

To illustrate the reconstruction ability of the models, Figure 3B&C shows examples from two models - one from the Pareto front, having good reconstruction performance (with $\beta = 4$, *latent-layer size* $= 32$, *learning rate* $= 10^{-3}$) and the other with a low one ($\beta = 4$, *latent-layer size* $= 128$, *learning rate* $= 10^{-5}$). In both panels, the top row shows examples from the original images, and the bottom row shows their corresponding reconstructions. In the case of low reconstruction performance, the reconstructed images are blurry, with only 'retinal' location preserved.

4.2. Behavioral Evaluation using CompOrth

Spatial Generalization – β-VAEs can generalize to unseen 'retinal' locations: Figure 4A-Left shows the mean generalization performance to unseen retinal positions, for all models from the Pareto front (mean performance in black). On average, the models show good ability to reconstruct words in positions where they were not seen during training. Except for three models on the Pareto front ($layer - size(ls) = 16$, orange and green lines, and $beta = 64$, brown line) all other individual models achieve mean accuracy above 90% for all spatial generalizations, also for vertical generalization (not shown) (Figure A.2). Figure 4A-Right further shows reconstruction examples: each of the plots shows 6 random samples (top) and their reconstruction (bottom) by a given model. Overall, the reconstruction is, qualitatively, similar to the original image, even for models with relatively low performance.

Length Generalization – β-VAEs fail to generalize to longer word lengths: Figure 4B-Left shows the reconstruction accuracy for all left-out word lengths, as measured with the Evaluator Model. Overall, reconstruction accuracy for short left-out word length is high, in particular for the four models with also better performance on spatial generalization. However, a qualitative inspection of random samples from these models (4B-Right, for examples) show that letter parts are nonetheless present in the reconstruction of images with a single letter, remnants from the longer words in the training data. In general, for longer word lengths, generalization performance decreases. In particular, generalization performance for words with five letters is lowest. A qualitative inspection shows that, indeed, the reconstruction of words with five letters contains in many cases only four letters (red rectangle).

Compositional Generalization: β-VAEs fail to generalize to unseen compositions of letter identity and letter position: Finally, the compositional test assessed the ability of the models to develop an abstract understanding of letter position. Overall, results show that the compositional test was most challenging for the models compared to the other two tests (Figure 4C-Left), with one model's performance approaching chance level. Figure 4C-Right illustrates the type of errors the models make (e.g., in red rectangles). For example, when presented with strings where 'B' was never present in the 5^{th} position in the training data, the models 'hallucinate' an 'A' in this position. Similar errors occur when 'B' or 'A' were omitted from other positions during training.

4.3. Neural Evaluation using Perturbation Experiments

We next studied to what extent beta-VAEs develop neural disentanglement of letter position and letter identity information. Figure 5A illustrates what such neural disentanglement could look like – it shows

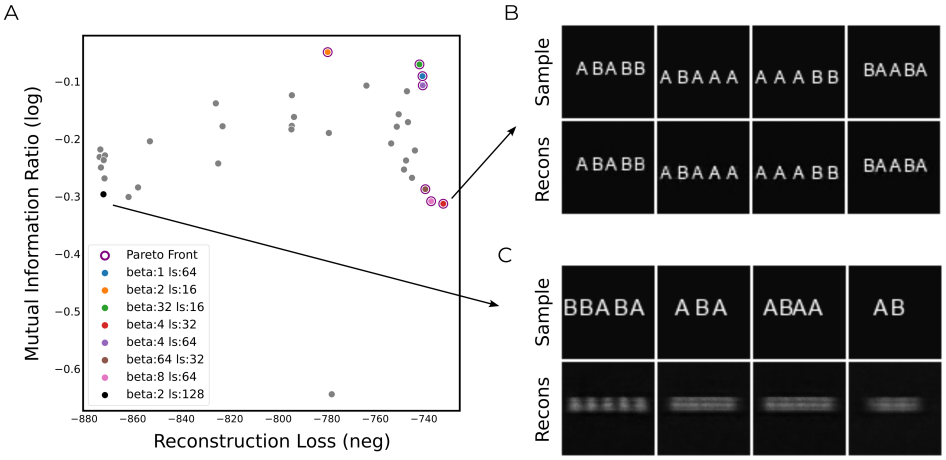


Figure 3: **Model Selection: (A)** We trained a large set of $\beta - VAEs$, with varying values of hyper-parameters, such as the beta value of the model or the size of their latent layer. An optimal $\beta - VAE$ would be one that achieves both good reconstruction *and* strong disentanglement. Here, we thus show the reconstruction loss against Mutual Information Ratio for all models. Each dot represents a model. Orange dots represents models from which samples were taken as examples in the right panels. **(B)** Reconstruction examples from the model with best reconstruction loss and MIG (β=4, Latent Size=32, Learning Rate=0.0001). **(C)** Reconstruction examples from a model with relatively poor reconstruction loss (β=2, Latent Size=128, Learning Rate=0.0001). Models marked with purple circles represent the Pareto Front.

a possible encoding scheme, where different units of the latent layer encode for different positions in the word, and different levels of activity encode for different level identities. Therefore letter identity is independently encoded of letter position.

To test whether such neural disentanglement emerges in the model, even partly, we conducted a perturbation experiment. That is, given an input image from the training set, we computed the neural activations at the latent layer of the model, and then separately for each unit, we systematically perturbed its activity to different levels. After each perturbation, we then reconstructed the image. The difference between the input image and the reconstructed image (after perturbation) is revealing about the information that the perturbed unit encodes. For example, if a model developed neural disentanglement of letter position and letter identity (Figure 5A), then perturbing one of the latent units can cause a replacement of one letter with another one, in only the perturbed position.

Figures 5B-D show examples from the perturbation experiments, from the model with the best reconstruction loss and strong performance on CompOrth ($\beta = 4$, *latent-size*$= 32$; red lines in Figure 4). Each panel corresponds to a latent unit; rows of each panel correspond to different samples (input images) and columns to different levels of perturbation. Unit 22 (Panel B) encodes vertical translation, partly entangled with letter identity. Unit 3 (Panel C) modulates word length, partly entangled as well with letter identity. Unit 23 (Panel D) mainly encodes letter identity, without spatial effects. Perturbation effects for all 32 units are shown in Figures A.3 & A.4.

However, analyzing all 32 latent units in the model, none fully disentangled letter identity and position (Figure 5A). No model on the Pareto-front achieved strong neural disentanglement. This is, in fact, consistent with their poor performance on CompOrth's Compositional-Generalization test – Limited neural disentanglement is consistent with poor compositionality and thus low performance on CompOrth.

4.4. The Relationship between Neural Disentanglement and Compositionality

While we haven't discovered strong neural disentanglement of identity and position in the previous section, a weak neural disentanglement might have nonetheless emerged in some of the models, which is hard to detect with mere perturbation experiments. Such weak disentanglement would possibly lead to a small, yet significant, improvement in performance on the Compositional-Generalization test in CompOrth.

We therefore next tested the hypothesis that neural disentanglement facilitates the separation of letter-positions and letter-identity information, and therefore, in turn, their composition. This predicts that models that achieve high neural disentangle-

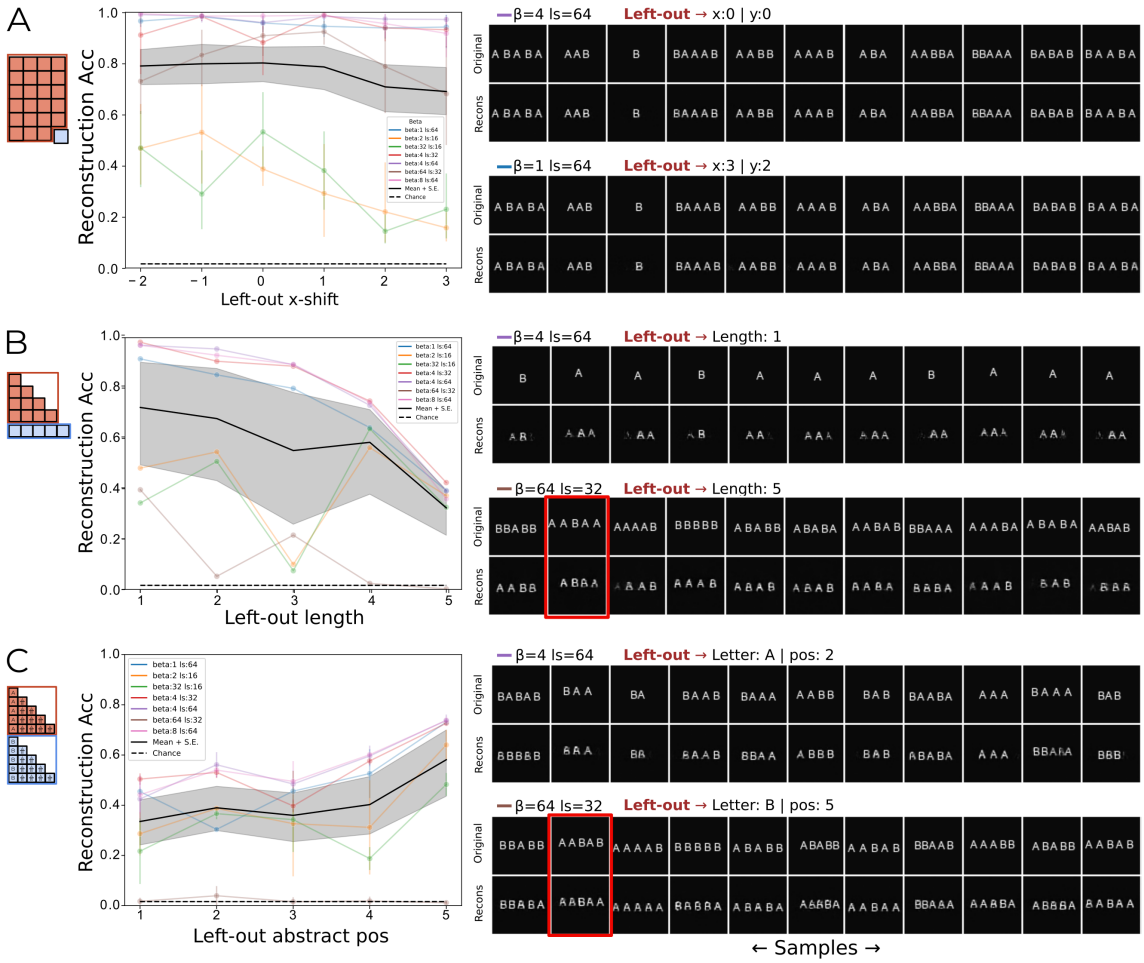


Figure 4: **Results on CompOrth tests.** Average Reconstruction Accuracy on the **(A)** Retinal-Position Test (for vertical axis see Figure A.2), **(B)** Word-Length Test, and **(C)** the Abstract-Position Test. Each of these tests evaluates the optimal models (Figure 3) on a left-out portion of the dataset, to assess its generalization performance, and in particular, its compositionality ability, as tested in panel C. For each model, dots represents the average across all the left-out splits and error bar represents the Standard Error. Black line shows the average across all the selected models with the Standard Error in gray shadow. Dashed lines mark the chance level for the classifier ($chance = 1/62$). On the right of each panel, several examples are shown for how the model reconstruct test images. Note the red marking on the images, which highlight the type of errors the model makes. Information about the generalization test is given at the top of each panel. For example, `Left-out → Letter: A | pos: 2` means that strings with the letter 'A' at the second position of the word were not seen during training.

ment, as measured by MIR, will achieve better performance on compositional generalization, as measured by the Compositional-Generalization test in CompOrth. To test this, we computed the correlation between the MIR and reconstruction accuracy on CompOrth for all models on the Pareto front. We found a weak correlation (Pearson coefficient $\rho = 0.13$). (Figure A.5). However, it was not statistically significant ($p - value = 0.26$).

5. Summary and Discussion

We introduced CompOrth, a novel benchmark for evaluating orthographic processing in visual models. The primary goal of CompOrth is to assess compositionality — the ability of a model to generalize to new combinations of letter identities and positions beyond the training set. This task is considered trivial for humans, so passing the CompOrth test is essential for a model to be regarded as achieving human-like performance.

We tested numerous VAEs, including β-VAEs, which encourage neural disentanglement. All models generalized well to unseen spatial locations

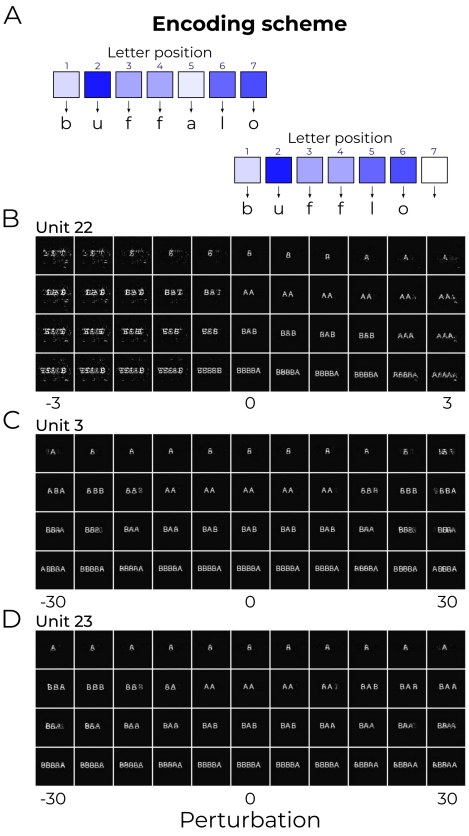


Figure 5: **Neural Perturbation Analyses (A)** Examples of a hypothetical encoding scheme: single neurons for positions, where the degree of activation indicates which letter is present in that position. **(B-D)** Perturbation results for example units from a model with strong performance on CompOrth ($\beta = 4$, *latent-size*$= 32$). Each row represents different samples (word images), columns represents different levels of perturbation.

and disentangled spatial factors in their latent layers. These results align with CNNs' inherent translation invariance—driven by weight sharing and pooling—allowing them to process patterns regardless of position (but see, Azulay and Weiss, 2019; Khemakhem et al., 2020). The match between the model's inductive bias and the dataset's spatial manipulations likely supported this disentanglement (Locatello et al., 2019).

However, all models failed CompOrth's compositional tests, failing to generalize to new letter-position combinations and longer word lengths. This shows models did not capture the underlying generative process—sampling combinations of letters in varying positions. Instead, results suggest the models rely on *memorizing* training data. This is observed when the model 'hallucinates' or adds letters, failing to reconstruct the input image.

The inductive biases of autoencoders—specifically the architecture's bottleneck and β-VAE regularization—could encourage models to learn underlying patterns through abstract representations rather than memorization. However, an overly constrained bottleneck can reduce reconstruction accuracy (Figure 3). Disentanglement (MIR) showed no correlation with generalization, consistent with prior work (Montero et al., 2021; Schott et al., 2021). The explored VAEs' inductive biases might not sufficiently align with the data's generative process, hindering the emergence of disentanglement (Locatello et al., 2019). While more data could theoretically compensate for weak priors (Welling, 2019; Goyal and Bengio, 2022), this is often infeasible, necessitating architectures with better-adapted inductive biases.

Future directions for orthographic processing include incorporating dual-route processing into computational models. Cognitive science distinguishes between habitual (System 1) and controlled (System 2) processing—fast and unconscious versus slow and attention-demanding. In orthography, behavioral and neuroimaging data suggest two distinct routes: lexical and sublexical, supported by memory-based and rule-based processing (e.g., Marshall and Newcombe, 1973; Coltheart et al., 2001; Jobard et al., 2003; Fiebach et al., 2002). Integrating these cognitive insights into inductive biases may improve out-of-distribution generalization (Goyal and Bengio, 2022). We suggest integrating dual-route architectures where one route relies on implicit symbol-based rules. Other future avenues include disentanglement via distributed equivariant operators (Bouchacourt et al., 2021), alternative notions of disentanglement (O'Neill et al., 2024), and vision transformers with contextual positional encoding (CoPE; Golovneva et al., 2024) for ordinal position.

We further hypothesized that neural disentanglement, which tends to emerge in β-VAEs with high values of β, would facilitate the separation of letter-positions and letter-identity information, and in turn, their composition. To test this, we conducted perturbation experiments with β-VAEs, to see whether some of the units disentangle identity and position information. Exploring all units in the model, we have not identified any such units, which is consistent with the failure of the models on CompOrth. However, a weak neural disentanglement of identity and position may have emerged in some of the models, unobserved by our perturbation experiments. We therefore tested whether there exists a correlation between MIR (a metric that quantifies neural disentanglement based on mutual information) and CompOrth performance across all our VAE models. We found no significant correlation between MIR and reconstruction accuracy on CompOrth.

The observed failure of the models in this study

is one more example of the shortcoming of artificial neural networks to dynamically and flexibly bind information, which might be distributively encoded in the network (Greff et al., 2020), even when neural disentanglement is explicitly optimized, as in β-VAEs. This binding problem affects the capacity of the models to achieve compositional ability by manipulating symbols (letters) and combining them in various, unbounded ways, as humans (Fodor and Pylyshyn, 1988; Lake and Baroni, 2018; Baroni, 2020). Similar limitations were also observed, for similar reasons, for shapes (Montero et al., 2021; Schott et al., 2021) and in language models (Delétang et al., 2022).

However, unlike existing benchmarks for testing compositionality in language models (e.g., Lake and Baroni, 2018), for orthographic processing and OCR, there are no existing targeted benchmarks. CompOrth therefore aims to fill this gap by providing means to evaluate compositionality in vision models.

Ethical statement

This paper presents work whose goal is to bridge closer the fields of Machine Learning and Psycholinguistics; being theoretical in nature, we believe that no societal risks need to be specifically highlighted.

Limitations

This study investigates the ability of a neural architecture to disentangle relevant information in the input for compositional generalization. One possible limitation is that the models explored here were trained on the CompOrtho dataset only. However, we note that while the models were trained from scratch on CompOrth, the challenges posed by the benchmark are also applicable to pre-trained models. These models can be refined and evaluated using the same approach (Figure 1A) to assess their compositional generalization capabilities.

Acknowledgments

B.B. received a travel grant from the University of Buenos Aires (UBAint Docentes 2023) that made the completion of this study possible. B.B. was also supported by funding from CONICET.

6. Bibliographical References

Aakash Agrawal, KVS Hari, and SP Arun. 2020. A compositional neural code in high-level visual cortex can explain jumbled word reading. *Elife*, 9:e54846.

Alireza Azadbakht, Saeed Reza Kheradpisheh, and Hadi Farahani. 2022. Multipath vit ocr: A lightweight visual transformer-based license plate optical character recognition. In *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 092–095. IEEE.

Aharon Azulay and Yair Weiss. 2019. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25.

Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. 2021. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3113–3122.

Alejandro Baldominos, Yago Saez, and Pedro Isasi. 2019. A survey of handwritten character recognition with mnist and emnist. *Applied Sciences*, 9(15):3169.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Jason JS Barton, Hashim M Hanif, Laura Eklinder Björnström, and Charlotte Hills. 2014. The word-length effect in reading: A review. *Cognitive neuropsychology*, 31(5-6):378–412.

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European conference on computer vision*, pages 178–196. Springer.

Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, and Yi-Zhe Song. 2021a. Towards the unseen: Iterative text recognition by distilling from errors. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14950–14959.

Ayan Kumar Bhunia, Aneeshan Sain, Amandeep Kumar, Shuvozit Ghose, Pinaki Nath Chowdhury, and Yi-Zhe Song. 2021b. Joint visual semantic reasoning: Multi-stage decoder for text recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14940–14949.

Diane Bouchacourt, Mark Ibrahim, and Stéphane Deny. 2021. Addressing the topological defects of disentanglement via distributed operators. *arXiv preprint arXiv:2102.05623*.

Thomas M Breuel, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. 2013. High-performance ocr for printed english and fraktur

using lstm networks. In *2013 12th international conference on document analysis and recognition*, pages 683–687. IEEE.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. 2017. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE.

Laurent Cohen, Stéphane Lehéricy, Florence Chochon, Cathy Lemer, Sophie Rivaud, and Stanislas Dehaene. 2002. Language-specific tuning of visual cortex? functional properties of the visual word form area. *Brain*, 125(5):1054–1069.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Mengmeng Cui, Wei Wang, Jinjin Zhang, and Liang Wang. 2021. Representation and correlation enhanced encoder-decoder framework for scene text recognition. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16*, pages 156–170. Springer.

Colin J Davis and Jeffrey S Bowers. 2004. What do letter migration errors reveal about letter position coding in visual word recognition? *Journal of Experimental Psychology: Human Perception and Performance*, 30(5):923.

Stanislas Dehaene, Laurent Cohen, José Morais, and Régine Kolinsky. 2015. Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, 16(4):234–244.

Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. 2005. The neural code for written words: a proposal. *Trends in cognitive sciences*, 9(7):335–341.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, et al. 2022. Neural networks and the chomsky hierarchy. *arXiv preprint arXiv:2207.02098*.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. 2022. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*.

Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7098–7107.

Christian J Fiebach, Angela D Friederici, Karsten Müller, and D Yves Von Cramon. 2002. fmri evidence for dual routes to the mental lexicon in visual word recognition. *Journal of cognitive neuroscience*, 14(1):11–23.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Francesco Fontanella, Francesco Colace, Mario Molinara, A Scotto Di Freca, and Filippo Stanco. 2020. Pattern recognition and artificial intelligence techniques for cultural heritage.

Benjamin Gagl, Fabio Richlan, Philipp Ludersdorfer, Jona Sassenhagen, Susanne Eisenhauer, Klara Gregorova, and Christian J Fiebach. 2022. The lexical categorization model: A computational model of left ventral occipito-temporal cortex activation in visual word recognition. *PLOS Computational Biology*, 18(6):e1009995.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2024. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*.

Ehsan Goodarzi, Mina Ziaei, and Edward Zia Hosseinipour. 2014. *Introduction to optimization analysis in hydrosystem engineering*, volume 25. Springer.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Anirudh Goyal and Yoshua Bengio. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068.

Jonathan Grainger and Carol Whitney. 2004. Does the huamn mnid raed wrods as a wlohe? *Trends in cognitive sciences*, 8(2):58–59.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. 2020. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*.

T Hannagan, A Agrawal, L Cohen, and S Dehaene. 2021. Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, 118(46):e2104779118.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. 2021. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

Mohit Jain, Minesh Mathew, and CV Jawahar. 2017. Unconstrained ocr for urdu using deep cnn-rnn hybrid networks. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 747–752. IEEE.

Gaël Jobard, Fabrice Crivello, and Nathalie Tzourio-Mazoyer. 2003. Evaluation of the dual route theory of reading: a metanalysis of 35 neuroimaging studies. *Neuroimage*, 20(2):693–712.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pages 2207–2217. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13094–13102.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pretraining. *arXiv preprint arXiv:2206.00311*.

John C Marshall and Freda Newcombe. 1973. Patterns of paralexia: A psycholinguistic approach. *Journal of psycholinguistic research*, 2:175–199.

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. 2021. The role of disentanglement in generalisation. In *International Conference on Learning Representations*.

Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. 2020. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 158–174. Springer.

Asma Naseer and Kashif Zafar. 2019. Meta features-based scale invariant ocr decision making using lstm-rnn. *Computational and Mathematical Organization Theory*, 25:165–183.

Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*.

Lukas Schott, Julius Von Kügelgen, Frederik Träuble, Peter Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. 2021. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304.

Masanori Suganuma, Mete Ozay, and Takayuki Okatani. 2018. Exploiting the potential of standard convolutional autoencoders for image

restoration by evolutionary search. In *International Conference on Machine Learning*, pages 4771–4780. PMLR.

JSH Taylor, Kathleen Rastle, and Matthew H Davis. 2013. Can cognitive models explain brain activation during word and pseudoword reading? a meta-analysis of 36 neuroimaging studies. *Psychological bulletin*, 139(4):766.

Zhaoyi Wan, Minghang He, Haoran Chen, Xiang Bai, and Cong Yao. 2020. Textscanner: Reading characters in order for robust scene text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12120–12127.

Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. 2021. From two to one: A new scene text recognizer with visual language modeling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14194–14203.

Max Welling. 2019. Do we still need models or just more data and compute. *University of Amsterdam*, 7.

James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy Behrens. 2022. Disentanglement with biological constraints: A theory of functional cell types. In *The Eleventh International Conference on Learning Representations*.

Ruijie Yan, Liangrui Peng, Shanyu Xiao, and Gang Yao. 2021. Primitive representation learning for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 284–293.

Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12113–12122.

Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision*, pages 135–151. Springer.

Haochen Zhang, Dong Liu, and Zhiwei Xiong. 2017. Cnn-based text image super-resolution tailored for ocr. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE.

Hui Zhang, Quanming Yao, Mingkun Yang, Yongchao Xu, and Xiang Bai. 2020. Autostr: efficient backbone search for scene text recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 751–767. Springer.

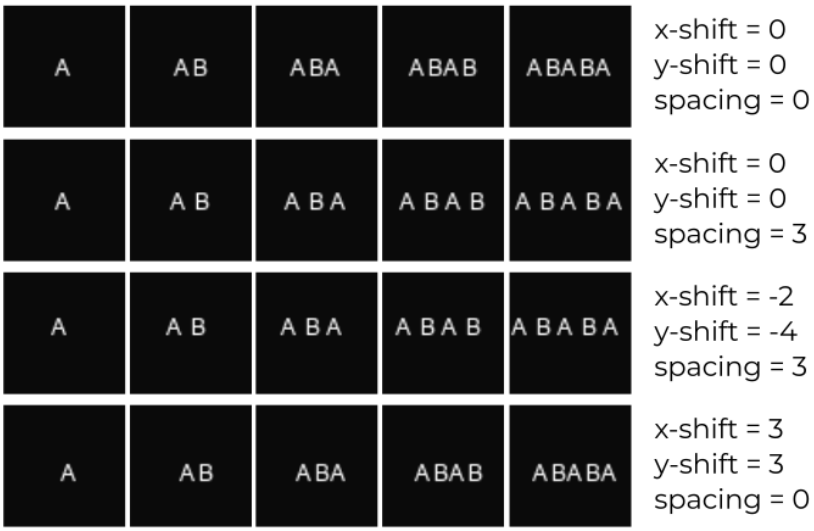


Figure A.1: **Examples from the generated dataset and model training.** All the images are comprised by a string of 1 to 5 letters, using only the uppercase characters A and B. To generate variations of this strings the spacing and the x and y position were modified.

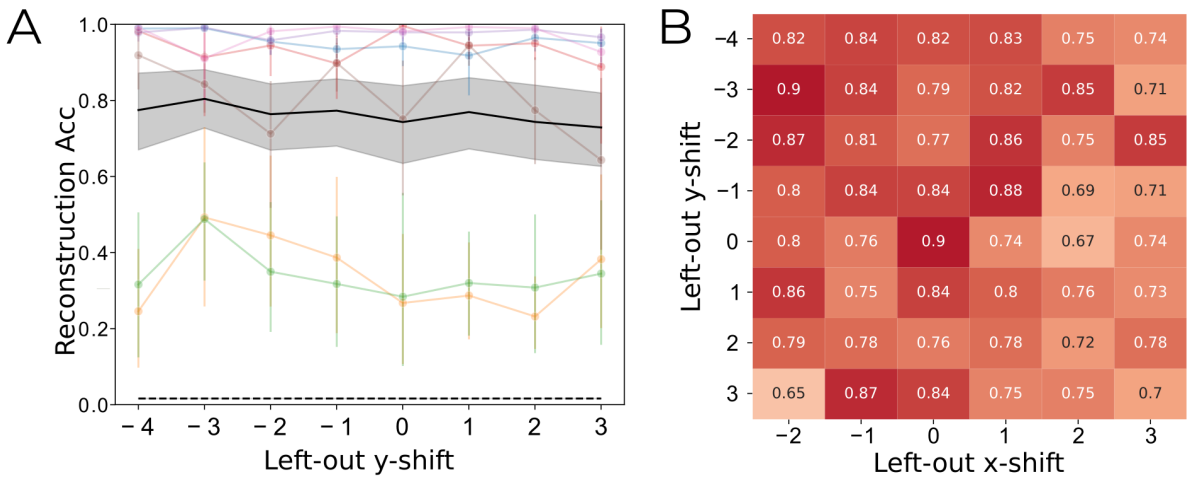


Figure A.2: **Reconstruction accuracy for Spatial Generalization test. (A)** averaged by model, across y-shift; and **(B)** average for each combination of x- and y-shift, across models.

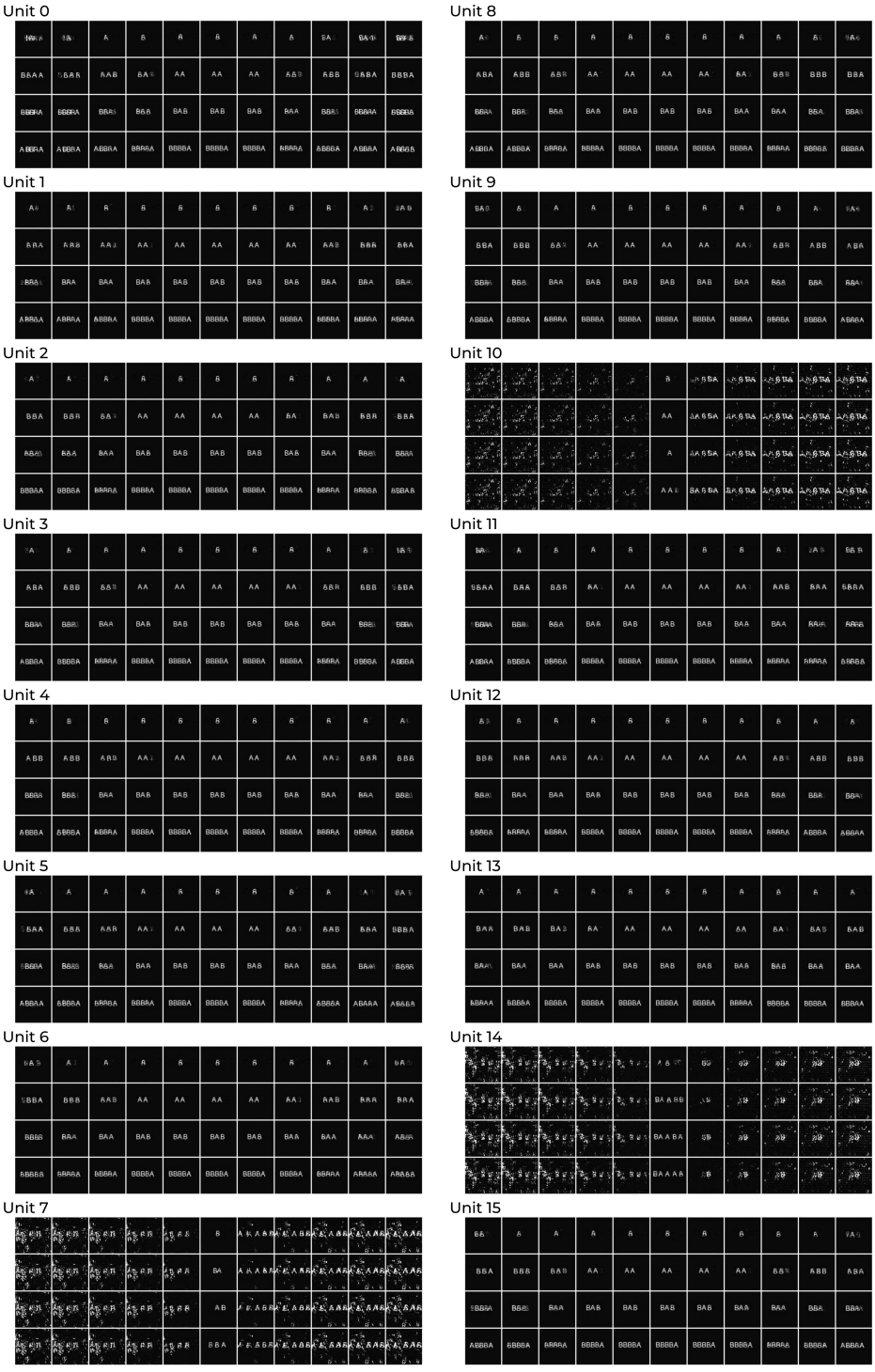


Figure A.3: **Neural Perturbation Analyses.** Perturbation results for example units from a model with strong performance on CompOrth ($\beta = 4$, $latent - size = 32$). First 16 neurons of the model.

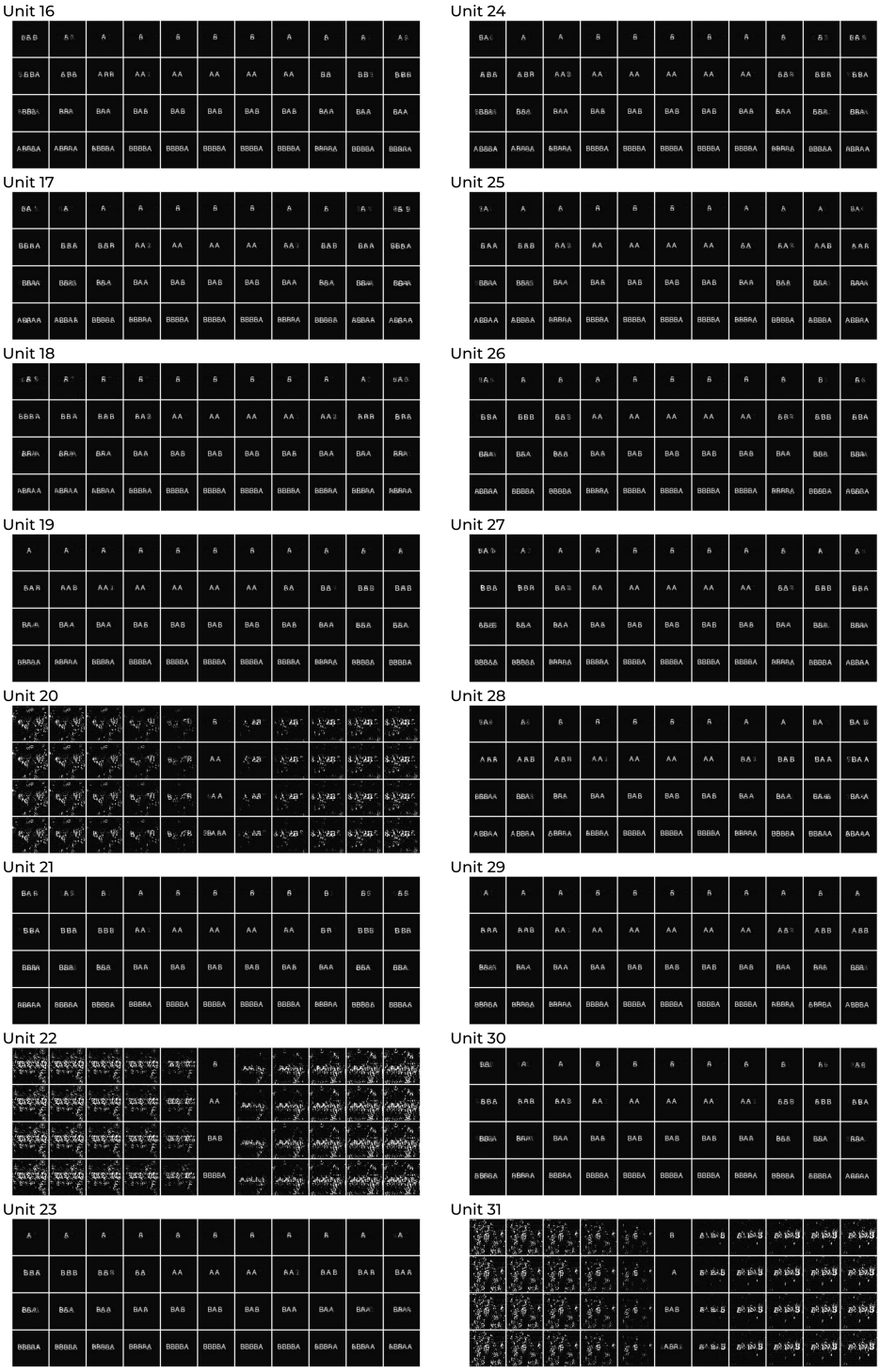


Figure A.4: **Neural Perturbation Analyses.** Perturbation results for example units from a model with strong performance on CompOrth ($\beta = 4$, $latent - size = 32$). Last 16 neurons of the model.

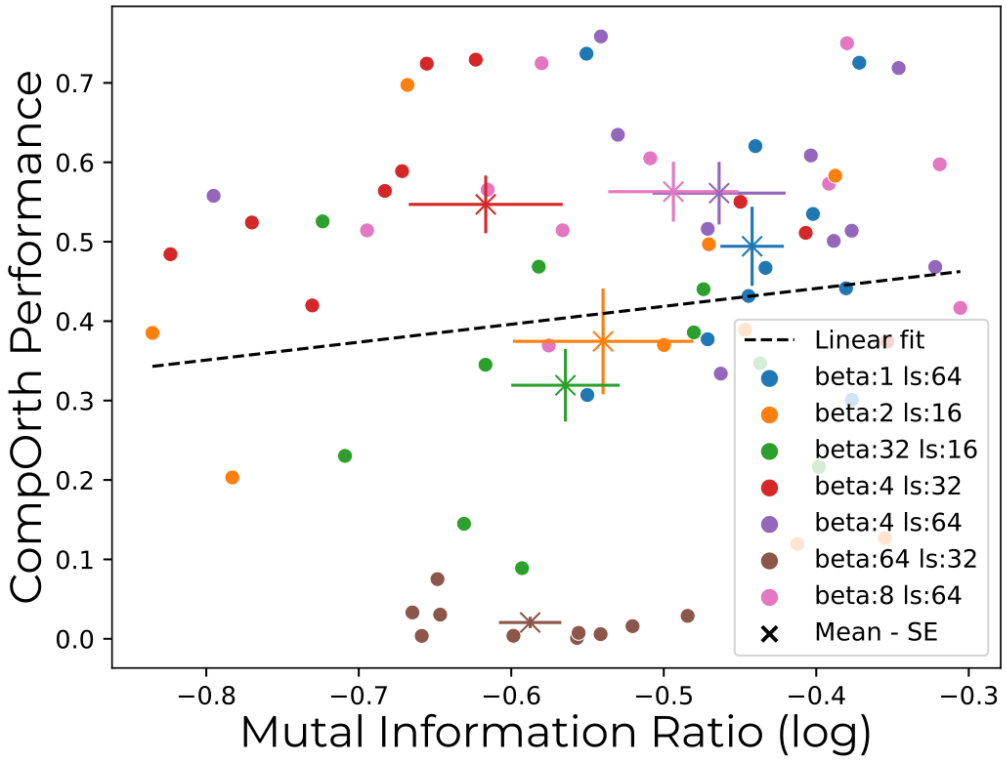


Figure A.5: **Mutual Information Ratio (MIR) vs Reconstruction Accuracy for Compositional Generalization analyses.** We computed MIR using the 5-letter word identity encoding (Figure 5A) across all Pareto Front models. Each model's results per test split are shown as colored circles, with crosses indicating their mean performance across splits. Error bars reflect SEM, and the dashed line shows the linear trend. Pearson correlation resulted in $\rho = 0.12$ and $p - val = 0.30$.

Plausibility as Commonsense Reasoning: Humans Succeed, Large Language Models Do not

Sercan Karakaş
University of Chicago
skarakas@uchicago.edu

Abstract

Large language models achieve strong performance on many language tasks, yet it remains unclear whether they integrate world knowledge with syntactic structure in a human-like, structure-sensitive way during ambiguity resolution. We test this question in Turkish prenominal relative-clause attachment ambiguities, where the same surface string permits high attachment (HA) or low attachment (LA). We construct ambiguous items that keep the syntactic configuration fixed and ensure both parses remain pragmatically possible, while graded event plausibility selectively favors High Attachment vs. Low Attachment. The contrasts are validated with independent norming ratings. In a speeded forced-choice comprehension experiment, humans show a large, correctly directed plausibility effect. We then evaluate Turkish and multilingual LLMs in a parallel preference-based setup that compares matched HA/LA continuations via mean per-token log-probability. Across models, plausibility-driven shifts are weak, unstable, or reversed. The results suggest that, in the tested models, plausibility information does not guide attachment preferences as reliably as it does in human judgments, and they highlight Turkish RC attachment as a useful cross-linguistic diagnostic beyond broad benchmarks.

Keywords: Turkish, relative clause attachment, world knowledge, plausibility, large language models

1. Introduction

Recent progress in large language models (LLMs) has produced systems that reach state-of-the-art results on a wide range of language tasks, including open-domain question answering, generation, summarization, creative writing and translation (Anil et al., 2023; OpenAI et al., 2024; Team et al., 2024; Li et al., 2025; Singh et al., 2025; Team et al., 2025; Finkelstein et al., 2026). In addition to this broad linguistic competence, LLMs often exhibit strong performance in demanding problem-solving workloads, such as software development and symbolic mathematics. In parallel, psycholinguistics and cognitively oriented computational linguistics have increasingly treated neural language models as explicit probabilistic theories of incremental comprehension. Surprisal-based accounts have long been used to explain processing difficulty in reading-time data and related measures. Foundational work showed that surprisal derived from symbolic or probabilistic grammars predicts human comprehension difficulty, including effects related to syntactic ambiguity resolution and expectation-based parsing (Hale, 2001; Levy, 2008); more recent work extends this approach to surprisal derived from neural language models (Arehalli et al., 2022). At the same time, a foundational issue remains open: whether these models carry out human-like reasoning that is logically compositional and systematically generalizes, or whether their apparent success primarily reflects the ability to exploit massive training data to reproduce high-probability surface regularities. Several diagnostic evaluations suggest that LLM behavior often reflects probabilistic pattern completion or shallow heuristics rather than fully systematic rule-based inference (McCoy et al., 2019; Yang et al., 2024). This appears to be especially visible under controlled compositional generalization settings where models often struggle to recombine familiar primitives in genuinely novel ways (Lake and Baroni, 2018; Ruis et al., 2020).

Furthermore, interest in using large language models as cognitive models of human language processing has grown rapidly in recent years, but the empirical foundation remains strongly English-centric: comparatively few studies have carried out systematic, theory-driven psycholinguistic evaluations in other languages, limiting the generalizability of current conclusions across typologically diverse settings (Futrell et al., 2019; Joshi et al., 2020; Hollenstein et al., 2021; Alves, 2025; Boeve and Bogaerts, 2025). Accordingly, we study Turkish, which is a morphologically rich language that is nonetheless less represented in widely used NLP training resources and evaluation benchmarks (Conneau et al., 2020; Joshi et al., 2020), and focus on a particularly informative phenomenon, which is relative-clause attachment ambiguity. In Turkish prenominal relative clause (RC) configurations, the clause can in principle modify either the higher noun (high attachment; HA) or the lower noun (low attachment; LA), yielding a classic hierarchical ambiguity under a fixed surface string. In the broader sentence-processing literature, ambiguity resolution in high- versus low-attachment configurations has often been characterized in terms of economy-driven, rightward-attachment preferences—Late Closure

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 309–328
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(Frazier and Fodor, 1978), Right Association (Kimball, 1973), and Recency (Gibson et al., 1996)—while cross-linguistic work has tested the scope and limits of these biases (Baccino et al., 2000; Fernández, 2003). At the same time, a large body of evidence shows that attachment preferences are not determined by structural heuristics alone: discourse and semantic context can substantially modulate parsing commitments, including in reduced relative-clause ambiguities (Spivey-Knowlton et al., 1993).

Building on this line of work, we focus on a more fine-grained cue than broad pragmatic *possibility*. Specifically, we isolate graded *world-knowledge plausibility*: we hold the syntactic environment constant while selectively biasing the plausibility of attachment to the higher vs. lower nominal, and we validate these contrasts via independent norming ratings. We then test whether such plausibility guides attachment in Turkish by combining (i) an online speeded forced-choice comprehension experiment with native Turkish speakers and (ii) a parallel evaluation of three LLMs using matched log-probability scoring. The two settings are aligned at the level of the underlying interpretive contrast, not at the level of task mechanics: humans make an explicit attachment choice after reading the sentence, whereas models are evaluated through their relative preference for matched HA versus LA continuations. Accordingly, we treat the human response and the model log-probability preference as parallel but non-identical proxies for attachment preference under the same plausibility manipulation. Our results demonstrate a robust human sensitivity to plausibility: attachment preferences shift substantially in the predicted direction when world-knowledge cues favor high attachment (HA) versus low attachment (LA). In contrast, the LLMs show markedly weaker and less consistent plausibility-based shifts, with their preferences often seeming to reflect structural attachment biases rather than the intended plausibility manipulation.

The rest of the paper is structured as follows. §2 reviews related work, and §3 introduces Turkish relative clause attachment. §4 presents the human experiment, while §4.1 outlines the LLM evaluation setup. We then present the human and model results, including the model-specific findings in §5.1, before discussing their implications for world-knowledge integration and the use of LLMs as cognitive models in §6.

2. Related Work

2.1. Attachment ambiguity and constraint interaction

Relative-clause attachment ambiguities have long served as a central testbed for theories of incremental parsing, because a single surface string can license multiple hierarchical structures. Early accounts emphasized economy- and locality-based heuristics such as Late Closure (Frazier and Fodor, 1978), Right Association (Kimball, 1973), and Recency (Gibson et al., 1996), which predict a general bias toward attaching incoming material to the most recently processed constituent. A broad cross-linguistic literature has evaluated the scope of these principles and documented substantial variability, motivating proposals that attachment is not governed by a single universal heuristic, but is instead sensitive to language-specific distributions and to interactions among multiple sources of information (Baccino et al., 2000; Fernández, 2003).

In parallel, constraint-based approaches argue that syntactic commitments are continuously shaped by probabilistic cues from semantics, discourse, and world knowledge, often well before disambiguating material arrives. Classic demonstrations show that supportive discourse and semantic context can dramatically reduce garden-path difficulty in reduced RCs (Spivey-Knowlton et al., 1993; Altmann and Steedman, 1988). Related work on "thematic fit" further indicates that event knowledge and plausibility can be quantified independently (e.g., via norming) and can guide online interpretation early in processing (McRae et al., 1998).

2.2. Prenominal RC attachment in Turkish

Turkish provides a particularly informative setting for attachment research because RCs are typically prenominal and because attachment ambiguities often arise in configurations with multiple potential hosts inside complex nominal structures. Early work highlighted that Turkish had been comparatively less explored in the RC-attachment literature and reported that attachment outcomes are strongly modulated by lexical-semantic properties of the potential hosts (e.g., animacy, semantic compatibility) and by properties of the complex NP environment, challenging accounts that rely on syntactic locality alone (Kırkıcı, 2004). Subsequent studies likewise emphasize the importance of controlling semantic relations within the nominal complex; when such factors are carefully balanced, Turkish comprehenders may show no stable baseline preference for one attachment site over the other, suggesting that apparent "preferences" can reflect uncontrolled semantic biases rather than

a fixed parsing strategy (Başer and Hohenberger, 2020).

Other work has asked more directly whether locality-style biases surface under particular timing or structural conditions. For example, experiments manipulating predicate proximity and related locality factors report patterns consistent with recency-driven attachment in at least some Turkish materials, while also indicating that additional structural factors can shift rates of high vs. low attachment (Akal, 2021). More recent online studies connect Turkish prenominal RC attachment to broader debates about ambiguity advantage and underspecification: under some accounts, readers may delay commitment to save effort, whereas race-based models predict different processing signatures for prenominal structures. Evidence from Turkish reading experiments has been used to argue against strong underspecification-based predictions in this domain, and plausibility has been employed as a controlled disambiguating cue in constructing attachment conditions (Logačev et al., 2022).

Our contribution, on the other hand, is to move beyond relatively coarse manipulations of pragmatic *possibility* and instead isolate a narrower, graded world-knowledge *plausibility* cue. We construct minimally contrasting Turkish materials in which the syntactic configuration is held constant while the plausibility of attachment is selectively biased toward the higher vs. lower nominal, and we validate these contrasts with independent norming ratings (cf. thematic-fit methodologies; McRae et al., 1998). This lets us ask whether attachment decisions track the intended plausibility gradient, rather than reflecting uncontrolled lexical associations.

2.3. Neural language models as cognitive models, and evaluation beyond English

A growing body of psycholinguistic and cognitively oriented computational work uses neural language models as probabilistic models of incremental comprehension, asking to what extent their surprisal estimates and internal representations align with human sentence-processing behavior (Futrell et al., 2019; Oh and Schuler, 2023; Arehalli et al., 2022). As mentioned before, in this view, surprisal derived from a model's next-word distribution provides a linking hypothesis to processing difficulty, with influential proposals and evidence showing that surprisal predicts reading-time patterns and ambiguity-resolution effects in humans (Hale, 2001; Levy, 2008). Recent work extends these ideas to modern neural architectures and explores syntactic ambiguity phenomena under LM-based predictors (Arehalli et al., 2022). At the same time, targeted "psycholinguistic diagnostics" highlight systematic mismatches between model behavior and human generalizations, especially for inferences that require robust integration of context, roles, or negation (Ettinger, 2020). Related lines of work further investigate when distributional models do (or do not) acquire pragmatic constraints that humans exploit in disambiguation (Davis and van Schijndel, 2020).

Importantly for the present paper, much LM-as-cognition evidence remains concentrated on English, and cross-linguistic generalizations are therefore less secure. This motivates testing typologically different languages to evaluate whether purportedly general cognitive conclusions about LM surprisal, ambiguity resolution, and cue integration persist under different structural priors and training-resource profiles. While recent work has begun to extend LM-as-cognition evaluations to Turkish—both via human–LLM comparative processing studies and via cross-linguistic surprisal validations that include Turkish—the overall evidence base remains comparatively sparse (de Varda and Marelli, 2022; Keleş and Dinçtopal Deniz, 2024; Karakaş, 2026). We aim to leverage this gap by comparing human RC attachment judgments and model attachment preferences under the same controlled, normed plausibility manipulation in Turkish. Our goal is not to equate the human task with the model evaluation procedure, but to test whether the same item-level plausibility contrast shifts attachment in a consistent direction across the two systems, or whether model behavior is instead dominated by surface regularities and base-rate biases.

3. Turkish relative clauses

Turkish relative clauses are typically *prenominal*: the RC precedes the noun it modifies, and relativization is expressed morphologically on the verb, commonly via the nominalizer/participle in RCs. In complex noun phrases with two potential nominal hosts, the same prenominal RC string can in principle modify either the higher noun (high attachment; HA) or the lower noun (low attachment; LA), yielding a hierarchical ambiguity under an otherwise fixed surface string.

(1) Birileri, balkon-da dur-an
someone-NOM balcony-LOC stand-REL
[aktris-in]$_{\text{LOW}}$ [hizmetçi-si-ni]$_{\text{HIGH}}$
actress-GEN servant-POSS-ACC
vur-du.
shoot-PST
'Someone shot the servant of the actress who was on the balcony.' (Kırkıcı 2004: 4–5)

In (1), the prenominal relative clause balkon-da dur-an ‘standing on the balcony’ can modify either noun inside the complex NP. Under low attachment, the RC modifies the lower noun *aktris* ‘actress’, yielding the reading ‘Someone shot the servant of the actress who was on the balcony.’ ’ Under high attachment, the RC modifies the higher noun hizmetçi ‘servant’, yielding the reading ‘Someone shot the servant (of the actress) who was on the balcony.’ Because both hosts are structurally available, the surface string remains fixed while the attachment site determines the intended referent of the RC.

4. Experiments

We recruited 102 native speakers of Turkish. To ensure data quality, we excluded 16 participants whose response latencies were extremely fast or extremely slow relative to the sample distribution. Specifically, for each participant i, we first computed their mean response time across trials:

(2)

$$\bar{t}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} t_{ij}$$

where t_{ij} is the response time on trial j and N_i is the number of trials completed by participant i.

We then computed the grand mean and standard deviation across participants:

(3)

$$\mu = \frac{1}{P} \sum_{i=1}^{P} \bar{t}_i, \qquad \sigma = \sqrt{\frac{1}{P-1} \sum_{i=1}^{P} (\bar{t}_i - \mu)^2}$$

Finally, we excluded participant i if their mean response time satisfied the following outlier criterion:

(4)

$$\bar{t}_i < \mu - 2\sigma \qquad \text{or} \qquad \bar{t}_i > \mu + 2\sigma$$

After these exclusions, 86 native speakers of Turkish remained for the online experiment, which was administered using the PCIbex/PennController platform (Zehr and Schwarz, 2018).

Materials and design. Previous work on Turkish (Başer 2018, 251) uses examples such as (5) to illustrate *relative-clause attachment ambiguity* inside a complex NP. In this configuration, the prenominal RC precedes two potential nominal hosts, the lower noun and the higher noun. Structurally, the RC can in principle modify either NP1 or NP2, so the surface string is compatible with two parses. Crucially, however, the contrast in this type of item is not a subtle plausibility manipulation. In (5), the RC *okula kaydedilen* ‘who was enrolled in the school’ can in principle modify either *müdür* ‘principal’ or *yeğen* ‘nephew’, but only the latter yields a semantically coherent interpretation. Attaching the RC to *müdür* is structurally available, yet semantically anomalous, since a principal is not normally interpreted as someone being enrolled in a school.

(5) Okul-a kaydedil-en müdür-ün
school-DAT enroll-PASS.REL principal-GEN
yeğen-i bahçe-de
nephew-POSS.3SG garden-LOC
oyna-yor-du.
play-PROG-PST
‘The nephew of the principal who was enrolled in the school was playing in the garden.’ (Başer 2018, 251)

In our design, by contrast, the critical manipulation differs in a theoretically important way: *both* candidate attachment analyses are constructed to remain pragmatically and semantically *viable*. Concretely, for each item, the relative clause (RC) can plausibly attach to either NP1 or NP2 without yielding a categorical anomaly or selectional clash in contrast to (5). Thus, the manipulation does not instantiate a binary contrast between a *possible* parse and an *implausible* parse. Rather, it keeps the underlying syntactic configuration constant and manipulates a *graded* asymmetry in *world-knowledge plausibility*: local event knowledge is used to make one attachment interpretation *more plausible* than its competitor, while preserving the acceptability of the alternative interpretation. Methodologically, this design supports a more fine-grained test of attachment processing. Because neither parse can be dismissed outright, any preference for HA vs. LA is less likely to reflect simple anomaly rejection and more likely to reflect sensitivity to subtle plausibility weighting during online interpretation. In turn, this allows us to ask whether human comprehenders and LLMs exploit probabilistic world knowledge as a soft constraint in attachment resolution, rather than relying only on cases in which one candidate parse is pragmatically untenable.

(6) Gazeteci-ler-den kaç-an manken-in
reporter-PL-ABL run.away-REL model-GEN
koruma-sı güçlü-ydü.
bodyguard-POSS strong-PST
‘The bodyguard of the model, who ran away from reporters, was very strong.’

In (6), both *manken* ‘model’ and *koruma* ‘bodyguard’ are structurally viable RC hosts, but world

knowledge makes the *model* a much more plausible agent of *kaç-* 'run away' than the bodyguard, thereby creating a targeted plausibility cue that directly bears on attachment choice. Lexical frequency was controlled by consulting corpus-based frequency estimates for critical nouns and verbs and avoiding systematic imbalances across conditions; animacy was matched as well.

The critical materials comprised 40 ambiguous Turkish prenominal RC items (20 HIGH-WK, 20 LOW-WK). Across conditions, the syntactic configuration was held constant so that *both* potential nominal hosts (N1 vs. N2) were structurally available; the manipulation targeted a narrower, graded *world-knowledge plausibility* cue rather than broad pragmatic *possibility*. Concretely, items were constructed so that attachment to one host would be strongly favored by local event plausibility ("who plausibly does what"), while attachment to the alternative host remained syntactically licensed and broadly possible. We validated this plausibility contrast with independent norming ratings.

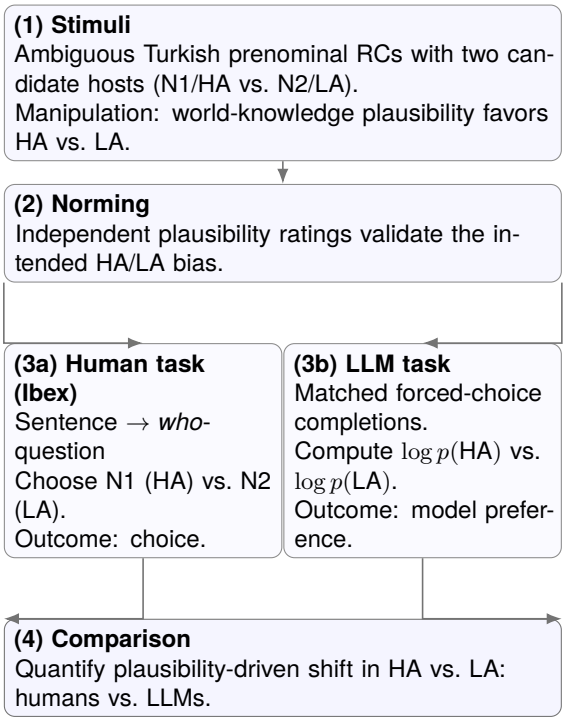


Figure 1: Procedure overview. After constructing syntactically matched Turkish RC attachment ambiguities and validating plausibility via norming, we evaluate (a) human attachment choices in a speeded forced-choice task and (b) LLM attachment preferences via log-probability scoring over matched HA/LA continuations.

Task and procedure. The experiment used a speeded forced-choice comprehension paradigm. On each trial, participants read an ambiguous RC sentence and then answered a *who*-question designed to force an attachment decision by selecting the intended RC host (N1 vs. N2). The model evaluation does not reproduce this task literally. Instead, it operationalizes attachment preference by comparing the relative probability of two short continuations that unambiguously instantiate the HA and LA readings. We therefore treat the human task and the model evaluation as parallel decision settings over the same underlying ambiguity, rather than as identical online measures of processing.

LLM experiments: log-probability scoring. In §4.1, we evaluate whether LLMs show plausibility-driven shifts over the same items by comparing the relative probability of matched HA and LA continuations. This does not constitute the same task as the human forced-choice experiment; rather, it provides a parallel model-side proxy for attachment preference under the same item manipulation.

(7)

$$s(\mathbf{y} \mid x) \;=\; \frac{1}{k}\sum_{t=1}^{k} \log P(y_t \mid x, y_1, \ldots, y_{t-1})\,.$$

We then define the log-probability preference:

(8)

$$\Delta \;=\; s(\mathbf{y}^{\mathsf{HA}} \mid x) \;-\; s(\mathbf{y}^{\mathsf{LA}} \mid x),$$

and take the model's predicted attachment to be HA iff $\Delta > 0$ (otherwise LA). This yields a categorical HA/LA outcome per item, which we analyze analogously to the human choices via logistic regression of HA on WK condition.

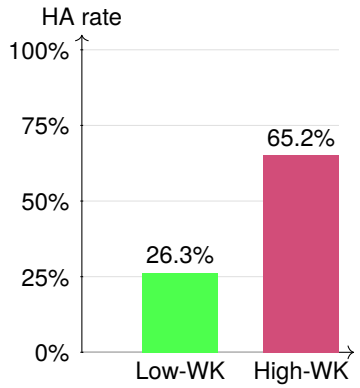
4.1. Models

We evaluate attachment preferences in autoregressive transformer language models by scoring the two matched HA/LA continuations described above using token-level conditional log-probabilities. We include two Turkish-specific checkpoints and one strong multilingual checkpoint:

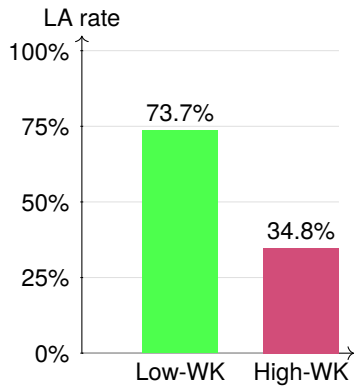
Ytu-ce-cosmos/turkish-gpt2. A Turkish GPT-2 style decoder-only language model used as a lightweight Turkish baseline. (Kesgin et al., 2024; ytu-ce-cosmos, 2024)

Duxx/DeepSeek-R1-Distill-Qwen-1.5B-Turkish. A Turkish-adapted reasoning-oriented model obtained by fine-tuning a distilled DeepSeek-R1 checkpoint (Qwen-1.5B backbone) on Turkish reasoning data. We use it as a compact model that may better reflect structured inference than a plain GPT-2 baseline. (Guo et al., 2025; duxx, 2025)

Qwen/Qwen3-30B-A3B-Instruct-2507. A multilingual mixture-of-experts instruction-tuned model (30.5B total parameters, with 3.3B activated per token). We include this model to test whether a



(a) HA by WK condition



(b) LA by WK condition

Figure 2: Human attachment rates by world-knowledge (WK) condition. Panel (a) shows HA rates; panel (b) shows the complementary LA rates (100–HA).

higher-capacity multilingual system shows more human-like sensitivity to world-knowledge plausibility in Turkish attachment. (Yang et al., 2025; Qwen, 2025)

Recent Turkish benchmarking places Qwen3-30B-Instruct among the stronger publicly reported multilingual models for Turkish. TurkBench evaluates 27 open-source models on 8,151 instances across 21 subtasks, and its leaderboard sorts systems by the *Avg* metric, that is, the overall average across task results; on this measure, Qwen3-30B-Instruct scores 73.4, compared with 78.6 for the top-ranked gpt-oss-120b (Toraman et al., 2026). Because several higher-scoring systems were not tractable in our setup, we use Qwen3-30B-Instruct as a strong multilingual comparison point.

5. Human Experiment Results

Figure 2 summarizes attachment choices by world-knowledge (WK) condition. In Low-WK contexts, where plausibility favors low attachment, participants selected high attachment (HA) on 26.3% of trials (panel a), corresponding to a low-attachment (LA) rate of 73.7% (panel b). In High-WK contexts, where plausibility favors HA, HA increased to 65.2% and LA decreased to 34.8%. This yields a 38.9 percentage-point shift in HA (65.2–26.3) and a mirror-image shift in LA (34.8–73.7), showing that graded plausibility cues robustly reweight attachment preferences even though both parses remain pragmatically possible in our materials.

To test whether WK reliably modulated attachment, we fit a logistic regression predicting HA (vs. LA) from WK condition. The effect was large and highly reliable ($\beta_{\text{WK}} = 1.65 \pm 0.11$, $z = 14.75$, $p < 10^{-50}$), corresponding to an odds ratio of $e^{1.65} \approx 5.2$ (a 5.2$\times$ increase in the odds of HA in High-WK relative to Low-WK). Collapsing across conditions, the overall HA rate was 45.7%, consistent with a modest baseline tendency toward low attachment.

Robustness checks converged with the regression results. A contingency analysis confirmed a strong association between WK condition and attachment choice ($\chi^2_{(1)} = 229.1$, $p = 1.7 \times 10^{-53}$). Moreover, across items, the strength of the plausibility manipulation as measured by independent norming ratings strongly tracked HA rates (Spearman $\rho = 0.85$, $p = 2 \times 10^{-6}$), indicating that graded world-knowledge plausibility reliably shapes attachment decisions in Turkish.

5.1. Model results

We evaluated model attachment preferences using the continuation-based log-probability comparison described in §4.1. For each item, we scored a high-attachment continuation and a low-attachment continuation and predicted HA when the HA continuation had higher mean per-token conditional log-probability (see (7)–(8)). Figure 3 summarizes HA choice rates by condition (blue: High-WK; orange: Low-WK).

The human pattern in Figure 3 shows a large, correctly directed plausibility effect: HA increases from 26.3% in Low-WK to 65.2% in High-WK. In contrast, the models show substantially weaker and less human-like modulation. Turkish GPT-2 exhibits a stable LA tendency: HA remains at 30% in both High-WK and Low-WK, with the Low-WK-only breakdown showing 70% LA and 30% HA in contexts intended to favor LA. DeepSeek-R1-Distill-Qwen-1.5B-Turkish shows only a small shift (HA 60% in High-WK vs. 50% in Low-WK), and its Low-WK-only breakdown is essentially balanced (50% HA vs. 50% LA). Qwen3 shows a different failure mode on the 20-item pilot (10 High-WK + 10 Low-WK): a strong overall HA bias and a reversed direction relative to the manipulation (HA 70% in High-WK but 90% in Low-WK), which is also visible

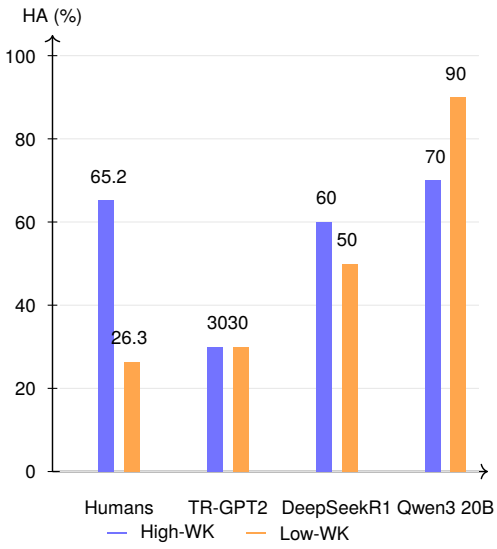


Figure 3: HA rates (%) by WK condition (High-WK vs. Low-WK) for humans and models.

in Figure 3 as only 10% LA in Low-WK items.

To summarize world-knowledge sensitivity across systems, we report the difference in HA rate across conditions: $\Delta\mathrm{HA} = \mathrm{HA}_{\text{High-WK}} - \mathrm{HA}_{\text{Low-WK}}$.

(9)

$$\begin{aligned}
\Delta\mathrm{HA}_{\text{Humans}} &= +38.9 \text{ pp} \quad (65.2 - 26.3) \\
\Delta\mathrm{HA}_{\text{TR-GPT2}} &= 0.0 \text{ pp} \quad (30 - 30) \\
\Delta\mathrm{HA}_{\text{DeepSeek}} &= +10.0 \text{ pp} \quad (60 - 50) \\
\Delta\mathrm{HA}_{\text{Qwen3-20}} &= -20.0 \text{ pp} \quad (70 - 90)
\end{aligned}$$

We tested condition effects on the categorical HA/LA outcomes using Fisher exact tests on the 2×2 contingency table (condition × attachment). Turkish GPT-2 shows no condition effect (6/20 HA in High-WK vs. 6/20 HA in Low-WK; $p = 1.0$). DeepSeek shows no reliable condition effect at this sample size (12/20 HA in High-WK vs. 10/20 HA in Low-WK; $p = 0.751$). Qwen3 also does not show a reliable categorical difference under Fisher's test on the 10+10 pilot (7/10 HA in High-WK vs. 9/10 HA in Low-WK; $p = 0.582$), despite the visibly reversed direction in Figure 3.

For Qwen3, we also analyzed the continuous log-probability margin to quantify how strongly the model prefers one attachment over the other. For each item i, we define the HA–LA margin as:

(10)

$$m_i = s(\mathbf{y}^{\text{HA}} \mid x_i) - s(\mathbf{y}^{\text{LA}} \mid x_i).$$

Larger m_i indicates a stronger preference for HA. Comparing margins across conditions reveals a statistically detectable effect, but in the wrong direction: Low-WK items are more HA-favoring than High-WK items. Concretely, the mean difference satisfies:

(11)

$$\mathbb{E}[m_{\text{Low-WK}} - m_{\text{High-WK}}] \approx 5.74 \text{ nats},$$

with a bootstrap 95% interval approximately $[1.15, 9.96]$. A Welch t-test gives $p = 0.0266$ and a Mann–Whitney U test gives $p = 0.0211$. Thus, even when the condition effect is detectable at the margin level, it reflects increased HA preference in Low-WK contexts, consistent with the reversed pattern in Figure 3.

Overall, the tests indicate that the models do not reproduce the strong, correctly directed plausibility-driven shift observed in humans. Instead, Turkish GPT-2 shows a rigid LA bias with no modulation, DeepSeek shows weak and unreliable modulation at this sample size, and Qwen3 shows a strong HA bias with a reversed condition effect on the pilot set.

6. Discussion

Our results reveal a sharp human–model dissociation in how graded world knowledge is used to resolve Turkish RC attachment ambiguity. Humans show a large, correctly directed plausibility effect: HA rises from 26.3% in Low-WK to 65.2% in High-WK (a +38.9 percentage-point shift), with the complementary LA plot showing the mirror-image decrease. By contrast, all tested LMs exhibit weak or qualitatively different behavior. In the abstract baselines, Turkish GPT-2 shows a stable LA bias (HA 30% in both conditions), while DeepSeek-R1-Distill-Qwen-1.5B-Turkish shows only a small shift (HA 60% vs. 50%). Qwen3, despite being a much stronger general-purpose model, shows a strong overall HA bias and, on our 20-item subset, a reversed WK effect (HA 70% in High-WK vs. 90% in Low-WK), yielding negative WK sensitivity.

A natural interpretation is that LLMs can store substantial commonsense and factual knowledge, yet deploy it unreliably in incremental ambiguity resolution where syntactic structure and plausibility must be integrated under tight constraints. At the same time, the comparison should be interpreted at the level of attachment preference rather than as a direct process-level alignment. Human participants performed a speeded comprehension task with an explicit question, whereas models were evaluated through relative preference for matched continuations. The value of the comparison, therefore, lies in asking whether the same plausibility manipulation shifts interpretation in the same direction across humans and models, not in claiming that the two systems were probed with an identical online measure.

This aligns with recent cognitive-science arguments that linguistic proficiency and human-like

conceptual reasoning can dissociate in LLMs, motivating careful use of targeted behavioral diagnostics rather than relying on broad benchmark success alone (Mahowald et al., 2024). In our setting, the plausibility manipulation is deliberately subtle: both parses remain pragmatically possible, so the task is not to reject an anomalous interpretation but to reweight two licensed structures using graded event knowledge. Humans do so robustly, but the models show much weaker and less consistent plausibility-driven shifts, suggesting that plausibility information is not being used as reliably as it is in human attachment resolution.

Reconciling with "LLMs are good at commonsense" results. At first glance, our findings may appear in tension with results showing strong LLM performance on commonsense and world-knowledge benchmarks, including large technical reports documenting broad benchmark strength for frontier models (Grattafiori et al., 2024), and recent multilingual physical-commonsense evaluations where state-of-the-art LLMs perform well in aggregate (Chang et al., 2025). There is also growing evidence that *scaffolding* methods can substantially improve commonsense QA (e.g., guided knowledge generation) (Wei et al., 2024). We argue that these successes are compatible with our dissociation because they often probe *knowledge access under explicit QA framing* (and sometimes with additional guidance), whereas Turkish RC attachment requires *knowledge deployment in real-time structure building*. In other words, passing a commonsense benchmark does not guarantee that the model will (i) retrieve the relevant event knowledge, (ii) align it with syntactic roles, and (iii) use it to *reweight* two grammatically licensed parses when neither interpretation is outright anomalous. Our paradigm targets exactly this "commonsense-in-use" requirement and shows that, relative to humans, current LLMs do not robustly integrate graded plausibility with syntactic attachment. In that sense, our study supports the broader conclusion that LLM commonsense reasoning remains unreliable: not necessarily because the knowledge is absent, but because it is not consistently *applied* when the task demands subtle, structurally constrained disambiguation.

Besides, our Turkish findings are consistent with a growing line of work arguing that attachment ambiguities remain a sensitive probe of LLM inductive biases. Recent multilingual studies report that models often default to local attachment and show limited responsiveness to language-specific or structure-sensitive patterns, with strong effects of evaluation framing and prompt format (Lee et al., 2025). Work focusing specifically on semantic bias in RC attachment similarly suggests that LLMs can miss or inconsistently apply plausibility cues that humans use, with variability across models and setups (Scheinberg et al., 2025). Hence, these findings support a picture in which LLM behavior on ambiguity resolution is not well predicted by their general benchmark strength, but is shaped by a mixture of base-rate structural tendencies, data-driven lexical associations, and sensitivity to how the disambiguation question is posed.

6.1. Why might Turkish be especially challenging?

One possibility is that Turkish prenominal RCs place heavier demands on anticipating structure before the head noun, and the relevant plausibility information is distributed across morphosyntax and lexical roles (Özge et al., 2015). If training data provides weaker or noisier supervision for these configurations (relative to high-resource Indo-European patterns), models may fall back on simpler heuristics that do not incorporate graded plausibility in a human-like way. Another possibility is that instruction-tuned models tend to produce overly peaked choice distributions even when multiple continuations are viable, which can interact with our aligned scoring protocol and yield a rigid bias (as in Qwen3) (Zhang et al., 2024). Distinguishing these explanations requires systematically varying continuation format, question framing, and the amount and position of plausibility-supporting context.

Furthermore, tokenization matters especially for morphologically rich and agglutinative languages such as Turkish, where many cues relevant to interpretation are carried by suffix chains and where subword tokenizers can fragment roots and affixes in ways that are only weakly aligned with linguistic units. In Turkish, prior work shows that tokenizer choice can substantially affect model behavior and downstream performance, and that simple morphological-level tokenization does not automatically yield better language modeling than subword approaches (Toraman et al., 2023; Bayram et al., 2025a,b). Complementary evidence on Turkish LLMs likewise reports measurable differences attributable to tokenization granularity, reinforcing that tokenizer design can alter what a model treats as an atomic cue and how efficiently it represents Turkish wordforms (Kaya and Tantuğ, 2024).

7. Conclusion

We presented a controlled, cross-population test of how graded world-knowledge plausibility shapes relative-clause attachment in Turkish prenominal RC ambiguities. Using normed materials in which *both* parses remained pragmatically possible, we

showed that native Turkish speakers robustly integrated event plausibility in attachment resolution: HA increased by 38.9 percentage points from Low-WK to High-WK. In contrast, the three autoregressive LLMs evaluated with a continuation-based log-probability comparison showed substantially weaker, less consistent, and in one case reversed sensitivity to the same plausibility manipulation. Overall, the model results suggest that graded plausibility did not guide attachment preferences as reliably as it did in human judgments.

These findings are consistent with the possibility that, in the tested models, plausibility information is not incorporated as reliably when attachment resolution requires tightly constrained integration of world knowledge with hierarchical structure. More broadly, the results suggest that strong overall benchmark performance does not by itself guarantee human-like cue integration in psycholinguistic tasks, and that attachment ambiguities under subtle plausibility manipulations can provide a useful diagnostic for comparing human and model behavior in language understanding.

Limitations

First, the stimulus set is modest, which limits power to detect small plausibility effects and reduces confidence in fine-grained cross-model comparisons. Second, the model suite is necessarily selective and excludes the very strongest Turkish-capable systems, so our conclusions are restricted to the tested checkpoints. Finally, we compare categorical human choices to model preferences rather than directly matching incremental time-course measures; future work may add stronger online proxies such as region-wise surprisal profiles to more closely link models to human processing. Finally, future work should also investigate whether corpus evidence can help clarify baseline attachment tendencies in Turkish and better contextualize the experimental results.

Acknowledgments

I thank Chris Kennedy for valuable discussions in his *Linguistics and LLMs* class. I am also grateful to Muharrem Taha Aydın and Mustafa Baki Varol for their help with the design, discussion of the results, and feedback. Finally, I thank the anonymous reviewers for their thoughtful comments and suggestions.

8. Bibliographical References

Taylan Akal. 2021. Recency preference in ambiguous relative clause attachment in Turkish. *Journal of Language and Linguistic Studies*, 17(Special Issue 1):139–159.

Gerry Altmann and Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition*, 30(3):191–238.

Diego Alves. 2025. Benchmarking language model surprisal for eye-tracking predictions in Brazilian Portuguese. In *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pages 7–17, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313. Association for Computational Linguistics.

Thierry Baccino, Marica De Vincenzi, and Remo Job. 2000. Cross-linguistic studies of the late closure strategy: French and italian. In Marica De Vincenzi and Vincenzo Lombardo, editors, *Cross-Linguistic Perspectives on Language Processing*, volume 25 of *Studies in Theoretical Psycholinguistics*, pages 89–118. Springer, Dordrecht.

Zeynep Başer and Annette Hohenberger. 2020. Is there a particular RC attachment preference in Turkish? negotiating the effects of semantic factors. *Journal of Psycholinguistic Research*, 49(4):511–539.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, and Savaş Yıldırım. 2025a. Tokenization standards for linguistic integrity: Turkish as a benchmark.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, Savaş Yıldırım, and Demircan Çelik. 2025b. Tokens with meaning: A hybrid tokenization approach for nlp.

Zeynep Başer. 2018. *Syntactic Priming of Relative Clause Attachment in Monolingual Turkish Speakers and Turkish Learners of English*. Ph.D. thesis, Middle East Technical University.

Sam Boeve and Louisa Bogaerts. 2025. A systematic evaluation of Dutch large language models' surprisal estimates in sentence, paragraph and book reading. *Behavior Research Methods*, 57:266. Published 18 Aug 2025; article number 266.

Tyler A. Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Condez, Andiswa Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anna Vacalopoulou, Annette Kuuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Ashvanth. S, Ayodeji Joseph Lana, Barkin Kinay, Bashar Alhafni, Benedict Cibalinda Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska Đurić, Bola Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Alakuş Çınar, Cade J. Kanoniakapueo Kane, Can Udomcharoenchaikit, Catherine Arnett, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Glory Nwokolo, Cristina España-Bonet, Cynthia Amol, DaeYeop Lee, Dana Arad, Daniil Dzenhaliou, Daria Pugacheva, Dasol Choi, Daud Abolade, David Liu, David Semedo, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dhyuthy Krishna Kumar, Diogo Glória-Silva, Diogo Tavares, Divyanshu Goyal, DongGeon Lee, Ebele Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elena Tutubalina, Elias Herranen, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Olabisi Itiola, Faran Taimoor Butt, Fathima Thekkekara, Fatima Haouari, Filbert Aurelian Tjiaranata, Firas Laakom, Francesca Grasso, Francesco Orabona, Francesco Periti, Gbenga Kayode Solomon, Gia Nghia Ngo, Gloria Udhehdhe-oze, Gonçalo Martins, Gopi Naga Sai Ram Challagolla, Guijin Son, Gulnaz Abdykadyrova, Hafsteinn Einarsson, Hai Hu, Hamidreza Saffari, Hamza Zaidi, Haopeng Zhang, Harethah Abu Shairah, Harry Vuong, Hele-Andra Kuulmets, Houda Bouamor, Hwanjo Yu, Iben Nyholm Debess, İbrahim Ethem Deveci, Ikhlasul Akmal Hanif, Ikhyun Cho, Inês Calvo, Inês Vieira, Isaac Manzi, Ismail Daud, Itay Itzhak, Iuliia, Alekseenko, Ivan Belashkin, Ivan Spada, Ivan Zhelyazkov, Jacob Brinton, Jafar Isbarov, Jaka Čibej, Jan Čuhel, Jan Kocoń, Jauza Akbar Krito, Jebish Purbey, Jennifer Mickel, Jennifer Za, Jenny Kunz, Jihae Jeong, Jimena Tena Dávalos, Jinu Lee, João Magalhães, John Yi, Jongin Kim, Joseph Chataignon, Joseph Marvin Imperial, Jubeerathan Thevakumar, Judith Land, Junchen Jiang, Jungwhan Kim, Kairit Sirts, Kamesh R, Kamesh V, Kanda Patrick Tshinu, Kätriin Kukk, Kaustubh Ponkshe, Kavsar Huseynova, Ke He, Kelly Buchanan, Kengatharaiyer Sarveswaran, Kerem Zaman, Khalil Mrini, Kian Kyars, Krister Kruusmaa, Kusum Chouhan, Lainitha Krishnakumar, Laura Castro Sánchez, Laura Porrino Moscoso, Leshem Choshen, Levent Sencan, Lilja Øvrelid, Lisa Alazraki, Lovina Ehimen-Ugbede, Luheerathan Thevakumar, Luxshan Thavarasa, Mahnoor Malik, Mamadou K. Keita, Mansi Jangid, Marco De Santis, Marcos García, Marek Suppa, Mariam D'Ciofalo, Marii

Ojastu, Maryam Sikander, Mausami Narayan, Maximos Skandalis, Mehak Mehak, Mehmet İlteriş Bozkurt, Melaku Bayu Workie, Menan Velayuthan, Michael Leventhal, Michał Marcińczuk, Mirna Potočnjak, Mohammadamin Shafiei, Mridul Sharma, Mrityunjaya Indoria, Muhammad Ravi Shulthan Habibi, Murat Kolić, Nada Galant, Naphat Permpredanun, Narada Maugin, Nicholas Kluge Corrêa, Nikola Ljubešić, Nirmal Thomas, Nisansa de Silva, Nisheeth Joshi, Nitish Ponkshe, Nizar Habash, Nneoma C. Udeze, Noel Thomas, Noémi Ligeti-Nagy, Nouhoum Coulibaly, Nsengiyumva Faustin, Odunayo Kareemat Buliaminu, Odunayo Ogundepo, Oghojafor Godswill Fejiro, Ogundipe Blessing Funmilola, Okechukwu God'spraise, Olanrewaju Samuel, Olaoye Deborah Oluwaseun, Olasoji Akindejoye, Olga Popova, Olga Snissarenko, Onyinye Anulika Chiemezie, Orkun Kinay, Osman Tursun, Owoeye Tobiloba Moses, Oyelade Oluwafemi Joshua, Oyesanmi Fiyinfoluwa, Pablo Gamallo, Pablo Rodríguez Fernández, Palak Arora, Pedro Valente, Peter Rupnik, Philip Oghenesuowho Ekiugbo, Pramit Sahoo, Prokopis Prokopidis, Pua Niau-Puhipau, Quadri Yahya, Rachele Mignone, Raghav Singhal, Ram Mohan Rao Kadiyala, Raphael Merx, Rapheal Afolayan, Ratnavel Rajalakshmi, Rishav Ghosh, Romina Oji, Ron Kekeha Solis, Rui Guerra, Rushikesh Zawar, Sa'ad Nasir Bashir, Saeed Alzaabi, Sahil Sandeep, Sai Pavan Batchu, SaiSandeep Kantareddy, Salsabila Zahirah Pranida, Sam Buchanan, Samuel Rutunda, Sander Land, Sarah Sulollari, Sardar Ali, Saroj Sapkota, Saulius Tautvaisas, Sayambhu Sen, Sayantani Banerjee, Sebastien Diarra, SenthilNathan. M, Sewoong Lee, Shaan Shah, Shankar Venkitachalam, Sharifa Djurabaeva, Sharon Ibejih, Shivanya Shomir Dutta, Siddhant Gupta, Silvia Paniagua Suárez, Sina Ahmadi, Sivasuthan Sukumar, Siyuan Song, Snegha A., Sokratis Sofianopoulos, Sona Elza Simon, Sonja Benčina, Sophie Gvasalia, Sphurti Kirit More, Spyros Dragazis, Stephan P. Kaufhold, Suba. S, Sultan AlRashed, Surangika Ranathunga, Taiga Someya, Taja Kuzman Pungeršek, Tal Haklay, Tasi'u Jibril, Tatsuya Aoyama, Tea Abashidze, Terenz Jomar Dela Cruz, Terra Blevins, Themistoklis Nikas, Theresa Dora Idoko, Thu Mai Do, Tilek Chubakov, Tommaso Gargiani, Uma Rathore, Uni Johannesen, Uwuma Doris Ugwu, Vallerie Alexandra Putra, Vanya Bannihatti Kumar, Varsha Jeyarajalingam, Varvara Arzt, Vasudevan Nedumpozhimana, Viktoria Ondrejova, Viktoryia Horbik, Vishnu Vardhan Reddy Kummitha, Vuk Dinić, Walelign Tewabe Sewunetie, Winston Wu, Xiaojing Zhao, Yacouba Diarra, Yaniv Nikankin, Yash Mathur, Yixi Chen, Yiyuan Li, Yolanda Xavier, Yonatan Belinkov, Yusuf Ismail Abayomi, Zaid Alyafeai, Zhengyang Shan, Zhi Rui Tam, Zilu Tang, Zuzana Nadova, Baber Abbasi, Stella Biderman, David Stap, Duygu Ataman, Fabian Schmidt, Hila Gonen, Jiayi Wang, and David Ifeoluwa Adelani. 2025. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020. Interaction with context during recurrent neural network sentence processing. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020)*. Conference paper.

Andrea de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144, Online only. Association for Computational Linguistics.

duxx. 2025. DeepSeek-R1-Distill-Qwen-1.5B-Turkish. Hugging Face model card. Accessed 2026-02-09.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Eva M. Fernández. 2003. *Bilingual Sentence Processing: Relative Clause Attachment in English and Spanish*, volume 29 of *Language Acquisition and Language Disorders*. John Benjamins, Amsterdam/Philadelphia.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. Translategemma technical report.

Lyn Frazier and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*, 6(4):291–325.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Edward Gibson, Neal J. Pearlmutter, Elizabeth Canseco-Gonzalez, and Gregory Hickok. 1996. Recency preference in the human sentence processing mechanism. *Cognition*, 59(1):23–59.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tin-

dal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen

Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*.

Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. 2021. Multilingual language models predict human reading behavior. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 106–123, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sercan Karakaş. 2026. Clause-internal or clause-external? testing turkish reflexive binding in adapted versus chain of thought large language models.

Yiğit Bekir Kaya and A. Cüneyd Tantuğ. 2024. Effect of tokenization granularity for turkish large language models. *Intelligent Systems with Applications*, 21:200335.

Onur Keleş and Nazik Dinçtopal Deniz. 2024. A comparative study with human data: Do LLMs have superficial language processing? In *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, Mersin, Türkiye. IEEE. IEEE Xplore document 10600807; proceedings held 15–18 May 2024 in Mersin, Türkiye.

H. Toprak Kesgin, M. Kaan Yuce, Eren Dogan, M. Egemen Uzun, Atahan Uz, H. Emre Seyrek, Ahmed Zeer, and M. Fatih Amasyali. 2024. Introducing cosmosgpt: Monolingual training for turkish language models.

John Kimball. 1973. Seven principles of surface structure parsing in natural language. *Cognition*, 2(1):15–47.

Bilal Kırkıcı. 2004. The processing of relative clause attachment ambiguities in turkish. *Turkic Languages*, 8. Preprint PDF available online.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

So Young Lee, Russell Scheinberg, Amber Shore, and Ameeta Agrawal. 2025. Who relies more on world knowledge and bias for syntactic ambiguity resolution: Humans or LLMs? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3484–3498, Albuquerque, New Mexico. Association for Computational Linguistics.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Ethan Li, Anders Boesen Lindbo Larsen, Chen Zhang, Xiyou Zhou, Jun Qin, Dian Ang Yap, Narendran Raghavan, Xuankai Chang, Margit Bowler, Eray Yildiz, John Peebles, Hannah Gillis Coleman, Matteo Ronchi, Peter Gray, Keen You, Anthony Spalvieri-Kruse, Ruoming Pang, Reed Li, Yuli Yang, Emad Soroush, Zhiyun Lu, Crystal Xiao, Rong Situ, Jordan Huffaker, David Griffiths, Zaid Ahmed, Peng Zhang, Daniel Parilla, Asaf Liberman, Jennifer Mallalieu, Parsa Mazaheri, Qibin Chen, Manjot Bilkhu, Aonan Zhang, Eric Wang, Dave Nelson, Michael FitzMaurice, Thomas Voice, Jeremy Liu, Josh Shaffer, Shiwen Zhao, Prasanth Yadla, Farzin Rasteh, Pengsheng Guo, Arsalan Farooq, Jeremy Snow, Stephen Murphy, Tao Lei, Minsik Cho, George Horrell, Sam Dodge, Lindsay Hislop, Sumeet Singh, Alex Dombrowski, Aiswarya Raghavan,

Sasha Sirovica, Mandana Saebi, Faye Lao, Max Lam, TJ Lu, Zhaoyang Xu, Karanjeet Singh, Marc Kirchner, David Mizrahi, Rajat Arora, Haotian Zhang, Henry Mason, Lawrence Zhou, Yi Hua, Ankur Jain, Felix Bai, Joseph Astrauskas, Floris Weers, Josh Gardner, Mira Chiang, Yi Zhang, Pulkit Agrawal, Tony Sun, Quentin Keunebroek, Matthew Hopkins, Bugu Wu, Tao Jia, Chen Chen, Xingyu Zhou, Nanzhu Wang, Peng Liu, Ruixuan Hou, Rene Rauch, Yuan Gao, Afshin Dehghan, Jonathan Janke, Zirui Wang, Cha Chen, Xiaoyi Ren, Feng Nan, Josh Elman, Dong Yin, Yusuf Goren, Jeff Lai, Yiran Fei, Syd Evans, Muyang Yu, Guoli Yin, Yi Qin, Erin Feldman, Isha Garg, Aparna Rajamani, Karla Vega, Walker Cheng, TJ Collins, Hans Han, Raul Rea Menacho, Simon Yeung, Sophy Lee, Phani Mutyala, Ying-Chang Cheng, Zhe Gan, Sprite Chu, Justin Lazarow, Alessandro Pappalardo, Federico Scozzafava, Jing Lu, Erik Daxberger, Laurent Duchesne, Jen Liu, David Güera, Stefano Ligas, Mary Beth Kery, Brent Ramerth, Ciro Sannino, Marcin Eichner, Haoshuo Huang, Rui Qian, Moritz Schwarzer-Becker, David Riazati, Mingfei Gao, Bailin Wang, Jack Cackler, Yang Lu, Ransen Niu, John Dennison, Guillaume Klein, Jeffrey Bigham, Deepak Gopinath, Navid Shiee, Darren Botten, Guillaume Tartavel, Alex Guillen Garcia, Sam Xu, Victoria MönchJuan Haladjian, Zi-Yi Dou, Matthias Paulik, Adolfo Lopez Mendez, Zhen Li, Hong-You Chen, Chao Jia, Dhaval Doshi, Zhengdong Zhang, Raunak Manjani, Aaron Franklin, Zhile Ren, David Chen, Artsiom Peshko, Nandhitha Raghuram, Hans Hao, Jiulong Shan, Kavya Nerella, Ramsey Tantawi, Vivek Kumar, Saiwen Wang, Brycen Wershing, Bhuwan Dhingra, Dhruti Shah, Ob Adaranijo, Xin Zheng, Tait Madsen, Hadas Kotek, Chang Liu, Yin Xia, Hanli Li, Suma Jayaram, Yanchao Sun, Ahmed Fakhry, Vasileios Saveris, Dustin Withers, Yanghao Li, Alp Aygar, Andres Romero Mier Y Teran, Kaiwei Huang, Mark Lee, Xiujun Li, Yuhong Li, Tyler Johnson, Jay Tang, Joseph Yitan Cheng, Futang Peng, Andrew Walkingshaw, Lucas Guibert, Abhishek Sharma, Cheng Shen, Piotr Maj, Yasutaka Tanaka, You-Cyuan Jhang, Vivian Ma, Tommi Vehvilainen, Kelvin Zou, Jeff Nichols, Matthew Lei, David Qiu, Yihao Qian, Gokul Santhanam, Wentao Wu, Yena Han, Dominik Moritz, Haijing Fu, Mingze Xu, Vivek Rathod, Jian Liu, Louis D'hauwe, Qin Ba, Haitian Sun, Haoran Yan, Philipp Dufter, Anh Nguyen, Yihao Feng, Emma Wang, Keyu He, Rahul Nair, Sanskruti Shah, Jiarui Lu, Patrick Sonnenberg, Jeremy Warner, Yuanzhi Li, Bowen Pan, Ziyi Zhong, Joe Zhou, Sam Davarnia, Olli Saarikivi, Irina Belousova, Rachel Burger, Shang-Chen Wu, Di Feng, Bas Straathof, James Chou, Yuanyang Zhang, Marco Zuliani, Eduardo Jimenez, Abhishek Sundararajan, Xianzhi Du, Chang Lan, Nilesh Shahdadpuri, Peter Grasch, Sergiu Sima, Josh Newnham, Varsha Paidi, Jianyu Wang, Kaelen Haag, Alex Braunstein, Daniele Molinari, Richard Wei, Brenda Yang, Nicholas Lusskin, Joanna Arreaza-Taylor, Meng Cao, Nicholas Seidl, Simon Wang, Jiaming Hu, Yiping Ma, Mengyu Li, Kieran Liu, Hang Su, Sachin Ravi, Chong Wang, Xin Wang, Kevin Smith, Haoxuan You, Binazir Karimzadeh, Rui Li, Jinhao Lei, Wei Fang, Alec Doane, Sam Wiseman, Ismael Fernandez, Jane Li, Andrew Hansen, Javier Movellan, Christopher Neubauer, Hanzhi Zhou, Chris Chaney, Nazir Kamaldin, Valentin Wolf, Fernando Bermúdez-Medina, Joris Pelemans, Peter Fu, Howard Xing, Xiang Kong, Wayne Shan, Gabriel Jacoby-Cooper, Dongcai Shen, Tom Gunter, Guillaume Seguin, Fangping Shi, Shiyu Li, Yang Xu, Areeba Kamal, Dan Masi, Saptarshi Guha, Qi Zhu, Jenna Thibodeau, Changyuan Zhang, Rebecca Callahan, Charles Maalouf, Wilson Tsao, Boyue Li, Qingqing Cao, Naomy Sabo, Cheng Leong, Yi Wang, Anupama Mann Anupama, Colorado Reed, Kenneth Jung, Zhifeng Chen, Mohana Prasad Sathya Moorthy, Yifei He, Erik Hornberger, Devi Krishna, Senyu Tong, Michael, Lee, David Haldimann, Yang Zhao, Bowen Zhang, Chang Gao, Chris Bartels, Sushma Rao, Nathalie Tran, Simon Lehnerer, Co Giang, Patrick Dong, Junting Pan, Biyao Wang, Dongxu Li, Mehrdad Farajtabar, Dongseong Hwang, Grace Duanmu, Eshan Verma, Sujeeth Reddy, Qi Shan, Hongbin Gao, Nan Du, Pragnya Sridhar, Forrest Huang, Yingbo Wang, Nikhil Bhendawade, Diane Zhu, Sai Aitharaju, Fred Hohman, Lauren Gardiner, Chung-Cheng Chiu, Yinfei Yang, Alper Kokmen, Frank Chu, Ke Ye, Kaan Elgin, Oron Levy, John Park, Donald Zhang, Eldon Schoop, Nina Wenzel, Michael Booker, Hyunjik Kim, Chinguun Erdenebileg, Nan Dun, Eric Liang Yang, Priyal Chhatrapati, Vishaal Mahtani, Haiming Gang, Kohen Chia, Deepa Seshadri, Donghan Yu, Yan Meng, Kelsey Peterson, Zhen Yang, Yongqiang Wang, Carina Peng, Doug Kang, Anuva Agarwal, Albert Antony, Juan Lao Tebar, Albin Madappally Jose, Regan Poston, Andy De Wang, Gerard Casamayor, Elmira Amirloo, Violet Yao, Wojciech Kryscinski, Kun Duan, and Lezhi L. 2025. Apple intelligence foundation language models: Tech report 2025.

Pavel Logačev, Özgur Aydın, and Aylin Müge Tuncer. 2022. Absence of evidence for underspecification in prenominal relative clause attachment. Manuscript (version dated January 8, 2022); includes Turkish eye-tracking and self-paced reading.

Kyle Mahowald, Anna Ivanova, Idan Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Ken McRae, Michael J. Spivey-Knowlton, and Michael K. Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in online sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Byung-Doh Oh and William Schuler. 2023. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1915–1921, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu,

Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Duygu Özge, Theodoros Marinis, and Deniz Zeyrek. 2015. Incremental processing in head-final child language: Online comprehension of relative clauses in turkish-speaking children and adults. *Language, Cognition and Neuroscience*.

Qwen. 2025. Qwen3-30B-A3B-Instruct-2507. Hugging Face model card. Accessed 2026-02-09.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. 2020. A benchmark for systematic generalization in grounded language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Russell Scheinberg, So Young Lee, and Ameeta Agrawal. 2025. Missing the cues: LLMs' insensitivity to semantic biases in relative clause attachment. *Proceedings of the Linguistic Society of America*, 10(1):5902.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh,

Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card.

Michael J. Spivey-Knowlton, John C. Trueswell, and Michael K. Tanenhaus. 1993. Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 47(2):276–309. PsycNet record: https://psycnet.apa.org/record/1994-04330-001.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma,

Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Çağrı Toraman, Ahmet Kaan Sever, Ayse Aysu Cengiz, Elif Ecem Arslan, Görkem Sevinç, Mete Mert Birdal, Yusuf Faruk Güldemir, Ali Buğra Kanburoğlu, Sezen Felekoğlu, Osman Gürlek, Sarp Kantar, Birsen Şahin Kütük, Büşra Tufan, Elif Genç, Serkan Coşkun, Gupse Ekin Demir, Muhammed Emin Arayıcı, Olgun Dursun, Onur Gungor, Susan Üsküdarlı, Abdullah Topraksoy, and Esra Darıcı. 2026. Turkbench: A benchmark for evaluating turkish large language models.

Çağrı Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource*

Language Information Processing, 22(4). Also available as arXiv:2204.08832.

Xiao Wei, Haoran Chen, Hang Yu, Hao Fei, and Qian Liu. 2024. Guided knowledge generation with language models for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Haoran Yang, Hongyuan Lu, Wai Lam, and Deng Cai. 2024. Exploring compositional generalization of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 16–24, Mexico City, Mexico. Association for Computational Linguistics.

ytu-ce-cosmos. 2024. turkish-gpt2. Hugging Face model card. Accessed 2026-02-09.

Jeremy Zehr and Florian Schwarz. 2018. Penncontroller for internet based experiments (ibex).

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. 2024. Forcing diffuse distributions out of language models.

Benchmarking Source-Sensitive Reasoning in Turkish: Humans and LLMs under Evidential Trust Manipulation

Sercan Karakaş[1], **Yusuf Şimşek**[2]
[1]University of Chicago, skarakas@uchicago.edu
[2]Fırat University, ysimsek@firat.edu.tr

Abstract

This paper investigates whether *source trustworthiness* shapes Turkish evidential morphology and whether large language models (LLMs) track this sensitivity. We study the past-domain contrast between *-DI* and *-mIş* in controlled cloze contexts where the information source is overtly external, while only its perceived reliability is manipulated (High-Trust vs. Low-Trust). In a human production experiment, native speakers of Turkish show a robust trust effect: High-Trust contexts yield relatively more *-DI*, whereas Low-Trust contexts yield relatively more *-mIş*, with the pattern remaining stable across sensitivity analyses. We then evaluate 10 LLMs in three prompting paradigms (open gap-fill, explicit past-tense gap-fill, and forced-choice A/B selection). LLM behavior is highly model- and prompt-dependent: some models show weak or local trust-consistent shifts, but effects are generally unstable, often reversed, and frequently overshadowed by output-compliance problems and strong base-rate suffix preferences. The results provide new evidence for a trust-/commitment-based account of Turkish evidentiality and reveal a clear human–LLM gap in source-sensitive evidential reasoning.

Keywords: Turkish evidentiality, trustworthiness, LLMs, cloze task, human–LLM comparison, benchmarks

1. Introduction

Evidentiality refers to the linguistic encoding of information source, that is, how a speaker indicates the basis on which a proposition is presented (e.g., direct perception, inference, or report) (Willett, 1988; Dendale and Tasmowski, 2001; de Haan, 2001; Plungian, 2001; Aikhenvald, 2004; Boye, 2012; Ünal and Papafragou, 2020). In many languages, evidential markers systematically signal whether the speaker witnessed an event, inferred it from available evidence, or learned it from someone else. Because evidentiality lies at the interface of semantics, pragmatics, and discourse, it has been central to research on speaker commitment, reliability, and perspective in natural language.

A useful perspective on evidentiality comes from language acquisition and "thinking for speaking": as children learn a language, they learn to attend to distinctions that the language regularly encodes (Brown and Lenneberg, 1954; Slobin, 1996). In evidential languages, this can make knowledge source (e.g., direct experience, inference, report) especially salient early on, since speakers must repeatedly track and express it in discourse. For Turkish, this suggests that evidential morphology is not just grammatical, but part of a broader cognitive routine linking source monitoring, memory, and speaker stance (Aksu-Koç and Slobin, 1986). Turkish offers a particularly important case for evidentiality research because past-domain morphology is closely tied to distinctions in information source and epistemic stance (Aksu-Koç and Slobin, 1986; Underhill, 1976; Johanson, 2003; Göksel and Kerslake, 2005; Kornfilt, 1997). A well-known contrast is between the suffix *-DI*, traditionally associated with direct evidence or stronger speaker commitment, and the suffix *-mIş*, which is commonly associated with indirect evidence, such as inference or report (Aksu-Koç and Slobin, 1986; Underhill, 1976; Johanson, 2003; Gül, 2009; Şener, 2011; Göksel and Kerslake, 2005).

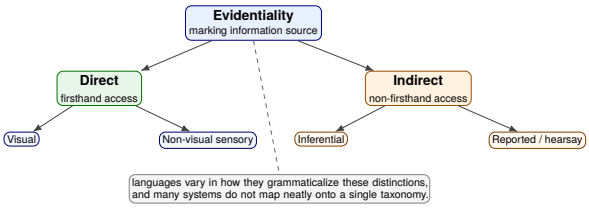


Figure 1: Information-source marking, adapted from typological overviews of evidential systems (Willett, 1988; Aikhenvald, 2004).

A minimal illustration is given in (1)–(2): in (1), the use of *-DI* is classically associated with a context in which the speaker directly witnessed the event, whereas in (2), the use of *-mIş* suggests that the speaker did not directly witness the event but instead inferred it or learned it from another source. At the same time, a substantial body of work notes that the actual distribution and interpretation of these forms are sensitive to discourse context, information structure, and speaker stance, so the traditional labels are best understood as useful descriptive generalizations rather than rigid one-to-one mappings between form and meaning

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 329–346
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

(Gül, 2009; Şener, 2011).

(1) Ahmet Ali'yi gör-dü.
Ahmet Ali-ACC see-PST.DIR
'Ahmet saw Ali (witnessed/direct evidence).'

(2) Ahmet Ali'yi gör-müş.
Ahmet Ali-ACC see-PST.INDIR
'Ahmet apparently/reportedly saw Ali (hearsay/inference).'

The *-DI*/*-mIş* contrast offers a well-studied case for investigating how languages encode information source, speaker stance, and pragmatic inference (Aksu-Koç and Slobin, 1986; Johanson, 2003; Göksel and Kerslake, 2005; Gül, 2009; Şener, 2011; Kandemirci et al., 2023; Ataman-Devrim et al., 2025), and it is also central to developmental work on how children learn to track and express knowledge source in discourse (Aksu-Koç and Slobin, 1986; Slobin, 1996). In addition, evidential distinctions are highly relevant for NLP tasks such as translation, summarization, and source-sensitive interpretation.

1.1. Why evidentiality matters for LLM research?

Evidentiality is also increasingly relevant to LLM research, because many of the core reliability problems in modern text generation are, in effect, problems about *source status*: whether a claim is grounded in direct input evidence, retrieved documents, parametric memory, inference, or unsupported generation. Recent work on factuality and hallucination has shown that fluent model outputs may be unfaithful to available evidence or false in world knowledge terms, motivating finer-grained analyses of how models represent and communicate the basis of their claims (Maynez et al., 2020; Lin et al., 2022; Manakul et al., 2023; Min et al., 2023; Bang et al., 2025; Liu et al., 2026). In parallel, work on calibration and self-evaluation suggests that models can sometimes track uncertainty or answerability, but this ability is uneven and sensitive to task format and available context (Kadavath et al., 2022). At the same time, frontier and open models are improving rapidly in reasoning, long-context processing, multilinguality, and agentic performance, as reflected in recent reports (Singh et al., 2025; Yang et al., 2025). These gains make source-sensitive evaluation more, not less, important: as models become stronger and more fluent, the central question is increasingly not only whether they can produce plausible language, but whether they can reliably represent the evidential basis of what they say.

From this perspective, evidentiality offers a linguistically grounded framework for studying how systems encode (or fail to encode) distinctions among *witnessed*, *inferred*, and *reported* information. This is especially important in settings that explicitly require provenance, such as retrieval-augmented generation and citation-based QA, where models must connect claims to supporting evidence and where attribution quality itself has become a major evaluation target (Lewis et al., 2020; Nakano et al., 2022; Gao et al., 2023; Li et al., 2024). More broadly, evidential categories provide a principled way to ask whether LLMs merely optimize surface plausibility or whether they track source-sensitive epistemic distinctions that are central to human communication. Besides, this line of work is also important because direct *human–LLM* comparisons for Turkish remain relatively scarce, especially for source-sensitive semantic/pragmatic phenomena such as evidentiality. Much of the recent Turkish LLM literature has focused on model development, benchmarking, and engineering issues (Bayram et al., 2025; Er et al., 2025; Isbarov et al., 2025; Umutlu et al., 2025; Wicaksono et al., 2025; Bayram et al., 2026; Karakaş and Şimşek, 2026; Toraman et al., 2026; Uğur et al., 2026), rather than tightly controlled comparisons between human judgments and model preferences on linguistically targeted contrasts. Emerging Turkish-focused human–LLM studies suggest that such comparisons can reveal substantial mismatches between surface-form preferences and human-like contextual reasoning in several linguistic domains (Keleş and Dinçtopal Deniz, 2024, 2025; Karakaş, 2026).

2. Related Work

Turkish evidentiality: traditional descriptive and grammatical accounts. Turkish has long occupied a central place in the evidentiality literature because its past-domain morphology is closely tied to distinctions in information source and speaker stance (Underhill, 1976; Aksu-Koç and Slobin, 1986; Johanson, 2003; Kornfilt, 1997; Göksel and Kerslake, 2005). A traditional descriptive contrast distinguishes *-DI* and *-mIş*: *-DI* is often introduced as the form associated with direct evidence (or stronger speaker commitment), while *-mIş* is associated with indirect evidence, especially inferential and reportative uses (Underhill, 1976; Johanson, 2003; Göksel and Kerslake, 2005; Aksu-Koç and Slobin, 1986). This contrast has also been influential in acquisition and discourse-based work, where Turkish evidential morphology is treated as a system that helps speakers track and communicate how information was obtained (Aksu-Koç and Slobin, 1986).

At the same time, the Turkish literature shows that the traditional mapping is not a simple one-to-one encoding. A central debate is whether *-DI* is a true direct evidential, or primarily a past

tense form whose "directness" and stronger commitment effects arise through discourse-pragmatic inference in canonical contexts (Gül, 2009; Şener, 2011). Likewise, analyses of *-mIş* emphasize sensitivity to evidential source, discourse configuration, and speaker perspective, rather than a single fixed meaning (Gül, 2009; Johanson, 2003). For this reason, recent work often treats *direct*/*indirect* as useful descriptive labels while allowing context-sensitive meaning and pragmatic strengthening.

For the present study, we adopt Giannakidou and Mari's framework on veridicality, nonveridicality, commitment, and evidential bias (Giannakidou, 1998, 2006; Giannakidou and Mari, 2016, 2018, 2021). This framework is useful because it separates notions often conflated in evidentiality research: (i) information source, (ii) speaker commitment, and (iii) the structure of the speaker's information state. It also motivates treating source trustworthiness as a factor shaping evidential interpretation and commitment (Boscaro et al., 2025).

Let $M = M_i(w, t)$ be speaker i's information state at world w and time t. A compact way to encode the relevant distinction is to contrast veridical and nonveridical states:

(3) $\textsc{Veridical}_M(p)$ iff all worlds in M are p-worlds: $\forall w \in M\,[p(w)]$.

(4) $\textsc{Nonveridical}_M(p)$ iff M contains both a p-world and a $\neg p$-world: $\exists w \in M\,[p(w)]$ and $\exists v \in M\,[\neg p(v)]$.

The formula in (3) states that the speaker's information state supports only p-worlds (full support for p), while the formula in (4) states that the information state remains compatible with both p and $\neg p$ possibilities. In this perspective, evidential and modal expressions can encode an evidence-based bias toward p without requiring full veridical commitment.

This distinction is useful for Turkish evidentiality because it allows us to ask not only which source type is invoked, but also whether that source is treated as strong enough to support a more veridical commitment profile. In our trust-based design, source trustworthiness is therefore modeled as a factor that can shift speakers between stronger commitment and weaker, evidence-biased interpretations, rather than as a purely descriptive source label.

3. Methods

To examine how Turkish evidential morphology responds to contextual *trustworthiness*, we used two complementary experimental paradigms with two corresponding datasets. First, we designed a controlled cloze-style task that can be administered to both human participants and LLMs. Second, we constructed a larger dataset for LLM-only evaluation in order to test generalization over a broader set of contexts. In all tasks, the target is a *cloze* (fill-in-the-blank) completion: participants or models infer and produce the missing final word of a sentence from the preceding context. This shared format provides a direct measure of context-sensitive morphological preferences in production and enables human–model comparison within a single task setup.

3.1. Dataset 1 for Human and LLM Comparison

Dataset 1 consists of 60 original items, partitioned into three categories: high-trust, low-trust, and filler sentences. All items were manually authored by the researchers. High-trust sources correspond to announcement channels that are generally considered reliable in everyday life, whereas low-trust sources typically consist of channels that are difficult to trust. These sources are provided in Appendix A. During construction, we aimed to make the target completion as uniquely and unambiguously recoverable as possible, thereby minimizing cases with multiple equally plausible completions, while also balancing potential confounds such as lexical frequency and sentence length across conditions. In the cloze task, participants were asked to complete the final blank with the most natural and contextually appropriate continuation.

(5) Belediyenin SMS uyarısına göre sular .....
'According to the municipality's SMS alert, the water .....'

(6) Yan binadaki teyzenin dediğine göre sular .....
'According to the lady next door, the water .....'

Across conditions, the source of information is kept *overtly external* through explicit attribution frames (e.g., *X'e göre* 'according to X'; cf. (5)–(6)). The manipulation targets the *perceived reliability* of that external source: High-Trust items present the proposition as coming from institutionally authoritative channels (e.g., an official municipal SMS notification), whereas Low-Trust items present it as coming from informal or weakly accountable channels (e.g., a statement by a neighbor). To verify that this manipulation isolates trust rather than idiosyncratic item properties, we ran a separate norming study with a broad participant pool. Participants consistently rated High-Trust sources as more credible than Low-Trust sources, and items that failed to show the intended separation were revised or removed.

3.2. Dataset 2 for LLM-Only Evaluation

Dataset 2 is an expanded item set containing all experimental items from Dataset 1 together with additional items, and was designed exclusively for large language model evaluation. It preserves the same structural properties and manipulation logic as Dataset 1, while increasing coverage and statistical stability by substantially expanding the number of scenarios and lexical realizations. The expanded set contains 200 items in total. Dataset 2 was *not* administered to human participants; instead, it was designed to test model behavior on a broader item base without increasing participant burden, and to assess whether patterns observed in the human-scale dataset replicate and generalize when the trust manipulation is instantiated across a wider range of contexts.

4. Experiments

4.1. Human Experiment

The human experiment received prior approval from the University of Chicago Institutional Review Board (IRB26-0198). All participants provided informed consent before participation. We recruited 75 unique participants who self-identified as native speakers of Turkish, were at least 18 years old, and were residing in Turkey at the time of participation. The experiment was implemented and administered online using the PCIbex/PennController platform (Zehr and Schwarz, 2018). The main study produced 4,500 trials in total: 1,500 High-Trust trials, 1,500 Low-Trust trials, and 1,500 filler trials. For each trial, we recorded the participant's completion response. Our primary analyses target the critical High-Trust vs. Low-Trust contrast, so we analyze 3,000 trials after excluding fillers. All responses were analyzed in de-identified form. This balanced design allows a controlled test of how the trust manipulation affects production preferences.

4.2. LLM Experiments

In this study, LLMs were tested using two different datasets and ten models (Team et al., 2025; neuralwork; NovusResearch; mradermacher; Yang et al., 2025; Trendyol; TURKCELL; Bezir et al., 2024; ocaklisemih; cenkersisman). and three prompts.The models evaluated in this study consist of Turkish-focused and multilingual large language models. gpt2-turkish-10M, WiroAI-turkish-LLM-8B, Turkcell-LLM-7B, Trendyol-LLM-7B, Gemma-2-9B-IT-TR, and Novus-7B-TR are models that have been either directly trained on Turkish or fine-tuned with Turkish data. In contrast, Gemma-3 and Qwen3-32B are trained on large-scale multilingual data and are capable of processing Turkish through their multilingual capacities. Some models, on the other hand, are built on a multilingual foundation and later fine-tuned with Turkish instruction data. This diversity allows us to compare the effects of Turkish-focused and multilingual training approaches on sensitivity to evidentiality morphology. During the study, the models were asked to fill in the blank format that was created (5)–(6). Three prompts were used to perform this task; these prompts are given in Appendix B. While the models were expected to generate the word in the space indicated by the space in the first two prompts, in the third prompt two different words were given as options, one of which is the word conjugated with "-DI" (7), and the other is the word conjugated with "-mIş" (8). In the third prompt, the models were expected to choose the most appropriate word between these two options according to the flow of the sentence. While generating the model outputs, the temperature was set to 0.1 and the top-p to 1; an additional max token setting of 8 was added to the third prompt.

(7) Getir-il-di
bring-PASS-PST.DIR
'was brought (-DI form)'

(8) Getir-il-miş
bring-PASS-PST.INDIR
'apparently/ reportedly was brought (-mIş form)'

5. Human experiment results

5.1. Production: trust robustly shifts *-DI* vs. *-mIş*

Table 1 shows strict-coding counts for critical trials (High+Low; fillers excluded). Descriptively, High-trust contexts produce more DI completions, whereas Low-trust contexts produce more mIş completions. The Other category is similar across conditions (High: 28.4%; Low: 29.5%), suggesting that the trust manipulation primarily redistributes responses *within* the evidential morphology space rather than increasing off-target productions.

Condition	DI	mIş	Other
High trust	**727** (48.5%)	347 (23.1%)	426 (28.4%)
Low trust	548 (36.5%)	**509** (33.9%)	443 (29.5%)

Table 1: Strict coding counts by condition (critical trials only; $n = 1500$ per condition). Percentages are within condition.

Figure 2 plots the *within-condition* proportions of DI vs. mIş after excluding Other. In the High-trust condition, DI $= 727/(727 + 347) = 0.677$ and mIş $= 0.323$. In the Low-trust condition, DI $= 548/(548 + 509) = 0.518$ and mIş $= 0.482$.

Two robustness checks yield the same qualitative pattern. (i) Under lenient last-token coding (al-

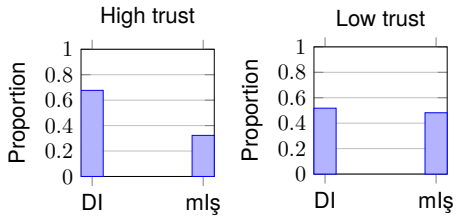


Figure 2: Human production (strict coding): within-condition proportions of *-DI* vs. *-mIş* among DI/mIş responses only.

lowing two-word VP completions), the association remains strong ($\chi^2(1) = 52.685$, $p = 3.92 \times 10^{-13}$; OR $= 1.859$, 95% CI $[1.571, 2.199]$). (ii) A content-controlled stratified analysis with matched content keys gives a pooled OR $= 2.029$, 95% CI $[1.654, 2.490]$ (CMH $\chi^2(1) = 46.789$, $p = 7.91 \times 10^{-12}$), with no strong evidence of heterogeneity across strata ($p = 0.137$).

6. The First Experiment with LLMs

Experiment I evaluates LLM behavior in a prompting-based gap-fill task using 200 trust-manipulated Turkish cloze items. Each item was presented to the model three times, and the final label was determined by majority vote across the three outputs. Table 2 summarizes the model-level results on the prompted task, reporting (i) the percentage of *usable* outputs (i.e., generations classifiable as DI or mIş), (ii) run-level DI/mIş counts in High- and Low-Trust conditions, (iii) the within-condition DI share, and (iv) Fisher exact tests for the High-vs.-Low trust contrast. A first important result is that the amount of usable evidential data varies substantially across models, ranging from 64.5% (Gemma-3-27B-IT) to 0.0% (Trendyol-LLM-7B). This means that, in the prompting setup, a large portion of outputs may be off-target and therefore cannot be interpreted as an evidential choice.

Among models with substantial usable DI/mIş outputs, the trust effect is not uniform. Gemma-3-27B-IT shows a clear and statistically reliable shift in the expected direction: the DI share decreases from 98.5% in High-Trust contexts to 88.9% in Low-Trust contexts (ΔDI $= +9.6$ pp, $p = 7.44 \times 10^{-5}$), indicating relatively more mIş in Low-Trust items. Gemma-2-9B-IT-TR shows a similar but smaller significant pattern (98.8% vs. 91.9%; ΔDI $= +6.9$ pp, $p = 0.006$). In contrast, Orbita-v0.1 shows a significant effect in the *opposite* direction, with a higher DI share in Low-Trust than in High-Trust contexts (81.5% vs. 67.5%; ΔDI $= -14.0$ pp, $p = 0.018$).

The remaining models mostly show null, unstable, or weak effects. Qwen3-32B, Novus-7B-TR, and WiroAI-TR-8B exhibit small and non-significant differences ($p = .371$, $.556$, and $.696$, respectively). GPT-OSS-20B-Astro shows a larger numerical reverse shift (ΔDI $= -20.4$ pp) but does not reach conventional significance ($p = .058$). For GPT2-TR-10M and Turkcell-LLM-7B, all usable outputs are DI, so no informative DI/mIş comparison can be made; Trendyol-LLM-7B yields no usable outputs. Overall, Experiment I suggests that in an open-generation prompting setup, trust-sensitive evidential behavior is highly model-dependent and often overshadowed by prompting/compliance limitations and a strong general bias toward DI.

7. The Second Experiment with LLMs

Experiment II evaluates LLM behavior in an *explicit past-tense* generation setup using the same 200 trust-manipulated Turkish cloze items. Unlike Experiment I, the prompt here explicitly instructed models to produce a *past-tense* completion. As in Experiment I, each item was sampled three times. To obtain a single interpretable evidential label per item, we assigned a final item-level label by majority vote across the three outputs (`tur1-tur3`) among outputs classifiable as DI or mIş (ties excluded).

Table 3 summarizes the model-level results for the explicit-past prompting condition, reporting (i) the percentage of *usable* items (i.e., items receiving a majority label classifiable as DI or mIş), (ii) item-level DI/mIş counts in High- and Low-Trust conditions, (iii) the within-condition DI share, and (iv) Fisher exact tests for the High-vs.-Low trust contrast. A first key result is that usability remains strongly model-dependent even with explicit past-tense instructions, ranging from 96.0% (Gemma-3-27B-IT) and 91.0% (Gemma-2-9B-IT-TR) to 0.0% (Trendyol-LLM-7B). Thus, explicit past-tense prompting substantially improves the proportion of usable evidential outputs for some models, but not for all.

Among models with substantial usable DI/mIş outputs, Gemma-3-27B-IT shows the clearest and statistically reliable trust-sensitive pattern in the predicted direction. Its DI share decreases from 97.9% in High-Trust contexts to 84.5% in Low-Trust contexts (ΔDI $= +13.4$ pp, $p = 1.51 \times 10^{-3}$), while the number of mIş labels increases from 2 items (High) to 15 items (Low). This indicates a relative increase in mIş under Low-Trust contexts. Gemma-2-9B-IT-TR also shows high usability and a small numerical shift in the expected direction (100.0% vs. 97.8%; ΔDI $= +2.2$ pp), but the difference is not statistically significant ($p = .243$). Orbita-v0.1 shows a larger numerical shift in the expected direction (ΔDI $= +18.1$ pp), but its low usability (22.0%) and non-significant test ($p = .327$) limit the strength of this result.

The remaining models mostly show null, unsta-

Model	Usable	H-DI	H-MIŞ	L-DI	L-MIŞ	DI% (H)	DI% (L)	ΔDI (pp)	p
Gemma-3-27B-IT	64.5	195	3	168	21	98.5	88.9	+9.6	7.44×10^{-5}
Gemma-2-9B-IT-TR	47.2	158	2	113	10	98.8	91.9	+6.9	0.006
Orbita-v0.1	40.3	83	40	97	22	67.5	81.5	-14.0	0.018
Qwen3-32B	24.7	54	3	81	10	94.7	89.0	+5.7	0.371
Novus-7B-TR	21.3	79	1	46	2	98.8	95.8	+2.9	0.556
WiroAI-TR-8B	16.8	54	5	40	2	91.5	95.2	-3.7	0.696
GPT-OSS-20B-Astro	11.3	29	11	26	2	72.5	92.9	-20.4	0.058
GPT2-TR-10M	2.8	10	0	7	0	100.0	100.0	+0.0	–
Turkcell-LLM-7B	1.2	5	0	2	0	100.0	100.0	+0.0	–
Trendyol-LLM-7B	0.0	0	0	0	0	–	–	–	–

Table 2: Experiment I (prompting-based gap-fill) LLM results. **Usable** = percentage of all generations classifiable as DI or mIş (i.e., DI/MIŞ%). DI% is computed as $\mathrm{DI}/(\mathrm{DI}+\mathrm{MI})$ within each trust condition. ΔDI is High minus Low (percentage points). p values are Fisher exact tests on {High, Low}×{DI, mIş}.

Model	Usable	H-DI	H-MIŞ	L-DI	L-MIŞ	DI% (H)	DI% (L)	ΔDI (pp)	p
Gemma-3-27B-IT	96.0	93	2	82	15	97.9	84.5	+13.4	1.51×10^{-3}
Gemma-2-9B-IT-TR	91.0	92	0	88	2	100.0	97.8	+2.2	0.243
Orbita-v0.1	22.0	19	6	11	8	76.0	57.9	+18.1	0.327
Qwen3-32B	32.5	24	3	35	3	88.9	92.1	-3.2	0.686
Novus-7B-TR	28.0	27	6	21	2	81.8	91.3	-9.5	0.449
WiroAI-TR-8B	24.5	24	3	20	2	88.9	90.9	-2.0	1.000
GPT-OSS-20B-Astro	27.5	17	11	19	8	60.7	70.4	-9.7	0.573
GPT2-TR-10M	4.5	3	0	6	0	100.0	100.0	+0.0	–
Turkcell-LLM-7B	2.0	3	0	1	0	100.0	100.0	+0.0	–
Trendyol-LLM-7B	0.0	0	0	0	0	–	–	–	–

Table 3: Experiment II (explicit past-tense generation) LLM results. Each item was generated three times (`tur1`–`tur3`), and the final item-level label was determined by majority vote among outputs classifiable as DI or mIş (ties excluded). **Usable** = percentage of all items (out of 200) receiving a majority DI or mIş label. DI% is computed as $\mathrm{DI}/(\mathrm{DI}+\mathrm{MI})$ within each trust condition. ΔDI is High minus Low (percentage points). p values are Fisher exact tests on {High, Low}×{DI, mIş} using item-level majority labels.

ble, or reverse-direction patterns. Qwen3-32B, Novus-7B-TR, WiroAI-TR-8B, and GPT-OSS-20B-Astro all show non-significant High-vs.-Low differences ($p = .686$, $.449$, 1.000, and $.573$, respectively), and several exhibit a numerical reverse shift (i.e., a higher DI share in Low-Trust than in High-Trust contexts). For GPT2-TR-10M and Turkcell-LLM-7B, all usable majority labels are DI, so no informative DI/mIş comparison can be made. Trendyol-LLM-7B produces no usable items. Overall, Experiment II shows that explicit past-tense prompting can improve evidential compliance and reveal a trust-sensitive DI/mIş pattern in some models (most clearly Gemma-3-27B-IT), but the effect remains highly model-dependent and is weak, unstable, or absent in most others.

7.1. Exploratory Head-Level Modulation of Trust-Cue Attention

As an exploratory mechanistic analysis, we examined attention-weight redistribution in *Gemma-2-9B-IT-TR* under the explicit past-tense prompting condition. We focused on the pre-generation decision state (`q = last_prompt`) and quantified attention directed to the trust-cue span, operationalized as the token sequence immediately preceding *göre*. Across heads, the strongest condition sensitivity was confined to a small subset of mid-to-late layers, with the single best layer at $L = 25$ and the highest-scoring layers concentrated in the range $L = 22$–29. However, the aggregate pairwise contrast was weak and directionally unstable (Δ_{pair} mean $= -0.0179$, 95% bootstrap CI $[-0.0513, 0.0170]$; $P(\Delta_{\text{pair}} > 0) = 46\%$). Under this operationalization, the results do not provide strong evidence that trust is robustly encoded via systematic reallocation of attention mass to the cue span in *Gemma-2-9B-IT-TR*; accordingly, we treat the attention maps as qualitative diagnostic evidence rather than as a stable mechanistic signature. The corresponding visualizations are provided in Appendix C.

Applying the same analysis to *Gemma-3-27B-IT* yielded a larger and more internally consistent

Model	DI% High	DI% Low	Δpp	Acc%	Abstain%	Fisher p	Holm p	3-run same%
Gemma-2-9B-IT-TR	60.0	44.0	16.0	58.0	0.0	0.033	0.268	97.5
Gemma-3-27B-IT	78.9	66.3	12.6	56.0	3.5	0.054	0.381	97.5
GPT-OSS-20B-Astro	94.9	84.0	10.9	55.3	0.5	0.019	0.173	92.0
Orbita-v0.1	91.0	84.0	7.0	53.5	0.0	0.199	0.994	96.0
Turkcell-LLM-7B	97.0	90.0	7.0	53.3	0.5	0.082	0.491	81.0
WiroAI-TR-8B	15.2	11.0	4.2	52.3	0.5	0.408	1.000	91.0
Qwen3-32B	81.0	77.0	4.0	52.0	0.0	0.603	1.000	95.0
Novus-7B-TR	99.0	98.0	1.0	50.5	0.0	1.000	1.000	99.0
Trendyol-LLM-7B	0.0	0.0	0.0	50.0	0.0	1.000	1.000	100.0
GPT2-TR-10M	0.0	–	–	0.0	99.5	–	–	99.5

Table 4: Experiment III (A/B forced-choice prompting) combined model results using item-level majority vote over three runs. DI% High/Low are computed after excluding abstentions (invalid outputs and tie/no-majority items). Δpp = DI% High minus DI% Low. Acc% uses the target mapping High→*-DI*, Low→*-mIş*. Fisher exact tests are computed on High/Low × {DI, mIş} after excluding abstentions; Holm p values correct for multiple testing across models. 3-run same% is the proportion of items with identical outputs across all three replicates.

pair-level contrast. Condition-sensitive modulation was again localized to a restricted subset of mid-to-late-layer heads, but with a clearer concentration in higher layers (best layer $L = 31$; top layers $L = 31, 33, 36, 37$). The aggregate pairwise effect was reliably non-zero under the present metric definition (Δ_{pair} mean $= -0.0983$, 95% bootstrap CI $[-0.1151, -0.0817]$; $P(\Delta_{\text{pair}} > 0) = 9\%$), indicating systematic condition-dependent redistribution of attention toward the trust-cue span. We interpret this as evidence that *Gemma-3-27B-IT* exhibits a more coherent internal sensitivity to trust-related contextual cues than *Gemma-2-9B-IT-TR* in this prompting setup. At the same time, this interpretation should remain qualified, since raw span-level attention mass can be affected by surface differences such as cue-length asymmetries across conditions. The relevant maps are shown in Appendix C.

8. The Third Experiment with LLMs

Experiment III evaluates LLM behavior in a *forced-choice* prompting setup, in which models are explicitly asked to choose between two candidate completions: a form (option A) and a form (option B). This design is more constrained than Experiments I–II because it removes open-ended lexical generation and directly targets the evidential choice itself. The same 200 trust-manipulated Turkish items are used (100 High-Trust, 100 Low-Trust), and each item is queried three times (`tur1-tur3`). Final item-level labels are determined by majority vote across the three runs. Outputs other than `A/B`, as well as ties/no-majority cases, are treated as abstentions and excluded from evidential-rate and Fisher-test calculations.

Table 4 reports per-model condition-wise DI rates (DI% High and DI% Low), the High-minus-Low difference (Δpp), target-mapping accuracy (High→, Low→), and abstention rate over all 200 items. Table 4 reports per-model Fisher exact tests (High-/Low × DI/MIS), Holm-corrected p values across models, and 3-run agreement (*same*%), which measures response stability across the three replicates.

A first important result is that most usable models show a *directionally trust-consistent trend*: DI% is higher in High-Trust than in Low-Trust contexts. The largest shift is observed for neuralwork/gemma-2-9b-it-tr ($\Delta\text{pp} = 16.0$), followed by google/gemma-3-27b-it (12.6 pp) and ocaklisemih/gpt-oss-20b-turkish-astrology-gguf (10.9 pp). Several other models also show smaller positive shifts (e.g., mradermacher/Orbita-v0.1-i1-GGUF, Turkcell-LLM-7b-v1, WiroAI/wiroai-turkish-llm-8b, Qwen/Qwen3-32B, NovusResearch/Novus-7b-tr$_v1$).

Trendyol-LLM-7b-base-v1.0 shows no shift (0.0 vs. 0.0). This pattern suggests that, under direct A/B comparison, many models weakly track the trust manipulation in the predicted direction.

At the same time, the results also show that *base-rate preferences remain strong* and often dominate behavior. Several models overwhelmingly prefer one option across both conditions (e.g., very high DI rates for Novus, Turkcell, Orbita, and GPT-OSS; near-zero DI rates for Trendyol and GPT2-TR-10M), which compresses the effect of trust and limits target-mapping accuracy. Consistent with this, Acc% values remain close to chance for most models (roughly 50–58%), even when the direction of the trust effect is correct. In other words, models can show a small High>Low DI shift without demonstrating robust context-sensitive evidential selection at the item level.

Statistical testing reinforces this conclusion. Some models yield nominally small Fisher exact p values (e.g., $p = .019$ for GPT-OSS and $p = .033$ for Gemma-2-9B-IT-TR; Table 4), but *none* remains significant after Holm correction across models. Thus, Experiment III provides evidence for a weak, directionally consistent trend in several systems, but not for a robust per-model trust effect after correction for multiple comparisons.

Finally, response stability across the three replicates is generally high (3-run same% often above 90%), indicating that model choices in this A/B paradigm are relatively deterministic once the prompt format is fixed. This is methodologically useful: compared with open-ended prompting, the forced-choice setup reduces output-format variability and improves comparability across models. However, the combination of high replicate agreement and near-chance Acc% for many systems suggests that the models are often making stable choices driven by global option preferences rather than reliably applying the trust cue in a human-like way.

9. Discussion

This study makes theoretical and methodological contributions to Turkish evidentiality and its evaluation in LLMs. Human results show a clear trust effect: more credible external sources favor *-DI*, while less credible sources favor *-mIş* (Section 5; Table 1; Figure 2). In contrast, current LLMs do not reliably reproduce this pattern, and their apparent evidential behavior varies with prompt format, output compliance, and response constraints (Tables 2, 3, 4).

The paper's central theoretical contribution is a direct experimental test of a *trustworthiness*-based perspective on Turkish evidentiality, inspired by the Giannakidou–Mari framework. To our knowledge, this is a novel application of that framework to Turkish *-DI*/*-mIş*. Rather than treating evidential morphology only as a categorical source-type marker, we test whether the credibility of an explicitly external source shifts evidential choice; the human data show that it does.

This matters because the source is explicitly external in both conditions ((e.g., X'e göre 'according to X')); the manipulation changes only the source's trustworthiness. The observed shift therefore supports an analysis in which Turkish evidential morphology is sensitive not just to source *type*, but also to source *quality* and the speaker's resulting commitment profile. In Giannakidou–Mari terms, the pattern is consistent with differences in support strength for the prejacent: more trustworthy sources favor stronger commitment (and relatively more *-DI*), whereas less trustworthy sources favor weaker, evidence-weighted commitment (and relatively more *-mIş*) (Giannakidou and Mari, 2016, 2018, 2021; Boscaro et al., 2025).

Our findings do not by themselves settle the long-standing question of whether *-DI* is lexically a direct evidential or a past tense form whose "directness" effects arise pragmatically. However, they do provide an important constraint on the debate. Because the manipulation keeps source attribution explicit and varies trustworthiness instead, the human pattern is difficult to reduce to a simple source-present/source-absent contrast. The results are more naturally captured by accounts that allow context-sensitive interactions among evidential morphology, speaker commitment, and discourse-pragmatic reasoning. In this sense, the present data strengthen the case for treating Turkish evidential choice as part of a broader inferential and commitment-sensitive system, rather than a purely rigid source-labeling mechanism.

What the LLM results may suggest. Across the three LLM experiments, the main result is a dissociation between *surface task performance under prompting constraints* and *human-like trust-sensitive evidential reasoning*. In Experiment I (open gap-fill prompting), many models produce a high proportion of unusable outputs, and even among usable outputs the trust effect is highly model-dependent. In Experiment II (explicit past-tense prompting), usability improves substantially for some models (especially Gemma variants), and Gemma-3-27B-IT shows the clearest trust-consistent shift, but the pattern remains weak, absent, or unstable in most models. In Experiment III (A/B forced-choice), response stability becomes high and several models show directionally trust-consistent trends, yet effects are generally modest, target-mapping accuracy remains near chance for many systems, and no per-model effect survives Holm correction. Thus, these results suggest that many LLMs can exhibit *format-dependent traces* of the expected pattern, but they do not reliably reproduce the human sensitivity to source trustworthiness as a stable, cross-paradigm property. This is precisely the kind of mismatch that a linguistically targeted, human–LLM comparison can reveal: a model may appear "reasonable" in one prompt format while still failing to encode the underlying contextual factor in a robust way. A related question is why explicit past-tense prompting changes model performance. We suggest that the main reason is task specification: in the open gap-fill setting, models must jointly infer a plausible lexical completion, tense, and evidential morphology, which leaves the task relatively underspecified. By explicitly requiring a past-tense completion in Experiment II, the prompt narrows the response space and makes

the *-DI*/*-mIş* contrast more directly relevant to the decision. In this sense, the improvement for some models is consistent with recent work showing that more specific instructions can reduce prompt sensitivity relative to underspecified prompts, even if they do not solve the deeper representational problem (Pecher et al., 2026). At the same time, our results also align with recent findings that morphological generalization remains difficult for LLMs in agglutinative languages such as Turkish, especially when models must produce the appropriate inflection rather than merely recognize it (Ismayilzada et al., 2025).

The paper also contributes a reusable evaluation paradigm for Turkish source-sensitive semantics/pragmatics by combining: (i) a human cloze dataset, (ii) an expanded LLM-only dataset, and (iii) three prompting regimes that vary constraint level (open generation, explicit tense generation, forced choice). This multi-paradigm design makes it possible to separate at least three sources of variation that are often conflated in LLM evaluations: *compliance/formatting behavior*, *base-rate morphological preference*, and *context sensitivity to the trust manipulation*. In particular, the contrast between Experiments I–III shows that forcing the response space can improve comparability and stability, but it can also expose strong default biases that limit genuine contextual adaptation. More broadly, the study argues for evaluating Turkish LLMs on controlled linguistic contrasts with human baselines, rather than relying only on broad benchmarks. A useful point of comparison is *TurBLiMP*, the first Turkish benchmark of linguistic minimal pairs, which evaluates LMs on 16 grammatical phenomena and has significantly expanded linguistically informed evaluation for Turkish (Başar et al., 2025). Our study is complementary to that line of work. Whereas TurBLiMP focuses on minimal-pair judgments over a broad range of grammatical phenomena, it does not target evidentiality. This matters because evidentiality is not only a morphosyntactic contrast in Turkish, but also a source-sensitive semantic–pragmatic phenomenon tied to speaker commitment, information source, and discourse reasoning. Our results therefore extend Turkish LLM evaluation into a domain that is central to Turkish grammar but not captured by existing minimal-pair benchmarks. For evidentiality and related phenomena, this is especially important because the relevant behavior is not merely lexical accuracy but context-dependent mapping between linguistic form and epistemic/discourse structure.

10. Conclusion

This paper investigated how *source trustworthiness* shapes Turkish evidential choice in the *-DI*/*-mIş* contrast, and whether current LLMs track this sensitivity. In a controlled human cloze experiment, we found a robust and replicable trust effect: participants produced relatively more *-DI* in High-Trust contexts and relatively more *-mIş* in Low-Trust contexts. This provides new empirical support for a trust-/commitment-based perspective on Turkish evidentiality and shows that the Giannakidou–Mari framework offers a productive theoretical lens for Turkish data.

Across three LLM experiments (open gap-fill prompting, explicit past-tense prompting, and forced-choice A/B prompting), model behavior was substantially more variable and strongly dependent on prompt format, output compliance, and base-rate suffix preferences. While some models showed weak or local trust-consistent trends, the effect was not robust across models or paradigms. Overall, the results reveal a clear human–LLM gap in source-sensitive evidential reasoning and motivate future work on more faithful evaluation and modeling of evidential and commitment-related meaning in Turkish.

Ethics Statement

This work investigated the interaction between contextual source trustworthiness and Turkish evidential morphology (*-DI* vs. *-mIş*) using both human cloze responses and LLM-based preference scoring. Human data were collected under informed-consent procedures and analyzed only at the aggregate level; no personally identifying information is reported, and all examples are presented in anonymized form. The human-subjects component of this study received Institutional Review Board (IRB) approval.

For the computational component, model outputs are treated as task-conditioned behavioral responses rather than as evidence of human-like semantic or cognitive representations. Because LLM preferences can be sensitive to evaluation protocol design (including prompting format and tokenization effects), our findings should not be overgeneralized to broad claims about model competence or human evidential reasoning. Finally, by focusing on Turkish, this study helps address the English-centric bias of much computational psycholinguistics and LLM evaluation.

Limitations

This study has several limitations. First, the strongest behavioral evidence comes from the human experiment on Dataset 1, whereas the larger Dataset 2 was used only for LLM evaluation; future work should test whether the same human

trust effect scales to the expanded item set. Second, LLM results are partly constrained by prompt compliance and output-format errors, especially in open-generation settings, which can reduce the amount of usable evidential data and make model comparisons harder to interpret. Third, although the forced-choice setup improves control and stability, it may introduce response biases (e.g., option or format preferences) that are not specific to evidential reasoning. The exploratory attention analysis is also limited in scope since it was conducted only for two relatively successful models from Experiment II and is intended as a preliminary qualitative diagnostic rather than a definitive mechanistic account.

Acknowledgements

We thank Anastasia Giannakidou for valuable input on the semantic and pragmatic theory and the interpretation of the results, Ming Xiang for suggestions on the human experimental design, and Christopher Potts and Craig Thorburn for their comments on future directions. We are also grateful to the anonymous reviewers for their helpful feedback and suggestions. Any remaining errors are our own.

11. Bibliographical References

Alexandra Y. Aikhenvald. 2004. *Evidentiality*. Oxford University Press, Oxford.

Ayhan Aksu-Koç and Dan I. Slobin. 1986. A psychological account of the development and use of evidentials in Turkish. In Wallace Chafe and Johanna Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*, pages 159–167. Ablex Publishing Corporation, Norwood, NJ.

Merve Ataman-Devrim, Gaye Soley, and Ayhan Aksu-Koç. 2025. Children's understanding of source reliability and knowledge generalizability from grammatical cues: Evidence from Turkish. *Journal of Child Language*, pages 1–23. First published online 11 December 2025.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: LLM hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish benchmark of linguistic minimal pairs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16495–16510, Suzhou, China. Association for Computational Linguistics.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, and Savaş Yıldırım. 2025. Tokenization standards for linguistic integrity: Turkish as a benchmark. *arXiv preprint arXiv:2502.07057*.

M. Ali Bayram, Ali Arda Fincan, Ahmet Semih Gümüş, Sercan Karakaş, Banu Diri, Savaş Yıldırım, and Demircan Çelik. 2026. Tokens with meaning: A hybrid tokenization approach for Turkish.

Abdullah Bezir, Furkan Burhan Türkay, and Cengiz Asmazoğlu. 2024. Wiroai turkish llm 8b. Hugging Face model card.

Marie Boscaro, Anastasia Giannakidou, and Alda Mari. 2025. Evidence type and trustworthiness: The view from social media. In *The Grammar of Interaction*, pages 98–131.

Kasper Boye. 2012. *Epistemic Meaning: A Crosslinguistic and Functional-Cognitive Study*. De Gruyter Mouton, Berlin.

Roger W. Brown and Eric H. Lenneberg. 1954. A study in language and cognition. *Journal of Abnormal and Social Psychology*, 49(3):454–462.

cenkersisman. gpt2-turkish-10m. Hugging Face model card. Accessed: 2026-01-20.

Ferdinand de Haan. 2001. The place of inference within the evidential system. *International Journal of American Linguistics*, 67(2):193–219.

Patrick Dendale and Liliane Tasmowski. 2001. Introduction: evidentiality and related notions. *Journal of Pragmatics*, 33(3):339–348.

Yakup Abrek Er, Ilker Kesen, Gözde Gül Şahin, and Aykut Erdem. 2025. Cetvel: A unified benchmark for evaluating language understanding, generation and cultural capacity of LLMs for Turkish. *arXiv preprint arXiv:2508.16431*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Anastasia Giannakidou. 1998. *Polarity Sensitivity as (Non)veridical Dependency*. John Benjamins, Amsterdam/Philadelphia.

Anastasia Giannakidou. 2006. Only, emotive factive verbs, and the dual nature of polarity dependency. *Language*, 82(3):575–603.

Anastasia Giannakidou and Alda Mari. 2016. Epistemic future and epistemic must: Nonveridicality, evidence, and partial knowledge. In Joanna Błaszczak, Anastasia Giannakidou, Dorota Klimek-Jankowska, and Krzysztof Migdalski, editors, *Mood, Aspect, Modality Revisited: New Answers to Old Questions*, pages 75–124. The University of Chicago Press, Chicago.

Anastasia Giannakidou and Alda Mari. 2018. A unified analysis of the future as epistemic modality: The view from greek and italian. *Natural Language & Linguistic Theory*, 36(1):85–129.

Anastasia Giannakidou and Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought*. University of Chicago Press, Chicago.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Routledge, London and New York.

Demet Gül. 2009. Semantics of Turkish evidential -(i)miş. In S. Ay, Ö. Aydın, İ. Ergenç, S. Gökmen, S. İşsever, and D. Peçenek, editors, *Essays on Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics, August 6–8, 2008*, pages 177–186. Harrassowitz, Wiesbaden.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838, Vienna, Austria. Association for Computational Linguistics.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Lars Johanson. 2003. Evidentiality in Turkic. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *Studies in Evidentiality*, pages 273–290. John Benjamins, Amsterdam/Philadelphia.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Birsu Kandemirci, Anna Theakston, Ditte Boeg Thomsen, and Silke Brandt. 2023. Does evidentiality support source monitoring and false belief understanding? a cross-linguistic study with Turkish- and english-speaking children. *Child Development*, 94(4):889–904.

Sercan Karakaş. 2026. Clause-internal or clause-external? Testing Turkish reflexive binding in adapted versus chain of thought large language models. *arXiv preprint arXiv:2602.00380*.

Sercan Karakaş and Yusuf Şimşek. 2026. From lemmas to dependencies: What signals drive light verbs classification? In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 220–227, Rabat, Morocco. Association for Computational Linguistics.

Onur Keleş and Nazik Dinçtopal Deniz. 2024. A comparative study with human data: Do LLMs have superficial language processing? In *2024 32nd IEEE Conference on Signal Processing and Communications Applications (SIU)*.

Onur Keleş and Nazik Dinçtopal Deniz. 2025. When men bite dogs: Testing good-enough parsing in Turkish with humans and large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 219–231, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Jaklin Kornfilt. 1997. *Turkish*. Routledge, London.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401*.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association*

for Computational Linguistics: ACL 2024, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xuannan Liu, Xiao Yang, Zekun Li, Peipei Li, and Ran He. 2026. Agenthallu: Benchmarking automated hallucination attribution of LLM-based agents. *arXiv preprint arXiv:2601.06818*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

mradermacher. Orbita-v0.1-i1-GGUF. Hugging Face model card. Accessed: 2026-01-20.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

neuralwork. gemma-2-9b-it-tr. Hugging Face model card. Accessed: 2026-01-20.

NovusResearch. Novus-7b-tr$_v$1. Hugging Face model card. Accessed: 2026-01-20.

ocaklisemih. gpt-oss-20b-Turkish-astrology-gguf. Hugging Face model card. Accessed: 2026-01-20.

Branislav Pecher, Michal Spiegel, Robert Belanec, and Jan Cegin. 2026. Revisiting prompt sensitivity in large language models for text classification: The role of prompt underspecification. *arXiv preprint arXiv:2602.04297*.

Vladimir A. Plungian. 2001. The place of evidentiality within the universal grammatical space. *Journal of Pragmatics*, 33(3):349–357.

Nilüfer Şener. 2011. *Semantics and Pragmatics of Evidentials in Turkish*. Ph.D. thesis, University of Connecticut, Storrs, CT.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch

Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. 2025. Openai gpt-5 system card.

Dan I. Slobin. 1996. From “thought and language” to “thinking for speaking”. In John J. Gumperz and Stephen C. Levinson, editors, *Rethinking Linguistic Relativity*, pages 70–96. Cambridge University Press, Cambridge.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Cagri Toraman, Ahmet Kaan Sever, Ayşe Aysu Cengiz, Elif Ecem Arslan, Görkem Sevinç, Sarp Kantar, Mete Mert Birdal, Yusuf Faruk Güldemir, Ali Buğra Kanburoğlu, Sezen Felekoğlu, Birsen Şahin Kütük, Büşra Tufan, Elif Genç, Serkan Coşkun, Gupse Ekin Demir, Muhammed Emin Arayıcı, Olgun Dursun, Onur Gungor, Susan Üsküdarlı, Abdullah Topraksoy, and Esra Darıcı. 2026. TurkBench: A benchmark for evaluating Turkish large language models. In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)*, pages 126–154, Rabat, Morocco. Association for Computational Linguistics.

Trendyol. Trendyol-LLM-7b-base-v1.0. Hugging Face model card. Accessed: 2026-01-20.

TURKCELL. Turkcell-LLM-7b-v1. Hugging Face model card. Accessed: 2026-01-20.

Özgür Uğur, Mahmut Göksu, Mahmut Çimen, Musa Yılmaz, Esra Şavirdi, Alp Talha Demir, Rumeysa Güllüce, İclal Çetin, and Ömer Can Sağbaş. 2026. Mecellem models: Turkish models trained from scratch and continually pre-trained for the legal domain. *arXiv preprint arXiv:2601.16018*.

Elif Ecem Umutlu, Ayse Aysu Cengiz, Ahmet Kaan Sever, Seyma Erdem, Burak Aytan, Busra Tufan, Abdullah Topraksoy, Esra Darıcı, and Cagri Toraman. 2025. Evaluating the quality of benchmark datasets for low-resource languages: A case study on Turkish. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 471–487, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Ercenur Ünal and Anna Papafragou. 2020. Relations between language and cognition: Evidentiality and sources of knowledge. *Topics in Cognitive Science*, 12(1):115–135.

Robert Underhill. 1976. *Turkish Grammar*. MIT Press, Cambridge, MA.

Darmawan Wicaksono, Hasri Akbar Awal Rozaq, and Nevfel Boz. 2025. Emotion recognition for low-resource Turkish: Fine-tuning BERTurk on TREMO and testing on xenophobic political discourse. *arXiv preprint arXiv:2505.12160*.

Thomas Willett. 1988. A cross-linguistic survey of the grammaticization of evidentiality. *Studies in Language*, 12(1):51–97.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Jeremy Zehr and Florian Schwarz. 2018. Penncontroller for internet based experiments (ibex).

A. Source Categories

Condition	Turkish	English
High	Resmi Gazete	Official Gazette
High	Kurumun resmi sayfası	Official website of the institution
High	Hastanenin astığı duyuru	Hospital notice board announcement
High	Belediyenin SMS uyarısı	Municipality SMS alert
High	Resmi mail	Official email
Low	Anonim telegram grubu	Anonymous Telegram group
Low	Yan apartmanda oturan komşu	Neighbor living in the adjacent building
Low	Diğer şirketteki çalışan	Employee at another company
Low	Yanlış haberler paylaşan Instagram grubu	Instagram group sharing false news
Low	Sokakta geçen biri	A random passerby on the street

Table 5: Examples of high- and low-trust source categories used in the experiment.

B. Prompt Templates

B.1. Turkish Prompts

Prompt 1:

Prompt 1 (TR)

Aşağıdaki cümlede "_____." ile belirtilmiş bir boşluk var.
Bu boşluğu cümlenin yapısının akışına en uygun fiille tamamla.

Cevap verirken kurallar:
- Sadece boşluğa gelecek kısmıyaz.
- Cümleyi tekrar etme.
- Açıklama yapma.
Tek kelime cevap ver ve fiil kullanarak tamamla
- Cevap vermeden geçme.

Cümle:
`{prefix}` _____.

Cevap:

Prompt 2:

Prompt 2 (TR)

Aşağıdaki cümlede "_____." ile belirtilmiş bir boşluk var.
Bu boşluğu cümlenin yapısının akışına ve geçmiş zamana göre tamamlanabilecek en uygun kelimeyle tamamla.

Cevap verirken kurallar:
- Sadece boşluğa gelecek kısmıyaz.
- Cümleyi tekrar etme.
- Açıklama yapma.
- Cevap vermeden geçme.

Cümle:
`{prefix}` _____.

Cevap:

Prompt 3:

Prompt 3 (TR)

Cümledeki boşluğu doldurmak için doğru seçeneği seç.

Cümle: `{sentence}`

Seçenekler:
A) `{di}`
B) `{mis}`

Kurallar:
- Sadece A ya da B yaz.
- Açıklama yapma.

Cevap:

B.2. English Prompts

Prompt 1:

Prompt 1 (EN)

There is a blank indicated by "_____." in the sentence below.
Fill this blank with the verb that best fits the flow and structure of the sentence.

Rules for your answer:
- Write only what should go in the blank.
- Do not repeat the sentence.
- Do not provide an explanation.
Answer with a single word and complete it using a verb.
- Do not skip answering.

Sentence:
`{prefix}` _____.

Answer:

Prompt 2:

Prompt 2 (EN)

There is a blank indicated by "_____." in the sentence below.
Fill this blank with the most appropriate word that can complete the sentence according to the flow/structure and past tense.

Rules for your answer:
- Write only what should go in the blank.
- Do not repeat the sentence.
- Do not provide an explanation.
- Do not skip answering.

Sentence:
`{prefix}` _____.

Answer:

Prompt 3:

Prompt 3 (EN)

Choose the correct option to fill the blank in the sentence.

Sentence: `{sentence}`

Options:
A) `{di}`
B) `{mis}`

Rules:
- Write only A or B.
- Do not provide an explanation.

Answer:

C. Exploratory Attention Maps

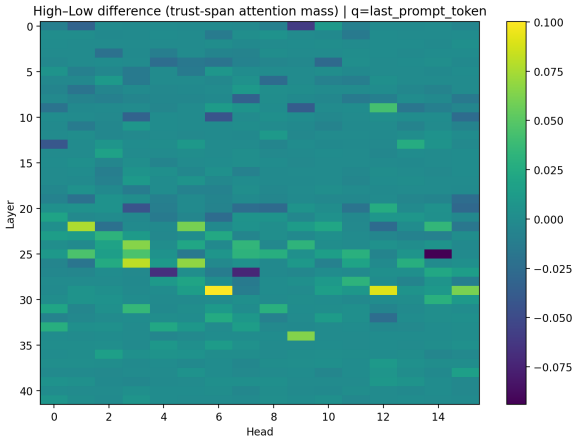


Figure 3: Gemma-2-9B-IT-TR (Experiment II): high–low difference in trust-span attention mass (query token: `last_prompt_token`). The pair-level effect is weak and inconsistent (Δ_{pair} mean $= -0.0179$, 95% bootstrap CI $[-0.0513, 0.0170]$; $P(\Delta_{\text{pair}} > 0) = 46\%$), so we treat this as a qualitative diagnostic rather than strong evidence of robust trust encoding.

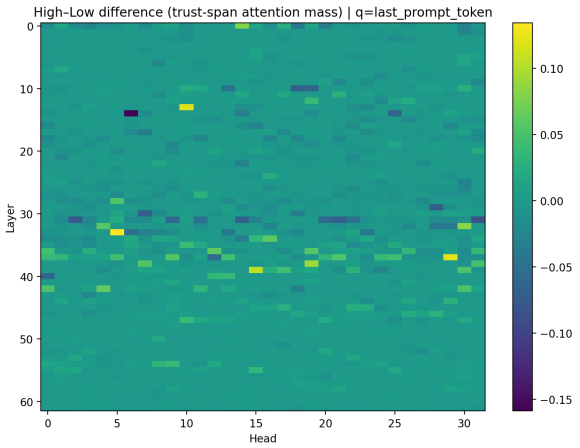


Figure 4: Gemma-3-27B-IT (Experiment II): high–low difference in trust-span attention mass (query token: `last_prompt_token`). The map shows localized modulation concentrated in a small cluster of mid-to-late layers and heads, consistent with the exploratory analysis in Section 7.

Toward Cognitive Alignment in Large Language Models: Integrating Linguistic Theory and Human Data

Wajdi Zaghouani
Northwestern University in Qatar
wajdi.zaghouani@northwestern.edu

Abstract

Large language models (LLMs) have reshaped natural language processing and reopened foundational questions about the relationship between computational systems, linguistic theory, and human cognition. While recent work shows that neural language models correlate with human behavioral and neural measures of language processing, predictive performance alone does not constitute cognitive adequacy. This position paper argues for a balanced and programmatic integration of cognitive science, linguistics, and NLP centered on the concept of cognitive alignment. We examine both the empirical promise and conceptual limitations of current LLM claims about human language processing. We then outline how experimental paradigms from psycholinguistics, formal insights from linguistic theory, and large-scale computational modeling can be combined to produce more interpretable, data-efficient, and cognitively grounded systems. We propose four concrete initiatives: a Cognitive Alignment Benchmark Suite, hybrid neuro-symbolic modeling challenges, population-level language simulation infrastructures, and interdisciplinary training programs. Together, these steps move toward a unified science of language grounded in empirical rigor, theoretical depth, and computational scalability.

Keywords: Cognitive modeling, computational psycholinguistics, large language models, linguistic theory, interdisciplinary research

1. Introduction

The rapid development of large language models has renewed a longstanding question: can computational systems serve as explanatory models of human language processing? While transformer-based architectures (Vaswani et al., 2017) achieve remarkable predictive performance, the relationship between predictive success and cognitive adequacy remains under-theorized. This tension reflects deeper questions about the nature of linguistic knowledge and the criteria for explanatory adequacy in cognitive science.

Cognitive science, linguistics, and NLP have historically pursued partially independent research programs. Cognitive science investigates representational format, learning constraints, and processing mechanisms. Linguistics develops formal accounts of structure, compositionality, and cross-linguistic variation. NLP prioritizes scalability, empirical coverage, and engineering performance. The emergence of LLMs creates both an opportunity and a risk: the opportunity to reconnect these traditions, and the risk of collapsing explanation into benchmark performance.

This paper advances a balanced thesis. Neural language models provide unprecedented tools for modeling large-scale linguistic structure and predicting human behavioral data. However, predictive correlation alone does not constitute mechanistic explanation. A principled research program must integrate experimental constraints, formal theory, and computational scale. We draw on foundational work in computational psycholinguistics (Levy, 2008), cognitive architectures (Anderson et al., 2004), and formal semantics (Montague, 1973) to articulate criteria for cognitive alignment that go beyond surface-level behavioral matching.

The cognitive modeling community is uniquely positioned to lead this integration. Unlike purely engineering-driven NLP research, cognitive modeling has always maintained commitment to psychological plausibility and theoretical grounding. Unlike purely theoretical linguistics, it embraces computational implementation and empirical validation. The challenge now is to scale these commitments to the era of large language models while preserving their scientific integrity. Importantly, substantial work already exists at the intersection of NLP and cognitive science, including shared tasks on eye-tracking prediction organized at previous editions of the CMCL workshop (Hollenstein et al., 2021, 2022) and large-scale multilingual reading corpora (Siegelman et al., 2022). The present paper builds on these foundations to propose a more systematic integration.

The structure of this paper proceeds as follows. Section 2 defines cognitive alignment as a multi-level constraint spanning behavioral, representational, and developmental dimensions. Section 3 offers a measured critique of current LLM claims. Section 4 outlines a programmatic integration across disciplines. Section 5 proposes concrete infrastructure initiatives. Section 6 discusses challenges and limitations. Section 7 concludes with reflections on the path forward.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 347–358
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Cognitive Alignment as a Multi-Level Constraint

We define cognitive alignment not as superficial behavioral similarity, but as convergence across three explanatory levels. This tripartite framework draws on Marr's levels of analysis (Marr, 1982) while extending it to address the specific challenges posed by neural language models. True cognitive alignment requires satisfaction of constraints at multiple levels simultaneously, not merely optimization of any single criterion.

2.1. Behavioral Alignment

Surprisal derived from probabilistic models correlates robustly with reading times and EEG responses (Hale, 2001; Levy, 2008). The surprisal theory of sentence processing posits that processing difficulty at each word is proportional to its negative log probability given context. This relationship has been confirmed across multiple languages and experimental paradigms (Smith and Levy, 2013; Frank et al., 2015; Wilcox et al., 2023).

Neural language model representations capture predictive neural processing signals. Studies using magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI) demonstrate that transformer activations correlate with neural responses in language-selective cortical regions (Schrimpf et al., 2021; Goldstein et al., 2022). Caucheteux and King demonstrate that GPT-2 activations predict brain activity with remarkable precision (Caucheteux and King, 2022). These findings establish that distributional models approximate certain aspects of incremental expectation.

However, such correlations remain level-neutral. As Lake et al. (2017) argue, human-like cognition requires structured inductive biases that support generalization beyond observed distributions. Correlation with processing difficulty does not entail isomorphism of representation. A model might achieve high brain-score correlations through mechanisms entirely unlike those deployed by the human brain. Furthermore, behavioral alignment on naturalistic stimuli does not guarantee alignment on carefully controlled experimental materials designed to isolate specific linguistic phenomena.

Recent work by Wilcox et al. (2020) demonstrates that while neural language models capture broad patterns of syntactic expectation, they fail on specific constructions that require hierarchical structure sensitivity. Similarly, Futrell et al. (2019) show that models trained on realistic language data are less robust, less sensitive, and less selective than humans in processing garden-path sentences, which are known to cause significant processing difficulty for human readers as well. Linzen et al. (2016) introduced systematic tests of subject-verb agreement across intervening material, revealing that models often fail where humans succeed. These findings underscore the need for fine-grained behavioral testing beyond aggregate correlation metrics.

Crucially, recent studies have revealed that the relationship between model scale and cognitive plausibility is more nuanced than previously assumed. Kuribayashi et al. (2024) show that instruction-tuned LLMs often yield worse psychometric predictive power for human reading behavior than base LLMs, indicating that alignment with human preferences does not entail alignment with human processing. Furthermore, the inverse scaling effect, in which surprisal from larger pre-trained language models provides a poorer fit to human reading times, represents a significant challenge to the assumption that better language models are necessarily more cognitively plausible (Oh et al., 2025). Kuribayashi et al. (2025) offer a more nuanced picture, demonstrating that when internal layers of larger LMs are examined rather than only final-layer outputs, larger models align with human sentence processing data as well as or better than smaller ones. This finding holds across behavioral (self-paced reading, gaze durations, maze processing times) and neurophysiological (N400 potentials) measures, suggesting that the cognitive plausibility of larger LMs has been underestimated.

2.2. Representational Alignment

Formal linguistics provides explicit hypotheses about the structure of linguistic representations. Construction Grammar emphasizes form-meaning pairings at multiple levels of abstraction (Goldberg, 1995, 2006). Abstract Meaning Representation offers graph-structured semantic representations that capture predicate-argument structure (Banarescu et al., 2013). FrameNet encodes frame-semantic relations that link linguistic expressions to conceptual structures (Baker et al., 1998).

While neural architectures may implicitly encode similar information, interpretability research remains essential. Mechanistic analysis of transformer components reveals attention heads specialized for specific syntactic relations (Clark et al., 2019; Voita et al., 2019). Causal probing techniques enable controlled intervention studies that distinguish correlation from causation (Vig et al., 2020; Elazar et al., 2021). However, systematic mapping to formal constructs remains incomplete.

The question of whether transformers implement symbolic computation implicitly or represent a fundamentally different computational paradigm remains open. Smolensky (1990) proposed tensor product representations as a bridge between connectionist and symbolic computation. More recent work explores whether attention mechanisms im-

plement variable binding (Manning et al., 2020). Hewitt and Manning (2019) introduced structural probes to test whether syntax trees are recoverable from neural representations, finding evidence for implicit hierarchical encoding.

Critically, representational alignment requires more than demonstrating that linguistic information is present in model activations. It requires showing that this information is organized in ways consistent with theoretical proposals and, crucially, that the model uses this information in the same way humans do during processing. This functional criterion remains largely unaddressed in current probing research. Ravfogel et al. (2020) demonstrate that probing accuracy can be high even when the probed information is not causally implicated in model behavior, highlighting the need for causal rather than merely correlational evidence.

From the perspective of brain data, recent advances using fMRI and representational similarity analysis techniques have begun mapping whether LLM embeddings are structurally aligned with brain activations. Doerig et al. (2025) show that LLM embeddings of scene captions successfully characterize brain activity evoked by viewing natural scenes, revealing alignment between LLM representations and high-level visual processing in the human brain. Using RSA on 23 mainstream LLMs, Ren et al. (2025) find that LLM-brain similarity correlates positively with pre-training data size and model scaling, and that alignment training can significantly improve LLM-brain correspondence. These studies suggest genuine points of convergence between representational formats in artificial and biological systems.

Furthermore, representations must support the productivity and systematicity characteristic of human language (Fodor and Pylyshyn, 1988). A model might encode syntactic structure in a way that supports parsing but fails to support novel combination. Testing for systematic generalization thus provides a crucial constraint on representational adequacy.

2.3. Developmental Alignment

Children acquire language from sparse, structured, socially embedded input. Bayesian models of word learning demonstrate that children leverage cross-situational statistics efficiently (Xu and Tenenbaum, 2007; Smith et al., 2014). Theory-driven frameworks for human-like learning highlight the role of inductive bias and active hypothesis testing (Lake et al., 2017; Tenenbaum et al., 2011).

Web-scale training regimes differ radically from child-directed input. The CHILDES database documents that children hear approximately 10 to 15 million words by age 10 (MacWhinney, 2000), orders of magnitude less than LLM training corpora. Furthermore, child-directed speech exhibits distinctive prosodic, lexical, and syntactic properties optimized for learning (Soderstrom, 2007).

Recent work has begun to explore whether models trained on developmentally realistic data can achieve human-like generalization. Warstadt et al. (2020) introduce the BLiMP benchmark for evaluating grammatical knowledge, finding that models require substantially more data than children to acquire comparable competence. The BabyLM Challenge (Warstadt et al., 2023) explicitly addresses this gap by constraining training data to approximate child language exposure, revealing both the difficulty of data-efficient learning and promising architectural directions.

Bridging the developmental gap requires curriculum design and learning constraints informed by developmental research. Work on curriculum learning (Bengio et al., 2009) and meta-learning (Finn et al., 2017) offers computational frameworks, but integration with developmental theory remains nascent. Understanding how children leverage multimodal, interactive, and socially scaffolded learning environments may prove essential for achieving data-efficient language acquisition in artificial systems.

The social dimension deserves particular emphasis. Children learn language in interaction with caregivers who provide contingent feedback, joint attention, and communicative scaffolding (Tomasello, 2003). Current LLMs learn from static text corpora stripped of these interactive dynamics. Active learning and curiosity-driven exploration suggest alternative training regimes grounded in intrinsic motivation rather than passive ingestion (Gottlieb et al., 2013; Oudeyer et al., 2007). Children do not simply observe language; they actively probe their environment through questions and experimentation. Incorporating such mechanisms may enhance data efficiency and generalization, and may be necessary for achieving human-like acquisition trajectories.

3. A Measured Critique of LLM Hype

Public discourse often equates LLM fluency with understanding. This conflation introduces both conceptual and scientific risks. A balanced assessment must acknowledge genuine achievements while maintaining appropriate skepticism about overclaims.

3.1. Prediction versus Explanation

Predictive models can approximate behavioral patterns without instantiating explanatory mechanisms. Shanahan (2024) caution against anthropomorphic descriptions that obscure mechanistic uncertainty. When we say a model “understands” or “knows,”

we risk importing unjustified assumptions about its internal states.

The distinction between prediction and explanation has deep roots in philosophy of science (Hempel, 1965). A model might predict human judgments with high accuracy while implementing mechanisms entirely unlike human cognition. Conversely, a model with lower predictive accuracy might better capture the causal structure of human language processing.

Recent debates about whether LLMs possess genuine semantic understanding illustrate these tensions. Bender and Koller (2020) argue that models trained solely on form cannot acquire meaning. Piantadosi and Hill (2022) counter that distributional information may suffice for certain semantic tasks. Resolving this debate requires precise operationalization of "understanding" and empirical tests that distinguish genuine comprehension from sophisticated pattern matching.

We suggest that the cognitive modeling community can contribute by developing precise behavioral signatures that differentiate alternative hypotheses about model cognition. Rather than debating understanding in the abstract, we should identify specific predictions that competing accounts make and design experiments to adjudicate between them.

3.2. Scale as Surrogate for Theory

LLMs achieve performance through scale. Yet scale cannot substitute for representational clarity. Baroni (2022) argues that grammatical abstraction remains indispensable for systematic compositionality. Typological diversity studies demonstrate that distributional similarity alone does not guarantee cross-linguistic generalization (Bender, 2011).

The emergence of apparently complex behaviors from simple learning objectives at scale has been termed "emergence" (Wei et al., 2022). However, recent work questions whether true qualitative transitions occur or whether apparent emergence reflects smooth scaling obscured by measurement artifacts (Schaeffer et al., 2023). Distinguishing genuine cognitive capacities from brittle shortcuts requires careful experimental design.

Furthermore, scaling laws that hold for English may not generalize to morphologically rich or low-resource languages. Wu and Dredze (2020) demonstrate substantial performance variation across languages even for multilingual models. This concern is equally relevant for cognitively oriented evaluation: the majority of reading time corpora and psycholinguistic benchmarks are in English, limiting the generalizability of cognitive alignment claims. Wilcox et al. (2023) address this gap by testing the predictions of surprisal theory across 11 languages from five language families, finding that surprisal is consistently predictive of reading times crosslinguistically, that contextual entropy also contributes, and that the surprisal-reading time relationship is linear. While these results offer the most robust cross-linguistic support for surprisal theory to date, they also highlight that model performance varies substantially across languages and scripts.

The cognitive modeling perspective emphasizes that human language acquisition is remarkably uniform across languages despite surface typological diversity, suggesting that human learners possess inductive biases that current models lack. Achieving truly general cognitive alignment may require architectural innovations beyond simple scaling, and evaluation across a broad typological sample is essential before claims of cognitive plausibility can be considered robust.

3.3. Benchmark Overfitting and Ecological Validity

Many evaluations rely on static benchmarks rather than dynamic processing measures. Eye-tracking corpora such as Dundee (Kennedy et al., 2003), GECO (Cop et al., 2017), and self-paced reading corpora such as Natural Stories (Futrell et al., 2021) provide ecologically valid constraints that go beyond accuracy metrics. These resources capture the temporal dynamics of language processing, revealing phenomena invisible to end-state accuracy measures. The Maze task has recently emerged as a complementary paradigm offering superior localization of processing effects, as demonstrated by Boyce and Levy (2023) who adapted the method for naturalistic passages using the Natural Stories corpus.

Benchmarks risk becoming targets rather than measures (Goodhart, 1984). As models are optimized for specific test sets, performance may inflate without corresponding improvements in genuine linguistic competence. McCoy et al. (2019) demonstrate that models achieving high accuracy on natural language inference benchmarks rely on superficial heuristics rather than robust reasoning.

Process-level evaluation using reading time prediction, N400 amplitude modeling, and garden-path recovery times offers more stringent tests of cognitive alignment. These measures constrain not just what the model gets right but how it processes linguistic input incrementally. The Provo corpus (Luke and Christianson, 2018) and the UCL corpus (Frank and Bod, 2011) provide valuable resources for such evaluation, but their adoption in NLP evaluation remains limited.

3.4. Societal and Ethical Dimensions

LLMs inherit biases from training data. Comprehensive auditing frameworks document systematic dis-

parities across demographic groups (Blodgett et al., 2020). Implicit association paradigms adapted from psychology reveal encoded stereotypes (Caliskan et al., 2017). These concerns extend naturally to the cross-linguistic dimension. Work on multilingual author profiling (Zaghouani and Charfi, 2018) reveals how language, dialect, and demographic variation interact in ways that both NLP and cognitive science must account for.

Research with low-resource languages raises ethical considerations regarding data sovereignty, community consent, and equitable benefit distribution (Bird, 2020). The computational resources required for LLM training concentrate power in wealthy institutions, potentially exacerbating existing inequalities in NLP research capacity. A balanced stance recognizes both empirical promise and conceptual limits while attending to broader societal implications. Cognitive alignment research can contribute by ensuring that models not only perform well but do so through mechanisms that are interpretable, fair, and aligned with human values.

4. Toward a Programmatic Integration

Building on the preceding critique, we outline a research program that leverages the complementary strengths of cognitive science, linguistics, and NLP.

4.1. Cognitive Science as Constraint Provider

Psycholinguistic paradigms offer fine-grained process-level data. Reading time studies using eye-tracking provide millisecond-resolution measures of incremental processing difficulty (Rayner, 2009). Event-related potential studies reveal distinct neural signatures for semantic versus syntactic violations (Kutas and Hillyard, 1980; Osterhout and Holcomb, 1992). These paradigms constrain models not just to produce correct outputs but to exhibit appropriate processing dynamics.

Lesioning studies of neural language models demonstrate how artificial systems can simulate selective impairments, creating bidirectional hypotheses between computational and neurocognitive research (Lakretz et al., 2021). By selectively ablating components and observing behavioral consequences, researchers can test causal hypotheses about representational organization.

Active learning and curiosity-driven exploration suggest alternative training regimes grounded in intrinsic motivation rather than passive ingestion (Gottlieb et al., 2013; Oudeyer et al., 2007). Children do not simply observe language; they actively probe their environment through questions and experimentation. Incorporating such mechanisms may enhance data efficiency and generalization.

Visual world paradigm studies provide real-time measures of referential processing (Tanenhaus et al., 1995). These paradigms could be adapted to evaluate whether model predictions align with human incremental interpretation in grounded contexts. The tight temporal coupling between linguistic input and visual attention provides a rich source of constraints on incremental processing that goes beyond what text-only measures can capture.

Memory and attention constraints from cognitive psychology also deserve consideration. Human working memory limitations shape sentence processing (Gibson, 1998), and models that incorporate similar constraints may better approximate human behavior. The interplay between memory retrieval and predictive processing remains an active area of investigation that computational models can help elucidate (Lewis and Vasishth, 2005).

4.2. Linguistics as Representational Scaffold

Formal theory supplies structured hypotheses about compositionality and semantic roles. Embedding structured representations such as AMR (Banarescu et al., 2013) or frame semantics (Baker et al., 1998) into neural architectures may enhance interpretability without sacrificing coverage.

Dependency grammar provides cross-linguistically applicable representations of syntactic structure (Tesnière, 1959; Mel'čuk, 1988). Universal Dependencies (Nivre et al., 2016) offers a standardized annotation framework enabling systematic cross-linguistic comparison.

Typological databases such as WALS (Dryer and Haspelmath, 2013) encode structural variation across languages, providing constraints for evaluating whether models capture universal versus language-specific properties. Models claiming cognitive alignment should perform consistently across typologically diverse languages, not just English. Construction grammar offers a framework for representing form-meaning pairings at multiple levels of abstraction (Goldberg, 1995), potentially offering better alignment with gradient human judgments and productivity patterns.

Discourse and pragmatic theory provide additional constraints often neglected in NLP evaluation. Coherence relations (Asher and Lascarides, 2003), information structure (Krifka, 2007), and pragmatic inference (Grice, 1975) all shape human language processing in ways that current benchmarks rarely assess.

4.3. NLP as Hypothesis Generator

Computational models can surface distributional regularities at scales beyond manual analysis.

Corpus-based approaches reveal statistical patterns that may inform theoretical proposals (Gries, 2006). The distributional hypothesis suggests that semantic similarity correlates with distributional similarity (Harris, 1954), a claim that neural embeddings operationalize at unprecedented scale.

Iterated learning frameworks demonstrate how population-level transmission shapes structure (Kirby et al., 2014). Mechanistic interpretability tools enable controlled intervention at representational levels (Geiger et al., 2021). Neuro-symbolic integration offers a path toward combining the strengths of neural and symbolic computation (Garcez and Lamb, 2020). Hybrid architectures that incorporate explicit symbolic reasoning may better capture the compositional structure of human language while retaining neural flexibility.

5. Infrastructure for Cognitive Modeling

We propose four initiatives aligned with cognitive and computational linguistics research priorities:

5.1. Cognitive Alignment Benchmark Suite

A comprehensive benchmark suite should integrate multiple data sources: eye-tracking corpora including Dundee (Kennedy et al., 2003) and GECO (Cop et al., 2017); self-paced reading data from the Natural Stories corpus (Futrell et al., 2021) and the UCL corpus (Frank and Bod, 2011); Maze task data providing superior localization (Boyce and Levy, 2023); ERP datasets targeting specific linguistic phenomena (Fedorenko et al., 2016); and child-directed speech corpora from CHILDES (MacWhinney, 2000). Recent large-scale multilingual eye-tracking resources, such as the Multilingual Eye-movement Corpus (MECO) covering 13 languages (Siegelman et al., 2022) and the ongoing MultiplEYE initiative (Jakobi et al., 2025), should also be incorporated to ensure cross-linguistic coverage.

Evaluation should span multiple metrics: surprisal correlation with reading times; psychometric predictive power for individual differences (Kuribayashi et al., 2024); representational similarity analysis comparing model activations to neural data; and incremental commitment measures capturing garden-path effects. Critically, the benchmark should include adversarial test sets designed to probe specific theoretical predictions, not just naturalistic data.

The CMCL shared tasks on eye-tracking prediction (Hollenstein et al., 2021, 2022) offer valuable precedents for standardized evaluation, and future benchmarks should build on these by incorporating a wider range of behavioral paradigms and languages.

5.2. Hybrid Modeling Shared Task

A shared task should require explicit integration of structured semantic representations with neural architectures. Participants would receive AMR-annotated data (Banarescu et al., 2013) and FrameNet annotations (Baker et al., 1998) alongside raw text.

Evaluation criteria should include compositional generalization on COGS (Kim and Linzen, 2020) and SCAN (Lake and Baroni, 2018) benchmarks; interpretability assessments using probing tasks and attention analysis; and behavioral alignment with human processing data. Submissions should include documentation of how structured knowledge is incorporated and used during inference. Such a shared task would incentivize development of architectures that bridge neural and symbolic paradigms rather than treating them as competing approaches. It would also generate valuable data on which integration strategies prove most effective for different aspects of cognitive alignment.

5.3. Multi-Agent Language Evolution Platform

Grounded in iterated learning theory (Kirby et al., 2014), a simulation platform should enable controlled experiments on language emergence and change. Agents would communicate to accomplish shared goals, with language emerging from interaction rather than being imposed externally.

The platform should support manipulation of population structure, transmission fidelity, and environmental complexity. Emergent communication systems could be analyzed using standard linguistic tools, enabling comparison with natural language typology. Such simulations provide controlled environments for testing hypotheses about language universals and historical change. Integration with reinforcement learning frameworks would enable exploration of how communicative utility shapes linguistic structure (Lazaridou et al., 2020). The platform could also support investigation of how social structure and network topology affect language evolution.

5.4. Cross-Disciplinary Training Programs

Sustainable progress requires researchers fluent in multiple traditions. Joint doctoral programs bridging computational linguistics, cognitive psychology, and formal linguistics would address this need. Core curricula should include probability theory and statistical modeling; formal syntax and semantics;

experimental methods in psycholinguistics; and machine learning and neural network architectures.

Shared repositories linking computational implementations, experimental materials, and theoretical analyses would lower barriers to interdisciplinary collaboration. Summer schools and workshops bringing together researchers from different traditions can foster communication and identify productive research directions. Funding agencies should incentivize interdisciplinary proposals, and review panels should include expertise spanning the relevant fields.

6. Challenges and Limitations

Several challenges complicate the proposed agenda. First, the three levels of alignment may conflict: a model achieving strong behavioral alignment might violate theoretical constraints on representation, or vice versa. Resolving such conflicts requires principled criteria for adjudicating trade-offs. We suggest that when conflicts arise, they should be treated as scientifically informative rather than merely problematic, potentially revealing inadequacies in current theories. Multi-objective optimization across potentially conflicting criteria requires careful methodological development, and the field has yet to establish widely accepted frameworks for weighing behavioral fit against representational fidelity or developmental plausibility.

Second, human language processing itself remains incompletely understood. Models cannot align with a target that is poorly specified. Ongoing empirical research must continue refining our understanding of human cognition alongside computational modeling efforts. The interaction between modeling and empirical research should be bidirectional, with models generating predictions that drive new experiments and experimental results informing model design.

Third, computational and experimental resources remain unevenly distributed. Many of the proposed initiatives require substantial investment in data collection, infrastructure, and interdisciplinary training. Ensuring equitable access and participation will require deliberate effort. In particular, the concentration of reading time datasets on English and a handful of European languages limits our ability to assess whether cognitive alignment generalizes across the world's languages. Recent efforts toward multilingual eye-tracking data collection, such as MECO (Siegelman et al., 2022) covering 13 languages and the MultiplEYE initiative (Jakobi et al., 2025) collecting data across numerous additional languages, represent important steps toward addressing this gap. However, many language families, scripts, and morphological typologies remain unrepresented. The concentration of NLP research and psycholinguistic experimentation on English continues to limit our understanding of how well current approaches generalize, and the computational resources required for LLM training further concentrate power in wealthy institutions.

Fourth, evaluation metrics for cognitive alignment remain underdeveloped. While we have proposed multiple criteria, integrating them into a coherent framework and validating that framework against human cognition presents ongoing challenges. Standardized evaluation protocols would enable systematic comparison across models and facilitate meta-analysis, but the diversity of data types (reading times, ERP amplitudes, fMRI patterns, acquisition trajectories) makes unified scoring non-trivial.

Finally, the relationship between cognitive alignment and practical utility is unclear. Systems optimized for engineering performance may differ substantially from cognitively aligned systems. Whether cognitive alignment enhances robustness, interpretability, and alignment with human values remains an empirical question. We hypothesize that cognitively aligned systems may exhibit better generalization and more predictable failure modes, but this requires systematic investigation.

7. Conclusion

Large language models mark a significant empirical advance. They demonstrate that distributional learning at scale can capture substantial linguistic structure. Correlations with human behavioral and neural data suggest genuine points of contact between artificial and natural language processing.

However, cognitive modeling requires more than predictive performance. A principled research program must integrate experimental constraint, formal theory, and computational scalability. Each tradition contributes essential perspectives: cognitive science provides process-level data and learning constraints; linguistics supplies representational hypotheses and cross-linguistic generalizations; NLP offers computational tools for modeling and analysis.

By pursuing cognitive alignment across behavioral, representational, and developmental levels, the field can move beyond hype toward cumulative scientific progress. The initiatives proposed here, including benchmark suites, shared tasks, simulation platforms, and training programs, offer concrete steps toward this goal.

The ultimate aim is not merely to build better language models but to understand language itself: how it is represented, processed, learned, and used. This understanding has implications beyond basic science, informing applications in education, clinical intervention, and human-computer interac-

tion. A cognitively aligned approach ensures that progress in language technology remains grounded in and accountable to our knowledge of human cognition. The path forward requires sustained collaboration across disciplinary boundaries, and we hope this position paper contributes to ongoing dialogue and inspires concrete initiatives advancing the vision of a unified science of language.

8. Limitations

This paper proposes a research agenda and is necessarily constrained in several ways. As a position paper, it does not include empirical validation of the proposed framework. The tripartite alignment scheme (behavioral, representational, developmental) is presented as a guiding framework rather than a formally specified evaluation metric, and operationalizing each level into concrete, standardized measures remains an open challenge.

The paper focuses primarily on text-based language processing and does not extensively address multimodal aspects of human language use, including gesture, prosody, and visual grounding, which are increasingly recognized as central to human communication. Human language processing is fundamentally situated in embodied, social contexts, and a complete account of cognitive alignment would need to incorporate these dimensions.

Furthermore, while we advocate for cross-linguistic evaluation, the examples and resources discussed in this paper still skew toward well-resourced languages, particularly English and European languages. Extending cognitive alignment research to typologically diverse languages, including morphologically rich languages, tonal languages, and languages with non-Latin scripts, is essential for testing the universality of proposed alignment criteria but faces significant practical barriers in terms of both data availability and computational infrastructure. The collection of psycholinguistic processing data (e.g., eye-tracking corpora, self-paced reading data, ERP recordings) in underrepresented languages remains a crucial bottleneck.

9. Ethics Statement

This paper proposes a research agenda and does not involve new human subjects data. The proposed research directions raise ethical considerations regarding data collection from human participants, which should follow established protocols including informed consent and appropriate compensation. Research with low-resource languages should adhere to community-engaged practices respecting data sovereignty and ensuring equitable benefit distribution.

Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

10. Bibliographical References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111(4), 1036–1060.

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of COLING-ACL*, 86–90.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., & Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop*, 178–186.

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic Structures in Natural Language*, 1–28. CRC Press.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of ACL*, 5185–5198.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3), 1–26.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of ICML*, 41–48.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of COLING*, 3504–3519.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of ACL*, 5454–5476.

Boyce, V., & Levy, R. (2023). A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics*, 2(1).

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1), 134.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. In *Proceedings of the BlackboxNLP Workshop*, 276–286.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2), 602–615.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2025). High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7(8), 1220–1234.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology.

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the ACL*, 9, 160–175.

Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., & Kanwisher, N. (2016). Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41), E6256–E6262.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML*, 1126–1135.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2), 3–71.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6), 829–834.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1–11.

Futrell, R., Wilcox, E., Mober, T., Qian, P., Ballesteros, M., Dyer, C., Gibson, E., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of NAACL-HLT*, 32–42.

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1), 63–77.

Garcez, A. d'A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876*.

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, 34, 9574–9586.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1–76.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Goldstein, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369–380.

Goodhart, C. A. (1984). Problems of monetary management: The UK experience. In *Monetary Theory and Practice*, 91–121. Springer.

Gottlieb, J., Oudeyer, P. Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11), 585–593.

Grice, H. P. (1975). Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, 41–58. Academic Press.

Gries, S. T. (2006). Corpus-based methods and cognitive semantics: The many senses of to run. In *Corpora in Cognitive Linguistics*, 57–99. Mouton de Gruyter.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL*, 1–8.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146–162.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. Free Press.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT*, 4129–4138.

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2021). CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 72–78.

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2022). CMCL 2022 shared task on multilingual and crosslingual prediction of human reading behavior. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Jakobi, D. N., Stegenwallner-Schütz, M., Hollenstein, N., et al. (2025). MultiplEYE: Creating a multilingual eye-tracking-while-reading corpus. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, 1–11. ACM.

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Kim, N., & Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of EMNLP*, 9087–9105.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114.

Krifka, M. (2007). Basic notions of information structure. In *Interdisciplinary Studies on Information Structure*, 6, 13–55.

Kuribayashi, T., Oseki, Y., & Baldwin, T. (2024). Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 1983–2005.

Kuribayashi, T., Oseki, Y., Ben Taieb, S., Inui, K., & Baldwin, T. (2025). Large language models are human-like internally. *Transactions of the Association for Computational Linguistics*, 13, 1743–1766.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of ICML*, 2873–2882.

Lakretz, Y., Desbordes, T., King, J. R., Crabbé, B., Oquab, M., & Dehaene, S. (2021). Can RNNs learn recursive nested subject-verb agreements? In *Proceedings of NAACL-HLT*, 3558–3569.

Lazaridou, A., Potapenko, A., & Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of ACL*, 7663–7674.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3), 375–419.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the ACL*, 4, 521–535.

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2), 826–833.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Lawrence Erlbaum Associates.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of ACL*, 3428–3448.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. SUNY Press.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In *Approaches to Natural Language*, 221–242. Springer.

Nivre, J., et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, 1659–1666.

Oh, B.-D., Zhu, H., & Schuler, W. (2025). The inverse scaling effect of pre-trained language model surprisal is not due to data leakage. In *Findings of the Association for Computational Linguistics: ACL 2025*, 1820–1827.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6), 785–806.

Oudeyer, P. Y., Kaplan, F., & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2), 265–286.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 7237–7256.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8), 1457–1506.

Ren, Y., Jin, R., Zhang, T., & Xiong, D. (2025). Do large language models mirror cognitive language processing? In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2988–3001.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, 36.

Schrimpf, M., et al. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.

Siegelman, N., Schroeder, S., Acartürk, C., et al. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, 54(6), 2843–2863.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302–319.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, 18(5), 251–258.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2), 159–216.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4), 501–532.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.

Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Vig, J., et al. (2020). Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 33, 12388–12401.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of ACL*, 5797–5808.

Warstadt, A., et al. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the ACL*, 8, 377–392.

Warstadt, A., et al. (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at CoNLL*, 1–34.

Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of CogSci*, 1707–1713.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11, 1451–1470.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, 120–130.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.

Zaghouani, W., & Charfi, A. (2018). AraP-Tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Author Index

Abel, Magdalena, 172
Agrawal, Aakash, 293
Akama, Haruka, 124
Alishahi, Afra, 60
Antonenko, Pavlo, 86

Baer-Henney, Dinah, 74
Barth, Raphael, 276
Belcher, Kate Rebecca, 99
Bianchi, Bruno, 286, 293
Briscoe, Ted, 48
Brisebois, Zoe, 250

Cardona Ruiz, Joaquin, 230
Chemla, Emmanuel, 293
Chen, Yaran, 10
Chen, Zesheng, 10
Crocker, Matthew, 99

Dehaene, Stanislas, 293
Delmagro, Agustin, 286
Demberg, Vera, 110
Donatelli, Lucia, 230

El-Naggar, Nadine, 48
Enevoldsen, Kenneth, 60

Fernandez Slezak, Diego, 286
Fokkens, Antske, 230
Frank, Stefan L., 110
Fraser, Alexander, 145

Gorzoni, Greta, 203
Gregori, Lorenzo, 221

Ilinykh, Nikolai, 25

Jang, Hyewon, 25

Kakolu Ramarao, Akhilesh, 74
Kallmeyer, Laura, 188
Kamienkowski, Juan E., 286
Karakas, Sercan, 309, 329
Kong, Lingchen, 86
Kuribayashi, Tatsuki, 48

Laforest, Frederique, 264
Lakretz, Yair, 293
Lappin, Shalom, 25
Lau, Jey Han, 25
Lenci, Alessandro, 35
Li, Luan, 10
Li, Na, 10
Loaiciga, Sharid, 25
Lombardi, Agnese, 35

Masini, Francesca, 203
Meyer, Selina, 172
Mielczarek, Lukas, 188
Moncla, Ludovic, 264
Mühlenbernd, Roland, 162
Müller-Eberstein, Max, 124

Nicenboim, Bruno, 60

Oseki, Yohei, 124
Østergaard, Sara Møller, 60

Paci, Walter, 221
Pannitto, Ludovica, 203
Panunzi, Alessandro, 221
Parker, Dan, 1
Piccirilli, Prisca, 145

Qiu, Mengyang, 250

Roth, Michael, 172

Schlotterbeck, Fabian, 276
Schulte im Walde, Sabine, 145
Shin, Jinnie, 86
Şimşek, Yusuf, 329
Škrjanec, Iza, 110
Sun, Siena, 250
Sun, Xin, 10

Tang, Kevin, 74
Taroni, Antoine, 264
Travi, Fermin, 286

Wang, Gan, 10
Wang, Shaonan, 10
Winther, Irene Elisabeth, 110

Wu, Yiwen, 10

Yoshida, Ryo, 124

Zaghouani, Wajdi, 347
Zhang, Shuchen, 10