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.

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.

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

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,

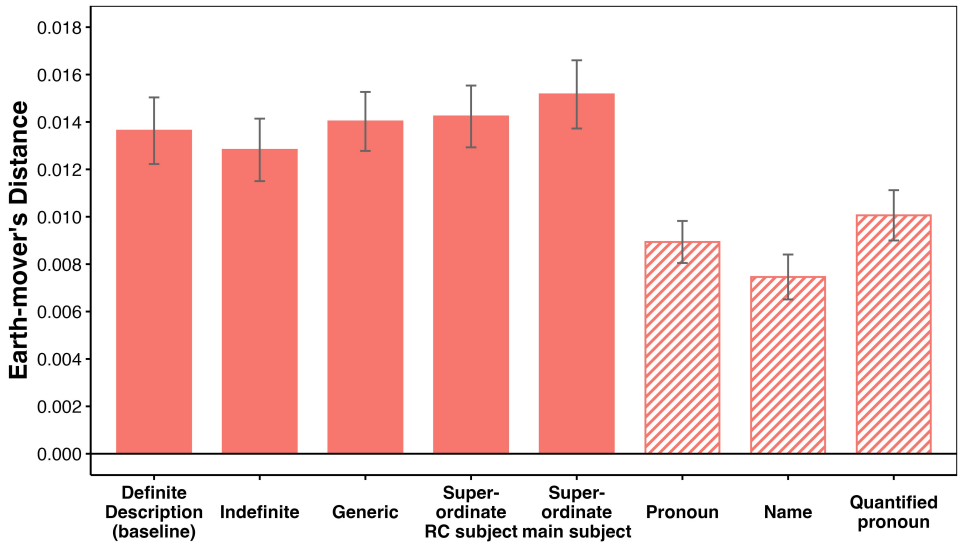

Figure 1: Earth Mover's Distance for each experimental manipulation. Positive values reflect greater redistribution of attention. Solid bars are conditions predicted to show an ORC/SRC asymmetry (higher EMD) due to encoding interference from referential similarity. Striped bars are conditions predicted 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.

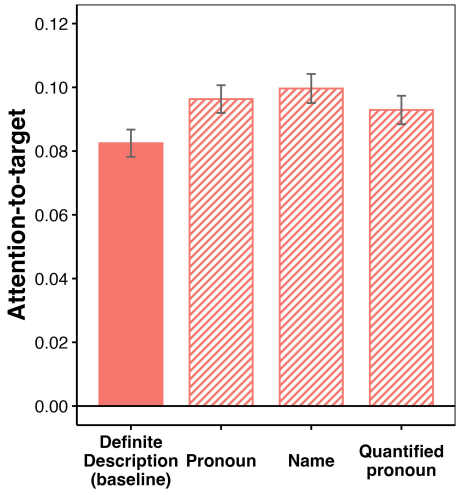

Figure 2: Attention to the target at the critical region for the baseline, pronoun, name, and quantified pronoun conditions. Higher values indicate greater attention to the target, reflecting more accurate retrieval. The solid bar (definite description) is predicted to show decreased attention to the 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

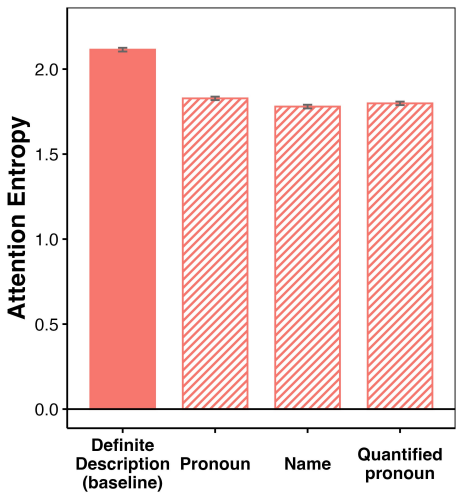

Figure 3: Attention entropy at the critical region for the baseline, pronoun, name, and quantified pronoun conditions. Higher values reflect more diffuse attention over the prior context, indicating greater uncertainty. The solid bar (definite description) is predicted to exhibit higher entropy due to 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.