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.

(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.

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.

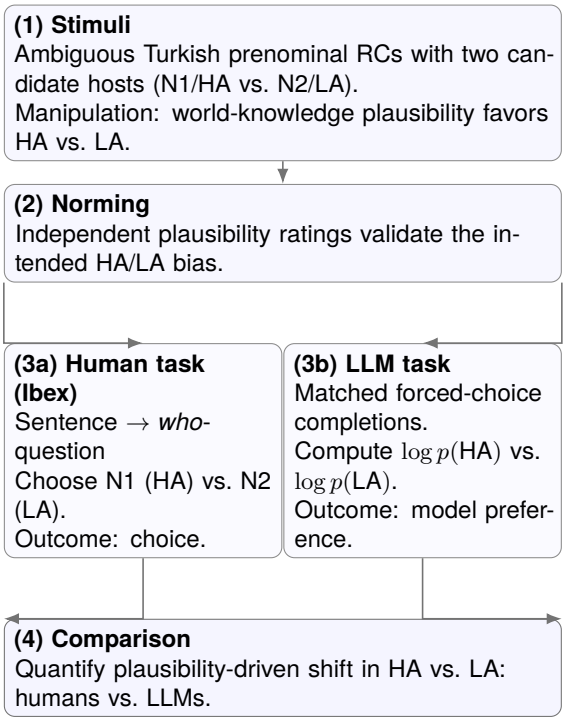


Figure 1: Procedure overview. After constructing syntactically matched Turkish RC attachment ambiguities and validating plausibility via norming, we evaluate (a) human attachment choices in a speeded forced-choice task and (b) LLM attachment preferences via log-probability scoring over matched HA/LA continuations.

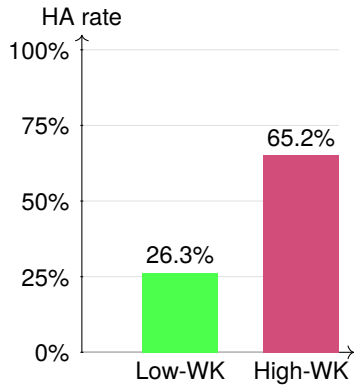
4.1. Models

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:

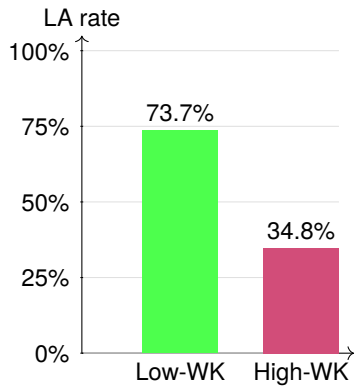
Ytu-ce-cosmos/turkish-gpt2. A Turkish GPT-2 style decoder-only language model used as a lightweight Turkish baseline. (Kesgin et al., 2024; ytu-ce-cosmos, 2024)

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

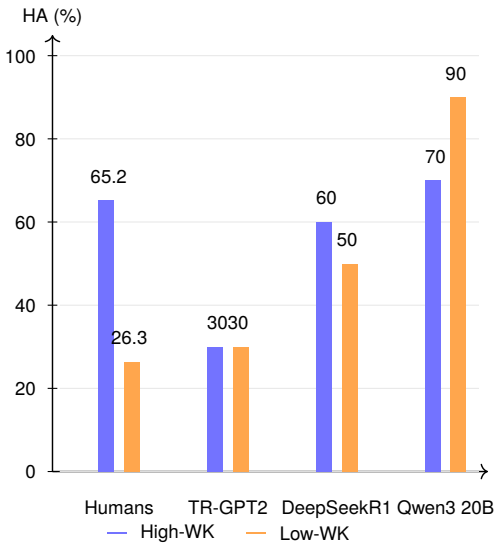


Figure 3: HA rates (%) by WK condition (High-WK vs. Low-WK) for humans and models.

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.