Human-Like Anaphor Resolution in Large Language Models

Keane Zhang, Varshini Chinta, Raj Sanjay Shah, Sashank Varma
{keane, vchinta6, rajsanjayshah, varma}@gatech.edu
Georgia Institute of Technology

Abstract

Anaphors are expressions that refer to other expressions, called *antecedents*. The process of connecting the two is called *resolution*. Cognitive science has identified multiple factors that affect the speed and success of anaphor resolution, including discourse structure, situation-model properties, and semantic factors. Here, we investigate whether these factors also affect anaphor resolution in five Large Language Models (LLMs) with open weights: GPT-2-XL, Llama-3.1-8B, Pythia-12B, Mistral-7B, and Mistral-24B. To model processing difficulty, we adopt the standard linking hypothesis that relates human reading times to model surprisal at the anaphor. As a second behavioral measure, we compare model accuracy to human accuracy on comprehension questions probing the antecedents of anaphors. The results show selective cognitive alignment: some LLMs exhibit human-like sensitivity to discourse prominence and distance-based factors in anaphor resolution, while showing weaker or absent sensitivity to semantic interference effects. These findings delimit the conditions under which LLMs approximate human anaphor resolution.

Keywords: anaphor resolution; situation models; Large Language Models; surprisal; cognitive alignment

Introduction

Large Language Models (LLMs) are deep neural networks with millions or billions of parameters trained on large text corpora. Transformer-based architectures, beginning with BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) and extending to more recent families such as GPT (OpenAI, 2025), Llama (Grattafiori et al., 2024), and Mistral (Mistral AI Team, 2025), are highly performant on language comprehension tasks (Wang et al., 2024). Beyond task performance, these models are increasingly evaluated as candidate models of human cognition (Ivanova, 2025; Piantadosi et al., 2024; Shah & Varma, 2025), and more specifically as models of human language processing (BabyLM Organizers, 2025; Li et al., 2024).

Much of the existing work evaluating LLMs as cognitive models has focused on phenomena at the word and sentence levels, examining model behavior in relatively local and decontextualized linguistic settings. This leaves open the question of whether LLMs capture discourse-level processes that unfold over extended text. Anaphor resolution is one such process: it requires maintaining and retrieving information across sentences, integrating linguistic input with a developing situation model, and resolving interference from competing referents. Modeling these demands goes beyond local prediction and instead requires sensitivity to connected text. In sequential language models, shifts in expected word probabilities, typically operationalized as surprisal, have been shown to track human reading times, providing a process-level link between model predictions and human comprehension behavior (Li et al., 2024; Wilcox et al., 2020).

Despite growing interest in cognitive alignment, comparatively little attention has been paid to text- and discourse-level phenomena. The current study addresses this gap by focusing on anaphor resolution. It focuses on *anaphors*, which are expressions that refer to other expressions, called *antecedents*. Pronouns are a familiar class of anaphors; more generally, language is rife with referential expressions. The process of connecting an anaphor to its antecedent during online comprehension is called *resolution*. This can be easy when an anaphor and its antecedent occur within the same sentence or within a few sentences of each other; in this case, resolution requires only searching working memory (Daneman & Carpenter, 1980). However, when reading longer texts, anaphors can be separated from their antecedents by many sentences. In this case, resolution requires using an anaphor as a cue to memory and attempting to retrieve its antecedent from the reader's situation model, which is the evolving representation of the overall meaning of the text (McNamara & Magliano, 2009; van Dijk & Kintsch, 1983).

Cognitive science research has investigated the factors that affect the speed and accuracy of anaphor resolution for humans. For example, the greater the number of sentences between an anaphor and its antecedent, the slower it is read, presumably because readers must "search" farther back in their memory for the text. Reading time is one measure; another is the accuracy of comprehension questions. To continue the example, the greater the number of sentences, the less accurate people are when answering a comprehension question about the antecedent, presumably because there is a greater chance that the resolution failed. *Here, we ask whether LLMs are sensitive to the same factors as human comprehenders during anaphor resolution.* To the extent that they do, LLMs gain credibility as candidate models of discourse-level language processing in cognitive science (Ivanova, 2025; Shah & Varma, 2025).

Code for reproducing all model evaluations, analyses, and figures is available at: `github.com/wristy/anaphor`.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

Cognitive Science Studies of Anaphor Resolution

The current study focuses on six factors identified by cognitive science as affecting the speed and accuracy of anaphor resolution.

The first factor is a thematic feature of texts: If the antecedent is *topicalized*, then it should feature more prominently in a reader's situation model, and therefore it should be more accessible when an anaphor to it is encountered. In this case, resolution should be relatively fast and accurate. There are various ways to topicalize an antecedent. For example, O'Brien (1987) manipulated whether or not it was focused by the title of the text. The second factor concerns a surface feature of text itself: The greater the *sentential distance* (i.e., number of sentences) between an anaphor and its antecedent, the slower and less accurate resolution is (Clark & Sengul, 1979). O'Brien (1987) orthogonally varied these two factors, antecedent topicality and sentential distance, and Experiment 1 runs LLMs on the materials of this study.

The third and fourth factors concern not surface distance but *contextual* distance. As readers comprehend a text, they build a model of the story world it describes, variously called the situation model, mental model, or world model (McNamara & Magliano, 2009; van Dijk & Kintsch, 1983). The third factor is the *spatial distance* (i.e., perceived physical distance) between the anaphor and antecedent in the reader's situation model (Morrow et al., 1987; Rinck & Bower, 1995). The fourth factor is the *temporal duration* (i.e., perceived elapsed time) between the two (Anderson et al., 1983; Varma & Janssen, 2019; Zwaan, 1996). The greater the spatial distance and the longer the temporal duration, the slower and less successful anaphor resolution becomes. The following text illustrates a relatively long temporal duration:

Antecedent: The mechanic kicked the metal chair.

*Duration (long): It took **2 hours** to drive back home and return with the critical tool.*

Anaphor: Afterwards, he felt foolish for having kicked the metal furniture when he had been frustrated.

Anderson et al. (1983) found that readers are relatively slow to read the anaphor and less accurate in answering a follow-up comprehension question about the antecedent:

What piece of metal furniture did the mechanic kick out of frustration?

compared to a text with a relatively short temporal duration:

*Duration (short): It took **10 minutes** to drive back home and return with the critical tool.*

The fifth and sixth factors concern the semantics of the situation models that readers build. The *semantic similarity* between the anaphor and antecedent is how much they resemble each other: the greater the similarity, the better the anaphor is as a cue to memory to retrieve the antecedent from the situation model. Returning to the example above, *chair* is a typical member of the *furniture* category (Rosch et al., 1976). Thus, the two have high semantic similarity, and this will speed resolution. By contrast, if the antecedent had been atypical (e.g., *stool*), this would have slowed resolution (Garrod & Sanford, 1977; Varma & Janssen, 2019). Next, consider if a second member of the *furniture* category had been present in the text, one that is not the antecedent:

Distractor: The mechanic walked past the wooden desk and headed to the door.

If it is a typical member of the same category, as in the example above, then this causes *semantic interference* in the search for the correct antecedent (i.e., *chair*) in the situation model, and this will slow resolution and undermine its accuracy. By contrast, if the non-antecedent distractor is atypical (e.g., *wooden shelf*), then this should have only a small deleterious effect (Corbett, 1984; Varma & Janssen, 2019).

Anaphor Resolution in LLMs

Process-level links between language models and human comprehension are often established using surprisal, which has been shown to correlate with human reading times across a range of syntactic and semantic phenomena (Hale, 2001; Levy, 2008). While surprisal-based analyses have been widely applied at the word and sentence levels, their use in evaluating discourse-level processes such as anaphor resolution remains limited (Li et al., 2024; Wilcox et al., 2020).

In early work, Hu et al. (2020) and Lee and Schuster (2022) evaluated the ability of early LLMs (e.g., BERT, GPT-2) to resolve reflexive pronominal references to antecedents within the same sentence, with mixed results even for this simple case. More promisingly, Pandit and Hou (2021) evaluated BERT's ability to resolve anaphors over increasing function of sentential distance. They identified attention heads at higher layers that bridge between anaphors and their antecedents, though these connections were weaker at longer distances (6 – 10 sentences). They also used a cloze procedure to probe, at the anaphor position, the model's best guess of the antecedent, finding low accuracy even at close sentential distances (i.e., ≥ 3 sentences).

Recent work has examined the ability of LLMs to resolve anaphors and coreference relations, either by prompting them directly or by adapting them to existing benchmarks. These studies show substantial variability across models, datasets, and prompt formulations, and general-purpose LLMs often underperform specialized coreference systems (Cambria, 2025; Liu et al., 2025). Moreover, reported gains can be sensitive to dataset artifacts, lexical heuristics, and recency biases, raising questions about whether high accuracy reflects robust discourse understanding or shallow statistical cues (Gan et al., 2024; Novák et al., 2025).

More importantly, high coreference accuracy does not imply human-like processing. Existing evaluation practices rely on inconsistent metrics and lack a standard framework for assessing LLM-generated responses, particularly across datasets and

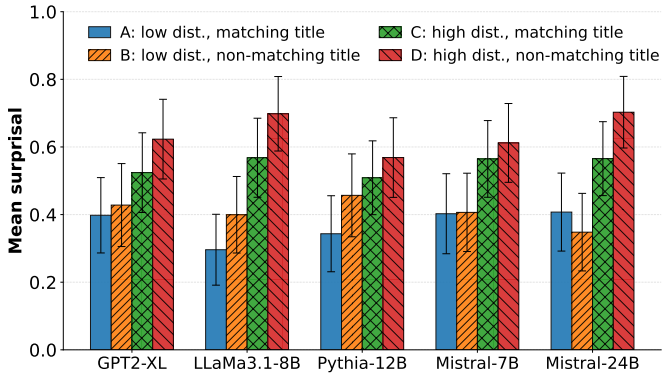

Figure 1: Mean surprisal on the anaphor as a function of antecedent topicality and sentential distance (Exp. 1). The error bars are standard errors computed across the 16 texts.

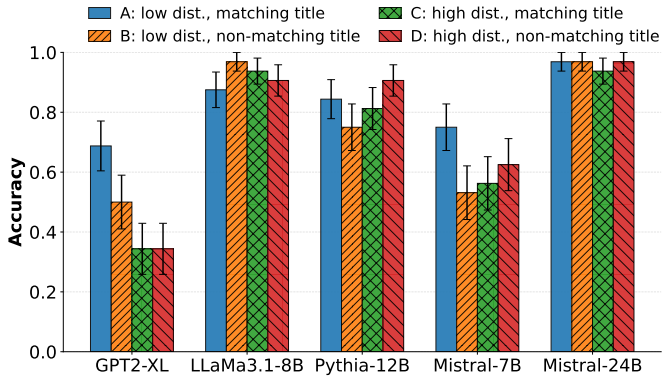

Figure 2: Mean comprehension question accuracy as a function of antecedent topicality and sentential distance (Exp. 1). The error bars are standard errors.

task formulations (Talukdar & Rahman, 2025). Standard NLP benchmarks do not manipulate or isolate the factors that cognitive science has shown to govern anaphor resolution, such as discourse prominence, distance-based accessibility, semantic similarity, or interference from competing referents. As a result, existing evaluations cannot distinguish between models that resolve anaphors using surface heuristics and those that exhibit sensitivity to the cognitive constraints shaping human comprehension (Ivanova, 2025; Shah & Varma, 2025).

Rather than proposing a new benchmark or optimizing coreference performance, the present study adopts classic cognitive science paradigms as the evaluation framework. We ask whether LLMs exhibit sensitivity to the same discourse, situational, and semantic factors that shape human anaphor resolution, using surprisal and comprehension accuracy as complementary behavioral proxies.

This framing allows us to evaluate cognitive alignment beyond correct referent identification, focusing instead on whether models are sensitive to the same processing factors that affect human readers.

Research Questions

We evaluate whether LLMs are sensitive to the six factors delineated above that have been shown to affect anaphor resolution in humans. In particular, we test whether resolution is facilitated when antecedents are more accessible, due to greater discourse prominence, shorter sentential, spatial, or temporal distance, higher semantic similarity, and reduced interference from competing referents.

We address this question using five open-weight LLMs (GPT-2-XL, LLaMa-3.1-8B, Pythia-12B, Mistral-7B, and Mistral-24B). To assess process-level sensitivity, we adopt the standard linking hypothesis that relates model surprisal at the anaphor to human reading times (Hale, 2001; Levy, 2008; Li et al., 2024; Wilcox et al., 2020). To assess resolution success, we also measure model accuracy on comprehension questions that probe the antecedents of anaphors after text processing.

Experiment 1

Experiment 1 investigated the effect of (1) a thematic feature, whether the antecedent is topicalized by the text's title, and (2) a textual feature, the sentential distance (i.e., number of intervening sentences) between anaphors and antecedents, on anaphor resolution time and accuracy.

Design and Materials

The materials were from O'Brien (1987). There were 16 texts, each occurring in four versions formed by orthogonally varying two factors. One factor was sentential distance (near, far), with the antecedent occurring either $M = 5.6$ or $M = 15.2$ sentences prior to the anaphor. The other factor was antecedent topicality (high, low), with the antecedent either focused by the title of the text or not. Thus, the four versions were:

A. near sentential distance, high antecedent topicality
B. near sentential distance, low antecedent topicality
C. far sentential distance, high antecedent topicality
D. far sentential distance, low antecedent topicality

Procedure and Dependent Measures

Anaphor Processing. The surprisal of a probabilistic model in predicting the next token i is $-\log_2(p_i)$ where p_i is the probability of i. A common linking hypothesis is that the greater a model's surprisal for the next word, the longer the predicted reading time (Hale, 2001; Levy, 2008; Li et al., 2024; Wilcox et al., 2020). The surprisal for a text/version was computed as the average surprisal across the tokens making up the anaphor. Because these values were highly variable across the 16 texts, we normalized them within each text using:

$$\frac{surprisal - \min(surprisal_j)}{\max(surprisal_j) - \min(surprisal_j)}$$

where $surprisal_j$ denotes the set of the four surprisals for versions A-D. Thus, the normalized surprisals ranged from

0 to 1. For each of the four versions A-D, we averaged the normalized surprisals across the 16 texts, resulting in four mean surprisals. We carried out this process for each LLM.

Comprehension Question Answering. For each of the 16 texts, there is a comprehension question asking for the antecedent of the anaphor. After an LLM processed a text, it was prompted with this question. We code the generated responses by adopting an "LLM-as-a-judge" paradigm using Gemini-2.5-flash-preview-09-2025 as the judge. Each question had a "gold" answer provided by us. The judge was prompted to first provide a brief justification comparing the candidate response to the gold answer and then output a binary accuracy label (1 = accurate, 0 = inaccurate). As with surprisal, for each of the four versions A-D, we averaged the accuracy score across the 16 texts, resulting in four mean accuracies. We repeated this process for each LLM.

Results and Discussion

Surprisal Prediction of Reading Times. Figure 1 shows, for each of the LLMs, the average surprisal for each of the four text versions. The prediction is that this should be lowest (i.e., that anaphor reading times should be fastest) when the title topicalizes the antecedent, leading readers to focus their attention on it during situation model construction, and when the sentential distance between the anaphor and antecedent is near. Thus, average surprisal should be lowest for version A, highest for version D, and intermediate for versions B and C. GPT-2-XL, Llama-3.1-8B, Pythia-12B, and Mistral-7B showed this pattern.

Comprehension Question Accuracy. Figure 2 shows, for each of the LLMs, the average accuracy for each of the four text versions. The predictions mirror those for surprisal, i.e., that accuracy should be highest when the title topicalizes the antecedent and when the sentential distance between the anaphor and antecedent is near. Thus, accuracy should be highest for version A, lowest for version D, and intermediate for the other versions B and C. GPT-2-XL showed this pattern. Among the other models, only Mistral-7B correctly ordered versions A and D. We note that because Llama-3.1-8B and Mistral-24B performed at ceiling on the comprehension questions, this might have limited our ability to detect differences between conditions.

Experiment 2

Experiment 2 investigated whether two contextual factors, spatial distance and temporal duration between anaphors and antecedents in readers' situation models, affect resolution time and accuracy. The details were the same as Experiment 1, except where otherwise noted.

Design and Materials

The materials were those of Experiment 1 of Varma and Janssen (2019). There were 19 texts, and each occurred in four versions formed by orthogonally varying the spatial distance (near, far) and temporal duration (short, long) between the anaphor and antecedent in the reader's situation model:

A. near spatial distance, short temporal duration
B. near spatial distance, long temporal duration
C. far spatial distance, short temporal duration
D. far spatial distance, long temporal duration

These factors were manipulated in the M = 15.1 (SD = 2.4) sentences that separated anaphors from their antecedents. The same categorical anaphor (e.g., *metal furniture*) and typical antecedent (e.g., *metal chair*) were used for each of the four versions A-D.

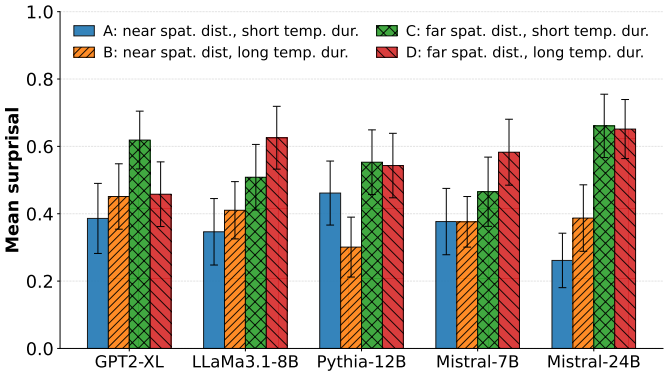

Figure 3: Mean surprisal on the anaphor as a function of spatial distance and temporal duration (Exp. 2). The error bars are standard errors.

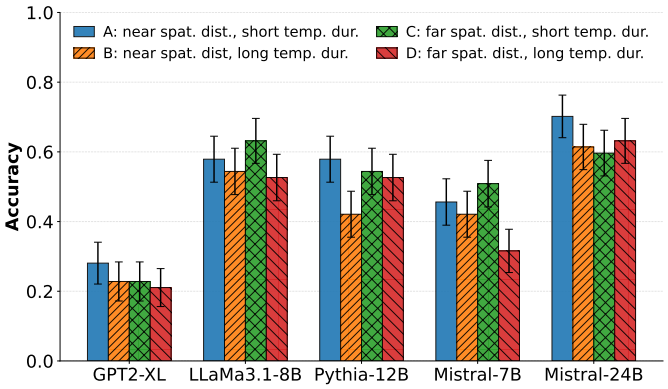

Figure 4: Mean comprehension question accuracy as a function of spatial distance and temporal duration (Exp. 2). The error bars are standard errors.

Results and Discussion

Surprisal Prediction of Reading Times. Figure 3 shows the average surprisal of each of the five models on each of the four text versions. The prediction is that this should be lowest (i.e., anaphor resolution fastest) for version A, highest (i.e., anaphor resolution slowest) for version D, and intermediate

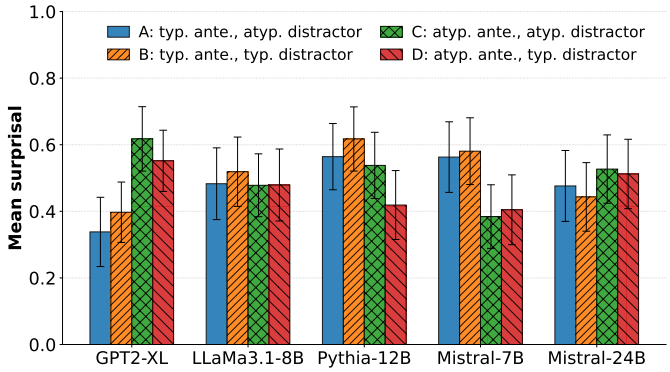

Figure 5: Mean surprisal on the anaphor as a function of the semantic overlap and semantic interference factors (Exp. 3). The error bars are standard errors.

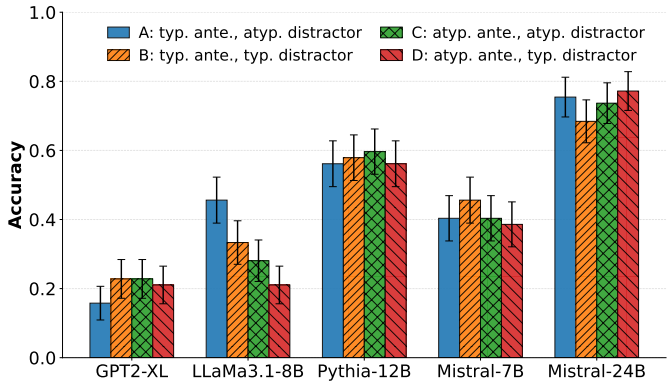

Figure 6: Mean comprehension question accuracy as a function of the semantic overlap and semantic interference factors (Exp. 3). The error bars are standard errors.

for versions B and C. Llama-3.1-8B and Mistral-7B show the predicted pattern.

Comprehension Question Accuracy. Figure 4 shows the average accuracies. The predictions parallel those for surprisal: the highest average accuracy is expected for version A, the lowest for version D, and intermediate values for versions B and C. Only GPT-2-XL shows this pattern. That said, all of the other models correctly order versions A and D.

Experiment 3

Experiment 3 investigated whether two semantic factors, the semantic overlap between the anaphor and antecedent and the semantic interference from non-antecedent distractors, affect resolution time and accuracy. Except where otherwise noted, the details are the same as previous experiments.

Design and Materials

The materials were those of Experiment 2 of Varma and Janssen (2019): 19 texts, each occurring in four versions formed by orthogonally varying the semantic overlap between the antecedent and anaphor (high, low) and the semantic interference caused by a non-antecedent distractor (low, high). The anaphors were categorical (e.g., *metal furniture*). Following prior work (Corbett, 1984; Garrod & Sanford, 1977; Varma & Janssen, 2019), high semantic overlap was operationalized by antecedents that were typical members of the category (e.g., *chair*) and low semantic overlap by antecedents that were atypical (e.g., *stool*). Low semantic interference was operationalized by non-antecedent distractors that are atypical members and high semantic interference by non-antecedent distractors that are typical. Thus, the four versions were:

A. typical antecedent, atypical distractor
B. typical antecedent, typical distractor
C. atypical antecedent, atypical distractor
D. atypical antecedent, typical distractor

The base texts were the version D (far spatial distance, long temporal duration) texts from Experiment 2. The four versions of each text shared the same categorical anaphor. Only the antecedent and the non-antecedent distractors were varied.

Results and Discussion

Surprisal Prediction of Reading Times. The average surprisal of the five models on the four text versions is shown in Figure 5. The prediction is that the anaphor resolution is fastest for version A, slowest for D, and intermediate for versions B and C. None of the models show the predicted pattern, although GPT-2-XL correctly orders versions A and D.

Comprehension Question Accuracy. The average accuracy of the five models on the four text versions is shown in Figure 6. The predictions mirror those for surprisal on the anaphor, with the highest average accuracy expected for version A, the lowest for version D, and intermediate values for versions B and C. Only Llama-3.1-8B shows the predicted pattern, although its overall accuracies are quite low.

General Discussion

Cognitive science studies have identified factors that affect the speed and success of anaphor resolution. This study examined whether the same factors affect anaphor resolution in LLMs. Positive results would constitute evidence that LLMs can serve as cognitive models of human anaphor resolution, and perhaps of language understanding more generally.

Experiment 1 investigated the effects of (1) antecedent topicality and (2) sentential distance. The remaining experiments focused on the reader's situation model of the text. Experiment 2 investigated the contextual effects of (3) spatial distance and (4) temporal duration; Experiment 3 investigated the semantic effects of (5) antecedent similarity and (6) interference from non-antecedent distractors. Two indices of model performance, surprisal on the anaphor and accuracy on a comprehension question about the antecedent, were compared with the corresponding human measures of reading time and accuracy. The models considered were GPT-2-XL, Llama-3.1-8B,

Factor	Measures: Surprisal	Measures: Accuracy
Topicality (Exp. 1)	All models	GPT2-XL
Sentential distance (Exp. 1)	All models	GPT2-XL (others partial)
Spatial distance (Exp. 2)	All models except Pythia-12B	None/weak effects
Temporal duration (Exp. 2)	LLaMa-3.1-8B Mistral-7B	None/weak effects
Antecedent similarity (Exp. 3)	GPT2-XL	LLaMa-3.1-8B (low overall accuracy)
Semantic interference (Exp. 3)	GPT2-XL	None/weak effects

Table 1: Summary of where LLMs show human-like directional patterns in anaphor resolution across discourse, situation-model, and semantic factors.

Pythia, Mistral-24B, and Mistral-7B. To varying degrees, the models showed the six effects documented in the literature, whether in their surprisal values or comprehension accuracies. The model with the greatest overall alignment to human anaphor resolution was perhaps Mistral-7B, although it failed to account for the effects of (5) antecedent similarity and (6) interference in Experiment 3. GPT-2-XL achieved surprisingly respectable overall alignment given that it is the oldest and smallest of the models.

A notable pattern across experiments is that LLMs more reliably exhibited human-like sensitivity to discourse prominence and distance-based factors than to semantic similarity and interference. This asymmetry is informative. Effects of sentential, spatial, and temporal distance reflect graded accessibility, which may be approximated by recency biases and attention dynamics in sequential language models. In contrast, semantic interference effects require structured competition among partially overlapping representations during memory retrieval, a process central to cognitive theories of anaphor resolution but not explicitly implemented in current LLM architectures. From this perspective, partial alignment does not undermine the present findings; rather, it helps localize where LLMs diverge from human discourse processing.

Thus, we have some evidence of cognitive alignment between LLMs, especially Mistral-7B and GPT-2-XL, and human performance, suggesting their potential utility as cognitive science models. Still, work remains to be done. For example, although the profiles of these models' comprehension question accuracies across passage versions were generally human-like, their absolute performance was quite low.

The largest limitation of the current study is the lack of items (i.e., texts), which precluded running statistical analyses to better establish the informal trends present in the model results. Here, we were limited by the relatively small size of cognitive science studies relative to ML studies and the general unavailability of materials and human datasets for classic studies of anaphor resolution. It would have been easy to generate multiple 'participants' by increasing the temperature parameter, running each model multiple times, and computing statistics across these samples. However, we question the validity of this approach. There is no reason to believe that such simulated "participants" vary in the same ways that humans do: in their working memory capacity, executive function, domain knowledge, reading skill, etc. For this reason, we limited ourselves to providing descriptive, summary statistics and cautiously interpreting the observed trends. We also recognize the limitations of our LLM-as-a-judge technique for scoring comprehension accuracy. Although it provides a reasonable heuristic, the binary scoring and reliance on an automatic judge introduce known biases and can be different from human judgments. It is best to treat the reported accuracies as approximations.

One goal for future research is to examine whether other factors that affect anaphor resolution performance in humans also affect LLMs. Some of these are relatively subtle, such as whether an antecedent and anaphor belong to the same event within a narrative text or whether they are in different events separated by a boundary (Thompson & Radvansky, 2016). Whether or not LLMs are sensitive to such factors is an important question for cognitive scientists evaluating their potential as models of human language understanding. Examination of a broad range of factors known to affect human anaphor resolution might also guide NLP researchers as they design the new benchmarks for measuring the coreference resolution abilities of LLMs (Gan et al., 2024; Manikantan et al., 2024).

References

Anderson, A., Garrod, S. C., & Sanford, A. J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. *The Quarterly Journal of Experimental Psychology A*, *35*, 427–440.

BabyLM Organizers. (2025, November). Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In L. Charpentier, L. Choshen, R. Cotterell, M. O. Gul, M. Y. Hu, J. Liu, J. Jumelet, T. Linzen, A. Mueller, C. Ross, R. S. Shah, A. Warstadt, E. G. Wilcox, & A. Williams (Eds.), *Proceedings of the first babylm workshop* (pp. 399–420). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.babylm-main.28

Cambria, E. (2025). Semantics processing. Springer.

Clark, H. H., & Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Memory & Cognition*, *7*, 35–41. https://doi.org/10.3758/BF03196932

Corbett, A. T. (1984). Prenominal adjectives and the disambiguation of anaphoric nouns. *Journal of Verbal Learning and Verbal Behavior*, *23*, 683–695.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Gan, Y., Poesio, M., & Yu, J. (2024, May). Assessing the capabilities of large language models in coreference: An evaluation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 1645–1665). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.145/

Garrod, S., & Sanford, A. J. (1977). Interpreting anaphoric relations: The integration of semantic information while reading. *Journal of Verbal Learning and Verbal Behavior*, *16*, 77–90.

Grattafiori, A., et al. (2024). The llama 3 herd of models. https://doi.org/10.48550/arXiv.2407.21783

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Proceedings of the 2nd Meeting of NAACL*.

Hu, J., Chen, S. Y., & Levy, R. (2020). A closer look at the performance of neural language models on reflexive anaphor licensing. *Proceedings of the Society for Computation in Linguistics 2020*, 323–333. https://aclanthology.org/2020.scil-1.39/

Ivanova, A. A. (2025). How to evaluate the cognitive abilities of llms. *Nature Human Behavior*.

Lee, S.-H., & Schuster, S. (2022). Can language models capture syntactic associations without surface cues? a case study of reflexive anaphor licensing in English control constructions. *Proceedings of the Society for Computation in Linguistics 2022*, 206–211. https://aclanthology.org/2022.scil-1.18/

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Li, A., Cai, T., Feng, X., Narang, S., Peng, A., Shah, R. S., & Varma, S. (2024). Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention. *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, 6069–6076.

Liu, X., Deng, S., Wang, M., Dong, Z., Dai, L., Li, J., & Nong, R. (2025). Enhancing coreference resolution with pretrained language models: Bridging the gap between syntax and semantics. https://doi.org/10.48550/ARXIV.2504.05855

Manikantan, K., Tapaswi, M., Gandhi, V., & Toshniwal, S. (2024). Identifyme: A challenging long-context mention resolution benchmark.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension [In B. Ross (Ed.), The psychology of learning and motivation].

Mistral AI Team. (2025, January). Mistral small 3 [Accessed: 2026-01-30]. https://mistral.ai/news/mistral-small-3

Morrow, D. G., Greenspan, S. L., & Bower, G. H. (1987). Accessibility and situation models in narrative comprehension. *Journal of Memory and Language*, *26*, 165–187.

Novák, M., et al. (2025). Findings of the fourth shared task on multilingual coreference resolution. *Proceedings of the CRAC Shared Task*. https://doi.org/10.18653/v1/2025.crac-1.9

O'Brien, E. J. (1987). Antecedent search processes and the structure of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 278–290.

OpenAI. (2025, August). Introducing gpt-5 [Accessed: 2026-01-30]. https://openai.com/index/introducing-gpt-5/

Pandit, O., & Hou, Y. (2021). Probing for bridging inference in transformer language models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4153–4163. https://doi.org/10.18653/v1/2021.naacl-main.327

Piantadosi, S. T., et al. (2024). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *28*, 844–856.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rinck, M., & Bower, G. H. (1995). Anaphora resolution and the focus of attention in situation models. *Journal of Memory and Language*, *34*, 110–131.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects and natural categories. *Cognitive Psychology*, *9*, 382–440.

Shah, R. S., & Varma, S. (2025). The potential and the pitfalls of using pre-trained language models as cognitive science theories. https://arxiv.org/abs/2501.12651

Talukdar, C., & Rahman, M. (2025). Coreference resolution in machine learning: A survey. *2025 IEEE Guwahati Subsection Conference (GCON)*. https://doi.org/10.1109/GCON65540.2025.11173320

Thompson, A. N., & Radvansky, G. A. (2016). Event boundaries and anaphoric reference. *Psychonomic Bulletin & Review*, *23*, 849–856.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

Varma, S., & Janssen, A. (2019). The structure of situation models as revealed by anaphor resolution. *Language Sciences*, *72*, 104–115.

Wang, Y., et al. (2024). Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. https://doi.org/10.48550/arXiv.2406.01574

Wilcox, E., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713.

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1196–1207.