2025. Proc Ling Soc Amer 10(1). 5902. https://doi.org/10.3765/plsa.v10i1.5902.

Missing the cues: LLMs' insensitivity to semantic biases in relative clause attachment

Russell Scheinberg, So Young Lee, & Ameeta Agrawal*

Abstract. We investigate whether large language models (LLMs) replicate English speakers' well-established preference for low attachment in relative clause (RC) ambiguities, and how they respond when semantic cues such as world knowledge and stereotypical associations (e.g., age or gender plausibility) conflict with this preference. Eight commercial LLMs spanning the Claude-3/3.5 and GPT-3.5/4o families were evaluated using structurally and semantically ambiguous stimuli, alongside items that introduced plausibility-based biases toward either high or low attachment. In the absence of disambiguating cues, all models showed a strong preference for low attachment, consistent with human tendencies in ambiguous contexts (i.e. no semantic bias cues). However, models varied in their sensitivity to semantic information: newer Claude-3.5 models frequently shifted toward high attachment when the LA interpretation was implausible, while GPT-based models rarely did so. Attachment preferences were also affected by prompt format, suggesting that LLMs do not consistently integrate syntactic and semantic information in a stable, human-like way. These findings highlight both convergence and divergence between LLMs and human sentence processing, offering insight into the limits of current pretraining paradigms in handling structural ambiguity and world knowledge.

Keywords. LLMs; Relative clause attachment ambiguity; Semantic bias; Prompt sensitivity

1. Introduction. In 2025, large language models (LLMs) process and generate language at 'superhuman' levels of fluency (Tedeschi et al. 2023). But just how *human* is their language? After all, LLMs have fundamentally different acquisition and processing mechanisms. Moreover, perhaps precisely *because* of the task-oriented performance metrics that drive commercial LLM development to excel in the 'benchmark race', developers may ironically overlook many aspects of the actual language that these models produce (Guo et al. 2024; Banerjee et al. 2024). This raises the question of whether the superfluency exhibited by modern LLMs reflects a convergence with human language processing patterns or a divergence from them. To begin addressing this question, we focus on a core feature of human language, ambiguity, which is often overlooked in LLM training (Liu et al. 2023). Specifically, in this paper, we investigate how LLMs resolve syntactic ambiguity and whether their resolution strategies align with human cognitive processing. We use relative clause (RC) attachment ambiguity as our test case.

When a RC follows a complex noun phrase ('DP1 of DP2') as in (1), RC attachment ambiguities occur.

(1) The teacher saw the classmate$_{DP1}$ of the toddler$_{DP2}$ who was teething.

a. the classmate$_{DP1}$ was teething (high attachment)

b. the toddler$_{DP2}$ was teething (low attachment)

* Authors: Russell Scheinberg, Portland State University (rschein2@pdx.edu), So Young Lee, Miami University(soyoung.lee@miamioh.edu), & Ameeta Agrawal, Portland State University (ameeta@pdx.edu)

In English, we call attachment to DP1 'high attachment' (HA) and attachment to DP2 'low attachment' (LA), reflecting their positions in the syntax tree (Figure 1).

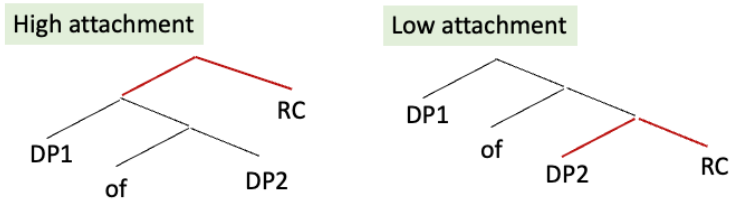

Figure 1. Syntactic Structures of DP1 Modification (left) and DP2 Modification (right) in English

One well-known finding is that English speakers typically prefer to attach the RC to DP2 (LA). This robust LA preference is often explained by the theories based on locality, such as a "recency" or "late closure" effect, which favors linking a new modifier to the nearest eligible noun (MacDonald et al. 1994; Carreiras & Clifton Jr 1993; Hemforth et al. 2015).

Yet this preference is not absolute. The tendnency to LA can be overridden by various prosodic, pragmatic, or semantic factors (e.g., MacDonald et al. 1994; Gilboy et al. 1995; Acuna-Farina et al. 2009; Hemforth et al. 2015). In particular, world knowledge can shift interpretation if LA results in an implausible meaning. For example, in (2), it is more plausible for a baby to be teething than a grandmother, so semantic considerations favor HA even though the locality preference would otherwise predict LA.

(2) The teacher saw the baby$_{DP1}$ of the grandmother$_{DP2}$ who was teething.

a. the baby$_{DP1}$ was teething (HA)

b. the grandmother$_{DP2}$ was teething (LA)

Thus, even if syntactic locality would favor a LA interpretation, semantic cues and world knowledge can override that syntactic attachment preferences.

How, then, do LLMs handle such ambiguities? While human language processing is shaped by the dynamic integration of cognitive constraints, social interaction, and embodied experience, LLMs are trained through self-supervised learning on massive text corpora, where they derive patterns from billions of words without explicitly modeling syntax, semantics, or real-world knowledge as distinct components. Instead, these elements become implicitly fused within the high-dimensional parameter space of the model, encoded in the learned weights of the neural network. Because LLMs lack explicit modular representations of linguistic structures and meaning, it remains unclear to what extent they replicate human-like processing and utilize structural and semantic information. Thus, we aim to explore how LLMs resolve RC ambiguities, with particular attention to their ability to leverage semantic information.

2. Background.

2.1. RC Attachment Preferences in Humans and Processing Theories. RC attachment has been a productive domain for investigating structural preferences in sentence comprehension. In constructions where a RC follows a complex noun phrase with two potential attachment sites (e.g., *the servant of the actress who*...), both HA and LA interpretations are grammat-

ically possible. This ambiguity provides a testing ground for identifying the factors that guide syntactic interpretation, including parsing strategies, memory constraints, and distributional patterns in the input.

Early research on English, including Frazier (1979), reported a robust preference for LA in RC attachment. This pattern was attributed to the principle of Recency or Late Closure, a syntactic parsing strategy proposed within the garden-path model of sentence processing. According to this principle, comprehenders tend to attach new material to the most recently processed constituent, thereby minimizing the number of syntactic nodes actively held in memory. This strategy was argued to reduce working memory demands and facilitate incremental parsing (Frazier 1979; Rayner et al. 1983). The LA preference has been consistently replicated in English across a range of experimental paradigms, including self-paced reading, eye-tracking, and offline interpretation tasks (e.g., MacDonald et al. 1994; Rayner et al. 1983).

While Recency or Late Closure accounts for the LA preference observed in English, its generalizability across languages has been called into question. In an early study on Spanish, Cuetos & Mitchell (1988) found a preference for *high* attachment, despite the surface word order of the ambiguous construction being identical to that of English. Spanish comprehenders tended to attach the RC to the first noun phrase (DP1) rather than to the more recent DP2. Subsequent cross-linguistic studies have revealed systematic variation in attachment preferences: languages such as English, Dutch, and Chinese generally favor LA, whereas others–including Spanish, French, and Korean–tend to favor HA. Still other languages, such as German and Portuguese, exhibit more variable or mixed patterns depending on structural or contextual factors (Grillo & Costa 2014; Lee et al. 2024). These findings suggest that structural ambiguity resolution may not be fully explained by Recency or Late Closure parsing strategies alone.

Another attempt to explain cross-linguistic variation in attachment preferences is the Tuning Hypothesis, which attributes these differences to distributional patterns in the input language (Mitchell et al. 1990, 1992). According to this account, comprehenders develop parsing preferences based on the frequency with which certain structures occur in their linguistic environment. For example, in languages where high-attaching RCs are more frequent, speakers are expected to show a corresponding preference for HA (Gibson et al. 1996; Desmet et al. 2006). However, frequency-based accounts alone cannot fully account for the observed variability. A growing body of research has shown that attachment preferences are also modulated by prosodic cues (Bergmann et al. 2008), semantic and pragmatic relationships (Gilboy et al. 1995; MacDonald et al. 1994), and lexical biases. These findings point to a more complex, multifactorial architecture for structural interpretation.

Among the additional factors that modulate attachment preferences, lexical and semantic information play a particularly important role. Numerous studies have shown that comprehenders are sensitive not only to syntactic structure and frequency, but also to the plausibility of specific noun-verb combinations and the thematic roles they imply. One crucial phenomenon illustrating this sensitivity is often referred to as semantic override: if an initially preferred syntactic parse results in an implausible interpretation (e.g., a baby driving a car), comprehenders typically revise the structure to yield a more plausible reading (Trueswell et al. 1994; MacDonald et al. 1994). In such cases, real-world knowledge and lexical semantics can override a default syntactic preference.

2.2. LLMs AND SYNTACTIC AMBIGUITY. Despite their impressive abilities in natural language understanding and generation, LLMs often display systematic gaps and divergences from human-like processing when it comes to resolving syntactic ambiguities.

For example, several recent studies have reported a consistent LA bias in LLMs' handling of RC ambiguities, even for languages where human comprehenders show a HA preference (Zhou et al. 2024; Cai et al. 2024; Lee et al. 2024). This suggests that LLMs' internalized parsing strategies – emergent from next-token prediction over massive unannotated text corpora – may fail to fully capture the mixture of structural, semantic, and pragmatic constraints that shape human sentence parsing.

Beyond attachment biases, LLMs appear less adept at performing semantic override. That is, when an otherwise default parse yields a semantically implausible interpretation (e.g., a baby driving a car), humans typically revise their parsing commitments, while LLMs more frequently persist with the preferred syntactic attachment despite strong world-knowledge signals (Amouyal et al. 2025; Lee et al. 2025).

While these initial findings have highlighted important divergences, our study contributes to this line of research by deeply investigating the conditions under which LLMs align or fail to align with human parsing behavior.

2.3. RESEARCH QUESTIONS. The evidence reviewed above highlights the complex interplay of syntactic, semantic, and lexical factors in human attachment preferences. However, it remains an open question whether LLMs – which are trained on vast linguistic input but lack explicit syntactic representations or grounded world knowledge – exhibit similar patterns. To investigate this, we examine how modern LLMs process English RC attachment by addressing two questions:

1. Do LLMs replicate the robust **LA** preference observed in English speakers?
2. When faced with implausible readings, do LLMs *override* their default attachment preference to preserve semantic coherence?

Section 3 details our experimental design, including model selection, stimulus construction, and evaluation metrics. Section 4 presents our findings, and Section 5 explores their implications for how LLMs handle syntactic and semantic information.

3. Experiments.

3.1. LANGUAGE MODELS. In our experiment, we evaluated a total of eight language models from two major developers, Anthropic and OpenAI. From Anthropic, we included five models: the three Claude-3 variants – Haiku, Sonnet, and Opus – released in early 2024, as well as the more recent Claude-3.5 versions of Haiku and Sonnet. From OpenAI, we tested three models: GPT-4o and its lightweight counterpart GPT-mini-4o, both released in 2024, and the final version of GPT-3.5-Turbo, released in early 2024. The properties of the models are summarized in Table 1.

This model selection allows us to examine two key dimensions of model variation: version updates and model size. Including both older and newer versions within a family (e.g., Claude-3 vs. Claude-3.5) enables us to assess whether recent advancements – often aimed at improving reasoning, efficiency, and generalization – also lead to more human-like sentence processing. Comparing smaller and larger models (e.g., GPT-mini-4o vs. GPT-4o; Haiku vs. Sonnet) further

allows us to investigate whether differences in model size influence syntactic ambiguity resolution. Together, this sampling strategy provides a principled basis for evaluating how architectural improvements and scale impact RC attachment preferences across commercially deployed LLMs.

Model ID	**Short Name**	**2024 Release**	**Size Category***
claude-3-haiku-20240307	Haiku-3	March 7	Medium
claude-3-opus-20240229	Opus-3	February 29	Very Large
claude-3-sonnet-20240229	Sonnet-3	February 29	Large
claude-3-5-haiku-20241022	Haiku-3.5	October 22	Medium
claude-3-5-sonnet-20241022	Sonnet-3.5	October 22	Large
gpt-3.5-turbo-0125	GPT-3.5	January 25	Large
gpt-4o-mini-2024-07-18	GPT-mini-4o	July 18	Medium
gpt-4o-2024-11-20	GPT-4o	November 20	Extremely Large

Table 1. Language Models Tested in This Study. *Size categories reflect relative model scale based on publicly available information and comparative benchmarks. Exact parameter counts are not disclosed for most commercial models.

3.2. STIMULI. Our stimuli manipulated one key factor, semantic congruency, and balanced across syntactic position. Table 2 shows a representative set of our stimuli.

Pos.	Bias	Example Sentence
Subj	Ambiguous	The sister $_{DP1}$ of the bride $_{DP2}$ [who became pregnant] wore a sweater.
	DP1	The bride $_{DP1}$ of the groom $_{DP2}$ [who became pregnant] wore a sweater.
	DP2	The groom $_{DP1}$ of the bride $_{DP2}$ [who became pregnant] wore a sweater.
Obj	Ambiguous	The client missed the sister $_{DP1}$ of the bride $_{DP2}$ [who became pregnant]
	DP1	The client missed the bride $_{DP1}$ of the groom $_{DP2}$ [who became pregnant]
	DP2	The client missed the groom $_{DP1}$ of the bride $_{DP2}$ [who became pregnant]

Table 2. Example sentences for each condition. DPs and RCs in the same color are semantically congruent, illustrating how bias is created through the interaction between DPs and RC.

First, to examine the default RC attachment preferences in LLMs, and whether world knowledge can override these structural preferences, we manipulated semantic congruency with three levels: i) ambiguous, where the RC was equally plausible for both DPs, ii) DP1-biased, where the RC was semantically plausible only for the first noun phrase (DP1), and iii) DP2-biased, where the RC applied only to the second noun phrase (DP2).

Our bias conditions leveraged real-world knowledge that would make a given RC more appropriate for one DP than the other (e.g., *the grandmother of the baby who was getting a driver's license*). Table 3 summarizes the bias categories used. While such biases are methodologically useful, we recognize the potential ethical implications of this in language research [1].

[1] Bias in language models is a major ethical concern, and "de-biasing" techniques aim to mitigate problematic associations (Gallegos et al. 2024). Our use of semantic biases here serves to examine how well LLMs integrate real-world knowledge into syntactic decisions. Nonetheless, we acknowledge the potential for inadvertently reinforcing stereotypes and have striven to select items that minimize harmful biases.

Bias Category	Description	Example
Age (n=43)	RC describes an age-specific activity.	*the grandmother of the baby who was getting a driver's license*
Role (n=34)	RC action matches a specific role/authority.	*the teacher of the student who flunked the whole class*
Gender (n=38)	RC describes trait associated with one gender.	*the uncle of the woman who was pregnant*
Logical Contradiction (n=10)	RC contradicts the properties of one DP.	*the doctor of the insomniac who slept 10 hours every day*

Table 3. Bias categories affecting RC attachment.

To introduce structural variation in the stimuli, we varied the position of the complex noun phrase (DP1 of DP2) across subject and object contexts. However, following Hemforth et al. (2015), who found no effect of syntactic position on attachment preferences in human comprehenders in English, we did not include position as a factor in our analysis.

3.3. PROCEDURE. We used direct prompting to test how models resolve RC attachment – whether to DP1 or DP2. [2] Our prompt in Figure 2 contained two questions sent in a single interaction: first, asking the model to identify the RC; and second, instructing the model to name the modified noun (DP1 or DP2) with a single-word answer to facilitate automated response parsing. The first question served as a validity check to ensure the model could identify the RC accurately, while the second question elicited the attachment decision. This two-step approach allowed us to distinguish between **misidentification** of the RC (a parsing error) and a **genuine** attachment decision.

Example Prompt

```
Read the sentence, then:
1) identify the relative clause in the sentence, and
2) with one word, which noun does the relative clause modify?
Answer without commentary.

"The child of the schoolteacher who was learning to count wore a hat."
```

Figure 2. Example prompt with instructions to identify the RC and its referent.

The prompt format was the same for all 750 total sentences (= 125 items $\times$ 6 conditions). We further discuss an alternative prompt design and its impact on model behavior in Section 5.3.

3.4. ANALYSIS. For every sentence, the model was instructed to (i) identify the RC and (ii) indicate which noun (DP1 or DP2) it modified. We removed eight responses where the RC was incorrectly identified (five included extra or missing words, three assigned the RC to an entirely

[2] Unlike human experiments where subjects respond to multiple questions, each LLM query contained only a single stimulus, eliminating priming or order effects.

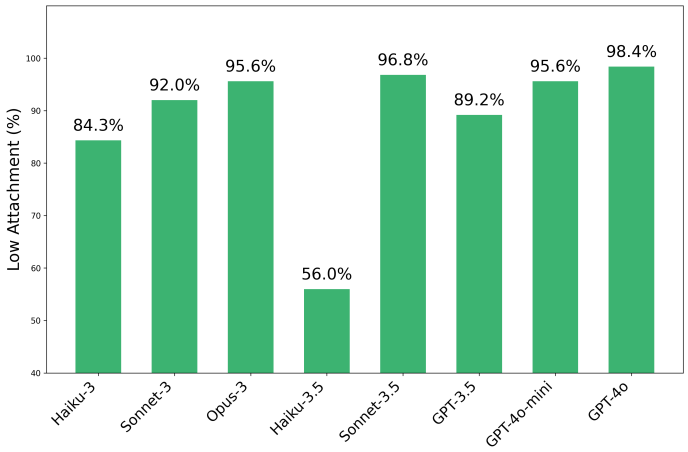

Figure 3. LA Rate in Ambiguous Condition by Model

different noun). The remaining responses were coded according to whether the RC attached high (DP1) or low (DP2).

All analyses used logistic mixed-effects models implemented in the `lme4` package in R (Bates et al. 2015), with the `bobyqa` optimizer to ensure stable convergence. Post-hoc pairwise comparisons were performed via the `emmeans` package, using Tukey corrections for multiple comparisons. For **ambiguous items**, we coded each response as 1 for HA and 0 for LA, then fit a model of the form $outcome \sim model + (1|set)$. For **biased items**, we manipulated two factors: (i) intended bias (`bias`: DP1 or DP2) and (ii) the language model (`model`). Each response was coded as 1 if the attachment matched the intended bias, and 0 otherwise. Our model was: $\texttt{response} \sim \texttt{bias} \times \texttt{model} + (1|\texttt{set})$.

Finally, we conducted a further investigation to determine whether any of the four semantic bias types (Age, Role, Gender, Logical Contradiction) influenced attachment. We again coded HA as 1 and LA as 0, and specified $\texttt{response} \sim \texttt{bias\_type} \times \texttt{model} + (1|\texttt{set})$. Some stimuli belonged to multiple bias categories: for example, a boy is arguably unlikely to be a midwife because of typical age, role and gender associated with that profession, so these results should be regarded as exploratory.

4. Results. This section presents our findings on RC attachment in LLMs, focusing on three core questions: (i) the extent to which each model exhibits a LA preference in structurally ambiguous sentences, (ii) whether the LA preference exhibited by the models is overridden in the presence of semantic implausibility, and (iii) which types of semantic or lexical biases are most likely to influence attachment decisions.

4.1. AMBIGUOUS CASES: LA PREFERENCE. Figure 3 shows the rate of LA responses by each model when resolving syntactically ambiguous RCs. GPT-4o exhibits the strongest LA preference (98.4%), followed by GPT-4o-mini (95.6%) and GPT-3.5 (89.2%), indicating a consistent trend across GPT variants. Similarly, Sonnet-3 (92.0%), Opus-3 (95.6%), and Sonnet-3.5 (96.8%) display strong preferences for LA. Haiku-3 shows a moderately lower LA rate (84.3%), while Haiku-3.5 diverges sharply from this pattern, with a substantially reduced LA rate of 56.0%. Overall, with the exception of Haiku-3.5, all models exhibit a pronounced LA preference, suggesting a general tendency to attach RCs to the most recent noun in the absence of semantic or

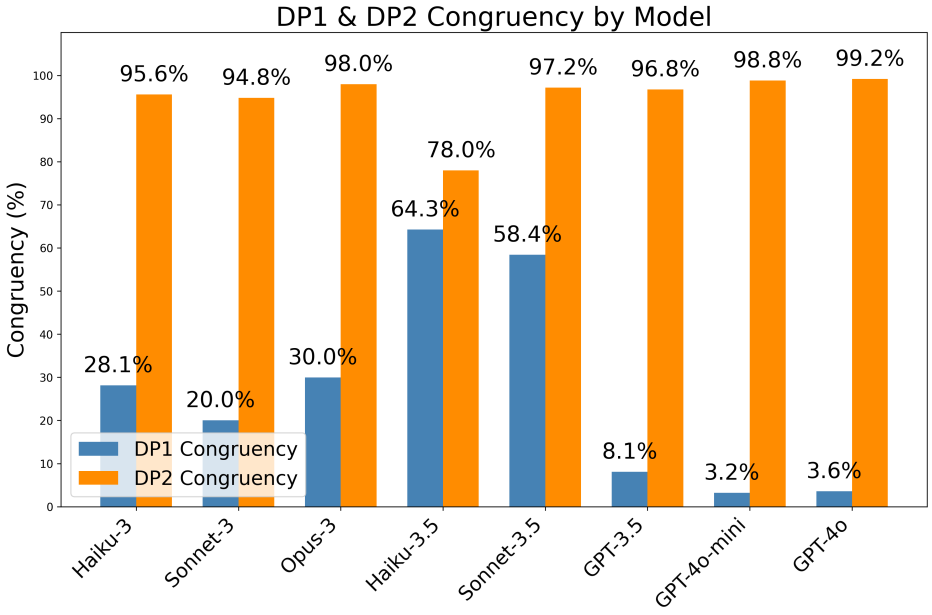

Figure 4. DP1 and DP2 congruency

structural cues.

The pairwise comparisons across models with Tukey adjustments reinforce these findings. Specifically, we observe that Haiku-3.5 diverges significantly from all other models ($p < 0.0001$), and that Haiku-3 diverges significantly from two models, Sonnet-3.5 and GPT-4o ($p < 0.0001$). This suggests that Haiku models follow distinct parsing trajectories. Interestingly, these are the 'smallest' of Anthropic's Claude family of models; they may favor higher attachment preferences potentially due to different training regimes.

4.2. Biased Cases: Semantic Override Effects. Next, we evaluated how models responded when plausibility cues favored a specific attachment site. In DP1-biased items, semantic coherence requires HA, whereas in DP2-biased items, LA is preferred. Figure 4 presents each model's congruency rate, i.e., the percentage of responses that aligned with the semantic bias.

A striking pattern emerges in the DP2-biased condition: nearly all models show extremely high congruency, with rates exceeding 94%, reflecting a strong preference for attaching the RC to the nearest noun. This trend is consistent with a default LA preference. Haiku-3.5 is the primary exception, with a notably lower DP2 congruency rate (78%). Pairwise comparisons confirmed statistically significant differences between Haiku-3.5 and all other models in this condition ($p < 0.0001$).

Turning to the DP1-biased items, Haiku-3.5 again stands out, showing the highest DP1 congruency rate (64.3%), followed by Sonnet-3.5 (58.4%). These two models are the newest members of the Claude family. The earlier Claude models – Haiku-3, Sonnet-3, and Opus-3 – show moderate DP1 congruency rates (ranging from 20-30%). In contrast, all GPT models (GPT-3.5, GPT-4o-mini, and GPT-4o) exhibit strikingly low DP1 congruency rates (3.2-8.1%), indicating that they rarely override their default LA preference even when semantic plausibility strongly favors HA. These differences may reflect distinct training regimes and architectural design choices across model families from different developers.

DP1 Congruency Rate (HA responses)				
Model	Age	Role	Gender	Contradiction
Haiku-3	36.0%	19.1%	27.6%	26.3%
Opus-3	31.0%	22.1%	38.7%	20.0%
Sonnet-3	26.7%	8.8%	26.3%	5.0%
Haiku-3.5	76.7%	56.7%	53.9%	75.0%
Sonnet-3.5	68.6%	45.6%	64.5%	35.0%
GPT-3.5	7.1%	6.0%	12.0%	5.0%
GPT-mini-4o	0.0%	8.8%	0.0%	10.0%
GPT-4o	3.5%	4.4%	3.9%	0.0%
DP2 Congruency Rate (LA responses)				
Model	Age	Role	Gender	Contradiction
Haiku-3	83.3%	85.3%	84.6%	79.3%
Opus-3	87.9%	89.7%	86.3%	88.3%
Sonnet-3	89.1%	89.7%	88.2%	88.1%
Haiku-3.5	51.9%	55.7%	65.8%	45.0%
Sonnet-3.5	75.6%	83.3%	77.6%	78.3%
GPT-3.5	94.1%	93.1%	90.3%	93.3%
GPT-mini-4o	98.1%	95.6%	97.8%	95.0%
GPT-4o	96.9%	98.0%	98.7%	100.0%

Table 4. Breakdown by bias type of DP1-congruent and DP2-congruent items. DP1 is computed using HA rates, while DP2 and ambiguous are computed using LA rates.

4.3. INFLUENCE OF BIAS TYPE. Although not the primary focus of this study, we conducted an exploratory analysis to examine how different types of semantic bias influenced attachment preferences across models. Bias types were categorized into four domains – Age, Role, Gender, and Logical Contradiction – based on the plausibility cues embedded in the stimuli (see Table 3). Table 4 reports DP1 and DP2 congruency rates by bias type.

In the DP1-biased condition, Haiku-3.5 and Sonnet-3.5 consistently showed higher HA rates. Haiku-3.5 was significantly more sensitive to **age** cues than four other models, while Sonnet-3.5 differed from three ($p < 0.0001$). For **gender**-biased items, Sonnet-3.5 showed greater sensitivity than GPT-4o, Sonnet-3, and Haiku-3, with Haiku-3.5 differing significantly only from GPT-4o. In the DP2-biased condition, Haiku-3.5 consistently showed lower congruency rates across all categories, though these differences were not statistically significant.

Taken together, these results reveal a consistent pattern: Haiku-3.5 and Sonnet-3.5 are more likely than other models to override LA preferences when age, gender, or role cues support HA, while most other models show limited sensitivity to such semantic biases.

5. Discussion.

5.1. CONVERGENCE AND DIVERGENCE FROM HUMAN PROCESSING PATTERNS. Our findings highlight both convergence with established human parsing preferences and divergence in how LLMs integrate semantic cues.

Convergence emerges with respect to our first research question: whether LLMs replicate the human preference for LA in structurally ambiguous RCs. All eight models tested show a robust LA preference, consistent with the well-documented human tendency toward late closure in En-

glish (Frazier 1979; Rayner et al. 1983). This suggests that large-scale pretraining on naturalistic corpora can yield syntactic behaviors aligned with those observed in English speakers.

However, the *strength* of this LA bias varies markedly across models. GPT-4o, GPT-4o-mini, Opus-3, and Sonnet-3.5 approach near-ceiling LA rates, while Haiku-3.5 stands out as a clear outlier with only 56% LA responses. Haiku-3 and GPT-3.5 fall into an intermediate range. These differences suggest that while LA is the default across architectures, the strength of the preference is model-dependent, potentially reflecting differences in training data, objectives, or inductive biases across the GPT and Claude model families.

In contrast, our second research question – whether models override syntactic defaults when semantic plausibility favors HA – reveals divergence from human behavior. Humans typically revise their initial parse when faced with semantically implausible interpretations (MacDonald et al. 1994; Trueswell et al. 1994). In DP1-biased items, where semantic plausibility clearly favors HA, most models nevertheless overwhelmingly produced LA responses. This is striking given that human comprehenders typically override default syntactic preferences when the resulting interpretation is implausible (MacDonald et al. 1994; Trueswell et al. 1994). The models' tendency to persist with LA responses – even when semantic cues strongly support HA – suggests a limited ability to integrate world knowledge or to revise an initial syntactic analysis. This rigidity is especially pronounced in the GPT-family models, which showed uniformly low congruency rates in DP1-biased conditions (ranging from 3.2-8.1%).

By contrast, some Claude-family models demonstrated greater flexibility. Notably, Haiku-3.5 achieved a DP1 congruency rate of 64.3%, and Sonnet-3.5 reached 58.4%, suggesting a partial ability to override structural preferences when semantic plausibility requires it. While the earlier Claude models (Haiku-3, Sonnet-3, and Opus-3) showed more moderate congruency (20–30%), they still outperformed all GPT models. These results suggest that while some LLMs exhibit human-like parsing flexibility – allowing semantic information to override default syntactic preferences – others remain more rigidly constrained by structural recency, failing to integrate plausibility cues effectively.

5.2. SENSITIVITY TO SPECIFIC SEMANTIC BIAS TYPES. Broadly, our findings suggest that while some LLMs are capable of integrating semantic cues to override structural preferences, their sensitivity varies across different types of semantic information. In particular, newer models in the Claude family (Haiku-3.5 and Sonnet-3.5) showed greater responsiveness to biases grounded in world knowledge, such as age, gender, and role plausibility, compared to other models. However, this sensitivity was not uniform: models differed in the specific semantic dimensions they appeared to track.

Our results suggest that Haiku-3.5 and Sonnet-3.5 may be especially sensitive to *age*-based biases, but differ in their responsiveness to *gender* and *role* cues. These divergences may reflect distinct "specializations" or training emphases in how semantic constraints are implicitly encoded. If such model-specific sensitivities are robust, they should in principle generalize across bias conditions. However, in DP2-biased items, both the semantic bias and the models' baseline syntactic preference favor LA, potentially masking any effect of semantic information. The absence of statistically significant differences in this condition may therefore reflect cue convergence rather than a lack of semantic sensitivity.

It will be interesting to see whether future model versions, which are often trained on expanded and refined datasets and with constantly improved training regimes, will continue to re-

flect this clear differential sensitivity within their patterns of semantic integration.

A/B Prompt from Lee et al. (2025)

```
I saw the daughter of the
woman who bought a dress.
'Who bought a dress?'
1.  the daughter 2.  the woman
```

Two-step Syntactic Prompt

Read the sentence, then: 1) identify the relative clause in the sentence, and 2) with one word, which noun does the relative clause modify? Answer without commentary.
'The child of the schoolteacher who was learning to count wore a hat.'

Figure 5. Prompt in the present study (left) and A/B prompt from (Lee et al. 2025) (right).

5.3. THE ROLE OF PROMPTING. In addition to RC attachment and semantic override, our findings raise a broader question about the role of prompt design in shaping LLM behavior. In psycholinguistics, considerable care is taken to ensure consistent presentation of stimuli, with the goal of isolating specific factors that influence interpretation. Analogously, in LLM-based studies, the framing of a prompt can dramatically alter the behavior of the model, even when the underlying sentence remains unchanged.

A companion study (Lee et al. 2025)[3] highlights this effect. Whereas the present study used a two-step "syntactic prompt", asking models first to identify the RC, then to determine which noun it modifies, Lee et al. (2025) employed a simpler "A/B" prompt that posed a direct question about the referent of the RC (e.g., "Who bought a dress?" with labeled options; see Figure 5).

As shown in Figure 6, prompt framing had a substantial impact on attachment decisions, particularly in semantically biased conditions. For example, Sonnet-3.5 produced HA responses 45% of the time with the syntactic prompt, but 72.6% with the A/B prompt. The effect was even more striking for GPT-4o: HA responses in DP1-biased items jumped from 4.8% (syntactic prompt) to 74.3% (A/B prompt). These results suggest that explicitly contrasting two labeled referents encourages models to focus on semantic plausibility, while a syntactic framing leads them to rely more heavily on structural cues.

What does this prompt-dependent variability tell us about how LLMs process language? From a theoretical perspective, these findings highlight a fundamental difference between human and model-based sentence processing. Human comprehenders integrate syntactic structure and semantic plausibility in a largely automatic and parallel fashion (Frazier 1979; Rayner et al. 1983). That is, the interpretation of an ambiguous sentence typically reflects a coordinated evaluation of both grammatical structure and world knowledge, regardless of how the question is posed. LLMs, by contrast, appear to lack a single, stable parsing mechanism. Instead, their interpretation strategies are highly sensitive to the surface framing of the prompt. When the task is presented in a way that explicitly contrasts referents (e.g., an A/B prompt), models tend to favor semantically coherent interpretations, resulting in increased HA responses in DP1-biased items.

[3] In that study, we tested earlier versions of Sonnet-3.5 (June 2025) and GPT-4o (May 2025) on a comparable set of English RC stimuli. The stimuli were first developed in the present study and then adapted to a multilingual design in Lee et al. (2025). While the exact items slightly differ, the structural patterns are matched, allowing meaningful comparison.

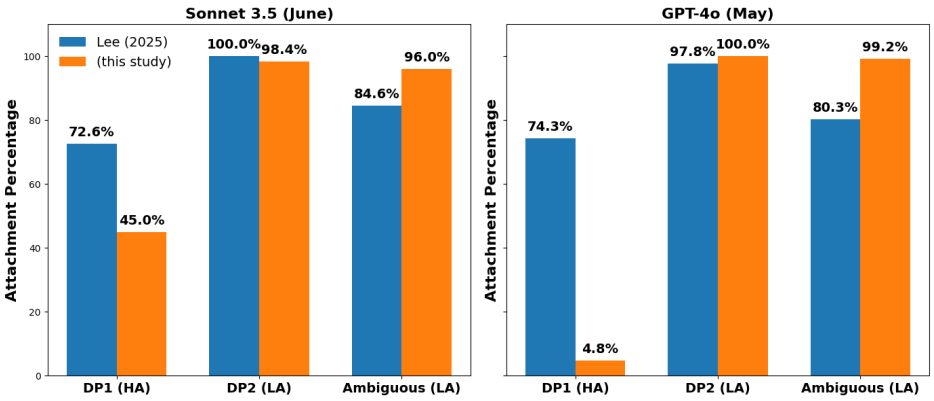

Figure 6. LA Rate in Ambiguous Condition by Model

However, when the prompt emphasizes syntactic analysis (e.g., identifying the RC before selecting the head noun), models are more likely to rely on structural heuristics such as recency, often defaulting to LA even when it leads to implausibility.

This contrast suggests that model behavior is sometimes driven less by an integrated syntactic-semantic parsing process and more by cue-based pattern matching shaped by prompt design. These findings align with prior work showing that LLM outputs can be steered toward syntactic or semantic interpretations depending on task framing (Kojima et al. 2023). While such flexibility may be advantageous in adapting to a variety of tasks, it underscores a key distinction from human comprehension: LLMs do not consistently integrate linguistic and contextual information unless the prompt structure is conducive to it doing so.

Overall, these results show that prompt design is not just a technical detail – it plays a central role in shaping how LLMs interpret sentences. Small changes in how a question is phrased can dramatically shift whether a model relies more on syntactic structure or semantic meaning. For this reason, prompt framing must be treated as an essential part of experimental design when evaluating how closely LLMs approximate human sentence processing.

6. Conclusion. This study examined how LLMs resolve RC attachment ambiguities in English, with a focus on two core questions: whether models replicate the human LA preference, and whether they override this preference when semantic cues favor HA.

Our findings show convergence in that while most models default to LA in structurally ambiguous cases, broadly mirroring the human late closure preference, the strength of this default varies substantially across architectures. We found divergence in that, in the presence of semantic bias towards DP1, GPT-family models exhibited near-ceiling LA rates, although Claude family models, particularly Haiku-3.5 and Sonnet-3.5, showed greater flexibility to override syntactic recency in favor of semantic coherence.

Crucially, our results highlight the central role of prompt design in shaping model behavior. Small changes in task framing – such as shifting from a two-step syntactic prompt to an A/B choice format – produced dramatic differences in attachment patterns. Unlike human comprehenders, whose interpretive strategies are relatively stable across task formats, LLMs exhibit striking prompt sensitivity, suggesting the absence of a unified parsing mechanism and a greater reliance on surface-level cues.

Taken together, these findings underscore both the potential and the limitations of LLMs as

models of human linguistic behavior. While some models approximate human-like structural preferences, their ability to integrate syntactic and semantic information remains uneven and highly context-dependent. From a scientific standpoint, this reinforces the need to treat LLMs as experimental subjects whose interpretive mechanisms must be inferred through carefully controlled behavioral probing (Griffiths et al. 2024; Ku et al. 2025). From an applied perspective, it raises questions about which aspects of human parsing are essential for successful communication, and which may be dispensable in artificial systems.

Although this study focused on a single phenomenon – RC attachment – it illustrates a broader challenge at the intersection of linguistics and AI: identifying which components of human language processing are necessary for modeling, understanding, and interacting with large-scale language technologies. As LLMs continue to evolve, linguistic theory will remain indispensable for both diagnosing model behavior and guiding the design of cognitively and communicatively robust systems.

References.

Acuna-Farina, Carlos, Isabel Fraga, Javier Garcia-Orza & Ana Pineiro and. 2009. Animacy in the adjunction of spanish rcs to complex nps. *European Journal of Cognitive Psychology* 21(8). 1137–1165. doi:10.1080/09541440802622824. https://doi.org/10.1080/09541440802622824.

Amouyal, Samuel Joseph, Aya Meltzer-Asscher & Jonathan Berant. 2025. When the lm misunderstood the human chuckled: Analyzing garden path effects in humans and language models. https://arxiv.org/abs/2502.09307.

Banerjee, Sourav, Ayushi Agarwal & Eishkaran Singh. 2024. The vulnerability of language model benchmarks: Do they accurately reflect true llm performance? *arXiv preprint arXiv:2412.03597* .

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. doi:10.18637/jss.v067.i01.

Bergmann, Anouschka, Meghan Armstrong & Kristine Maday. 2008. Relative clause attachment in english and spanish: A production study. In *Proceedings of speech prosody*, vol. 2008, 507–508.

Cai, Zhenguang G., Xufeng Duan, David A. Haslett, Shuqi Wang & Martin J. Pickering. 2024. Do large language models resemble humans in language use? *arXiv preprint arXiv:2303.08014* .

Carreiras, Manuel & Charles Clifton Jr. 1993. Relative clause interpretation preferences in spanish and english. *Language and Speech* 36(4). 353–372.

Cuetos, Fernando & Don C. Mitchell. 1988. Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in spanish. *Cognition* 30(1). 73–105. doi:https://doi.org/10.1016/0010-0277(88)90004-2. https://www.sciencedirect.com/science/article/pii/0010027788900042.

Desmet, Timothy, Constantijn De Baecke, Denis Drieghe, Marc Brysbaert & Wietske Vonk. 2006. Relative clause attachment in dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes* 21(4). 453–485.

Frazier, Lyn. 1979. *On comprehending sentences: Syntactic parsing strategies.* University of Connecticut.

Gallegos, Isabel O, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang & Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* 50(3). 1097–1179.

Gibson, Edward, Carson T Schütze & Ariel Salomon. 1996. The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research* 25. 59–92.

Gilboy, Elizabeth, Josep-MMaria Sopena, Charles Cliftrn & Lyn Frazier. 1995. Argument structure and association preferences in spanish and english complex nps. *Cognition* 54(2). 131–167. doi:https://doi.org/10.1016/0010-0277(94)00636-Y. https://www.sciencedirect.com/science/article/pii/001002779400636Y.

Griffiths, Thomas L, Jian-Qiao Zhu, Erin Grant & R Thomas McCoy. 2024. Bayes in the age of intelligent machines. *Current Directions in Psychological Science* 33(5). 283–291.

Grillo, Nino & Joao Costa. 2014. A novel argument for the universality of parsing principles. *Cognition* 133(1). 156–187. doi:https://doi.org/10.1016/j.cognition.2014.05.019. https://www.sciencedirect.com/science/article/pii/S0010027714001085.

Guo, Yanzhu, Guokan Shang & Chloe Clavel. 2024. Benchmarking linguistic diversity of large language models. https://arxiv.org/abs/2412.10271.

Hemforth, Barbara, Susana Fernandez, Charles Clifton Jr, Lyn Frazier, Lars Konieczny & Michael Walter. 2015. Relative clause attachment in german, english, spanish and french: Effects of position and length. *Lingua* 166. 43–64.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. https://arxiv.org/abs/2205.11916.

Ku, Alexander, Declan Campbell, Xuechunzi Bai, Jiayi Geng, Ryan Liu, Raja Marjieh, R Thomas McCoy, Andrew Nam, Ilia Sucholutsky, Veniamin Veselovsky et al. 2025. Using the tools of cognitive science to understand large language models at different levels of analysis. *arXiv preprint arXiv:2503.13401* .

Lee, So Young, Russell Scheinberg, Amber Shore & Ameeta Agrawal. 2024. Multilingual relative clause attachment ambiguity resolution in large language models. *arXiv preprint arXiv:2503.02971* .

Lee, So Young, Russell Scheinberg, Amber Shore & Ameeta Agrawal. 2025. Who relies more on world knowledge and bias for syntactic ambiguity resolution: Humans or llms? https://arxiv.org/abs/2503.10838.

Liu, Alisa, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith & Yejin Choi. 2023. We‘re afraid language models aren‘t modeling ambiguity. In Houda Bouamor, Juan Pino & Kalika Bali (eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing*, 790–807. Singapore: Association for Computational Linguistics. doi:10.18653/v1/2023.emnlp-main.51. https://aclanthology.org/2023.emnlp-main.51/.

MacDonald, Maryellen C, Neal J Pearlmutter & Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101(4). 676.

Mitchell, Don C, Fernando Cuetos & Martin MB Corley. 1992. Statistical versus linguistic determinants of parsing bias: Cross-linguistic evidence. In *Fifth annual cuny conference on human sentence processing*, .

Mitchell, Don C, Fernando Cuetos & Daniel Zagar. 1990. Reading in different languages: Is

there a universal mechanism for parsing sentences? .

Rayner, Keith, Marcia Carlson & Lyn Frazier. 1983. The interaction of syntax and semantics during sentence processing: eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior* 22(3). 358–374. doi:https://doi.org/10.1016/S0022-5371(83)90236-0. https://www.sciencedirect.com/science/article/pii/S0022537183902360.

Tedeschi, Simone, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova & Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In Anna Rogers, Jordan Boyd-Graber & Naoaki Okazaki (eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, 12471–12491. Toronto, Canada: Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.697. https://aclanthology.org/2023.acl-long.697/.

Trueswell, J.C., M.K. Tanenhaus & S.M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language* 33(3). 285–318. doi:https://doi.org/10.1006/jmla.1994.1014. https://www.sciencedirect.com/science/article/pii/S0749596X8471014X.

Zhou, Yuchen, Emmy Liu, Graham Neubig, Michael J. Tarr & Leila Wehbe. 2024. Divergences between language models and human brains. *arXiv preprint arXiv:2311.09308* .