# Do LLMs Understand Anaphoric Accessibility?

**Xiaomeng Zhu*** **Zhenghao Zhou*** **Simon Charlow** **Robert Frank**
Department of Linguistics
Yale University
{miranda.zhu, herbert.zhou, simon.charlow, robert.frank}@yale.edu

## 1 Introduction

In natural language, anaphora is influenced by the *accessibility* of antecedents (Karttunen, 1976):

(1) a. *Universal vs. Existential:* {#Every, A} man walked in the park. He dreamed.

b. *Negation:* It isn't true that John {doesn't own, #owns} a cow. He feeds it.

c. *Disjunction:* Either there is {no, #a} bathroom, or it is in a funny place.

In (1a), it is felicitous to use a singular pronoun to refer back to an existential quantifier, but not a universal quantifier, since the latter's discourse referent is not **accessible** outside its **scope** (roughly, its minimal tensed clause). Similarly, it is infelicitous to refer back to entities introduced within the scope of a single negation (1b), unless the negation is in the left clause of a disjunction (1c).

Discourse anaphora offers an intricate landscape for evaluating the semantic knowledge of large language models (LLMs) in a way that goes beyond truth conditions/entailment. Schuster and Linzen (2022) examined sensitivity to the scope of negation: an indefinite interpreted within the scope of negation should not introduce an entity that can be referred to. They found that while LLMs indeed exhibit such sensitivity, their performance is not systematic. Zhu and Frank (2024) extended their paradigm to allow for the evaluation of the semantic properties that govern discourse entity introduction and reference. However, these works only evaluated LLMs on simple sentences that only consider negation as the scope that interacts with discourse entities. This study investigates the extent to which LLMs represent the interactions between scope, logical operators, quantification, and anaphora in human-like ways. We investigate these questions by assessing LLMs' ability to judge the felicity of discourse anaphora in various contexts.

*Equal contribution.

## 2 Methodology

We examined four open-source models in the Llama3 family (1B, 3B, 8B, and 8B instruction-tuned) and two closed-source GPT3 models. To establish a human baseline, we also tested 104 participants on Prolific with a forced-choice task, which aligns with the evaluation metric on models.

Experimental stimuli were generated from a set of structural templates containing the target operators. We manually constructed 32 semantically plausible simple sentence frames with the help of GPT-4o and inserted them in the templates. This yields a set of 9816 experimental sentences.

Models are assessed using surprisal, i.e., the negative log probability assigned to target tokens conditioned on the previous context. Depending on the specific experimental condition, we apply one of three metrics to operationalize predictions from semantic theories. The **difference-of-difference (`DoD`) metric** is used to compare two pairs of sentences when only one is predicted to show a contrast. In this case, we expect that the difference in surprisal between one pair of sentences is greater than that between the other pair. The comparison has the following general form: $p(\text{sen}_a) - p(\text{sen}_{a'}) > p(\text{sen}_b) - p(\text{sen}_{b'})$. The **conditional probabilities of a sentence conditioned on another (`CondProb`)** metric is applied when we compare two sentences with the same context but different continuations: $p(\text{Cont}_1|\text{Context}) > p(\text{Cont}_2|\text{Context})$. The **syntactic log-odds ratio (`SLOR`)** metric (Lau et al., 2017) is applied when two sentences share neither the context nor the continuation. For this case, we compare the SLOR

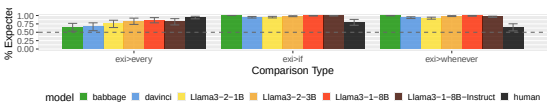


Figure 1: Accuracy results of Experiment 1.

scores of the two sentences. We take the model to accurately encode a comparison when the metrics stand in the expected inequality.

## 3 Experiments

**Universal vs. Existential Quantifiers** The scope of ∃ extends indefinitely in the discourse while the scope of ∀ is limited to its clause. Thus, we compare sentence pairs of the following form:

(2) a. *Intra*: {A (∃), Every (∀)} farmer worked in the field before he dreamed.

b. *Inter*: {A (∃), #Every (∀)} farmer worked in the field. He dreamed.

A `DoD` pattern mirroring human judgments is $p(\forall\text{-Intra}) - p(\forall\text{-Inter}) > p(\exists\text{-Intra}) - p(\exists\text{-Inter})$. Our results are displayed in Figure 1: all models showed above-chance accuracy in this comparison, which provides preliminary evidence that LLMs correctly model how these quantifiers' accessibility patterns differ. Similarly high performance is observed with donkey conditionals.

**Negation** Indefinites under a single negation do not license discourse anaphora:

(3) a. EXI: John owned a cow.

b. DN: It was not the case that John didn't own a cow.

c. NEG: #John didn't own a cow.

d. CONT: (In fact,) It was (just) away on the meadow.

For each licensing context $L$ (EXI or DN) and non-licensing context $N$ (NEG), the LLM should predict $p(\text{CONT}|L) > p(\text{CONT}|N)$.

As shown in the top two panels of Figure 2, all models succeed in preferring the EXI context over NEG, but three of the models struggle to favor DN over NEG. In particular, the two `Llama3-1-8B` models show a preference of NEG over DN, which is the reverse of what is expected. Human results, on the other hand, are high in `Exi>Neg` and exhibit a similar decrease from `Exi>Neg` to `DN>Neg`, but both are reliably above chance. As seen in the bottom panels, adding *in fact* does help to lift the accuracy for the `DN>Neg` comparison, but it also flips the direction of the `Exi>Neg` comparison. This suggests that LLMs rely on lexical cues rather than abstract semantic knowledge of the effects of negation on discourse anaphora. In contrast, human judgments stay above-chance, suggesting that humans prioritize the knowledge of accessibility over shallow lexical cues in making felicity judgments.

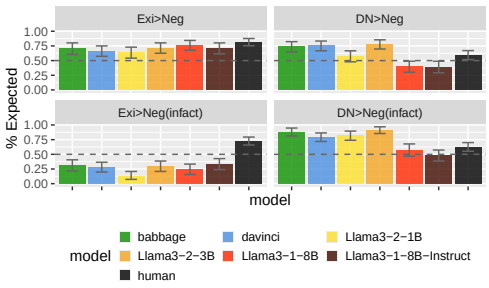


Figure 2: Experiment 2 results on negation.

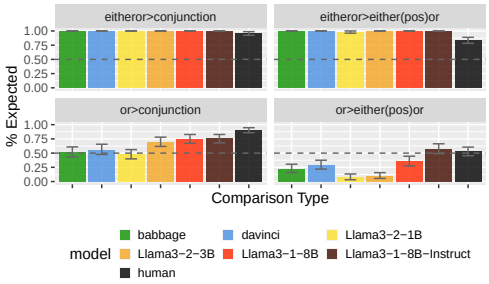


Figure 3: Experiment 3 results on disjunctions.

**Disjunction** Existentials within a first disjunct (with or without *either*) do not license anaphora in the second, while negative quantifiers in a disjunct (but not a conjunct) do (Evans, 1977).

(4) a. *EitherOr*: Either there is no manuscript, or it was hidden by John.

b. *Either(Pos)Or*: #Either there is a manuscript, or it was hidden by John.

c. *Or*: There is no manuscript, or it was hidden by John.

d. *Conjunction*: #There is no manuscript, and it was hidden by John.

We predict $\text{SLOR}(\{4a, 4c\}) > \text{SLOR}(\{4b, 4d\})$.

As seen in Figure 3, all LLMs show predicted inequalities with disjunctions with *either*. However, in disjunctions without *either*, the felicitous sentences are less favored than the infelicitous sentences without negation. This sharp contrast suggests that LLMs rely heavily on lexical cues.

## 4 Conclusion

Our results show that current LLMs capture accessibility contrasts with the *universal* and *existential* quantifiers within and across simple sentences. However, they rely on lexical cues in more complex contexts like *negation* and *disjunction*. We conclude that LLMs do not learn the constraints on semantic scope underlying anaphoric accessibility.

## References


Gareth Evans. 1977. Pronouns, quantifiers, and relative clauses (I). *Canadian Journal of Phil.*, 7(3):467–536.

Lauri Karttunen. 1976. Discourse referents. In James D. McCawley, editor, *Syntax and Semantics, volume 7*, pages 363–385. Academic Press, New York.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.

Xiaomeng Zhu and Robert Frank. 2024. LIEDER: Linguistically-informed evaluation for discourse entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13835–13850, Bangkok, Thailand. Association for Computational Linguistics.