Meaning Beyond Truth Conditions: Evaluating Discourse Level Understanding via Anaphora Accessibility

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

Abstract

We present a hierarchy of natural language understanding abilities and argue for the importance of moving beyond assessments of understanding at the lexical and sentence levels to the discourse level. We propose the task of *anaphora accessibility* as a diagnostic for assessing discourse understanding, and to this end, present an evaluation dataset inspired by theoretical research in dynamic semantics. We evaluate human and LLM performance on our dataset and find that LLMs and humans align on some tasks and diverge on others. Such divergence can be explained by LLMs' reliance on specific lexical items during language comprehension, in contrast to human sensitivity to structural abstractions. Dataset and code: .

1 Introduction

The success of modern large language models (LLMs) depends on their capacity for natural language understanding (NLU), i.e., the ability to extract the semantic information contained in a text. Systematic assessment of NLU abilities has been carried out using a diverse set of evaluation tasks, but few of them target whether LLMs accurately represent and update states of natural language discourse. Successful interpretation of discourse requires the ability to use pronominal expressions to refer to entities that have been introduced in a text.

The felicity of **pronominal anaphora**, i.e., using pronouns to refer back to discourse referents introduced earlier, is influenced by the semantic scope of the antecedent:

(1) {A, #Every} farmer worked in his field. He dreamed of the harvest.

Example (1) shows that an entity introduced by an existential quantifier is **accessible** in the same sentence, as well as in subsequent sentences. In contrast, entities introduced by universal quantifiers are only accessible to pronouns in the same sentence; anaphora is infelicitous otherwise. This is illustrated in Figure 1: the discourse referent is **subordinated** to the universal quantifier — that is, inaccessible outside its scope, which extends to the end of the first sentence in the sequence. This makes subsequent reference to *he* in the second sentence infelicitous.

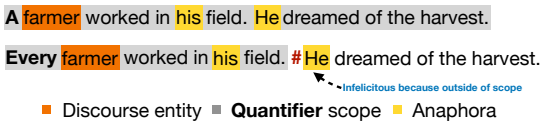

Figure 1: Quantifier scope and its impact on anaphora.

The process of introducing discourse referents is formalized in "dynamic" variants of formal semantics (e.g., Heim, 1983; Groenendijk and Stokhof, 1991; Kamp et al., 2010). In dynamic semantics, utterances precipitate changes in the discourse state, for example by introducing discourse referents. This gives rise to notions of discourse or textual scope which differentiate (e.g.) existential and universal quantifiers, in line with Figure 1.

Here, we focus on one aspect of discourse-level semantic knowledge, namely the fine-grained interactions between semantic scope and referent accessibility. We investigate whether LLMs demonstrate knowledge of the semantic scope properties of various quantifiers and logical connectives, and whether this knowledge is used to generate and update representations of discourse states in human-like ways.

Contribution We make the following contributions:

- In Section 2, we propose a hierarchy of levels of semantic understanding abilities, which

*Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8824–8842
July 27 - August 1, 2025

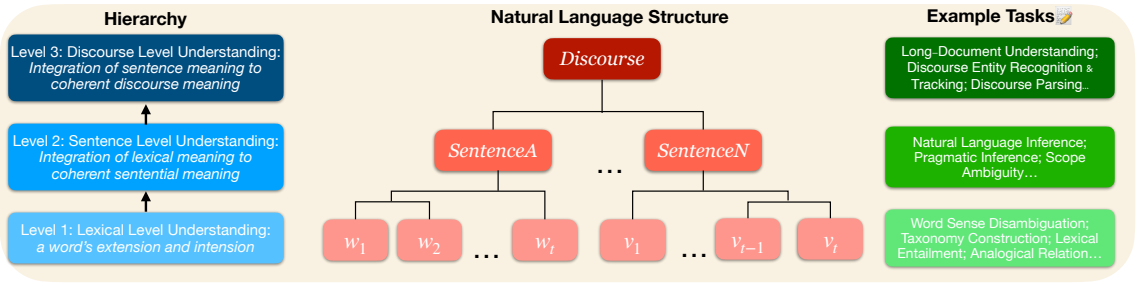

Figure 2: Proposed hierarchy of levels of semantic understanding abilities.

can serve as a guideline for characterizing the kinds of semantic knowledge that LLMs have.

- In Section 3, we propose an evaluation dataset covering discourse anaphora across a variety of linguistic constructions, all of which require sensitivity to the way in which the form of language determines the ways discourse states are implicitly updated in natural discourse.
- In Sections 4 through 7, we evaluate both LLMs and humans with our dataset, and uncover intriguing patterns where human and model behavior align and differ.

2 Levels of Semantic Understanding

Figure 2 illustrates three different levels of natural language understanding: (i) lexical level, (ii) sentential level, and (iii) discourse level. Semantic competence, we propose, requires knowledge of all of these. We discuss each one in detail and review existing work that has tried to evaluate LLM capacities at that level.

2.1 Lexical Level

We define lexical level understanding as **knowing the meaning of individual lexical items**. This requires knowledge of a word's extension (the objects in the world that a word picks out) and its intension (the objects it would pick out if the world were different). Such knowledge allows a competent speaker to make judgments of synonymy, antonymy, entailment and the like. In LLMs, lexical knowledge corresponds to vector representations of individual tokens.

Moskvoretskii et al. (2024) summarizes a range of Natural Language Understanding (NLU) tasks that assess lexical level understanding: Word Sense Disambiguation, Hypernym Discovery, Taxonomy Construction, Lexical Entailment, etc. Another test of lexical semantic understanding derives from the analogical reasoning tests explored by Mikolov et al. (2013), where word meaning is needed to complete analogies such as *man*:*king* as *woman*:*X*. All of these tasks rely on knowledge of word meaning that is independent of the effects on meaning that derive from the composition of words in phrases and sentences.

2.2 Sentence Level

On top of the building blocks provided by lexical understanding, sentence understanding is **the ability to integrate lexical meanings in phrases and to form coherent semantic representations for sentences**. Traditionally, sentence-level meaning is identified with truth conditions and encoded using a logical formalism with a rigorously defined semantics (e.g., Heim and Kratzer, 1998).

A model's capacity to encode the truth conditions of single sentences is implicated in important NLU tasks such as Natural Language Inference (NLI), which requires LLMs to form accurate meaning representations for two sentences and classify their logical relations as entailment, contradiction, or neutral (Williams et al., 2018). Similar evaluation tasks have been created for pragmatic inferences, targeting implicature and presupposition (Jeretic et al., 2020). These works investigate meaning representations of pairs of minimally different sentences, either with respect to logical relations or pragmatic relations, without the need to connect the two sentences in sequential order or track changes at the discourse level. Another type of work at the sentence level involves ambiguities, such as scope ambiguity (e.g., Kamath et al., 2024): a single sentence with multiple quantifiers might allow different interpretations given specific scopal arrangements between the quantifiers.

2.3 Discourse Level

We define discourse level understanding as **the ability to integrate the meaning of consecutive sentences into a unified discourse representation**. Discourse-level meaning requires moving beyond formalisms that express meaning as a static representation of truth conditions to dynamic formalisms in which meaning accrues via update to a contextual representation or state.

One type of task that probes discourse level understanding is discourse parsing (e.g., Maekawa et al. 2024), which evaluates the ability of a model to determine the relationships between sentences, such as *elaboration*, *attribution*, etc. While informative, this task requires the adoption of specific assumptions about the structure and categories that determine discourse relations.

An alternative, more theory-neutral evaluation considers the accumulation of information through a discourse. Li et al. (2021) examined the tracking of the state of individuals and situations across a text. They probed the internal representations of encoder-decoder transformers and found localizable, interpretable structures, supporting the claim that pretrained language models implicitly simulate entity tracking processes dynamically. Kim and Schuster (2023) extended the paradigm in Li et al. (2021) by removing the potential shortcuts that models can use in inferring the states of discourse entities. This line of work uses natural language to explicitly describe the initial state of a situation as well as each subsequent change in the state (e.g. *Box 1 contains the book. Box 2 contains the apple.... Move the book into Box 2...*), thereby functionally similar to the core idea of dynamic semantics. However, because of the simplicity of the language involved, this task did not probe sensitivity to the specific lexical items and syntactic structures that impact the evolution of discourse state.

Davis and Altmann (2021) took a different perspective and investigated the extent to which event representations propagate forward through the hidden states of recurrent neural networks (RNNs), with an emphasis on state change information induced by natural language lexical items. They used Representation Similarity Analysis (RSA, Kriegeskorte et al., 2008) to compare the relevant hidden states in RNNs and found that RNNs capture the extent to which subjects and objects change states as well as the temporal order between state changes. This is one first piece of evidence of the dynamic state changes at the discourse level in neural network models.

Another line of evaluation targets how processing each sentence in a discourse impacts the entities that can be discussed, the task of discourse entity recognition (Schuster and Linzen, 2022; Zhu and Frank, 2024). Schuster and Linzen examine sensitivity to the scope of negation at the discourse level: an indefinite in the scope of negation should not introduce an entity that can be referred to outside the negation's scope. They found that while LLMs indeed exhibit such sensitivity, their performance is not systematic. Zhu and Frank (2024) extended their paradigm by increasing the types of test items, which allows for the evaluation of the semantic properties that govern discourse entity introduction and reference. However, both Schuster and Linzen (2022) and Zhu and Frank (2024) only evaluated LLMs on sentences of a rather simple structure, such as *John owns a dog but Mark does not own a dog*, which only considers negation as the operator that interacts with discourse entities. This gap in the literature calls for a more comprehensive evaluation of **other operators** (such as existentials, universals, conditionals, and disjunctions) that interact with discourse entities, as in the present study.

3 Evaluating Discourse-level Meaning Representation: Case Study on Anaphora (In)accessiblity

As discussed in the previous section, existing work on the evaluation of LLMs' discourse level semantic understanding leaves unexplored the implications of the fine details of semantic composition and scope on the representation of discourse context. As we elaborate below, the scopal properties of quantifiers and logical connectives that are determined by sentence level semantic interpretation play a significant role in discourse level interpretation: depending on the semantic operator, they may license discourse entities only within their scope. We exploit such patterns of anaphora as a case study for diagnosing sensitivity to the structure-sensitive aspects of the discourse state-updating process. Thus, our work provides another way of studying LLMs' state-tracking ability, through attention to the linguistic details of the discourse as opposed to the world model consequences of the actions described in a discourse.

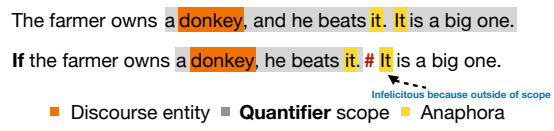

Figure 3: Illustration of anaphora accessibility in donkey conditionals.

3.1 Constructions

We consider three operators whose scope plays a significant role in licensing discourse anaphora: universal quantifiers, negation, and disjunction.

3.1.1 Universal Quantifiers

Every The first case of anaphora (in)accessiblity that we consider is the universal quantifier. We start with a simple example, which contrasts the behavior of sentences whose subjects involve the quantifiers *a* and *every*.

(2) a. EXISTENTIAL: A farmer worked in the field.
b. EVERY: #Every farmer worked in the field.
c. CONTINUATION: He dreamed of the harvest.

As shown in Figure 1, (2c) is felicitous following (2a), but not following (2b). This is because the semantic scope of the existential quantifier extends indefinitely to the right, but the pronoun *he* in (2c) is outside the scope of the universal quantifier in (2b).[*] In sum, the scope of universal quantifiers serves as a boundary for anaphoric accessibility. An LLM capable of discourse level understanding should therefore accurately represent the effects on the discourse context of examples like (2b) and reject the infelicitous continuation (2c).

Donkey Conditionals A more complex case of anaphora accessibility is known as "donkey conditionals" in the dynamic semantics literature (Kanazawa, 1994). In such cases, a discourse entity is introduced via an existential quantifier in the antecedent of a conditional. In such cases, the indefinite licenses pronouns in the conditional's consequent, but not in subsequent sentences. We consider 3 cases: two types of conditional sentences, namely *if* and *whenever* conditionals, and conjoined sentences with an existential object in the first conjunct.

(3) a. EXISTENTIAL (*Exi*): John owns a donkey, and he beats it.
b. CONDITIONAL (*Cond*): #If John owns a donkey, he beats it.
c. WHENEVER (*When*): #Whenever John owns a donkey, he beats it.
d. CONTINUATION (*Cont*): It is a big one.

Such cases can be assimilated to the quantifier cases discussed above, if we assume the conditional clauses implicity introduce a universal quantifier that is not directly tied to a lexical quantifier (see Figure 3). Assuming this to be the case, the pronoun *it* in (3d) is outside the scope of the implicit universal quantifier in (3b) and (3c), rendering the continuation (3d) infelicitous. The same continuation, however, is acceptable in (3a) for the same reasons as (2a). Thus, determining that this continuation sentence is infelicitous after (3b) and (3c) requires accurate processing of the context sentence in preparation for the continuation and subsequent integration, which is exactly what we define as understanding at the discourse level.

3.1.2 Negation

Negation is another logical connective that modulates anaphora accessibility—in general, it is impossible to refer back to discourse referents that are introduced within its scope. However, double negation is an exception (see Hofmann 2024 for discussion and references).

(4) a. EXISTENTIAL (*Exi*): The farmer owned a cow.
b. NEGATION (*Neg*): #The farmer didn't own a cow.
c. DOUBLENEGATION (*DN*): It was not the case that the farmer didn't own a cow.
d. CONTINUATION (*Cont*): (In fact,) It was (just) away on the meadow.

Consider the four conditions (4a-c) with negation, each followed by the same continuation (4d). As is analyzed by Hofmann and illustrated in Figure 4, the local context of the *cow* referent in DOUBLENEGATION is veridical, and the speaker is committed to the existence of *a cow* owned by *the farmer*. In other words, two negations cancel each other out. Thus, EXISTENTIAL is semantically equivalent to DOUBLENEGATION, and both of them license the anaphoric *it* in CONTINUATION. In contrast, no discourse referent of *a cow*

[*] Infelicitous examples are usually marked as # by linguistics conventions. However, we use # to indicate the infelicity of a sentence specifically in the context of the provided continuation.

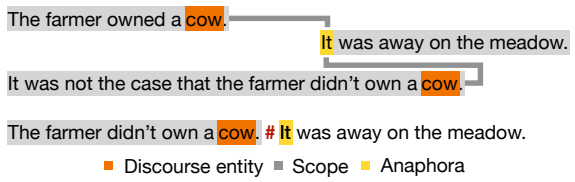

Figure 4: Illustration of anaphora accessibility in negation cases.

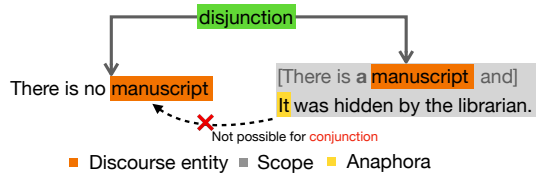

Figure 5: Illustration of anaphora accessibility in disjunction cases.

exists outside the scope of negation in NEGATION, which makes it an infelicitous context for the subsequent anaphora. Here, we examine whether LLMs know the semantic scope of negation and whether negation's inaccessibility can be reversed in double negation contexts.

3.1.3 Disjunction

Negation within disjunctions adds another layer of complexity to anaphora accessibility. Evans (1977) observes that discourse referents introduced through existentials within a first disjunct do not license anaphora in the second disjunct. Surprisingly, however, a discourse referent introduced with a negative quantifier in a first disjunct does. We see this contrast in the first two examples of (5):

(5) a. EITHERPOSOR: #Either there was a manuscript, or it was hidden by the librarian.

b. EITHEROR: Either there was no manuscript, or it was hidden by the librarian.

c. OR: There was no manuscript, or it was hidden by the librarian.

d. CONJUNCTION: #There was no manuscript, and it was hidden by the librarian.

(5c) demonstrates that the presence or absence of the lexical item *either* to introduce the disjunct does not have any impact on the discourse semantics. Finally, (5d) shows that negative quantifiers in conjunction do not have similar effects. These accessibility patterns are summarized in Figure (5).

3.2 Experiment Design

Model We investigated the performance of four open-source LLMs (`Llama3-2-1B`, `Llama3-2-3B`, `Llama3-1-8B` and `Llama3-1-8B-Instruct` (Dubey et al., 2024)), and two closed-source LLMs (GPT `babbage-002` and `davinci-002`) on our constructed dataset through the Huggingface transformer API (Wolf et al., 2019) and the OpenAI API, respectively. We also investigated the performance of GPT-4o (OpenAI et al., 2024) with the prompt-based method with no access to logits, and we report the results in Section 7.

Human Experiment To establish a human baseline for models' performance, we recruited 104 participants over Prolific. Each participant did 66 forced-choice trials, with 22 experimental items and 44 fillers. In each trial, participants were visually presented with 2 minimally different sentences on the screen, and they were asked to choose the more acceptable sentence from the pair. See Appendix A for more details on our experiment design. Human results are presented in the following sections along with language model performance.

Corpus Experimental stimuli were generated from a set of structural templates containing the target constructions. For each experiment, we manually constructed 32 semantically plausible simple sentence frames with the help of GPT-4o (OpenAI et al., 2024), following the example sentences shown in Section 3.1. Test sentences were then manually inspected by linguistics experts to ensure semantic plausibility and (un)acceptability. This yields a set of 9816 sentences in total. See Appendix B for more details on dataset construction.

Metrics We adopt the evaluation paradigm in Futrell et al. (2019) that considers LLMs as psycholinguistic subjects. That is, for each evaluated sentence, we take the surprisal (i.e., the negative log probability) assigned by the model to individual tokens, defined in Equation 1:

$$surprisal(w_i) = \log \frac{1}{P(w_i|w_1, ..., w_{i-1})} \quad (1)$$

The total probability the model assigns to a sentence or part of a sentence is obtained by taking the sum of $surprisal(w_i)$ for each target token w_i. The surprisal values serve as the base measurement for the analyses of each individual experiment described in the following sections.

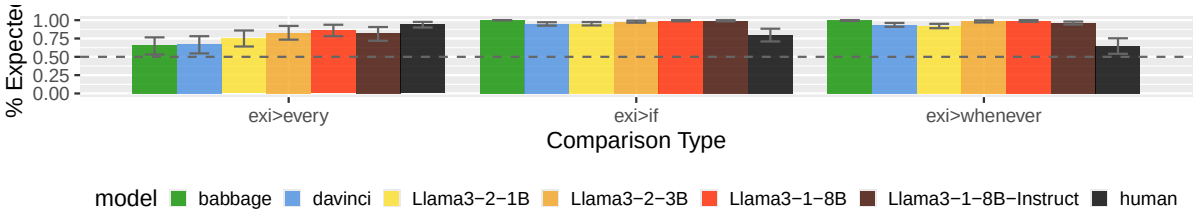

Figure 6: LLMs' performance on the comparisons involving existential vs. universal quantifiers. In the figures of this paper, $>$ signs indicate degrees of felicity from theoretical predictions. For example, `exi>every`, the label for the leftmost panel, means that EXISTENTIAL should be more felicitous than EVERY sentences in the relevant comparison. Such felicity preference is determined by whether models exhibit the inequality shown in equation (5).

4 Experiment 1: Universal

In this section, we present models' performance on anaphora accessibility with regard to the universal quantifier as discussed in Section 3.1.1. Here, we compare model performance with the theory-driven contrasts in Section 3. One could also directly compare LLM and human behavior, which we discuss in Appendix C.

In general, given different context sentences and the same continuation, we expect models to assign a higher conditional probability to the continuation given a context in which it is felicitous than another context in which it is infelicitous. In other words, we expect the following inequalities to hold if LLMs exhibit discourse level understanding abilities with regard to universal quantifiers.

$$p(Cont|Exi) > p(Cont|Every) \quad (2)$$
$$p(Cont|Exi) > p(Cont|Cond) \quad (3)$$
$$p(Cont|Exi) > p(Cont|When) \quad (4)$$

However, one problem about this measure is that it is too lenient – although continuations such as (2c) are infelicitous after (2b), it should become felicitous if *he* is instead embedded inside the scope of (2b), such as the contrast below.

(6) a. CROSSSEN: Every farmer worked in the field. #He dreamed of the harvest.

b. SINGLESEN: Every farmer worked in the field before he dreamed of the harvest.

Therefore, we would expect models to assign a higher probability to (6b) than (6a). Importantly, the contrast in example (6) does not exist for their counterparts with the existential quantifier—we would expect a smaller difference in probability between them if the LLMs that we tested have good discourse level understanding abilities. Thus, instead of using equations (2), (3), and (4) as our metric, we adopt the difference-of-difference metric with the general form shown in (5). We binarize the comparison of each trial by recording whether the inequality holds in the predicted direction.

$$p(\exists\text{-SINGLESEN}) - p(\exists\text{-CROSSSEN}) < p(\forall\text{-SINGLESEN}) - p(\forall\text{-CROSSSEN}) \quad (5)$$

Results As is shown in Figure 6, all models show above chance performance for the expected inequality in equation (5). Specifically, for the simple comparison between EXISTENTIAL and EVERY (leftmost panel in Figure 6), we found that the Llama family models that we tested achieved higher accuracy (around 75%) than `babbage` and `davinci` in the GPT family, while humans scored even higher at ceiling. In the other two comparisons where the universal quantifier is implicitly encoded through CONDITIONAL and WHENEVER, it is the LLMs that score at ceiling. In contrast, humans had lower accuracy but still performed above chance. This pattern indicates that the LLMs we examined know the scope of the discourse entity introduced within the universal quantifier and that it is infelicitous to refer back to such entities outside of the scope.

In addition to the continuation in (3d) that starts with the discourse bound pronoun *it*, for the comparisons `exi>if` and `exi>whenever`, we also con-

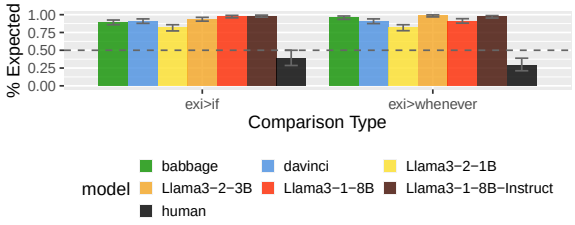

Figure 7: Model performance on *he*-continuations for `exi>if` and `exi>whenever`.

sidered a variant where the continuation starts with a uniformly felicitous *he* and the discourse bound pronoun *it* is in object position, such as *He also feeds it*. Given our framing of the anaphora accessibility task, there should not be a difference between *he*-continuations and *it*-continuations—they should both be infelicitous given a preceding CONDITIONAL or WHENEVER context. Results on this variant are shown in Figure 7. Interestingly, there is a striking contrast between human and models' performance. While models continue to exhibit the preference for EXISTENTIAL over CONDITIONAL and WHENEVER, humans actually prefer the universal counterparts for donkey conditionals, which is not predicted in the literature. We believe that this discrepancy could be due to an effect called *telescoping* (Roberts, 1989). The intuition is that humans have the tendency to interpret *he*-continuations as being subordinated under the scope of CONDITIONAL or WHENEVER, which makes *he*-continuations more felicitous than they should be. In comparison, *it*-continuations are less likely to be interpreted in a subordinated way. One potential factor that might explain the aforementioned tendency is subject bias: since *the farmer* is the subject of the context sentence, it is more saliently represented in the discourse. Therefore, humans are more likely to refer back to it in the continuation using *he* (Grosz et al., 1995). In sum, the models' success on this dataset shows their knowledge of the difference between universal and existential quantifiers.

5 Experiment 2: Negation

As discussed in Section 3.1.2, the second construction that we are interested in is negation. Following the reasoning there, we expect the following two inequalities to hold if the LLMs understand the semantic scope of negation:

$$p(Cont|Exi) > p(Cont|Neg) \quad (6)$$

$$p(Cont|DN) > p(Cont|Neg) \quad (7)$$

Since every pair of sentences we compare shares the continuation but not the context sentences, we apply the conditional probabilities metric: compare the summed surprisal on tokens in the CONTINUATION, with the concatenated context fed to the model as a preamble.

Results As shown in the top two panels of Figure 8, all models succeed in preferring the EXISTENTIAL context over NEGATION, but three of the

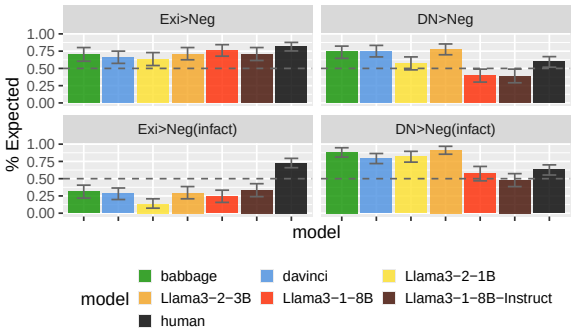

Figure 8: Model performance in Experiment 2.

models struggle to favor DOUBLENEGATION over NEGATION. In particular, the two `Llama3-1-8B` models show a preference of NEGATION over DOUBLENEGATION, 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. The most straightforward way to interpret these results is that the LLMs have trouble understanding that EXISTENTIAL is equivalent to DOUBLENEGATION in terms of their power in licensing subsequent anaphora to discourse referents introduced within their scopes. However, another hypothesis is that DOUBLENEGATION is dispreferred not because the LLMs failed to learn double negation elimination, but simply because DOUBLENEGATION sentences have a more complex (and presumably less frequent) structure than its EXISTENTIAL counterpart.

Influence of Specific Lexical Items To test this hypothesis, we considered a variant of the test sentences by adding the phrase *in fact* to the beginning of each continuation sentence and computed accuracy using the same inequalities as in (6) and (7). The intuition is that adding this phrase sets up a contrast relation that could help the models to process DOUBLENEGATION sentences to a larger degree than to process EXISTENTIAL ones. If the low accuracy that we observed for the `DN>Neg` comparison is due to lexical-level factors, we would expect an increase in accuracy in the variants. In contrast, if models failed to learn the difference between double negation and negation completely, the accuracy of the variants would remain low.

Results are shown in the bottom two panels of Figure 8. Compared to the base case, adding *in fact* does help to lift the accuracy for the `DN>Neg` comparison, as most models now have a stronger pref-

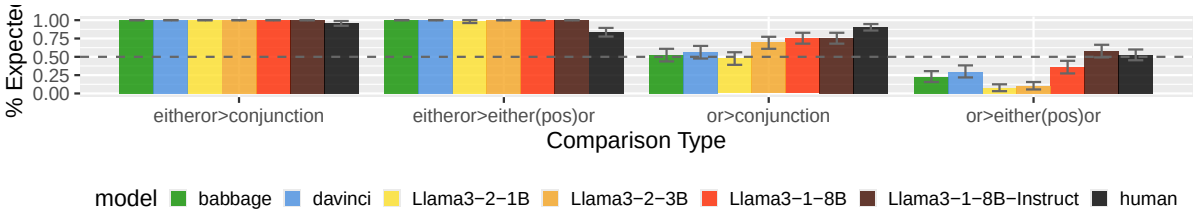

Figure 9: Model performance in Experiment 3.

erence of DOUBLENEGATION over NEGATION. However, adding *in fact* also flips the direction of the Exi>Neg comparison, as all models now favor NEGATION over EXISTENTIAL sentences. In contrast, human patterns remain stable regardless of the addition of *in fact*: they still show a clear preference for EXISTENTIAL and DOUBLENEGATION over NEGATION.

One way to interpret the flipped result is that the phrase *in fact* tends to co-occur with double negation sentences, thereby increasing the conditional probabilities of the continuation. Adding *in fact* to existential sentences makes the discourse less coherent to process, thereby lowering the accuracy in the Exi>Neg(infact) comparison. This results in the reversed DOUBLENEGATION>NEGATION>EXISTENTIAL ranking by language models, while human judgments remain consistent. Although adding *in fact* to the continuation does not change anaphora accessibility, the increase that we observed here suggests that LLMs are sensitive to the presence of specific lexical items and that their performance with respect to identifying the scope of negation is not systematic.

6 Experiment 3: Disjunction

In the last experiment, we test the constructions presented in Section 3.1.3 with respect to disjunction. Since the sentences that we compare share neither the context nor the continuation, we calculate the Syntactic Log-Odds Ratio score (SLOR) (Lau et al., 2017) on each sentence and compare the SLOR scores, which is defined as:

$$\text{SLOR}(s) = \frac{\log p_m(s) - \sum_{w \in s} \log p_u(w)}{|s|} \quad (8)$$

where for sentence s, $\log p_m(s)$ represents the log probability assigned by the model to the entire sentence (which is equivalent to summing up the surprisals for all tokens in s); $\log p_u(w)$ represents the unigram probability of each token w in the sentence; and $|s|$ represents the length of the sentence, which is the number of tokens in s. Intuitively, the SLOR score measures how much *additional* probability the model assigns to the sentence compared to the same bag-of-words, which in turn represents the well-formedness of the sentence, both syntactically and semantically. However, there is no standard on how to interpret the absolute values of the SLOR scores. In the current study, we obtain the estimation of the unigram probabilities by counting the frequency of the tokens from a fragment of the OpenWebText Corpus (Gokaslan and Cohen, 2019) obtained from the tokenizers of the Llama3 family and the GPT3 family, respectively.

Recall from Section 3.1.3 that OR and EITHEROR are felicitous, while CONJUNCTION and EITHERPOSOR are not. Translating the judgments to the metric, we expect the following four inequalities to hold if models exhibit discourse level understanding abilities.

$$\begin{aligned} \text{SLOR}(\text{OR}) &> \text{SLOR}(\text{CONJUNCTION}) \\ \text{SLOR}(\text{EITHEROR}) &> \text{SLOR}(\text{CONJUNCTION}) \\ \text{SLOR}(\text{OR}) &> \text{SLOR}(\text{EITHERPOSOR}) \\ \text{SLOR}(\text{EITHEROR}) &> \text{SLOR}(\text{EITHERPOSOR}) \end{aligned} \quad (9)$$

Results As shown in Figure 9, models achieved ceiling performance for all comparisons involving EITHEROR—they demonstrate a preference for this felicitous case over CONJUNCTION and EITHERPOSOR, which is consistent with human preferences. In contrast, the performance is around chance for the or>conjunction comparison, while humans show the predicted preference pattern to a larger extent than all LMs. Strikingly, models exhibit a preference for EITHERPOSOR over OR (rightmost panel), which is the reverse pattern of what we expect. Humans show no clear preference in this comparison. Overall, the pattern here repeats Experiment 2 in that LLMs' ability to differentiate contexts with different anaphora accessibility depends largely on lexical items and is not systematic—although EITHEROR and OR are

Exp.	Condition	GPT-4o	Other Models	Humans
Exp 1	exi>every	**0.859** ↓	0.766 ↓	0.941
	exi>if	0.567 ↓	**0.979** ↑	0.802
	exi>whenever	0.228 ↓	**0.966** ↑	0.647
Exp 2	Exi>Neg	**0.760** ↓	0.696 ↓	0.818
	DN>Neg	0.781 ↑	**0.604** ↑	0.594
	Exi>Neg(infact)	**0.771** ↑	0.266 ↓	0.729
	DN>Neg(infact)	0.812 ↑	**0.745** ↑	0.629
Exp 3	eitheror>conjunction	**1.000** ↑	**1.000** ↑	0.959
	eitheror>either(pos)or	**0.844** ↑	0.997 ↑	0.835
	or>conjunction	**0.875** ↓	0.628 ↓	0.906
	or>either(pos)or	0.242 ↓	**0.273** ↓	0.529

Table 1: A summary of the accuracy results of GPT-4o, humans, and the mean accuracies of the models tested in Section 4, 5, and 6. To compare logit-based results versus prompting with human results, for each condition, the model accuracy that is closer to (with smaller absolute difference) the human accuracy is bold. Arrows mark whether the models' accuracies are below or above the human baseline.

equivalent to each other, models' preference largely depends on whether there is *either* in the sentence.

7 Experiments on GPT-4o

In addition to the models examined in the previous sections, we are also interested in how GPT-4o (OpenAI et al., 2024), a currently state-of-the-art but closed-source model, performs on anaphora accessibility. Because this model does not provide access to log probabilities, we use prompting to get the model's judgments on the minimal pairs. We used the following prompt, which is maximally similar to the instructions in our human experiments, to minimize any instruction-related effects:

```
In this task, you will be presented with
two sentences. Your job is to select which
sentence in a pair is more acceptable by
only returning the index of the sentence:
1 or 2.

Sentence 1: {sent1}
Sentence 2: {sent2}

Which sentence is more acceptable?
```

The results are presented in Table 1. As a summary, we took the mean accuracies of the models examined in the previous sections and compared them with human accuracies. Arrows represent whether the models' accuracies are above or below the ones for humans, and for every condition, the accuracy that is closer to humans' accuracy is bold.* Overall, both logit-based and prompt-based methods result in qualitative alignment between humans' and models' preferences, as both humans and models showed above-chance preference for the expectation in most conditions. We do observe several exceptions: for the `exi>whenever` comparison, GPT-4o shows a strong preference for `whenever` sentences while other models and humans show the opposite; for the `Exi>Neg(infact)` condition, GPT-4o and humans showed a similar preference for the expected existential sentences, while the other models preferred the infelicitous ones. Finally, for the `or>either(pos)or` condition, all models preferred the infelicitous sentences, while humans showed around-chance preference.

The difference between logit-based and prompt-based evaluations could be one factor for the disparity we observed between GPT-4o and other models. Since prompting involves a text-based description of the task we want LLMs to perform, prompting exerts extra task burden on LLMs compared to getting the logits: there is an extra cost for LLMs to comprehend the task description and perform the task as it is understood. Hu and Levy (2023) have demonstrated that prompting is not a substitute for probability measurements in LLMs, where LLMs' metalinguistic judgments are inferior to quantities directly derived from representations. Webson and Pavlick (2022) also raised questions about whether LLMs truly understand their prompts. On the other hand, no metalinguistic judgments are needed for the pure logit-based metrics, which, in a sense, actually results in better alignment to what we aim to assess: LLMs' probability judgments on minimal pairs. The mechanisms underlying such qualitative mismatches between logit-based and prompt-based methods, as well as between models and humans, remain an open question for future investigation.

8 Conclusion

In this paper, we defined a hierarchy of semantic understanding abilities consisting of lexical, sentence, and discourse levels. Filling in the gap in the literature, we constructed an evaluation task of anaphora accessibility that allows for a fine-grained examination of the understanding abilities of LLMs. Results show that our task successfully identified places of convergence and divergence between model and human performance, where LLMs rely on specific lexical cues but humans don't. This work is one further step toward improving the discourse understanding abilities of LLMs.

*See Appendix D for the full accuracy results and statistics for all models.

Limitations

Running the Dataset in the most recent SOTA Models In the current study, we only tested our datasets with a limited range of LLMs. It would be interesting to see the performance of other state-of-the-art language models such as GPT-4.5 and the DeepSeek model family. Further, for the purposes of comparison it would be best to evaluate such models directly through the probabilities they assign, but this would require access to the logits the models assign to each token, something which is not available for closed-source models.

Evaluating More Subtle Constructions from Theoretical Predictions In addition to the three classes of quantifiers and logical connectives, there is a rich pool of linguistic constructions from the theoretical semantics literature that involve more complex scopal interactions that lead to other predictions about anaphora accessibility. An example is modal subordination (e.g., Roberts, 1989, where the scope of *if*-conditional sentence interacts with modal operators. There are few empirical studies on how humans process such sentences. Future work could further extend our dataset to incorporate a larger variety of constructions and acquire a human baseline.

Behavioral versus Mechanistic Level Evaluations In Section 2, we reviewed related works (Kim and Schuster, 2023; Li et al., 2021) that explicitly investigate the state or discourse entity tracking capability by probing the internal activation states of language models. The current study, despite investigating the discourse updates within natural language instead of simulating discourse updates, remains at the behavioral level and is empirical in nature. Developing methods that explicitly target models' internal representations that correlate with state-update behaviors would bring greater interpretability and could contribute to theory building. Future work could improve our understanding of the processing level details of models on the current dataset by importing techniques from mechanistic interpretability.

Acknowledgments

We would like to thank the anonymous reviewers from the ARR February 2025 cycles and the SCiL 2025 committee for their thoughtful comments, which helped us refine this paper. We would also like to thank the members of the Computational Linguistics at Yale Lab, the Yale Computation and Cognition Joint Lab, and the Yale Linguistics Department for helpful feedback. We thank the Yale Center for Research Computing for guidance and use of the research computing infrastructure, specifically the Grace cluster. Any errors are our own.

References

Alexander L. Anwyl-Irvine, Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1):388–407.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*, 50 edition. The MIT Press.

Forrest Davis and Gerry TM Altmann. 2021. Finding event structure in time: What recurrent neural networks can tell us about event structure in mind. *Cognition*, 213:104651.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

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.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and philosophy*, pages 39–100.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Irene Heim. 1983. On the projection problem for presuppositions. *Formal semantics–the essential readings*, pages 249–260.

Irene Heim and Angelika Kratzer. 1998. *Semantics in generative grammar*. Blackwell, Oxford.

Lisa Hofmann. 2024. Anaphoric accessibility with flat update. Manuscript, University of Stuttgart.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754.

Hans Kamp, Josef Van Genabith, and Uwe Reyle. 2010. Discourse representation theory. In *Handbook of Philosophical Logic: Volume 15*, pages 125–394. Springer.

Makoto Kanazawa. 1994. Weak vs. strong readings of donkey sentences and monotonicity inference in a dynamic setting. *Linguistics and philosophy*, 17:109–158.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249.

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.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Aru Maekawa, Tsutomu Hirao, Hidetaka Kamigaito, and Manabu Okumura. 2024. Can we obtain significant success in RST discourse parsing by using large language models? In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2803–2815, St. Julian's, Malta. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. TaxoLLaMA: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2331–2350, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan,

Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Craige Roberts. 1989. Modal subordination and pronominal anaphora in discourse. *Linguistics and philosophy*, 12:683–721.

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.

Jon Sprouse, Carson T Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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.

A Human Experiment

We tested a total of 11 comparison types (3 in Experiment 1, 4 each in Experiments 2 and 3) on human subjects. Each comparison type includes 32 sentence pairs. In each test trial, participants were presented with a pair of sentences in a multiple-choice format (see Figure 10 for the experimental interface) and were asked to click on the sentence that they found to be more acceptable. Each participant received 22 test items and 44 filler items, which sum to a total of 66 trials. The filler items were the same across participants and were selected from BLiMP (Warstadt et al., 2020) such that for each filler minimal pair, one of the sentences is strictly more acceptable than the other. Therefore, we also used filler items as attention checks. Participants who scored below 90% accuracy on the filler items were excluded from the final results. The experiment was also set up such that each test item was rated by at least 5 participants.

We used the Gorilla Experiment Builder (www.gorilla.sc) to create and host our experiment interface (Anwyl-Irvine et al., 2020), and participants were recruited through Prolific (www.prolific.com) under a university-approved IRB. We recruited a total of 104 native speakers of English without any language or vision-related disorders who also currently reside in the United States. 85 of them (81.73%) passed the filler check. Each participant filled out a consent form prior to completing the experiment. They each received a compensation of $3, which is equal to an hourly rate of $14.41.

B Dataset Construction

This section elaborates on the dataset construction procedure described in Section 3.2. We provide sample templates, frames, and experimental sentences and refer the readers to our GitHub Repository for the full dataset.

We created the dataset in two steps: (1) human curation of sentence templates; (2) template filling with content words inspired by GPT-4o (OpenAI et al., 2024). First, we follow the dynamic semantics literature in designing the sentence templates for the three types of constructions we identified in Section 3.1. The templates specify the kinds of constituents needed in each type of sentence, respectively. For example, an intransitive verb phrase is needed for sentences in Experiment 1. Here is the template for Experiment 1 that generates sentences in (2):

- $\exists$-SINGLESEN: A [Noun_1] [Verb_1_PP] before {he, she} [Verb_2] [Verb_2_PP].
- $\exists$-CROSSSEN: A [Noun_1] [Verb_1_PP]. {He, She} [Verb_2] [Verb_2] [Verb_2_PP].
- $\forall$-SINGLESEN: Every [Noun_1] [Verb1_PP] before {he, she} [Verb_2] [Verb_2_PP].
- $\forall$-CROSSSEN: Every [Noun_1] [Verb1_PP]. {He, She} [Verb_2] [Verb_2_PP].

Then, to create multiple sentences that share the same construction, we created 32 frames for each template based on the constituents needed. We used GPT-4o in this step as an inspiration for the content words that fill into the blanks of the template. For example, one of our prompts during the template filling stage was: "Please give me 10 action verbs that could be performed by a pirate". Not all of the verbs returned by GPT-4o are semantically natural in our contexts, so we performed manual selection among the results and picked out appropriate ones. Here is the set of sentences frames we used to fill in the templates above:

Finally, we include sample sentences for each condition of all three experiments:

- Experiment 1: Universal versus Existential
  - EXISTENTIAL-SINGLESEN: A farmer worked in the field before he dreamed of the harvest.
  - Existential-CrossSen: A farmer worked in the field. He dreamed of the harvest.
  - Universal-SingleSen: Every farmer worked in the field before he dreamed of the harvest.
  - Universal-CrossSen: Every farmer worked in the field. He dreamed of the harvest.
- Experiment 1: Donkey Conditionals
  - IF-SINGLESEN: If the farmer owns a cow, he beats it and it is a big one.

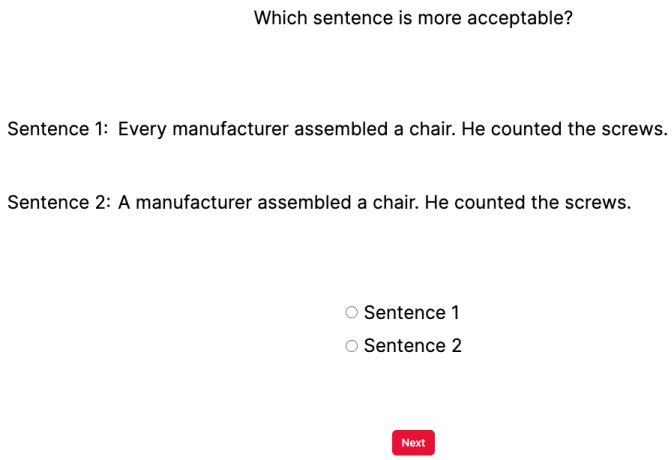

Figure 10: Experimental interface on Gorilla with an example test item where participants were expected to click on Sentence 2.

  - IF-CROSSSEN: If the farmer owns a cow, he beats it. It is a big one.
  - WHENEVER-SINGLESEN: Whenever the farmer owns a cow, he beats it and it is a big one.
  - WHENEVER-CROSSSEN: Whenever the farmer owns a cow, he beats it. It is a big one.
  - EXISTENTIAL-SINGLESEN: The farmer owns a cow, and he beats it and also feeds it.
  - EXISTENTIAL-CROSSSEN: The farmer owns a cow, and he beats it. He also feeds it.

- Experiment 2: Negation
  - DN: It was not the case that the farmer didn't own a cow. It was just away on the meadow.
  - DNINFACT: It was not the case that the farmer didn't own a cow. In fact, it was just away on the meadow.
  - NEG: The farmer didn't own a cow. It was away on the meadow.
  - NEGINFACT: The farmer didn't own a cow. In fact, it was away on the meadow.
  - EXISTENTIAL: The farmer owned a cow. It was away on the meadow.
  - EXISTENTIALINFACT: The farmer owned a cow. In fact, it was away on the meadow.

- Experiment 3: Disjunction
  - EITHEROR: Either there was no treasure, or it was guarded by the dragon.
  - EITHERPOSOR: Either there was a treasure, or it was guarded by the dragon.
  - OR: There was no treasure, or it was guarded by the dragon.
  - CONJUNCTION: There was no treasure, and it was guarded by the dragon.

C Linguistics Theories vs. Performance

One of the contributions of the current paper is that it proposes the use of dynamic semantics in evaluating the discourse-level understanding abilities of LLMs. It is important to note that dynamic semantics is a theory of **competence**: it characterizes the linguistic knowledge rather than how people use and process language. The theory of dynamic semantics, like other competence theories, has been developed on the basis of a range of empirical patterns discussed in the linguistics literature. While experimental investigation has suggested that the patterns of data used in linguistic theorizing are stable (Sprouse et al., 2013), the semantic theories we are considering here assume that the phenomena of interest are modulated by the structure of the sentences and do not take into account factors such as word frequency, lexical semantic content, or plausibility. Yet, such factors certainly do play a role in **performance**, e.g., in the task of assigning an interpretation to a sentence in real time in a specific context. Moreover, the evaluation that we have conducted on both LLMs and humans can be plausibly seen as an instance of linguistic performance:

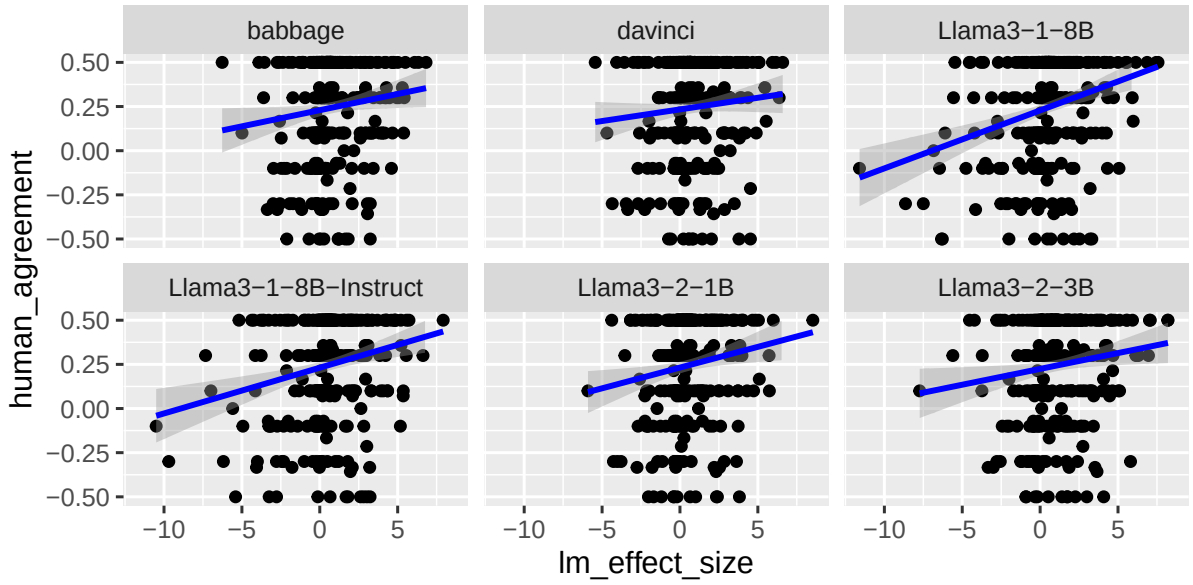

Figure 11: Correlation between human agreement score and LLM effect size for each LLM. Each point represents an experimental stimuli pair.

the responses we gathered from LLMs and humans are in response to the task of asking for acceptability judgments to specific sentences (Chomsky, 1965). The distinction between competence and performance might be able to explain some "imperfect" human performance patterns reported in earlier sections. For example, humans perform around chance for `or>either(pos)or` but almost at ceiling for `eitheror>either(pos)or`. Linguistic theories, on the other hand, predict that OR sentences are equally felicitous as EITHEROR ones, and EITHERPOSOR is not. It is unclear to us why the addition of *either* caused such a big shift in human performance—perhaps the absence of *either* makes some crucial scopal interpretation more likely—but this example emphasizes the distinction between linguistic theories and human performance.

This suggests that we might want to consider LLMs as models of performance rather than competence. If so, we can compare human performance and model performance with one another on our proposed benchmarks, as opposed to comparing each to a pattern of idealized competence. We take a first step in this direction by comparing two measures: human agreement score and LLM effect size. For each of our human stimuli pairs (which consists of an acceptable sentence and its unacceptable counterpart), we compute the human agreement score a where x_i is a binary value indicating whether the human judgment aligns with theoretical predictions, and n is the number of subjects who provided judgments on this stimuli pair.

$$a = \frac{1}{n}\sum_{i=1}^{n} x_i - 0.5, \quad \text{where } x_i \in \{0, 1\} \quad (10)$$

The agreement score a ranges between -0.5 and 0.5. A score that is closer to 0 indicates high variability across human responses, and a score that is farther away from 0 indicates that humans have high agreement with one another in terms of which of the sentence in the pair is more acceptable. We compute the effect size of LLM judgments by subtracting the side of our inequality metric that is expected to have a smaller value from the side that is expected to have a bigger value. The absolute value of the effect size indicates the extent to which the model prefers one sentence over the other, and the sign indicates which sentence is preferred (positive if models' preference is expected under linguistic theories and negative otherwise).

Figure 11 demonstrates the distribution of human agreement scores in relation to LLM effect size. We find that there is a positive correlation overall between human agreement scores and LLM effect sizes ($\beta = 0.02, p < 0.001$). This means that for items where LLMs display a clearer preference, humans also tend to agree more with each other. This can be taken to suggest that LLMs are indeed modeling the variability found in human performance, though much remains to be explored here. We leave this area for future work.

D Full Results and Statistics

For all 11 conditions across the three experiments reported in Section 4 through 7, we include the accuracies and 95% confidence intervals for all models, compared with the human results. All results are summarized in Table 3.

Noun$_1$	**Verb$_1$_PP**	**Verb$_2$**	**Verb$_2$_PP**
farmer	worked in the field	dreamed	of the harvest
pirate	sailed across the ocean	laughed	as the waves crashed around
kid	played in the park	shouted	with joy
student	studied in the library	stared	at the pages
doctor	waited in the hallway	spoke	to the nurse
photographer	stood by the fountain	waited	for the shot
surgeon	operated in the hospital	cried	from the weight of responsibility
editor	signed at the desk	sighed	at the workload
chef	cooked in the kitchen	danced	to the background music
hunter	hid in the bushes	listened	for footsteps
player	rested on the bench	cheered	for the team
police	stood by the intersection	waved	at the cars
researcher	hesitated before an experiment	wandered	through the notes
worker	stretched during a break	stumbled	on the floor
teacher	paused before answering	sat	back to consider the question
customer	browsed through the store	paused	at the shelf
athlete	trained at the gym	breathed	heavily to maintain focus
professor	lectured in the classroom	wrote	equations on the board
administrator	paced in the office	gestured	while speaking
server	hurried across the dining room	smiled	at the guests
director	paused between takes	nodded	at the actors
designer	sketched on a tablet	persisted	to perfect the lines
traveler	wandered through the streets	jumped	over the puddle
developer	coded in a cafe	typed	rapidly to meet the deadline
manufacturer	assembled a chair	counted	the screws
manager	relaxed after the meeting	stretched	to relieve tension
instructor	balanced on one leg	fell	onto the mat
lawyer	waved before speaking	rose	from the chair
warrior	trained in the courtyard	crawled	under the beam
entrepreneur	leaned against the wall	rested	to regain energy
pilot	sat in the helicopter	focused	on the controls
rider	biked on the street	sang	to enjoy the ride

Table 2: Sentence frames for Experiment 1, consisting of a subject noun (Noun$_1$), an initial event (Verb$_1$_PP), and a second event (Verb$_2$, Verb$_2$_PP).

Table 3: Full results and statistics of the experiments.

Condition	Model	Accuracy (95%CI)	Human Accuracy
exi>every	Llama3-1-8B	0.859 (0.766, 0.938)	0.941 (0.900, 0.971)
	Llama3-1-8B-Instruct	0.828 (0.734, 0.922)	
	Llama3-2-1B	0.750 (0.641, 0.844)	
	Llama3-2-3B	0.828 (0.734, 0.922)	
	babbage	0.656 (0.546, 0.766)	
	davinci	0.672 (0.547, 0.782)	
	GPT-4o	0.859 (0.766, 0.938)	
exi>if	Llama3-1-8B	0.994 (0.984, 1.000)	0.802 (0.709, 0.884)
	Llama3-1-8B-Instruct	0.994 (0.984, 1.000)	
	Llama3-2-1B	0.953 (0.928, 0.975)	
	Llama3-2-3B	0.981 (0.966, 0.994)	
	babbage	1.000 (1.000, 1.000)	
	davinci	0.950 (0.925, 0.972)	
	GPT-4o	0.567 (0.530, 0.608)	
exi>whenever	Llama3-1-8B	0.991 (0.978, 1.000)	0.647 (0.541, 0.741)
	Llama3-1-8B-Instruct	0.963 (0.941, 0.981)	
	Llama3-2-1B	0.922 (0.891, 0.950)	
	Llama3-2-3B	0.984 (0.972, 0.997)	
	babbage	0.997 (0.991, 1.000)	
	davinci	0.938 (0.909, 0.963)	
	GPT-4o	0.228 (0.197, 0.261)	
Exi>Neg	Llama3-1-8B	0.760 (0.677, 0.844)	0.818 (0.765, 0.876)
	Llama3-1-8B-Instruct	0.708 (0.615, 0.792)	
	Llama3-2-1B	0.635 (0.552, 0.729)	
	Llama3-2-3B	0.708 (0.615, 0.792)	
	babbage	0.708 (0.625, 0.802)	
	davinci	0.656 (0.562, 0.750)	
	GPT-4o	0.760 (0.677, 0.844)	
DN>Neg	Llama3-1-8B	0.396 (0.302, 0.490)	0.594 (0.524, 0.659)
	Llama3-1-8B-Instruct	0.385 (0.292, 0.490)	
	Llama3-2-1B	0.573 (0.469, 0.667)	
	Llama3-2-3B	0.781 (0.698, 0.854)	
	babbage	0.740 (0.646, 0.823)	
	davinci	0.750 (0.667, 0.834)	
	GPT-4o	0.781 (0.688, 0.854)	
Exi>Neg(infact)	Llama3-1-8B	0.240 (0.156, 0.333)	0.729 (0.665, 0.794)
	Llama3-1-8B-Instruct	0.333 (0.240, 0.427)	
	Llama3-2-1B	0.135 (0.073, 0.198)	
	Llama3-2-3B	0.292 (0.198, 0.385)	
	babbage	0.312 (0.219, 0.406)	
	davinci	0.281 (0.198, 0.375)	
	GPT-4o	0.771 (0.677, 0.854)	
DN>Neg(infact)	Llama3-1-8B	0.573 (0.469, 0.667)	0.629 (0.553, 0.700)
	Llama3-1-8B-Instruct	0.479 (0.385, 0.583)	
	Llama3-2-1B	0.823 (0.740, 0.896)	
	Llama3-2-3B	0.917 (0.854, 0.969)	

	babbage	0.885 (0.823, 0.938)	
	davinci	0.792 (0.708, 0.865)	
	GPT-4o	0.812 (0.729, 0.885)	
eitheror>conjunction	Llama3-1-8B	1.000 (1.000, 1.000)	0.959 (0.924, 0.982)
	Llama3-1-8B-Instruct	1.000 (1.000, 1.000)	
	Llama3-2-1B	1.000 (1.000, 1.000)	
	Llama3-2-3B	1.000 (1.000, 1.000)	
	babbage	1.000 (1.000, 1.000)	
	davinci	1.000 (1.000, 1.000)	
	GPT-4o	1.000 (1.000, 1.000)	
eitheror>either(pos)or	Llama3-1-8B	1.000 (1.000, 1.000)	0.835 (0.776, 0.888)
	Llama3-1-8B-Instruct	1.000 (1.000, 1.000)	
	Llama3-2-1B	0.984 (0.961, 1.000)	
	Llama3-2-3B	1.000 (1.000, 1.000)	
	babbage	1.000 (1.000, 1.000)	
	davinci	1.000 (1.000, 1.000)	
	GPT-4o	0.844 (0.781, 0.898)	
or>conjunction	Llama3-1-8B	0.750 (0.672, 0.820)	0.906 (0.859, 0.947)
	Llama3-1-8B-Instruct	0.758 (0.680, 0.828)	
	Llama3-2-1B	0.477 (0.391, 0.555)	
	Llama3-2-3B	0.695 (0.609, 0.773)	
	babbage	0.523 (0.445, 0.609)	
	davinci	0.562 (0.477, 0.648)	
	GPT-4o	0.875 (0.812, 0.930)	
or>either(pos)or	Llama3-1-8B	0.359 (0.281, 0.445)	0.529 (0.453, 0.600)
	Llama3-1-8B-Instruct	0.578 (0.484, 0.656)	
	Llama3-2-1B	0.078 (0.039, 0.133)	
	Llama3-2-3B	0.102 (0.055, 0.156)	
	babbage	0.227 (0.164, 0.297)	
	davinci	0.297 (0.219, 0.367)	
	GPT-4o	0.242 (0.172, 0.312)	