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

Anonymous ACL submission

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.

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.

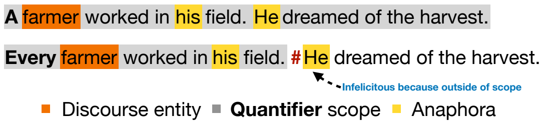

A farmer worked in his field. He dreamed of the harvest.
Every farmer worked in his field. #He dreamed of the harvest.
Infelicitous because outside of scope
■ Discourse entity ■ Quantifier scope ■ Anaphora

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

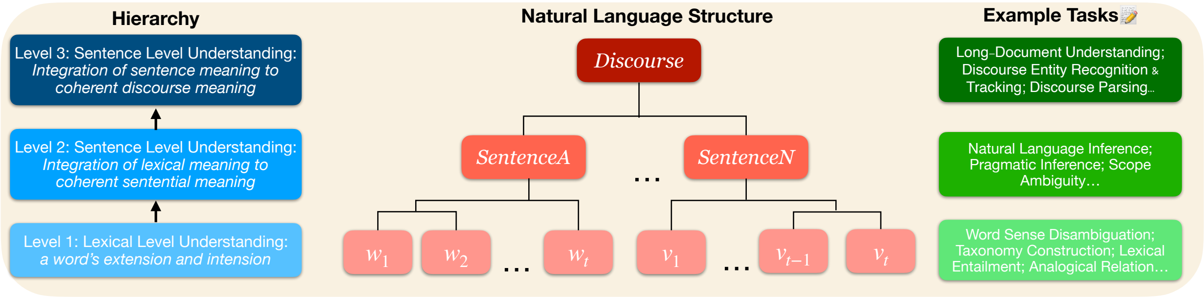

Hierarchy
Level 3: Sentence Level Understanding: Integration of sentence meaning to coherent discourse meaning
Level 2: Sentence Level Understanding: Integration of lexical meaning to coherent sentential meaning
Level 1: Lexical Level Understanding: a word's extension and intension
Natural Language Structure
Discourse
SentenceA
…
SentenceN
w_1
w_2
…
w_t
v_1
…
v_{t-1}
v_t
Example Tasks
Long-Document Understanding; Discourse Entity Recognition & Tracking; Discourse Parsing...
Natural Language Inference; Pragmatic Inference; Scope Ambiguity…
Word Sense Disambiguation; Taxonomy Construction; Lexical Entailment; Analogical Relation…

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

- 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 6, 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) summarize 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 derives 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 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) examine 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, the focus of the current work.

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 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 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 scope that interacts with discourse entities. This gap in the literature calls for a more comprehensive evaluation of **other scopes** (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.

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

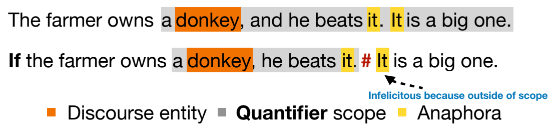

The farmer owns a donkey, and he beats it. It is a big one.
If the farmer owns a donkey, he beats it. # It is a big one.
Infelicitous because outside of scope
Discourse entity
Quantifier scope
Anaphora

Figure 3: Illustration of anaphora accessibility in donkey conditionals.

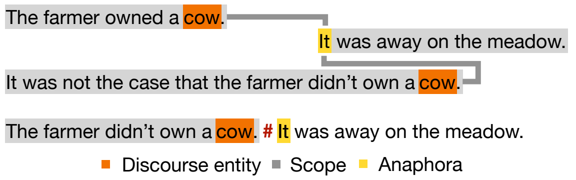

The farmer owned a cow.
It was away on the meadow.
It was not the case that the farmer didn't own a cow.
The farmer didn't own a cow. # It was away on the meadow.
Discourse entity
Scope
Anaphora

Figure 4: Illustration of anaphora accessibility in negation cases.

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).[1] 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 in why (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 anaphora *it* in Continuation. In contrast, no discourse referent of *a cow* 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

[1] 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.

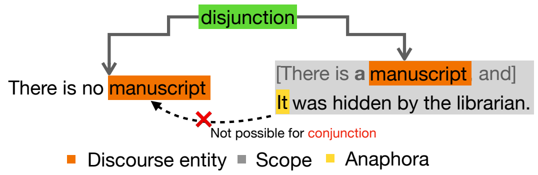

disjunction
There is no manuscript
[There is a manuscript and]
It was hidden by the librarian.
Not possible for conjunction
Discourse entity
Scope
Anaphora

Figure 5: Illustration of anaphora accessibility in disjunction cases.

(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.

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.[2] We ran inference using an NVIDIA A100 GPU with 32GB of memory allocated.

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 experimental sentences in total.

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.

4 Experiment 1: Universal

In this section, we discuss models' performance on anaphora accessibility with regard to the universal quantifier as discussed in Section 3.1.1.

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.

$$p(Cont|Exi) > p(Cont|Every) \quad (2)$$

$$p(Cont|Exi) > p(Cont|Cond) \quad (3)$$

$$p(Cont|Exi) > p(Cont|When) \quad (4)$$

[2] We were also not able to examine more recent OpenAI LLMs such as GPT-4o because the API for these models does not support access to the log probabilities. However, the perspective and evaluation tasks we propose in this paper are still helpful in informing the discourse-level semantic understanding of state-of-the-art LLMs.

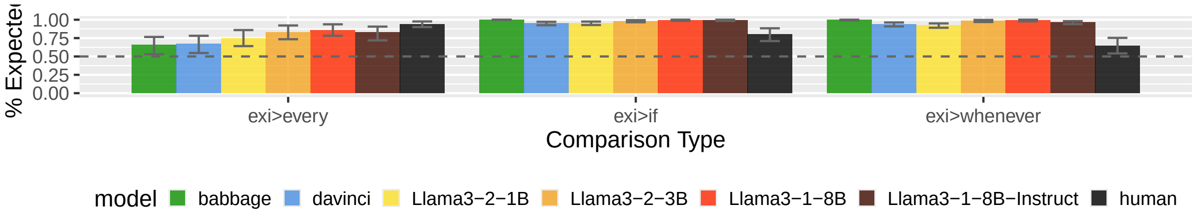


Figure 6: LLMs' performance on the comparisons involving existential vs. universal quantifiers. In the figures of this paper, > signs indicate degrees of felicity. For example, `exi>every`, the label for the leftmost panel, means that EXISTENTIAL should be more felicitous than EVERY sentences in the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

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 vest.

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.

$$\begin{gathered} p(\exists\text{-S}\textsc{ingle}\text{S}\textsc{en}) - p(\exists\text{-C}\textsc{ross}\text{S}\textsc{en}) \\ < \\ p(\forall\text{-S}\textsc{ingle}\text{S}\textsc{en}) - p(\forall\text{-C}\textsc{ross}\text{S}\textsc{en}) \end{gathered} \tag{5}$$

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, LLMs continue to score at ceiling. In contrast, humans had lower accuracy but still performed above chance. This pattern indicates that the LLMs 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.

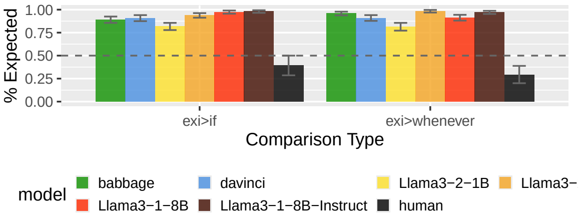


Figure 7: Model performance on *he*-continuations for `exi>if` and `exi>whenever`.

In addition to the continuation in (3d) that starts with *it*, for the comparisons `exi>if` and `exi>whenever`, we also considered a variant where the continuation starts with *he*, 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. Another potential factor that might contribute to the human performance difference between *he*- and *it*-continuations 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*. In sum, the models' success on this dataset shows their knowledge of the difference between universal and existential quantifiers.

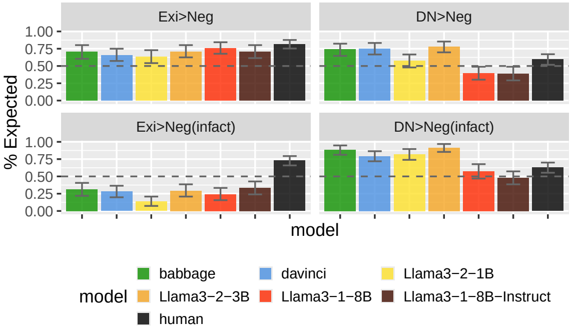


Figure 8: Model performance in Experiment 2.

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:

$$p(Cont|Exi) > p(Cont|Neg) \tag{6}$$

$$p(Cont|DN) > p(Cont|Neg) \tag{7}$$

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 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 helps the models to better 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 preference 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. 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.

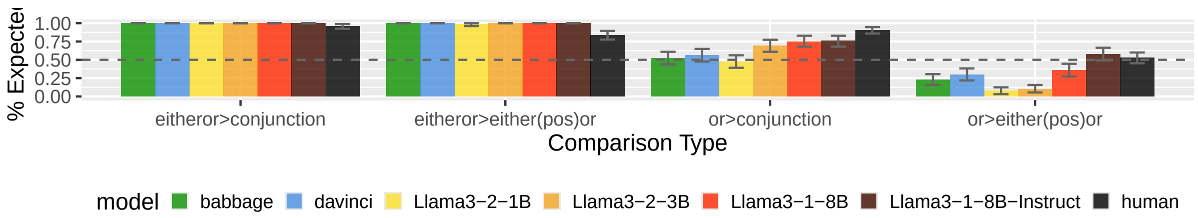


Figure 9: Model performance in Experiment 3.

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:

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

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-word, 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.[3]

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.

$$\texttt{SLOR}(\textsc{Or}) > \texttt{SLOR}(\textsc{Conjunction}) \quad (9)$$

$$\texttt{SLOR}(\textsc{EitherOr}) > \texttt{SLOR}(\textsc{Conjunction}) \quad (10)$$

$$\texttt{SLOR}(\textsc{Or}) > \texttt{SLOR}(\textsc{EitherPosOr}) \quad (11)$$

$$\texttt{SLOR}(\textsc{EitherOr}) > \texttt{SLOR}(\textsc{EitherPosOr}) \quad (12)$$

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

7 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.

[3] See the GitHub link for the unigram probability results.

Limitations

Running the Dataset in SOTA Models In the current study, we only tested our datasets with a limited range of LLMs. It would be interesting to see the performances of state-of-the-art language models such as GPT-4o and the DeepSeek model family. The main reason impeding us from testing our dataset on the latest models is that we require access to the logits the models assign to each token, which are not available for the closed-source models. Future studies could consider an alternative version of conducting such evaluation with prompting-based methods.

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.

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.

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.

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.

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.

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.

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.

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.

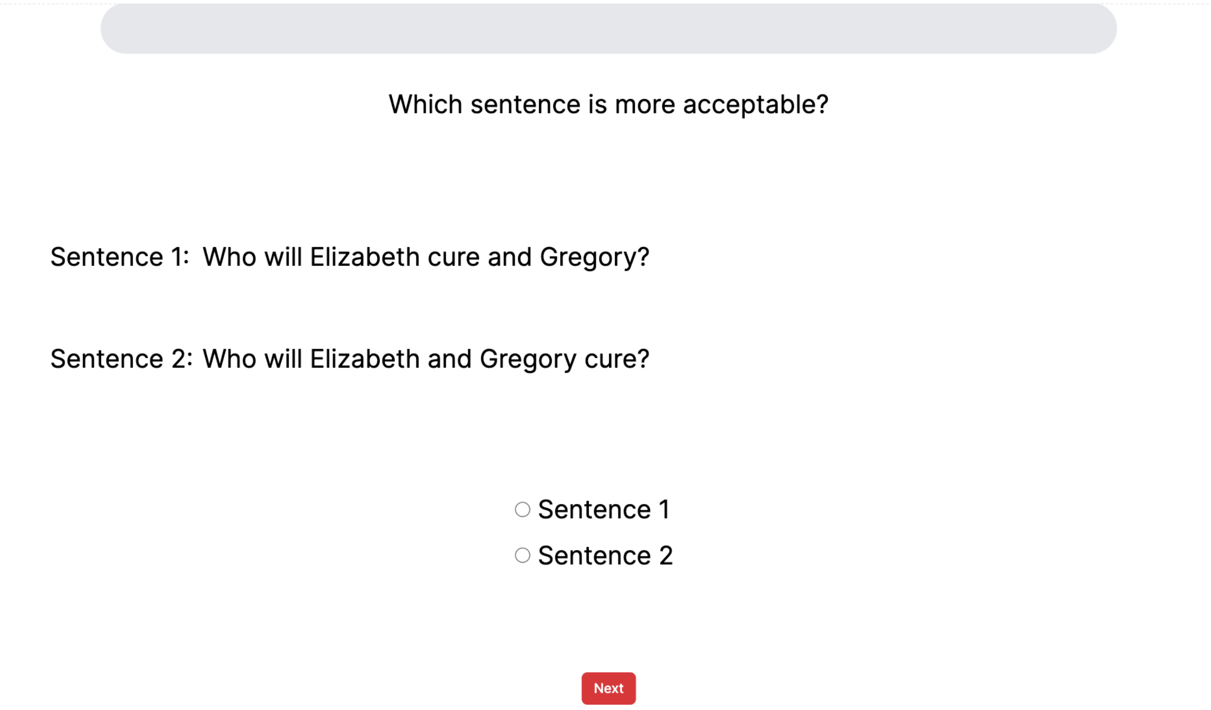
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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 sums 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 com-



Which sentence is more acceptable?

Sentence 1: Who will Elizabeth cure and Gregory?

Sentence 2: Who will Elizabeth and Gregory cure?

○ Sentence 1

○ Sentence 2

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

pensation of $3, which is equal to an hourly rate of $14.41.