LIEDER: Linguistically-Informed Evaluation for Discourse Entity Recognition

Xiaomeng Zhu
Department of Linguistics
Yale University
miranda.zhu@yale.edu

Robert Frank
Department of Linguistics
Yale University
bob.frank@yale.edu

Abstract

Discourse Entity (DE) recognition is the task of identifying novel and known entities introduced within a text. While previous work has found that large language models have basic, if imperfect, DE recognition abilities (Schuster and Linzen, 2022), it remains largely unassessed which of the fundamental semantic properties that govern the introduction and subsequent reference to DEs they have knowledge of. We propose the Linguistically-Informed Evaluation for Discourse Entity Recognition (LIEDER) dataset that allows for a detailed examination of language models' knowledge of four crucial semantic properties: EXISTENCE, UNIQUENESS, PLURALITY, and NOVELTY. We find evidence that state-of-the-art large language models exhibit sensitivity to all of these properties except NOVELTY, which demonstrates that they have yet to reach human-level language understanding abilities.

1 Introduction

One central component of language understanding is the ability to recognize entities in text. A large body of research in Natural Language Processing focuses on the task of Named Entity Recognition, where a system must identify whether a noun phrase (NP), typically a proper name, refers to a known individual of a certain semantic class (Li et al., 2022). The recognition of discourse entities (DEs), in contrast, involves identifying not only the occurrence of known entities but also novel ones that are introduced within a text. The recognition of DEs takes place at two sites: introduction sites and reference sites. Introduction refers to the first time where an entity appears in a discourse. Reference sites are subsequent mentions of an entity that has been previously introduced.

As humans, not only are we able to recognize DEs at both of these sites, but we also have knowledge of **how** to coordinate the introduction and subsequent reference to entities using appropriate linguistic means. For example, we know that the introduction of DEs is typically done using indefinite NPs such as *a man* in '*A man walked into the room*.' We also know that subsequent mentions often involve definite NPs like *the man* in '*The man sat down*.'

DE recognition is an important component of more complex semantic understanding tasks such as coreference resolution. Coreference relationships cannot be established between entities that have not been introduced into the discourse.

(1) a. John owns a dog. The dog is cute.

b. John doesn't own a dog. #The dog is cute.

For example, in (1a), the NP *the dog* in the second sentence and *a dog* in the first sentence refer to the same entity. This is not the case in (1b) because no entities have been introduced in the first sentence, which makes the continuation in (1b) infelicitous. Therefore, before establishing coreference relationships, language models first need to perform DE recognition,

Schuster and Linzen (2022) present an evaluation suite for DE recognition that focuses on the question of whether language models are sensitive to the linguistic context in which DEs are introduced. They find that transformer-based language models do not always demonstrate a clear preference for referring to entities that have been properly introduced into the discourse. While this work provides important insight into LM abilities with discourse reference, it does not engage directly with the underlying linguistic properties responsible for DE introduction and reference. As a result, it does not provide a means of assessing more precisely what LMs know about the linguistic encoding of discourse reference.

Semantics research has established properties of definite and indefinite NPs from which their use in

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13835–13850

introducing and referring to entities follows, four of which are particularly relevant here: EXISTENCE, UNIQUENESS, PLURALITY, and NOVELTY. We will define and discuss these properties in detail in Section 3. A good language model (LM) should reflect knowledge of all of these properties. In this paper, we provide a novel dataset, which builds on Schuster and Linzen's work, that provides a method of testing these properties directly.[1] Our results, across a number of state-of-the-art (SOTA) large language models (LLMs), provide evidence for knowledge of EXISTENCE, UNIQUENESS, and PLURALITY (all conditions on the use of definite NPs to refer to DEs), but difficulty with NOVELTY (a condition on the introduction of DEs by indefinite NPs) unless information about distinctiveness is made explicit. In addition, we find that transformer LMs, unlike humans, show strong sensitivity to linear distance in establishing DE reference. Taken together, these results suggest that SOTA LLMs do not reach human-level language understanding abilities.

2 Assessing Discourse Entity Recognition

Schuster and Linzen (2022) (henceforth SL) develop an evaluation suite that probes LLM performance on DE recognition. Such evaluation was centered on the ability of indefinite phrases to introduce DEs that can be referred to by subsequent occurrences of definites. They identified pairs of contexts that differ in their ability to support DE introduction.[2] The simplest case, and the one we focus on in our experiments, is *affirmative-negation*: indefinites in the object position of affirmative sentences introduce DEs that can be referred to by a following definite NP, but they do not in the object of sentences with negation. This is seen in the following example, where we use F to represent a felicitous completion, and I to represent an infelicitous completion:

(2) CONTEXT: John owns a dog but he doesn't own a cat.
   a. F: His dog follows him everywhere.
   b. I: His cat follows him everywhere.

SL argue that if a language model is sensitive to the difference between contexts which do and do not introduce DEs (such as the two conjuncts in the first sentence in (2)), we would expect the following inequality to hold for the probabilities assigned to felicitous (F) and infelicitous (I) continuations:

$$p(\text{F}|\text{CONTEXT}) > p(\text{I}|\text{CONTEXT})$$

They found that the models they examined (which included GPT-2 variants and GPT-3) showed above chance performance on distinguishing F and I continuations in the context of *affirmative-negation*, with GPT-3 showing the most human-like performance.

SL further explored the systematicity of these contrasts, where a contrast is counted as systematic only when the model correctly predicts felicity on all possible orderings of the operators and nouns in the conjoined clauses that comprise the context:

(3) a. Bob owns a dog but he doesn't own a cat.
   b. Bob owns a cat but he doesn't own a dog.
   c. Bob doesn't own a cat but he owns a dog.
   d. Bob doesn't own a dog but he owns a cat.

The relative order of the affirmative and negative sentences should not impact the felicity of subsequent definite descriptions. The continuation *His dog follows him everywhere* is felicitous after either (3a) or (3c). Similarly, the continuation *His cat follows him everywhere* remains felicitous after (3b) or (3d). However, SL find that the models are much less successful under this more demanding measure; even GPT-3 systematically distinguishes felicitous and infelicitous continuations only barely above 50% of the time, with other models showing lower performance. Schuster and Linzen do not, however, identify what underlies the models' failure in systematic performance on this task.

While intriguing, these results do not tell us what specifically causes difficulty for LLMs in this task, and how it relates to semantic properties of definite and indefinite NPs. To evaluate language models' DE recognition abilities, we believe that it is important to decompose models' performance in a more granular manner. We turn now to a paradigm that builds on this previous work to do precisely this.

3 Criteria for Discourse Entity Recognition

Informed by theoretical research in natural language semantics, we propose that a thorough evaluation of DE recognition and reference abilities should examine language models' knowledge of four fundamental semantic properties: EXISTENCE,

[1] All code, data, and results are available at https://github.com/xiaomeng-zhu/LIEDER.

[2] For a review of the pairs of contexts introduced by Schuster and Linzen (2022), see Appendix B.

UNIQUENESS, PLURALITY, and NOVELTY. We will define and discuss each of them in order.

Existence As SL argued, a language model with human-level understanding abilities should only use definite descriptions to refer to entities that have been introduced into the discourse. We define this requirement as EXISTENCE (Russell, 1905). For example, given the context *John doesn't own a dog*, a language model should recognize that the continuation *The dog barks at night* is infelicitous because of the non-existence of a dog DE.

Uniqueness A language model should use a singular definite description to refer to a previously introduced entity only when the referent is unique relative to the discourse. We will call this requirement UNIQUENESS (Russell, 1905; Heim and Kratzer, 1998). For example, given the context *John owns a dog and Mark owns a dog too*, the model should consider the continuation *The dog barks at night* as infelicitous. Since more than one dog has been introduced into the discourse, the singular definite phrase is not licensed. On the other hand, if the context is *John does not own a dog but Mark owns a dog*, UNIQUENESS is satisfied, so the same continuation should be judged to be felicitous.

Plurality A language model should use a plural definite description only if the set of DEs contains more than one individual of the relevant sort, a requirement we call PLURALITY (Landman, 1989). Notice that UNIQUENESS and PLURALITY cannot be satisfied or violated at the same time – referent expressions that require uniqueness (i.e., singular definites) are infelicitous after contexts that satisfy PLURALITY. In contrast, referent expressions that require plurality (i.e., plural definites) are infelicitous after contexts that satisfy UNIQUENESS. For example, the context *John owns a dog and Mark owns a dog too* satisfies PLURALITY since there are two dog DEs, and therefore supports a plural but not singular continuation (i.e., *The dogs bark*/**dog barks*).

Novelty The last requirement concerns the use of indefinite NPs: a language model should recognize that an occurrence of an indefinite noun phrase is associated with the introduction of a new entity into the discourse. Following Heim (1982), we will call this requirement NOVELTY. In the context sentence *John owns a dog and Mark owns a dog too*, this means that two distinct dogs are introduced as DEs.

Expression	Requirements
Indefinites	NOVELTY
Singular definites	EXISTENCE, UNIQUENESS
Plural definites	EXISTENCE, PLURALITY

Table 1: Expressions used for introducing and referencing DEs and their corresponding requirements.

Table 1 summarizes relevant expression types and the corresponding requirements that a language model should know in order to correctly introduce and refer to DEs. In the next section, we will describe our evaluation dataset and show how it evaluates model performance with respect to the four requirements.

4 The LIEDER Dataset

The Linguistically-Informed Evaluation of DE Recognition (LIEDER) evaluation dataset adapts the structure of SL's paradigm: a context example is provided that consists of two conjoined clauses, each containing an indefinite NP with the same head noun. This is followed by a continuation, test sentence containing a definite description. As in SL, we vary the conjoined clauses as to whether they introduce DEs or not (affirmative or negative). However, in order to evaluate the four linguistic properties defined in the previous section, we make two innovations: (i) we allow zero, one, or both of the conjoined clauses to include negation and thereby fail to introduce a DE; and (ii) we allow the definite description in the continuation to be either singular or plural. Example items in our dataset are shown in Table 2.

The *Context type* column indicates which sides of the conjunction introduce DEs. For example, `pos_neg` indicates that a DE is introduced in the first conjunct (`pos`), and no DEs are introduced in the second conjunct (`neg`). Since we consider all four context types with both singular and plural continuations, there are 8 different context-continuation combinations. These 8 combinations are then crossed with 16 distinct base sentence pairs for the context and continuation, resulting in 128 examples in total.[3]

[3] SL consider other pairs of sentential operators that differ in whether they introduce DEs, specifically *managed-failed* and *know-doubt*. The LIEDER dataset also includes contexts that combine these sentential operators, which sums to 384 examples in total. See the Appendix for results regarding these other operator contrasts.

Context type	Context	Singular Continuation	Plural Continuation
pos_neg	John owns a dog but Mark doesn't own a dog.	The dog is very cute.	#The dogs are very cute.
neg_pos	John doesn't own a dog but Mark owns a dog.	The dog is very cute.	#The dogs are very cute.
pos_pos	John owns a dog and Mark owns a dog too.	#The dog is very cute.	The dogs are very cute.
neg_neg	John doesn't own a dog and Mark doesn't own a dog either.	#The dog is very cute.	#The dogs are very cute.

Table 2: Example contexts and continuations in the LIEDER dataset. Infelicitous continuations are marked with #.

Different combinations of context type and continuation result in differences in felicity of our two-sentence discourse. The singular continuation is felicitous only when the context type is either pos_neg or neg_pos since these contexts introduce exactly one DE. In contrast, the plural continuation is felicitous only when the type is pos_pos. This means that of the eight context-continuation pairs, three are felicitous and five are infelicitous.

Success in the LIEDER dataset requires that a model accurately distinguish felicitous from infelicitous context-continuation pairs. Because metalinguistic judgments elicited from language models may not reflect the full extent of the model's knowledge (Hu and Levy, 2023), we instead compare felicity using the probabilities the model assigns to the continuation given the context. We assume that the probability a model assigns to a felicitous case should be greater than the probability it assigns to an infelicitous one. With 3 felicitous pairs and 5 infelicitous ones, this means we have 15 informative probability comparisons in total. These are depicted in Table 3.

Importantly, success in each of these comparisons can be tied to the linguistic requirements described in Section 3 that are involved in the introduction of and reference to DEs. For example, if the model assigns higher probability to a continuation with a singular definite in a neg_pos context as compared to a neg_neg context, this provides evidence for the model's awareness of the EXISTENCE requirement that singular definite NP imposes; otherwise, the singular should be possible in this context. On the other hand, if the model assigns higher probability to a singular definite in a pos_neg context as compared to a pos_pos context, this indicates that the model is aware of the UNIQUENESS requirement imposed on singular definiteness and the NOVELTY condition imposed on indefinites, since otherwise the pos_pos context could be taken to introduce only a single DE. In addition, if the model assigns higher probability to a singular definite in a neg_pos context than it does to a plural definite in the same context, this provides evidence of sensitivity to the PLURALITY requirement on plural definites, as the context introduces only a single entity. Because of the correspondences between the example types and the linguistic requirements, LIEDER can therefore be used to assess the details of the knowledge of DEs in a language model.

Comparison Type	Requirement	Section
p(sg|pos_neg)>p(sg|pos_pos)	UNIQUENESS, NOVELTY	5.1.1
p(sg|neg_pos)>p(sg|pos_pos)	UNIQUENESS, NOVELTY	5.1.1
p(sg|neg_pos)>p(sg|neg_neg)	EXISTENCE	5.1.1
p(sg|pos_neg)>p(sg|neg_neg)	EXISTENCE	5.1.1
p(pl|pos_pos)>p(pl|pos_neg)	PLURALITY	5.1.2
p(pl|pos_pos)>p(pl|neg_pos)	PLURALITY	5.1.2
p(pl|pos_pos)>p(pl|neg_neg)	EXISTENCE, PLURALITY	5.1.2
p(sg|pos_neg)>p(pl|pos_neg)	PLURALITY	5.1.3
p(sg|pos_neg)>p(pl|neg_pos)	PLURALITY	5.1.3
p(sg|pos_neg)>p(pl|neg_neg)	EXISTENCE, PLURALITY	5.1.3
p(sg|neg_pos)>p(pl|neg_pos)	PLURALITY	5.1.3
p(sg|neg_pos)>p(pl|pos_neg)	PLURALITY	5.1.3
p(sg|neg_pos)>p(pl|neg_neg)	EXISTENCE, PLURALITY	5.1.3
p(pl|pos_pos)>p(sg|pos_pos)	UNIQUENESS, NOVELTY	5.1.3
p(pl|pos_pos)>p(sg|neg_neg)	EXISTENCE	5.1.3

Table 3: Comprehensive list of all comparison types, the requirements they test, and the experiment that tests for them.

5 Experiment 1: Applying LIEDER

Models We investigated the performance of five open-source (Llama 2-7B, 13B, and 70B (Touvron et al., 2023), Llama 3-8B and 70B (Meta AI, 2024)) and two closed-source LLMs (GPT babbage-002 and davinci-002) on LIEDER through the Huggingface transformer API (Wolf et al., 2019) and the OpenAI API respectively.[4]

Metric To perform the probability comparisons discussed in Section 4, we provided the model with the context of each test item and calculated the total log probability for the entire continuation. For each comparison type, we compare the log probability of a felicitous continuation with an infelicitous one, and judge the model as accurate if the felicitous

[4]The number of parameters in babbage-002 and davinci-002 are not publicly available. We were also not able to examine more recent LLMs released by OpenAI because the API for these models does not support access to the log probabilities of prompts.

probability is higher. We then calculate the accuracy for each comparison type over the 16 items of each type.

Human Judgments In addition to evaluating large language models with our dataset, we also conducted an experiment to elicit human judgments. Specifically, participants were asked to provide a rating on the acceptability of a continuation given a context using a continuous slider with possible values ranging from 1 to 7. Each participant provided ratings for 16 experimental sentences (one for each context type-continuation pair), each with different lexical content. This means that comparisons of items were done across subjects. Using the Prolific platform, we recruited 80 participants who were native English speakers with perfect or corrected vision and without language disorders. Acceptability ratings on the same sentences were averaged, and the resulting averages were compared within an item to compute accuracies that could be compared to the language model results. See Appendix A for our experimental interface and more details on methodology.

5.1 Results

Because all continuations in LIEDER involve definite noun phrases, we refer to singular definite continuations simply as singular continuations and plural definites as plural continuations. Discussion of the results will be broken down based on the plurality of the continuations involved in the comparisons.

5.1.1 Singular Continuations

Singular continuations are felicitous after contexts where one and only one relevant DE is introduced (those that satisfy both EXISTENCE and UNIQUENESS), namely pos_neg and neg_pos. On the other hand, they are infelicitous for contexts that are of the type pos_pos (violating UNIQUENESS) and neg_neg (violating EXISTENCE). As a result, if the language models we are examining reach human-level understanding abilities, we would expect that the probabilities associated with the two felicitous cases are greater than the two infelicitous cases respectively, which gives us four comparison types.

The results for these comparisons are shown in Figure 1, respectively in the four panels. As the first two panels in the figure show, all models (and humans) have ceiling or near ceiling performance on dispreferring singular continuations that follow contexts where no DEs have been introduced (neg_neg). In other words, for a given singular continuation, they have a strong preference for contexts where one and only one DE is introduced over ones where none are introduced. Such preference indicates that all models know EXISTENCE.

In contrast, in the last two panels in Figure 1, we see that model accuracies are uniformly lower when the infelicitous context is pos_pos as compared to pos_neg or neg_pos (i.e., two relevant DEs are introduced). This drop in accuracy suggests that the language models do not consider pos_pos to be worse than pos_neg or neg_pos in licensing the same singular definite continuations. As a side note, a greater number of model parameters does not necessarily translate into higher performance. In fact, increasing parameter count in the Llama 2 series yields ever worse performance in panel 4.

Why might pos_pos contexts be confusing for large language models? With respect to the linguistic requirements on DE introduction and reference, there are two possible answers to this question:

Hypothesis 1

During training, the models have successfully learned the EXISTENCE requirement, but they failed to learn UNIQUENESS.

Hypothesis 2

During training, the models have successfully learned both the EXISTENCE and UNIQUENESS requirements but fail to recognize that two distinct DEs have been introduced in pos_pos contexts, resulting in difficulties in distinguishing the infelicitous pos_pos from felicitous pos_neg and neg_pos. To put it in another way, they fail at the NOVELTY requirement.

At this stage, we lack evidence that supports one hypothesis over another. Experiment 2 focuses on teasing these two hypotheses apart.

From the last two panels in Figure 1, we can see that accuracy of all LMs for p(sg|neg_pos)>p(sg|pos_pos) is uniformly higher than p(sg|pos_neg)>p(sg|pos_pos) across models, while human performance differs very little. We suggest that the source of this contrast is a preference for singular definites in the context of neg_pos over pos_neg, both of

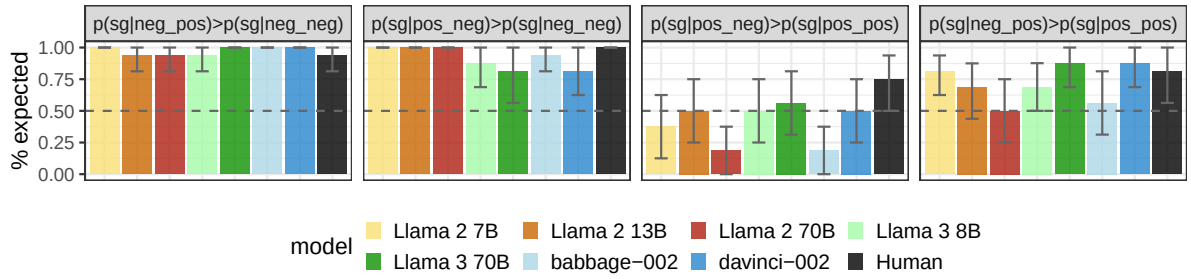

Figure 1: Results for singular continuations by model and comparison type. The dotted lines indicate chance performance and the error bars indicate bootstrapped 95% confidence intervals.

which ought to be felicitous contexts. Figure 2 evalulates this claim, illustrating the percentage of test examples where models and humans consider neg_pos to a better context for singular definite continuations than pos_neg. All models exhibit this preference for over half of the examples. Interestingly and perhaps surprisingly, humans also demonstrate a distance effect, showing a preference for neg_pos over pos_neg contexts for singular definites on a majority of examples. Thus, the presence of DISTANCE sensitivity need not be interpreted to be a deficiency of LM performance, but perhaps is a reflection of patterns in human language use. Nonetheless, though humans show a DISTANCE effect, they demonstrate systematic awareness of the UNIQUENESS and NOVELTY requirement, leading to 75% accuracy on p(sg|pos_neg)>p(sg|pos_pos), where no models except for Llama 3 70B achieve greater than chance performance.

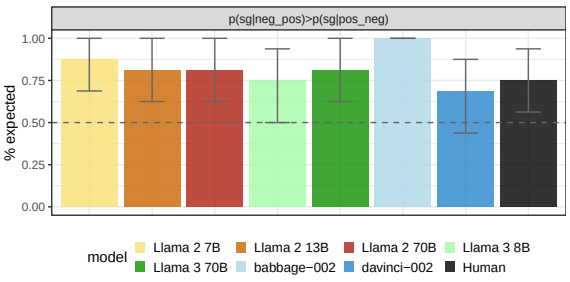

Figure 2: Preference for neg_pos over pos_neg by model.

The presence of this distance effect provides a possible explanation of the failure in systematic behavior that SL observed in their work. A reanalysis of their data (which involved only pos_neg and neg_pos cases) finds that the position of the DE-introducing sentence impacts model accuracy (Figure 3; $p = 0.0161$).[5]

5.1.2 Plural Continuations

Plural continuations are felicitous only following pos_pos contexts. Hence, we would expect the probability of a plural continuation given pos_pos to be greater than those given pos_neg, neg_pos, and neg_neg.

Results for these three comparisons are shown in Figure 4. All models (and humans) exhibit near-ceiling accuracy, which demonstrates that out of the four possible contexts, they consider pos_pos to be the best one prior to plural continuations, consistent with human judgments. The lower accuracy of Llama 3 70B compared to all other models again supports the observation from Figure 1 that larger models do not always perform better than smaller ones in terms of the properties we identified in LIEDER. Regardless, the models' ceiling performance on plural continuations serves as evidence that they have learned both the existence and the plurality requirements for plural definite descriptions.

5.1.3 Comparing Singluar and Plural Continuations

We finally compare across singular and plural continuations. There are 8 such comparison types in total. We will focus on three of them that are particularly informative, which are shown in Figure 5. See Appendix B for results of all comparisons.

For p(sg|neg_pos)>p(pl|neg_pos) and p(sg|pos_neg)>p(pl|pos_neg), all of the LLMs achieve near-ceiling accuracy, consistent

[5]We applied a linear mixed-effect model using the lme4 (Bates et al., 2015) library in R with a main effect of DISTANCE and a random effect of ITEM, collapsing across different models. We also examined the effect of DISTANCE with respect to other sentence types, where DISTANCE is also significant. See Appendix C for the corresponding plots and significance testing.

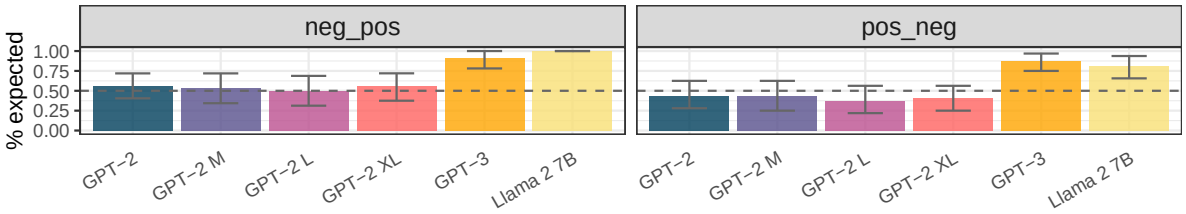

Figure 3: Decomposition of results for *affirmative-negation* type sentences in Schuster and Linzen (2022) by DISTANCE. Data for GPT-2, GPT-2 M, GPT-2 L, GPT-2 XL, and GPT-3 are retrieved from their GitHub Repository.

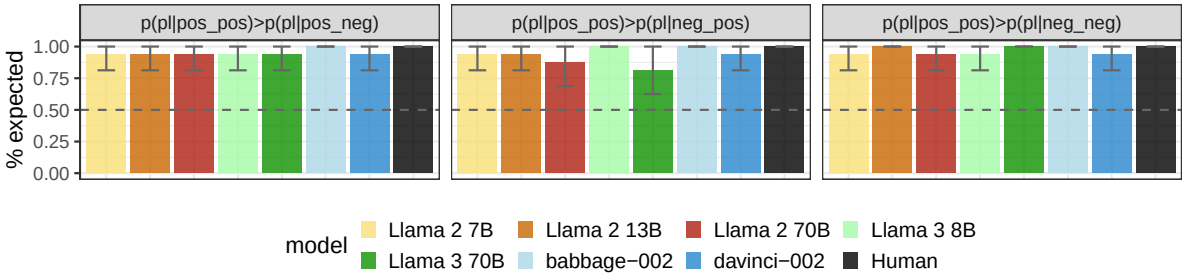

Figure 4: Results for plural comparisons by model and comparison type.

with human judgments. The high performance on these two comparisons suggests a preference for singular continuations over plural ones when one and only one relevant entity has been introduced into the discourse using a singular definite description. This preference reflects models' knowledge of UNIQUENESS, which speaks against **Hypothesis 1** that we proposed in Section 5.1.1. If the models know UNIQUENESS, why do they perform at chance for the comparison `p(sg|pos_neg)>p(sg|pos_pos)` in Figure 1? The only possibility left is **Hypothesis 2**: they take a singular continuation following a `pos_pos` context to be possible because they do not recognize that two distinct DEs have been introduced.

The `p(pl|pos_pos)>p(sg|pos_pos)` comparison provides further support for `pos_pos` being a difficult context for the models. We saw above that all models consider `pos_pos` to be the best context preceding plural continuations. Hence, the problem must be that somehow the models assign an incorrectly high probability to `p(sg|pos_pos)` because of failure to enforce the NOVELTY condition with two indefinites. We test this hypothesis further in Experiment 2 (and Experiment 3 in the Appendix).

6 Experiment 2: Facilitating Novelty

As described in Section 4, `pos_pos` contexts introduce two DEs using two instances of the same indefinite description. For example, in the context sentence *John owns a dog and Mark owns a dog*, two dogs have been introduced using the same indefinite description *a dog*. However, language models might have difficulty understanding that these are two different dogs because not only do they need to recognize DEs through distinct occurrences of the indefinite description *a dog*, but they must also know the NOVELTY condition to consider these distinct occurrences as distinct DEs.

6.1 Dataset

One way to test if LLMs fail to recognize two different DEs in `pos_pos` contexts is to use lexical cues that make explicit the distinctness of the first and second entities. If performance relative to `pos_pos` contexts increases when the distinction is explicit, then there is evidence that the LLMs fail to recognize the distinction in the implicit case, where the presence of multiple DEs results from the NOVELTY condition on indefinites alone. Accordingly, we make the following modification to our dataset: for each context of the type `pos_pos`, we add the adjective "different" to the second indefinite description:

(4) a. *Implicit*: John owns a dog and Mark owns a dog.

b. *Explicit Novelty*: John owns a dog and Mark owns a different dog.

Other contexts and continuations are kept the same.

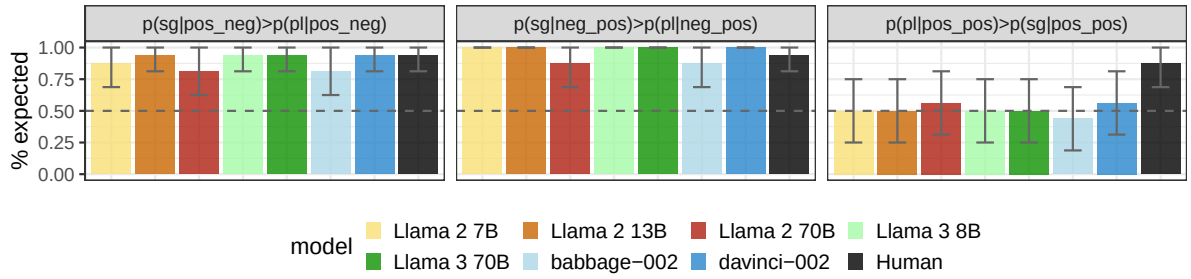

Figure 5: Results by model and comparison type for comparisons across singular and plural continuations.

6.2 Results

Results for the felicity comparisons are shown in Figure 6. To quantify the effects of our manipulations, we fit a linear mixed-effect model with two fixed effect predictors: VERSION (a categorical variable with two levels: *Implicit* or *Explicit Novelty*) and COMPARISON TYPE, and ITEM as a random effect. This model gives a significant coefficient for VERSION ($p < 0.001$) indicating that accuracy increases significantly from *Implicit* to *Explicit Novelty*. This increase supports **Hypothesis 2**, i.e., difficulty with the NOVELTY condition. When the distinctness of the two indefinites is made lexically explicit, the models can better recognize that two distinct DEs are introduced in `pos_pos` contexts. In Appendix D, we show that another way of cueing the creation of multiple DEs by explicitly supplying information about plurality achieves the same effect.

7 Discussion

In Experiment 1, we saw that all LLMs displayed clear knowledge of EXISTENCE and PLURALITY. They inarguably consider singular continuations to be bad after contexts that introduce no DEs and plural continuations to be good only when more than one DE has been introduced. However, they failed to show clear knowledge of UNIQUENESS, as indicated by their at-chance performance for the comparison type `p(sg|pos_neg)>p(sg|pos_pos)`. We believe that such underperformance is not due to a lack of knowledge of UNIQUENESS because the results in Figure 5 support knowledge of the UNIQUENESS requirement. Instead, we believe that the problem lies in NOVELTY. Specifically, the models cannot identify that `pos_pos` contexts introduce two distinct DEs.

We suspect that the `pos_pos` context might be particularly challenging because the two DEs are introduced using identical indefinite NPs (i.e., the two occurrences of *a dog* in *John owns a dog and Mark owns a dog too*). One might imagine that it is often sufficient to associate unique NPs with distinct DEs. For example, given the sentence *John owns a cat and a dog*, the information that *a cat* and *a dog* are two distinct DEs is lexically encoded. If a model adopts such a lexically dependent strategy for DE introduction, then the `pos_pos` cases from Experiment 1 are exactly the expected point of failure, as such cases require sensitivity to NOVELTY in order to succeed. Indeed, as shown in Experiment 2, supplying explicit information about distinctiveness does improve model performance.

Taken together, these results all point to language models' mastery of EXISTENCE, UNIQUNESS, and PLURALITY and a lack of knowledge on NOVELTY when minimal information is given.

Distance Effect Results from Experiment 1 also show a clear effect of DISTANCE. In our reanalysis of SL, LMs did a better job of recognizing DEs when the entities are introduced closer to the continuation that refers to them. Within results from our own evaluation suite, there is also a clear preference for introducing DEs closer to the definite description when the singular continuations are equally felicitous. It seems as if as the sentence unfolds, the status of a DE as one that can be referred to gradually decreases. In other words, there is a greater cost associated with referring to a DE that is introduced earlier in time than those that are introduced later.

8 Related Work

The current paper contributes to the body of literature that examines the semantic knowledge of neural language models. Kim et al. (2019) assessed LLM comprehension of functional words, including a subtask of differentiating indefinite and defi-

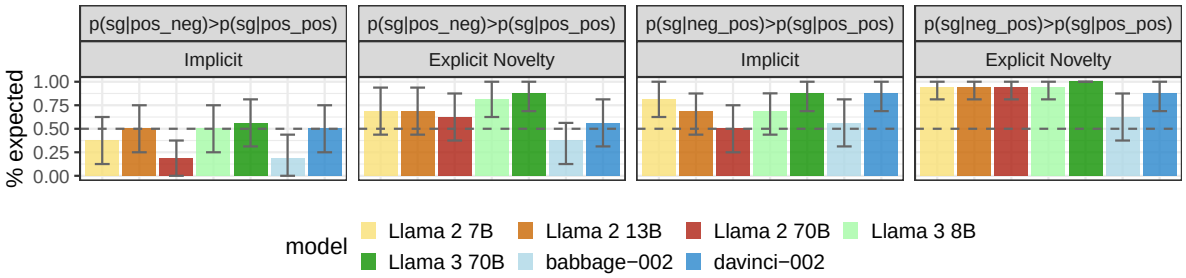

Figure 6: Experiment 2 results by model, version, and comparison type for singular continuations.

nite determiners. Their stimuli were constructed by swapping *a/an* with *the* in a sentence. Such manipulation implicitly encodes the properties identified in LIEDER: using *the* in the place of *a* often violates UNIQUENESS and EXISTENCE, and NOVELTY could also be violated vice versa.

On the discourse level, aside from Schuster and Linzen (2022) from which the current paper draws inspiration, Upadhye et al. (2020) examined contexts with different biases concerning the DE that will be mentioned next. They found that unlike humans, GPT-2 (Radford et al., 2019) and Transformer-XL (Dai et al., 2019) are not sensitive to the manipulation of contexts when predicting the entity that will be mentioned next. Loáiciga et al. (2022) built probing models to investigate whether pretrained representations encode information about an entity being newly introduced or having been mentioned before (which they call discourse-new/-old respectively). They found that while a high F1 score can be achieved for the classification of discourse-new vs. discourse-old, models struggle to locate the entities within a sequence, a finding that is in line with our result that models lack understanding of the NOVELTY requirement. Lastly, Kim and Schuster (2023) studied LLM ability to accurately represent the states of discourse entities across long narratives and observed that only LLMs that have been pretrained on code exhibited non-trivial entity tracking abilities.

The LIEDER dataset also adds to the body of work that uses linguistic insights in the development of semantic benchmarks for LLMs. This work includes COGS (Kim and Linzen, 2020), ReCOGS (Wu et al., 2023), and SLOG (Li et al., 2023) on compositional generalization and IMPPRES (Jeretic et al., 2020), NOPE (Parrish et al., 2021), Kim et al. (2021), and $(QA)^2$ (Kim et al., 2023) that assess models' ability in handling presuppositions.

9 Conclusion

In this paper, we proposed the Linguistically-Informed Evaluation for Discourse Entity Recognition (LIEDER) dataset. Our paradigm allows for a detailed examination of language models' knowledge of four properties of definite and indefinite NPs concerning their ability to introduce and refer to DEs. These properties are EXISTENCE, UNIQUENESS, PLURALITY, and NOVELTY. We demonstrated that despite mastering EXISTENCE, UNIQUENESS, and PLURALITY, the LLMs we examined lack understanding of the NOVELTY requirement. In spite of this deficiency, we showed that language models of the Llama 2, Llama 3, and GPT series reflected the human preference of referring to DEs that are introduced closer to their reference point, which we label an effect of DISTANCE. We recognize that given the fast-paced development of LLM research, it is highly likely that the performance presented in the current paper will be surpassed by future generations of LMs. However, the success of LIEDER in helping to identify language models' deficiency in DE recognition highlights the importance of linguistic considerations in evaluating the strengths and weaknesses of future language models.

Ethics Statement

Limitations The current study only focuses on English, which has overt determiners for indefinite and definite NPs. There are other languages that do not have determiners equivalent to *a* and *the* in English. For example, Mandarin Chinese makes use of demonstratives that can serve similar purposes as English determiners. Our evaluation paradigm is thus limited in that it cannot be directly used to evaluate DE recognition on language models trained on other languages without considering language-specific properties relating to DE introduction and

reference.

Risks All participants in the human experiment were recruited through Prolific under the approval of the Yale University IRB. At the beginning of the experiment, they were presented with consent forms that indicated the potential risks associated with participation, and only those who consented were allowed to proceed with the task. The risk was minimal. All identifier information has been removed from the data to guarantee anonymity. Participants received compensation equivalent to $12/hr, which is around 70% higher than the federal minimum wage.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell, Malden, MA.

Irene Roswitha Heim. 1982. *The semantics of definite and indefinite noun phrases*. University of Massachusetts Amherst.

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.

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.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. (QA)^2: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. 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 3932–3945, Online. Association for Computational Linguistics.

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.

Fred Landman. 1989. Groups, i. *Linguistics and philosophy*, pages 559–605.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A structural generalization benchmark for semantic parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Sharid Loáiciga, Anne Beyer, and David Schlangen. 2022. New or old? exploring how pre-trained language models represent discourse entities. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 875–886, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Bertrand Russell. 1905. On denoting. *Mind*, 14:479–493.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shiva Upadhye, Leon Bergen, and Andrew Kehler. 2020. Predicting reference: What do language models learn about discourse models? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 977–982, Online. 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*.

Zhengxuan Wu, Christopher D. Manning, and Christopher Potts. 2023. ReCOGS: How Incidental Details of a Logical Form Overshadow an Evaluation of Semantic Interpretation. *Transactions of the Association for Computational Linguistics*, 11:1719–1733.

A Human Experiments

A total of 80 participants were recruited through Prolific under the approval of the Yale University IRB. Screener conditions were set such that all participants were native English speakers with perfect or corrected vision and without language disorders.

We used the Gorilla experimental platform to present a context sentence for a period of 300ms per word, and then presented the continuation sentence, again for 300ms per word. Participants then moved the dot on the slider scale appearing below to indicate their judgment. Figure 7 demonstrates this experimental interface.

Each participant received 26 trials in total which were composed of 2 practice trials and 16 target trials with 8 control trials that appeared in random order. In the practice trials, they were instructed to move the slider all the way to the right if they thought the continuation was perfectly acceptable, and all the way to the left if clearly unacceptable. In the 16 target trials, they were exposed to the same set of stimuli as the language models corresponding to 16 different sentence frames (i.e. ITEM). The filler trials were the same for each participant where the judgments for each of them were either strictly felicitous or infelicitous. Each participant received compensation that averaged $12.20/hr.

Among the initial 80 participants we recruited, data from four participants was excluded because they failed to answer 7 of the 8 control items correctly.

B Supplementary Plots for Experiment 1 and 2

All results reported in the experiments were concerning *affirmative-negation* sentences. SL identify four pairs of operators where the operators in each pair differ in the way they modulate DE introduction. They are *affirmative-negation*, *affirmative-modal*, *know-doubt*, and *managed-failed*, which is exemplified in Table 4. The full LIEDER dataset includes conjoined sentences of all of these types except for *affirmative-modal*. This decision is based on our judgment that it is easier to get a wide-scope reading of indefinites when they are embedded in modals than when they are embedded in negations.

(5) John wants to own a dog and Mark owns a dog. The dog is cute.

For example, according to the design in LIEDER, (5) is of type `neg_pos`, and it is intended that the second conjunct introduces a discourse entity but not the first one. However, it is easy to get the reading that there is a specific dog that Mark wants to own, thus making the singular definite infelicitous. To avoid complexities like this, we decided to focus our analysis on the other three contrast types instead.

In the rest of this section, we show results for two other types of sentences: *know-doubt* and *managed-failed*. Note that we did not elicit human judgments for *know-doubt* sentences as some of these sentences are too long to format on the Gorilla interface.

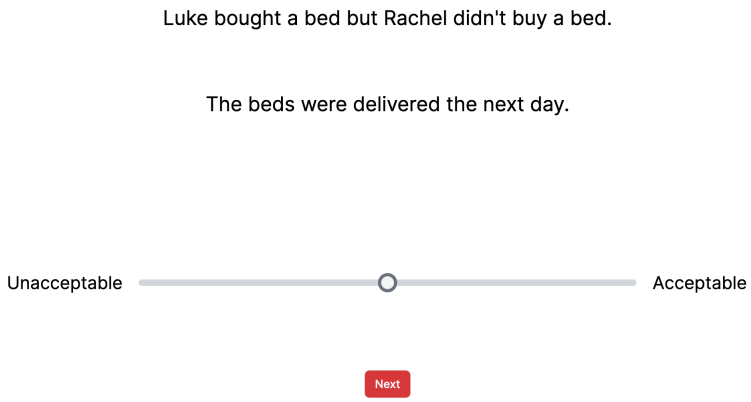

Figure 7: Experimental Interface on Gorilla. The first sentence on the screen is a target item of the pos_neg category Since the continuation is plural, it is expected to be unacceptable if following the context.

Operator Type	pos	neg
affirmative-negation	John owns a dog.	John doesn't own a dog.
affirmative-modal	John owns a dog.	John wants to own a dog.
know-doubt	I know that John owns a dog.	I doubt that John owns a dog.
managed-failed	John managed to adopt a dog.	John failed to adopt a dog.

Table 4: Four pairs of sentential operators introduced by SL. The pos column indicates cases where DEs are introduced. The neg column indicates cases where DEs are not introduced. All operator types are included in LIEDER except for *affirmative-modal*.

B.1 Experiment 1

Figure 8 corresponds to singular continuations, Figure 9 plural continuations, and Figure 10 for singular and plural comparisons.

B.2 Experiment 2

Figure 11 shows the increase from *Implicit* to *Explicit Novelty* for all three types of sentences. The accuracy increase from *Implicit* to *Explicit Novelty* is significant ($p < 0.001$) under the same linear mixed-effect model specified in Section 6 that collapses across language models and sentence type.

C Effects of DISTANCE in SL

In Section 5.1.1, we showed the effect of DISTANCE on *affirmative-negation* sentences. Here in Figure 12, we provide a comprehensive plot showing the effect of distance on all four types of sentences. The same linear mixed-effect model was applied. The main effect of DISTANCE is still significant ($p < 0.001$).

D Experiment 3: Plural Indefinites

We conducted a third experiment, where we introduced a third type of conjunct besides `pos` and `neg` which we call `two`. In the `two` conjunct, a plural indefinite is used as an explicit cue that there is more than one relevant entity in the discourse. Consider the following distinction:

(6) a. *Implicit*: John owns a dog and Mark owns a dog too. (`pos_pos`)

b. *Explicit Plurality*: John owns two dogs and Mark doesn't own a dog. (`two_neg`)

Both (6a) and (6b) involve the introduction of two dogs into the discourse. However, in (6a), the NOVELTY condition is necessary to conclude the existence of two distinct dogs, one owned by John and the other owned by Mark. In (6b), the fact that there are two dogs is directly encoded in the phrase *two dogs*. Hence, if our hypothesis about the models' difficulties with the `pos_pos` condition is correct, this way of directly supplying information about plurality in this way will increase models' preference for singular definites in contexts where only one DE as compared to contexts in which multiple discourse references are introduced. In other words, we expect there to be an increase in accuracy from `p(sg|pos_neg)>p(sg|pos_pos)` comparisons to `p(sg|pos_neg)>p(sg|two_neg)`, and similarly from `p(sg|neg_pos)>p(sg|pos_pos)` to `p(sg|neg_pos)>p(sg|neg_two)`.

New Combinations	Felicity
`p(sg|two_neg)`	infelicitous
`p(sg|neg_two)`	infelicitous
`p(pl|two_neg)`	felicitous
`p(pl|neg_two)`	felicitous

Table 5: Additional context-continuation combinations in Experiment 3 and their corresponding felicity judgments.

D.1 Dataset

We make the following changes to the *Implicit* dataset in Experiment 1, resulting in a new dataset that we will call *Explicit Plurality* henceforth. For each sentence where exactly one DE is introduced (i.e., `pos_neg` and `neg_pos`), we add another one where the conjunct of type `two` replaces `pos`. This results in four more context-continuation combinations, given in Table 5, of which two are felicitous and two are infelicitous.

D.2 Results

Figure 13 shows the results of comparisons of the probability of singular continuations in contexts that should evoke a single discourse referent as compared to those that should evoke two. Columns 1 and 3 are of the category *Implicit*, whereas columns 2 and 4 belong to *Explicit Plurality*. All of the LLMs show a clearer dispreference for contexts that evoke multiple discourse referents when such evocation is done by a single plural indefinite as compared to two singular indefinites. Model failure to recognize two distinct DEs from `pos_pos` is again supported by experimental data.

E Results Under A Different Metric

In Section 5.1.1 and 5.1.2, we compared the probabilities that the models assign to the same continuation given two different contexts, one felicitous and the other infelicitous. There are two ways to operationalize such comparisons according to SL.

The first one is to use a direct metric which we adopted in all of the plots presented above.

(7) a. F: John owns a dog but Mark doesn't own a dog.

b. I: John owns a dog and Mark owns a dog as well.

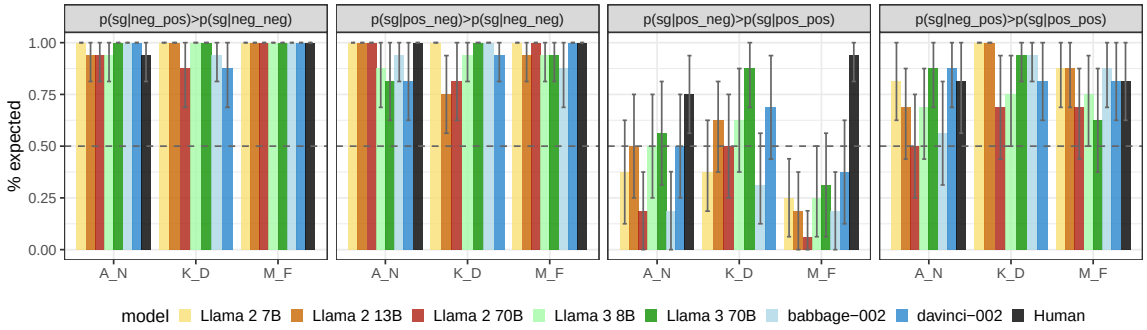

Figure 8: Results of all three sentence types from Experiment 1 - Singular Continuations. A_N, K_D, M_F stand for *affirmative-negation*, *know-doubt*, and *managed-failed* respectively.

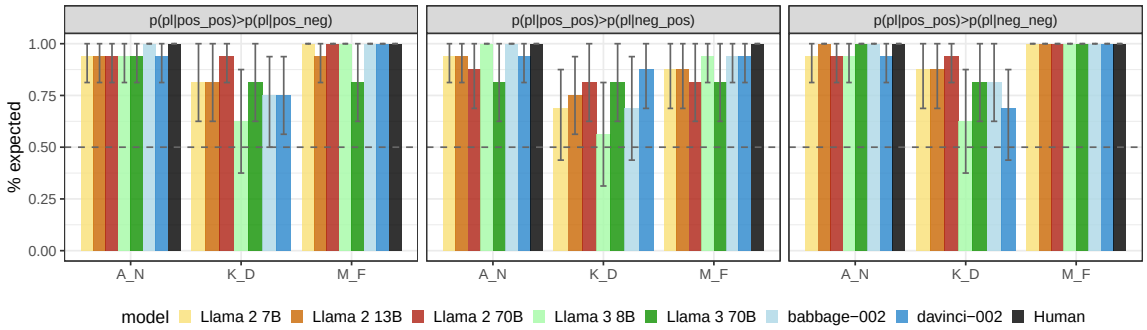

Figure 9: Results of all three sentence types from Experiment 1 - Plural Continuations.

TARGET: The dog is very cute.

Using (7) as an example, we expect the following inequality to hold:

$$p(\text{TARGET}|\text{F}) > p(\text{TARGET}|\text{I}). \quad (1)$$

Thus, we can compute accuracy for a given continuation with respect to a pair of environments, one felicitous and the other infelicitous, by measuring the proportion of times this inequality is satisfied.

However, as SL note in their Experiment 1, this metric can be problematic given that the two probabilities in the inequality are essentially drawn from different distributions, so it is possible that the probabilities are underestimated – if the language model considers that, say, given the context of F (pos_neg) in (7), there is some other continuation that is highly likely. Thus, the probability of *The dog is very cute* given this context can be smaller than its infelicitous pos_pos counterpart, although the language model may consider pos_pos to be less acceptable.

To solve this issue, Schuster and Linzen (2022) proposed a second metric which introduces control examples involving the non-coreferential continuation such as the following.

(8) CONTNONREF: It's not a big deal.

Using CONTNONREF, we now compare two fractions (2) and (3). Specifically, (2) is expected to be greater than (3).

$$\frac{p(\text{CONT}|\text{F})}{p(\text{CONT}|\text{F}) + p(\text{CONTNONREF}|\text{F})} \quad (2)$$

$$\frac{p(\text{CONT}|\text{I})}{p(\text{CONT}|\text{I}) + p(\text{CONTNONREF}|\text{I})} \quad (3)$$

Results for singular, plural, and singular vs. plural continuations using the relative metric are shown in Figure 14, Figure 15, and Figure 16 respectively. These results are qualitatively the same as the ones under the direct metric that we presented in the main body of the paper.

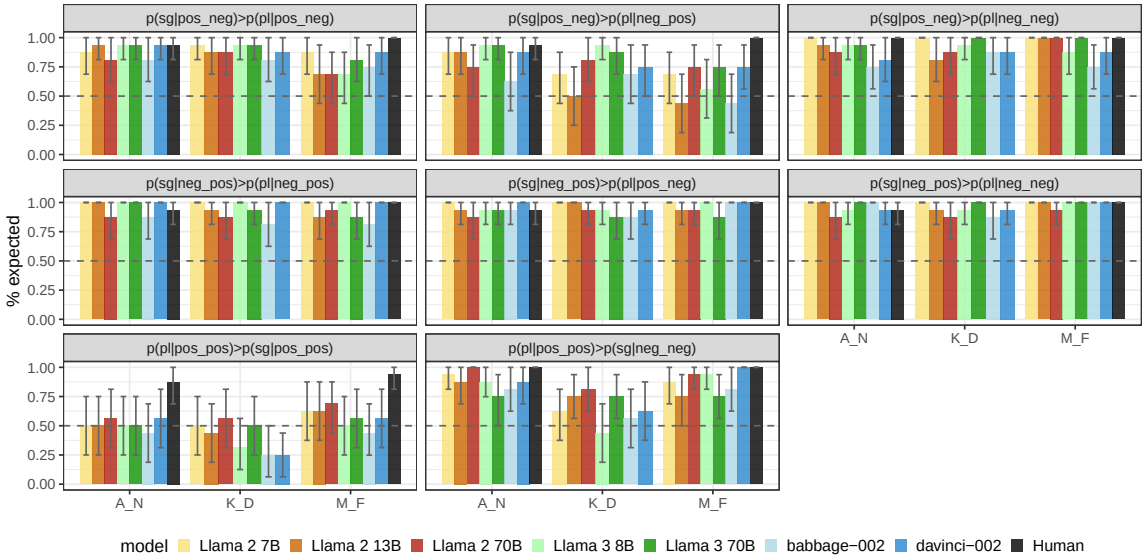

Figure 10: Results of all three sentence types from Experiment 1 - Singular and Plural Comparisons.

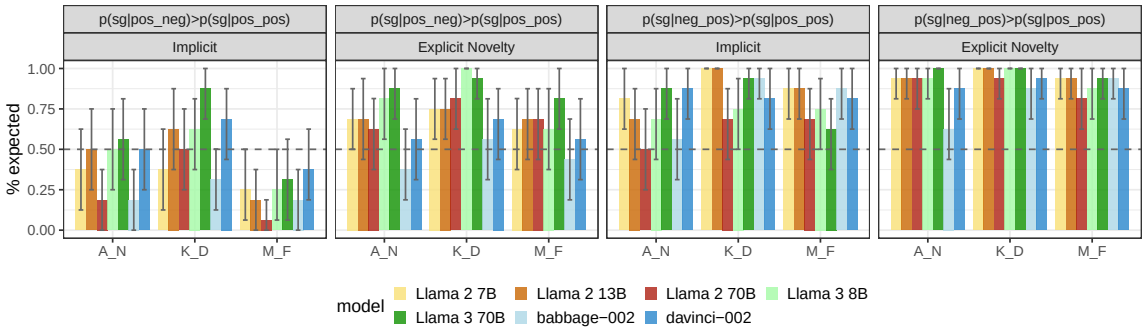

Figure 11: Results of all three sentence types from Experiment 2 - Implicit vs. Explicit Novelty.

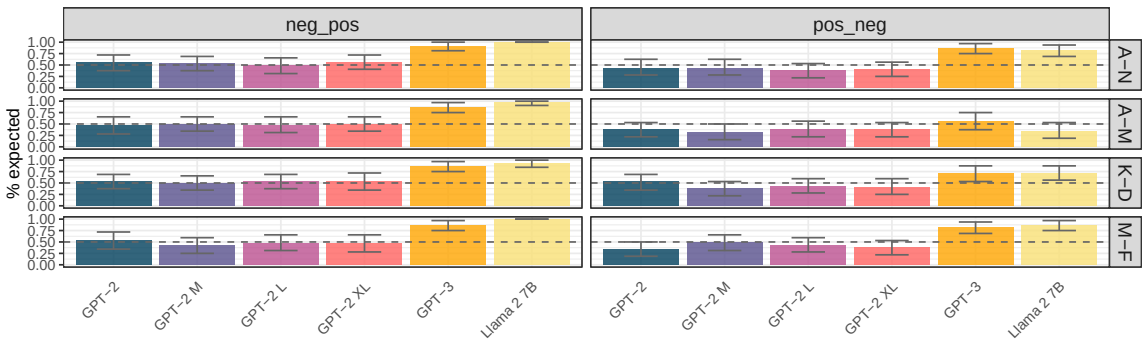

Figure 12: Decomposition of results in SL by distance and sentence type.

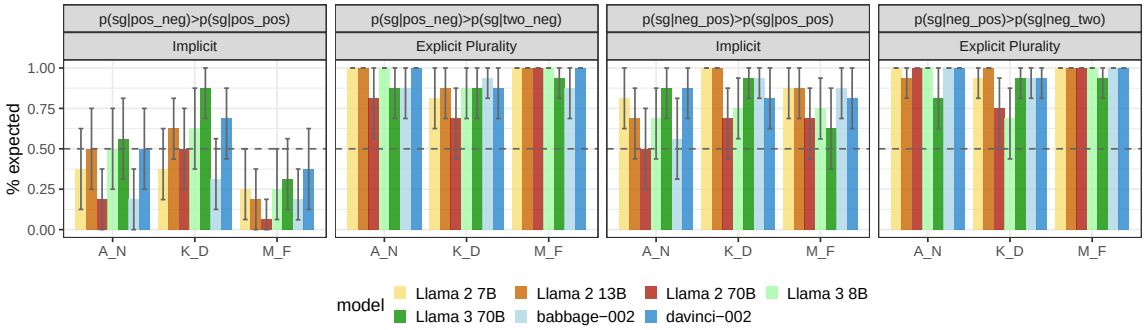

Figure 13: Results of all three sentence types from Experiment 3 - Implicit vs. Explicit Plurality.

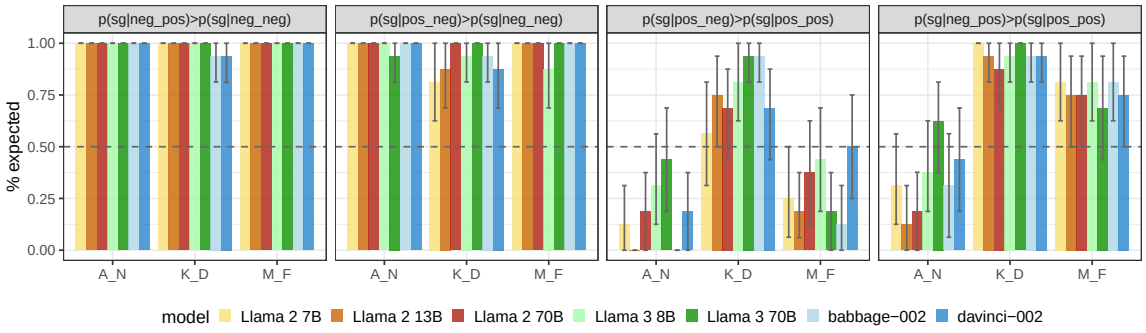

Figure 14: Results for singular continuations by model and comparison type under the relative metric.

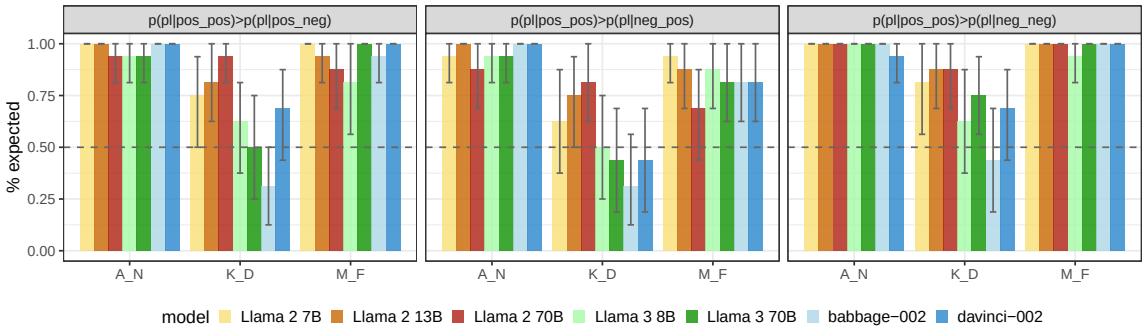

Figure 15: Results for plural continuations by model and comparison type under the relative metric.

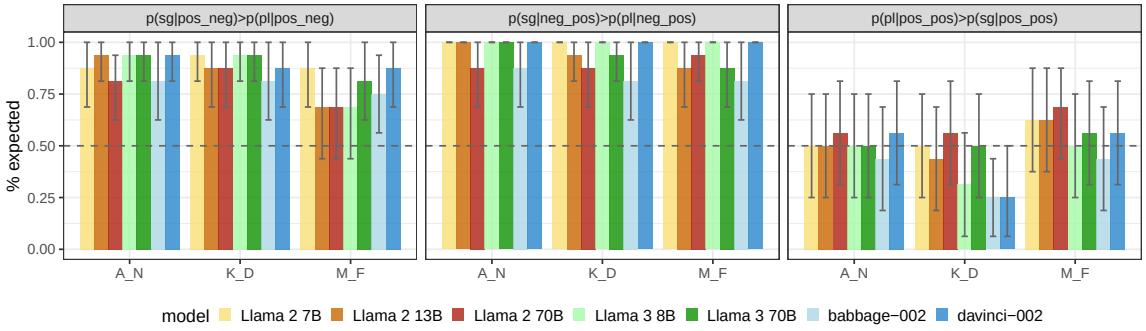

Figure 16: Results for comparisons across singular and plural continuations by model and comparison type under the relative metric.