Plural Interpretive Biases: A Comparison Between Human Language Processing and Language Models

Jia Ren
University of Massachusetts Amherst
jiaren@umass.edu

Abstract

Human communication routinely relies on plural predication, and plural sentences are often ambiguous (see, e.g., Scha, 1984; Dalrymple et al., 1998a, to name a few). Building on extensive theoretical and experimental work in linguistics and philosophy, we ask whether large language models (LLMs) exhibit the same interpretive biases that humans show when resolving plural ambiguity. We focus on two lexical factors: (i) the *collective bias* of certain predicates (e.g., size/shape adjectives) and (ii) the *symmetry bias* of predicates. To probe these tendencies, we apply two complementary methods to premise–hypothesis pairs: an embedding-based heuristic using OpenAI's `text-embedding-3-large/small` (OpenAI, 2024, 2025) with cosine similarity, and supervised NLI models (`bart-large-mnli`, `roberta-large-mnli`) (Lewis et al., 2020; Liu et al., 2019; Williams et al., 2018a; Facebook AI, 2024b,a) that yield asymmetric, calibrated entailment probabilities. Results show partial sensitivity to predicate-level distinctions, but neither method reproduces the robust human pattern, where neutral predicates favor entailment and strongly non-symmetric predicates disfavor it. These findings highlight both the potential and the limits of current LLMs: as cognitive models, they fall short of capturing human-like interpretive biases; as engineering systems, their representations of plural semantics remain unstable for tasks requiring precise entailment.

1 Introduction

Plural sentences permit multiple readings. For example, *the boys lifted the table* allows a *collective* reading (acting together) or a *distributive* reading (each acted separately). Even without context, human listeners show robust preferences, making plurality a rich testbed for evaluating whether language models track the same interpretive pressures.

We ask two questions. As *cognitive models*, do LLMs exhibit human-like interpretive biases in out-of-the-blue contexts? As *engineering systems*, do they represent plural semantics robustly enough for tasks requiring precise entailment? If such biases emerge, they may be encoded in linguistic distributions; if not, it shows the limits of text-only training.

We focus on two tendencies: the *collective bias*, where predicates vary in supporting collective over distributive readings, and the *symmetry bias*, where reciprocals differ in favoring symmetric interpretations. To probe these, we apply two methods to premise–hypothesis pairs: (i) cosine similarity with OpenAI's `text-embedding-3-large/small`, a simple but symmetric and uncalibrated proxy for entailment, and (ii) NLI models (`bart-large-mnli`, `roberta-large-mnli`), which provide asymmetric, probabilistic entailment judgments. We treat both as entailment-strength signals and compare their agreement and fit to human data.

2 Background

2.1 Collective Bias

Pluralities like *the students*, *John and Mary* are widely used in natural language. However, the semantics and pragmatics of predicating properties on plural entities is a complex issue. The complexity comes from the ambiguity of plural predications (Beck and Sauerland, 2000; Beck, 2001; Landman, 1989a,b; Link, 1983; Scha, 1984; Schwarzschild, 1996).

For example, for a sentence *the boys lifted the table*, a plural entity *the boys* is involved. The sentence allows for various interpretations. The first and most intuitive reading of the sentence is that all the boys lifted the table together. The property of table lifting applies to the plural entity *the boys* as a whole. This is commonly referred to as *the*

Proceedings of the Second Workshop on the Bridges and Gaps between Formal and Computational Linguistics (BriGap-2), pages 97–105
September 24, 2025, Licensed under the Creative Commons Attribution 4.0 International License

collective reading of plural predications. Another possible reading of the sentence is that the boys each lifted the table. The property of table lifting applies to each atom of the plural entity *the boys*. This is commonly referred to as *the distributive reading* of plural predications. In addition to the collective and distributive readings, there are also many intermediate readings. For example, the sentence is also true in a scenario where the boys were separated into groups, each group of boys lifted the table together.

Plural predication sentences are inherently ambiguous. However, this ambiguity does not hinder the efficiency or effectiveness of human communication. Rather than causing confusion, certain interpretation is usually prominent.

Experimental work has shown that collective readings are generally easier to access than distributive readings (Frazier et al., 1999; Dotlačil and Brasoveanu, 2021). However, this preference is not uniform: many special cases reveal a weak collective bias. For instance, Dotlačil and Brasoveanu (2021) find that the preference for collective interpretations disappears in cases of lexical distributivity, where the distributive meaning is encoded directly in the predicate. A well-documented example arises with adjectives of size and shape, which strongly promote distributive interpretations (Quine, 1960; Schwarzschild, 2011; Scontras and Goodman, 2017; Syrett, 2015; Maldonado, 2012; Zhang, 2013). Syrett (2015) show that this bias emerges early, in children as young as three. For example, the adjective *large* strongly favors a distributive reading: when interpreting *the boxes are big*, the most natural construal is that each box is big. By contrast, predicates such as *heavy* allow both collective and distributive interpretations: *the boxes are heavy* may mean that each box is heavy, or that the boxes are heavy as a group, even if no individual box is particularly heavy.

Why the preference arises remains an open question in the literature. One line of explanation attributes the interpretive bias to lexical semantics, certain predicates are argued to be semantically incompatible with collective readings due to their scalar or gradable nature, as in the case of size adjectives like *big* or *tall* (Schwarzschild, 2011; Maldonado, 2012; Zhang, 2013). Another line of research suggests that the preference is shaped by pragmatic reasoning or contextual factors; for instance, when interpreting size predicates, comparisons are naturally drawn at the level of each object. However, when the discourse context is set up appropriately, collective readings can emerge even for predicates that are otherwise known to strongly favor distributive interpretations (Scontras and Goodman, 2017). Scontras and Goodman (2017) collect natural-occurring examples of plural predications from the British National Corpus. For frequent plural sentences, Scontras and Goodman (2017) tests people's judgment of the salient interpretations of these sentences. The authors also manipulate the contexts of the same plural sentences and show that contexts can influence how salient the distributive reading is, thus refuting the lexical views mentioned above.

2.2 Symmetric Bias

The second generalization is that certain predicates evoke a symmetric bias, making the salient interpretations of plural sentences stronger compared to those without a symmetric bias (Beck, 2001; Dalrymple et al., 1998b; Gleitman et al., 1996; Poortman et al., 2018). For example, for the sentence *John, Mary and Bill knew each other*, the most salient reading is that John knew Mary, Mary knew John, John knew Bill, Bill knew John, Mary knew Bill and Bill knew Mary. In other words, every person knew every other person. The salient interpretation is symmetric between the atoms of the individual. For the sentence *John, Mary and Bill were hitting each other*, the most salient reading is not as strong as the one described just now. The most salient reading is that every person was either hitting or was being hit by some other person. The reading is weaker than the reading for the *knew* sentence. The reading is not symmetric between atoms of the plural.

To explain the difference in the strength of salient readings, many works focus on the use of contextual and lexical information for the selection between different readings (Dalrymple et al., 1998b; Sabato and Winter, 2012; Mari, 2013). In an influential paper, Dalrymple et al. (1998b) introduces a principle named the *Strongest Meaning Hypothesis*. The principle predicts that the strongest possible interpretation will be salient in case of ambiguity. According to the principle, the predicate *hit* was more non-symmetric to *know*. In a hitting event, a person most likely either hit someone or was hit by someone, but not both. In a knowing event, a person can both know someone and be known by someone. Thus, in the reciprocal sentences mentioned above, *John, Mary and Bill knew* *each other* has a reading which is stronger than *John, Mary and Bill were hitting each other*.

Some predicates exhibit a higher degree of symmetry than others, a generalization supported by various strands of empirical and theoretical work. Gleitman et al. (1996) observes that predicates like *be similar* are interpreted more symmetrically than predicates like *love* or *help*, suggesting that conceptual representations influence perceived symmetry. Winter (2018) provides a formal semantic account of such variability, arguing that reciprocal alternations reflect systematic differences in predicate symmetry, with certain predicates favoring reciprocal interpretations more naturally. Complementing these perspectives, Poortman et al. (2018) propose the Maximal Typicality Hypothesis, showing through experimental evidence that the interpretation of reciprocal expressions depends on how typically symmetrical a predicate is perceived to be, with more symmetrical predicates leading to stronger reciprocal inferences.

More specifically, Poortman et al. (2018) investigates how verb concepts influence the interpretation of plural reciprocal sentences in Dutch and Hebrew. Building on prior work, they first examine Hebrew data to evaluate the Strongest Meaning Hypothesis (SMH). Contrary to this prediction, their results show that participants often opt for weaker interpretations. Poortman et al. (2018) argue that this pattern reflects the sensitivity of reciprocal quantification to the underlying verb concept, and propose the Maximal Typicality Hypothesis. According to the hypothesis, a reciprocal sentence is most acceptable in a “core situation”, one that is both maximally extensive and maximally typical for the verb concept—and may also be acceptable in supersets of that situation, but not in others. They conducted two experiments in Dutch, one typicality ranking task assessing symmetry preferences across different verbs, and a truth-value judgment task with plural sentences using those verbs. The findings reveal systematic variation in how many patients are typically associated with each agent across verb types, and this variation significantly affects reciprocal interpretation. The stronger the verb’s bias toward non-symmetric scenarios, the more likely participants are to adopt a weaker reciprocal reading.

Collectively, these studies support the view that symmetry is a graded and conceptually grounded property of predicates, with consequences for both interpretation and grammatical alternation.

2.3 Goals

Prior experimental work has not examined how LLMs resolve these plural ambiguities. Using the two-method framework above (embedding similarity vs. NLI probabilities), we ask whether model signals reflect human biases for distributivity and symmetry. Our main questions are:

1. In out-of-the-blue contexts, do model-based entailment signals reflect the human *collective bias*?

2. In out-of-the-blue contexts, do model-based entailment signals reflect the human *symmetry bias*?

3. Do the two methods agree—cosine similarity vs. NLI $p(\text{entailment})$—on which readings are preferred, and where do they diverge?

We operationalize these questions by applying cosine similarity with OpenAI embeddings and by estimating $p(\text{entailment} \mid P, H)$ with `bart-large-mnli` and `roberta-large-mnli`, then comparing model outputs with prior human data (Scontras and Goodman, 2017; Poortman et al., 2018).

3 Data Collection

3.1 Collective Bias

To test the collective bias, we use the same dataset from Scontras and Goodman (2017). The authors selected the 40 most frequent combinations of the form "the nouns were adjective" from the British National Corpus, ensuring ecological validity by using naturally occurring language patterns. Participants were asked to judge what each sentence meant on a slider bar, with one end representing the paraphrase "the nouns each were adjective" (distributive interpretation) and the other end representing "the nouns together were adjective" (collective interpretation). This dataset is particularly valuable for our purposes because it provides a systematic comparison of human interpretive preferences across a range of predicate types, allowing us to assess whether language models capture the same semantic distinctions that guide human comprehension. The full list of sentences is provided in the appendix.

Participants showed a wide range of ratings across the 40 sentences, as shown in Figure 1. The figure displays the collective endorsement rate with

95% confidence intervals for each sentence tested. The x-axis shows the 40 sentences, and the y-axis indicates the proportion of responses toward the collective end of the slider bar. The results reveal systematic variation in how strongly participants favored collective versus distributive interpretations, with some sentences (e.g., "results disappointing") showing high collective endorsement and others (e.g., "classes small") showing low collective endorsement.

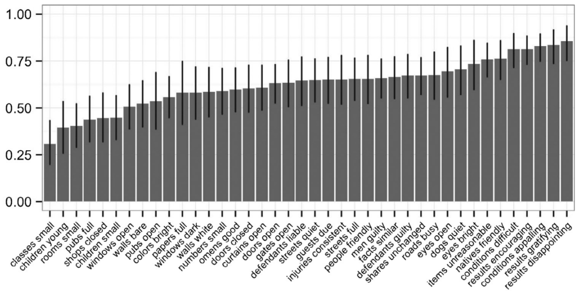

Figure 1: Results of Scontras and Goodman (2017). Collective endorsement rates with 95% confidence intervals for 40 sentences of the form "the nouns were adjective." Higher values indicate stronger preference for collective interpretations.

In our experiments, we adapt these same sentences as input to large language models to examine whether similar interpretive biases emerge in model predictions.

3.2 Symmetric Bias

To test the symmetric bias, we use the same dataset from Poortman et al. (2018). In the original paper, the authors tested 18 Dutch verbs among Dutch speakers. The 18 verbs are categorized into three types based on their patient preference.

- Type 1 (neutral): envy, know, understand, admire, miss, hate
- Type 2 (non-symmetric-preference): pinch, hit, caress, stab, shoot, grab
- Type 3 (strong non-symmetric-preference): kiss, dress, kick, lash out, bite, lick

Each verb was embedded in a sentence of the form "A, B and C Verb each other", where A, B, and C were random proper names. Participants were asked to perform truth value judgment tasks for these sentences under two types of scenarios: one depicting a symmetric action and the other a non-symmetric action. From a generalized linear mixed model (GLMM) logistic regression analysis, it was observed that in the symmetric scenarios, sentences with neutral verbs were rated significantly better than non-symmetric verbs, and non-symmetric verbs were rated as significantly better than strongly non-symmetric verbs. In the non-symmetric scenarios, the reverse pattern was observed.

In our experiments, we adapt these same sentences as input to large language models to examine whether similar interpretive biases emerge in model predictions.

4 Experiment 1

In Experiment 1, we ask whether an embedding-based metric recovers two human tendencies in plural interpretation: the collective–distributive preference and symmetry effects. We compute cosine similarity between bare and explicitly marked paraphrases using sentence embeddings from OpenAI's `text-embedding-3-large` and `text-embedding-3-small`.

4.1 Method

Experiment 1a: collective bias In this experiment, we examine the semantic similarity between two types of plural sentences: (i) bare plural sentences (Sentence 1), which lack explicit distributive or collective markers, and (ii) marked plural sentences (Sentence 2), which contain overt markers indicating distributive or collective interpretations. Examples of the tested sentences are as below.

1. Sentence 1: *the classes were small.*
2. Sentence 2 (distributive): *the classes each were small.*
3. Sentence 2 (collective): *the classes together were small.*

We use OpenAI's `text-embedding-3-large` and `text-embedding-3-small` to compute sentence embeddings and evaluate how similarly the two sentence types are represented.

Experiment 1b: symmetric bias In this experiment, we examine the semantic similarity between two types of plural sentences: (i) bare plural reciprocal sentences (Sentence 1), which lack explicit symmetric markers, and (ii) marked symmetric sentences (Sentence 2), which contain overt markers indicating symmetric interpretations. Examples of the tested sentences are as below.

1. Sentence 1: *the children knew each other.*
2. Sentence 2 (symmetric): *every child knew every other child.*

We use OpenAI's `text-embedding-3-large` and `text-embedding-3-small` to compute sentence embeddings and evaluate how similarly the two sentence types are represented.

We embed two sentences as vectors $\mathbf{u}, \mathbf{v}$ in the same semantic space and compute their cosine similarity,

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \in [-1, 1].$$

The embeddings for each pair of sentences are passed into a cosine similarity function, which returns a similarity score that is first linearly rescaled to [0, 1], then passed through a sigmoid transformation to smooth the scale. Cosine similarity is a measure used to quantify how similar two vectors are, regardless of their magnitude. Because cosine is scale-invariant and bounded, it is a convenient, single-number proxy for semantic relatedness. Some recent discussions on the application of the methods can be found in Steck et al. (2024) and You (2025), to name a few. It calculates the cosine of the angle between the two vectors in a multi-dimensional space, which reflects their orientation rather than their length. The resulting value ranges from –1 to 1, where 1 indicates that the vectors are pointing in the same direction (i.e., they are very similar), 0 means they are orthogonal (i.e., unrelated), and –1 means they are diametrically opposed. In natural language processing and information retrieval, cosine similarity is commonly used to compare text documents represented as word or sentence embeddings, allowing for efficient comparison of semantic content.

Our stimuli come in minimally different paraphrase sets that make the target interpretation explicit. For each bare sentence (e.g., *the classes were small*), we compare its similarity to a distributive paraphrase (*the classes each were small*) versus a collective paraphrase (*the classes together were small*). If a model encodes the *collective bias* that humans show for size/shape predicates, the bare sentence should be closer (higher cosine) to the distributive paraphrase than to the collective one. Analogously, for reciprocals, we compare a bare reciprocal (e.g., *A, B and C knew each other*) to stronger, fully symmetric paraphrases versus weaker, non-symmetric paraphrases. If the model encodes a *symmetry bias*, the bare reciprocal should sit closer to the fully symmetric paraphrase. Cosine similarity thus provides a simple, model-agnostic diagnostic that turns these preferences into ranked distances.

We use OpenAI's dedicated embedding models rather than hidden states from general-purpose LMs for three practical reasons. (i) Task fit: these models are trained explicitly to produce sentence embeddings whose geometry reflects semantic similarity, making cosine a meaningful signal out of the box. (iii) Sensitivity analysis: using two sizes (`*-large` and `*-small`) lets us check whether conclusions depend on embedding capacity: convergent patterns across sizes increase confidence that findings are not an artifact of a single representation. We still analyze limitations below: cosine is symmetric ($\cos(P, H) = \cos(H, P)$) and uncalibrated, so it cannot by itself model directional entailment—hence our complementary NLI experiment.

4.2 Result

We compared similarity scores across model conditions using paired t-tests, Wilcoxon signed-rank tests, and OLS regressions with item fixed effects. These analyses test whether mean differences between conditions are reliably different from zero while accounting for within-item variation. The results show that for the large model, distributive sentences were judged slightly more similar to their base forms than collective sentences (Mean Diff $\approx 0.01, p < .05$), whereas the small model showed no significant distributive–collective difference. In contrast, collective scores from the small model were systematically higher than those from the large model (Mean Diff ≈ 0.03, $p < 10^{-10}$), a large and robust effect. Overall, the large model appears sensitive to subtle distributive–collective contrasts, confirms the similarity between the bare sentences and their distributive/collective marked counterparts.

Figure3 ranks the cosine similarity scores of collective sentences in the large model from low to high. It serves as a language model analogue to Figure1. Comparing Figure1 and Figure3, we see that although both humans and the language model display a gradient of bias, the specific patterns of bias are not the same. The x-axis, which corresponds to item numbers, highlights the relative ranking of sentences, and this ranking for humans differs substantially from that of the language model. This

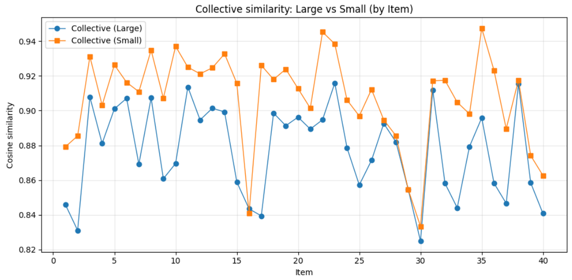

Figure 2: Comparison of similarity scores of the collective condition across the large and small embedding models.

indicates that while the large language model is sensitive to collective bias, its behavior diverges markedly from human judgments.

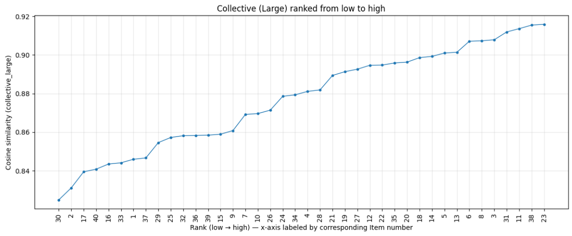

Figure 3: Comparison of similarity scores between collective and distributive conditions in the large embedding model.

Figure 4 presents the average scores of the large and small models across sentences with three verb types: neutral, non-symmetric, and strongly non-symmetric. The large model shows a steady rise across the three types, with the highest average score for strongly non-symmetric items. This indicates that the more non-symmetric a verb is, the greater the similarity between the original sentence and its symmetric paraphrase. However, this pattern contradicts the results observed in human experiments. By contrast, the small model exhibits a flatter trend, showing only minimal improvement between neutral and non-symmetric cases and even a slight decrease for strongly non-symmetric items—again diverging from human results. Overall, these findings demonstrate a clear difference between human judgments and model behavior.

Interim summary In this section, we used a cosine similarity task to test collective and symmetric biases in OpenAI's `text-embedding-3-large/small`. The large model shows slight collective–distributive contrasts, but overall neither model reproduces the collective bias observed in humans.

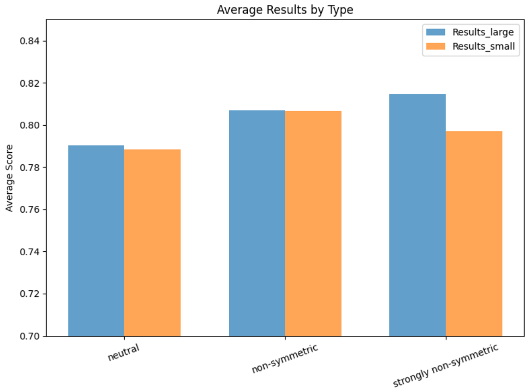

Figure 4: Comparison of average similarity scores for each verb type across the large and the small models.

5 Experiment 2

Whereas Experiment 1 employed a cosine similarity task with sentence embeddings, which measures the degree of semantic closeness between original sentences and their variants, Experiment 2 relies on supervised NLI models that explicitly calculate entailment probabilities between a premise and a hypothesis. In this way, the two experiments complement each other: cosine similarity offers an indirect, gradient measure of interpretive bias, while NLI provides a direct, categorical assessment of whether one interpretation is supported by another. Each pair consisted of a base sentence (premise) and a variant reflecting a collective/ symmetric interpretation (hypothesis). To evaluate whether the hypothesis was entailed by the premise, we employed two supervised NLI models: `bart-large-mnli` and `roberta-large-mnli`, both fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018b). The results show that, unlike human participants, language models do not show the collective/symmetric bias observed in human language use.

5.1 Method

We complement the cosine similarity experiments with a second paradigm based on supervised Natural Language Inference (NLI) models, specifically `bart-large-mnli` and `roberta-large-mnli`. In this setup, we calculate entailment probabilities between the same sentence pairs tested in Experiment 1. Namely, sentences and their corresponding collective, distributive or symmetric paraphrases. Whereas cosine similarity captures geometric closeness in embedding space without reference to specific inference relations, the NLI framework explicitly asks whether one sentence (the premise) entails another (the hypothesis). This difference is crucial: cosine similarity measures general semantic similarity, while NLI probes whether the model recognizes logical inference patterns such as symmetry.

The logic of the NLI experiments is as follows. We take each sentence as the premise and the collective/distributive or symmetric paraphrase as the hypothesis, and then use the supervised NLI models to compute the probability that the hypothesis is entailed. If a model assigns high entailment probability to the symmetric hypothesis, this suggests that it encodes a collective/distributive or symmetric bias for that predicate. Thus, these experiments go beyond the embedding-based cosine similarity approach by directly testing whether models treat symmetric interpretations as logically following from collective descriptions. Together, the two approaches provide complementary perspectives: cosine similarity reveals gradient semantic affinities, while NLI directly assesses whether symmetric readings are licensed as inferences.

We select `bart-large-mnli` and `roberta-large-mnli` because both are strong, widely used supervised NLI models that have been fine-tuned on the Multi-Genre Natural Language Inference (MNLI) dataset, which covers a broad range of sentence types and inference relations. `roberta-large-mnli` represents a transformer model trained with a robust masked-language-modeling objective, while `bart-large-mnli` combines an encoder–decoder architecture with denoising pretraining, making it particularly effective for classification tasks like NLI. Using these complementary models allows us to test whether our findings hold across different architectures, ensuring that observed patterns are not idiosyncratic to a single model design.

5.2 Result

Figure 5 reveal strikingly different patterns between the two models. For BART, there is a strong positive correlation between collective and distributive entailment probabilities ($r = 0.713$): items that score higher in the collective condition also tend to score higher in the distributive condition. The scores are the probability assigned to entailment, i.e., how strongly the model believes the collective/distributive interpretation logically follows from the original sentence. The regression line lies close to the 45-degree reference, suggesting that BART treats the two conditions as related and often raises both probabilities together. By contrast, for RoBERTa, the relationship is essentially flat ($r = 0.061$). The regression slope is close to zero, and the points are scattered broadly around the vertical axis, indicating that collective scores have little predictive value for distributive scores. This divergence suggests that BART encodes a stronger link between collective and distributive interpretations, whereas RoBERTa treats them as largely independent dimensions. Moreover, RoBERTa strongly favors the distributive interpretations, while BART, in contrast, treats the two interpretations as positively correlated without systematically preferring one over the other.

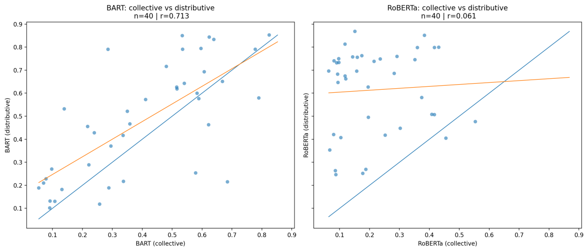

Figure 5: Comparison of the probabilities for collective/distributive pairs of BART and RoBERTa.

Figure 5 presents the RoBERTa counterpart to Figures 1 and 3, showing the collective scores ranked from low to high. While the model exhibits a gradient of scores, the ranking of items diverges considerably from the human rankings reported in Scontras and Goodman (2017). Notably, the pattern resembles the results obtained from the cosine similarity task, suggesting that the two approaches capture similar model-internal preferences rather than human-like biases.

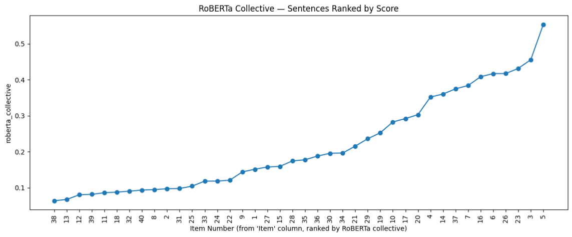

Figure 6: Sentences ranked by their RoBERTa collective scores.

Figure 7 compares the average symmetric scores of BART and RoBERTa across three verb types: neutral, non-symmetric, and strongly non-symmetric. BART consistently assigns high symmetric scores across all categories, with means ranging from 0.78 to 0.90, suggesting a strong tendency to treat symmetric paraphrases as entailed regardless of verb type. RoBERTa, in contrast, yields substantially lower scores (around 0.33–0.41), but exhibits a clearer distinction between categories: symmetric scores rise for strongly non-symmetric verbs relative to neutral and non-symmetric verbs. Again, as in the cosine similarity tasks, the human inference results are not shown here, which predict that neutral verbs should have the highest probability of entailment, while strongly non-symmetric verbs should have the lowest. BART flattens the differences almost entirely, while RoBERTa reverses the expected trend.

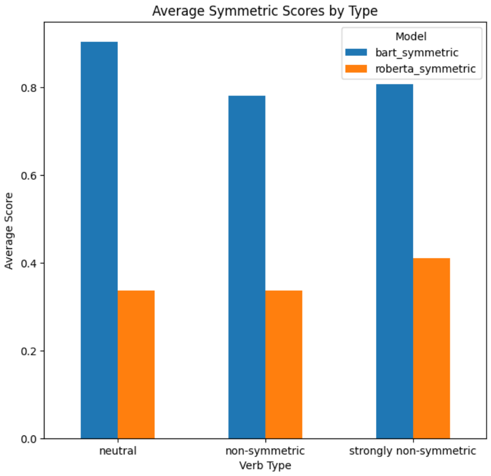

Figure 7: Average symmetric scores by verb types.

Interim summary In this section, we used an NLI task to test collective and symmetric biases. The models show slight sensitivity to plural-interpretation contrasts, but their behavior remains very different from human processing. Overall, the results are similar to those from the cosine similarity task, with neither method reproducing the robust biases observed in human judgments.

6 Limitations

Cosine similarity and NLI are imperfect proxies for human interpretation—the former symmetric and uncalibrated, the latter shaped by model-specific biases—so our results show alignment with human judgments but not competence. Stimulus design was limited to single agent–patient pairs (e.g., the boys with the table), reducing contextual variability to which human interpretation is sensitive. More broadly, by abstracting from discourse, world knowledge, and prosody, our study offers only coarse approximations. Future work should use richer behavioral methods and newer open-source models to assess alignment more fully.

7 Conclusion

This study asked whether large language models exhibit the same interpretive pressures that guide human comprehension of plural sentences. Using cosine similarity and NLI models, we probed collective and symmetry biases in out-of-the-blue contexts. The results show partial sensitivity to predicate-level distinctions, but neither method reproduced the robust human pattern—neutral verbs favoring entailment and strongly non-symmetric verbs disfavoring it. As cognitive models, LLMs therefore fall short of capturing human-like biases; as engineering systems, their representations of plural semantics remain unstable for tasks requiring precise entailment. These findings mark the limits of text-only training and point to future work in which we plan to incorporate visual cues, alongside richer context and more nuanced evaluation metrics, to better align model semantics with human judgments.

References

Sigrid Beck. 2001. Reciprocals Are Definites. *Natural Language Semantics*, 9(1):69–138. Publisher: Springer.

Sigrid Beck and Uli Sauerland. 2000. Cumulation is needed. *Natural Language Semantics*, 8(4):349–371.

Mary Dalrymple, Makoto Kanazawa, Yookyung Kim, Sam Mchombo, and Stanley Peters. 1998a. Reciprocal expressions and the concept of reciprocity. *Linguistics and Philosophy*, 21(2):159–210.

Mary Dalrymple, Makoto Kanazawa, Yookyung Kim, Sam Mchombo, and Stanley Peters. 1998b. Reciprocal Expressions and the Concept of Reciprocity.

Jakub Dotlačil and Adrian Brasoveanu. 2021. The representation and processing of distributivity and collectivity: ambiguity vs. underspecification. *Glossa: a journal of general linguistics*, 6(1):1–22.

Facebook AI. 2024a. `FacebookAI/roberta-large-mnli` — model card. `https://huggingface.co/FacebookAI/roberta-large-mnli`. RoBERTa fine-tuned on MultiNLI.

Facebook AI. 2024b. `facebook/bart-large-mnli` — model card. `https://huggingface.co/`

facebook/bart-large-mnli. BART fine-tuned on MultiNLI.

Lyn Frazier, Jeremy M. Pacht, and Keith Rayner. 1999. Taking on semantic commitments, ii: Collective versus distributive readings. *Cognition*, 70(1):87–104.

Lila R Gleitman, Henry Gleitman, Carol Miller, and Ruth Ostrin. 1996. Similar, and similar concepts.

Fred Landman. 1989a. Groups i. *Linguistics and Philosophy*, 12(6):559–605.

Fred Landman. 1989b. Groups ii. *Linguistics and Philosophy*, 12(6):723–744.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*.

Godehard Link. 1983. The Logical Analysis of Plurals and Mass Terms: A Lattice-theoretical Approach. page 22.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Violeta Vázquez Rojas Maldonado. 2012. *The Syntax and Semantics of Purépecha Noun Phrases and the Mass/Count Distinction*. Ph.d. dissertation, New York University, New York, NY.

Alda Mari. 2013. Each other, asymmetry and reasonable futures. *Journal of Semantics*, 31(2):209–261.

OpenAI. 2024. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Introduces text-embedding-3-large and text-embedding-3-small.

OpenAI. 2025. Vector embeddings – openai api documentation. https://platform.openai.com/docs/guides/embeddings. Model pages: text-embedding-3-large, text-embedding-3-small.

Eva B. Poortman, Marijn E. Struiksma, Nir Kerem, Naama Friedmann, and Yoad Winter. 2018. Reciprocal expressions and the Maximal Typicality Hypothesis. *Glossa: a journal of general linguistics*, 3(1).

Willard Van Orman Quine. 1960. *Word and Object*. MIT Press, Cambridge, MA.

Sivan Sabato and Yoad Winter. 2012. Relational domains and the interpretation of reciprocals. *Linguistics and Philosophy*, 35(3):191–241.

Remko J.H. Scha. 1984. *Distributive, Collective and Cumulative Quantification*, pages 131–158. De Gruyter Mouton, Berlin, Boston.

Roger Schwarzschild. 1996. *Pluralities*. Studies in Linguistics and Philosophy. Springer Netherlands.

Roger Schwarzschild. 2011. Stubborn distributivity, multiparticipant nouns and the count/mass distinction. In *Proceedings of the 39th Annual Meeting of the North East Linguistic Society (NELS 39)*, pages 661–678, Amherst, MA. GLSA, University of Massachusetts.

Gregory Scontras and Noah D. Goodman. 2017. Resolving uncertainty in plural predication. *Cognition*, 168:294–311.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is Cosine-Similarity of Embeddings Really About Similarity? *arXiv preprint arXiv:2403.05440*. Analysis showing cosine similarities can be arbitrary in linear embedding models.

Kristen Syrett. 2015. Mapping properties to individuals in language acquisition. In *Proceedings of the 39th Annual Boston University Conference on Language Development (BUCLD 39)*, Somerville, MA. Cascadilla Press.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018a. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018b. 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. Association for Computational Linguistics.

Yoad Winter. 2018. Symmetric predicates and the semantics of reciprocal alternations. *Semantics and Pragmatics*, 11(1):1.

Kisung You. 2025. Semantics at an Angle: When Cosine Similarity Works Until It Doesn't. *arXiv preprint arXiv:2504.16318*. Informal conceptual and empirical survey of cosine similarity limitations and alternatives.

Niina Ning Zhang. 2013. *Classifier Structures in Mandarin Chinese*. Mouton de Gruyter, Berlin.

Appendix

All plots, experimental data, and analysis scripts used in this paper are openly available at the following repository: https://github.com/ziaren/plurals-human-lm