Analysis of the Neglect-Zero Effect in Large Language Models

Jin Tanaka[1,2], **Daiki Matsuoka**[1,2], **Ryoma Kumon**[1,2], **Hitomi Yanaka**[1,2,3]
[1]The University of Tokyo, [2]RIKEN, [3]Tohoku University
{jt141214, daiki.matsuoka, kumoryo9, hyanaka}@is.s.u-tokyo.ac.jp

Abstract

We investigate the extent to which the language processing of LLMs resembles human cognitive processes, focusing on a human cognitive bias called the *neglect-zero effect*. This effect refers to the human tendency to ignore *zero-models*, which are configurations that render a proposition vacuously true by virtue of an empty set. We focus on two types of inferences driven by the neglect-zero effect, and examine how LLMs process these inferences by comparing their behavior with that in an inference that does not involve the neglect-zero effect. For this purpose, we employ a paradigm based on *structural priming*, where recent exposure to a preceding sentence (the *prime*) facilitates the processing of a subsequent sentence (the *target*) due to their structural similarity. We prepare primes to force LLMs to consider the zero-model, and analyze whether they also consider it in the target. The results suggest that the neglect-zero effect may not occur in the LLMs analyzed in this study. Our code is available at https://github.com/ynklab/neglect_zero.

1 Introduction

In recent years, the performance of large language models (LLMs) has improved dramatically, and their way of processing language superficially resembles that of humans to a significant degree. However, the mechanisms underlying their language processing remain unclear, and there is growing interest in how similar LLMs' language processing is to that of humans (Niu et al., 2025). Motivated by this interest, this study focuses on the *neglect-zero effect* (Aloni, 2022), a human cognitive bias in language processing. This effect is the human tendency to ignore *zero-models*, which are configurations where a proposition is rendered vacuously true due to an empty set.

It has been hypothesized that the neglect-zero effect underlies several linguistic phenomena that are difficult to explain within conventional theories (Aloni, 2022), and this hypothesis has already been experimentally verified with human subjects (Klochowicz et al., 2025). However, it remains underexplored whether LLMs also exhibit this effect. This study aims to address this question by applying an experimental procedure used for humans to LLMs.

Specifically, we compare the behavior of LLMs in two types of inferences, which are considered to stem from the neglect-zero effect, against their behavior in an inference called *scalar implicature*, where the neglect-zero effect does not occur. For this verification, we utilize *structural priming*, a psychological phenomenon in humans. Structural priming refers to the tendency for the processing strategy of a natural language sentence (*the target*) to become similar to that of a preceding sentence (*the prime*) when the two have the same syntactic or semantic structures. Since structural priming arises from shared structures across sentences, leveraging this cognitive tendency enables us to investigate whether a common underlying mechanism drives the processing of the inference in the prime and the target. Given that the existence of structural priming has been experimentally demonstrated in LLMs regarding other linguistic phenomena (Jumelet et al., 2024), appropriately adapting the methodology of Klochowicz et al. (2025) into prompts is expected to elucidate whether the neglect-zero effect also exists in LLMs.

The overall results of our experiment provide negative evidence for the existence of the neglect-zero effect in LLMs. In particular, we find that the Gemma-3 series and GPT-5 nano tend not to exhibit the neglect-zero effect in the inference employed in our research. Conversely, Gemma-3-27B and Llama-4-Scout exhibit a certain degree of sensitivity to zero-models, but in a manner different from humans.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1052–1064

2 Background

2.1 Neglect-Zero Effect

In this study, we focus on two types of inferences that are considered to be driven by the neglect-zero effect: *non-empty-scope strengthening in superlative quantifiers* (ESQ) and *distributive inference* (DIS). We provide an example of each inference below.

(1) ESQ
   a. {Fewer than three / At most two} circles are red.
   b. ⇝ There are some red circles.

(2) DIS
   a. Each circle is red or blue.
   b. ⇝ There is a red circle and a blue circle.

It is worth emphasizing that these inferences are not instances of logical entailment, in that they are not derived from the literal meaning of the sentences. Regarding (1), the premise (1a) literally means that the number of red circles in the set is less than three. Since this does not rule out the possibility that no red circles exist, the truth of (1a) does not logically guarantee the truth of (1b). The same holds for (2). Given that the premise (2a) literally means that every circle is colored either red or blue, it is true even in situations where all circles are red or all circles are blue. Thus, the conclusion (2b) is not necessarily true when (2a) is true. Therefore, (1) and (2) are not derived from the literal meaning alone, but rather involve a certain additional factor.

Crucially, the conclusions of these types of inferences become valid once we ignore the situations where the premise is vacuously true due to an empty set: namely, the situations with no red circles for (1a), and the situations where all circles are red/blue for (2a). These "empty" configurations are referred to as *zero-models*, and the human cognitive tendency to systematically ignore them is called the *neglect-zero effect* (Aloni, 2022). Graphical examples of (non-)zero-models for (1) and (2) are given in Figure 1. Aloni (2022) hypothesizes that this bias arises because humans tend to prefer concrete representations to abstract ones, which are generally more cognitively demanding, so as to reduce cognitive load during sentence processing.

Given this motivation, the primary question we address in this study is whether LLMs process ESQ and DIS through a common mechanism. In addition, we aim to confirm that the common mechanism, if it exists, indeed corresponds to the neglect-zero effect, by investigating whether it is distinct from the mechanism underlying an inference unrelated to the neglect-zero effect. Here, as an instance of such an inference, we consider *upper-bounded scalar implicature* (UPP), which is exemplified below.

(a) Zero-model (b) Non-zero-model

Figure 1: (a) is a zero-model of (1) and (2), and (b) is a non-zero-model of (1) and (2).

(3) UPP
   a. Some of the circles are red.
   b. ⇝ Not all of the circles are red.

Although UPP is not derived from literal meanings alone, like ESQ and DIS, it differs from the two because it is driven by a mechanism distinct from the neglect-zero effect. More specifically, UPP is an instance of *scalar implicature*, where scalar expressions such as "some" and numerals induce an inference excluding potential alternatives that belong to a certain predefined scale (Horn, 1984). In (3), we can assume a scale ⟨some, all⟩ because a simple declarative sentence with "all" entails its counterpart with "some." That is, "all" can be considered stronger than "some." Here, given that "some" is used in (3a), the hearer can infer that the speaker does not intend to convey the stronger expression "all," because they should have used it if all the circles were indeed red. The upshot is that UPP results from the hearer's inference about the speaker's intentions, unlike ESQ and DIS. Hence, by comparing UPP and ESQ/DIS, we can distinguish whether the observed behavior is due to a mechanism specific to the neglect-zero effect or due to a general mechanism for dealing with non-literal meaning.

2.2 Structural Priming

Structural priming is a human cognitive tendency to process a sentence in the same way as its preceding sentence if they are similar in terms of their semantic or syntactic structures (Bock, 1986; Pickering

and Ferreira, 2008). The sentence processed first is called the *prime*, and the subsequent sentence is called the *target*.

Structural priming is considered to result from the human tendency to reuse a mechanism during sentence processing. From this perspective, structural priming has been employed as indirect evidence for the existence of a common mechanism underlying different linguistic inferences. For example, Bott and Chemla (2016) have demonstrated that structural priming is observed in humans when sentences that induce scalar implicature are presented in both the prime and the target. On the other hand, structural priming was not observed between inferences hypothesized to be explained by the neglect-zero effect and those explained by scalar implicature (Meyer and Feiman, 2021).

Following the same line of reasoning, if the underlying principle of ESQ and DIS is the neglect-zero effect, while that of UPP is scalar implicature, it follows that structural priming should occur between ESQ and DIS, and should not occur between the ESQ/DIS group and UPP. Based on this idea, Klochowicz et al. (2025) hypothesized that if the prime suppresses the neglect-zero effect, specifically by forcing the subject to consider a zero-model of ESQ or DIS, the subject will be more likely to consider the zero-model of ESQ in the target as well. We illustrate this point further using examples (1) and (2), repeated below.

(1) a. {Fewer than three / At most two} circles are red.
    b. $\rightsquigarrow$ There are some red circles.

(2) a. Each circle is red or blue.
    b. $\rightsquigarrow$ There is a red circle and a blue circle.

Let us consider a simple task in which the subject judges whether sentences like (1a) or (2a) describe a given situation appropriately. First, we use a stimulus where the situation is a zero-model and the prime sentence is (1a). If the subject judges the description in the prime as appropriate, the cognitive act of considering the zero-model occurs. Selecting the zero-model as the answer here suppresses the neglect-zero effect in the subsequent target. Therefore, when the subject processes the target stimulus, provided that the target sentence is subject to inferences arising from the neglect-zero effect, we expect that the effect of considering the zero-model (i.e., suppression of the neglect-zero effect) in the prime continues. Consequently, if the target situation corresponds to a zero-model, we can expect a higher probability of the model judging it as true.

Although the above procedure is designed for humans, we believe that it is also applicable to LLMs. For instance, Jumelet et al. (2024) confirmed a syntactic structural priming effect on two dative constructions in LLMs, suggesting that LLMs are sensitive to structural priming. Thus, we can expect that the priming-based experimental framework of Klochowicz et al. (2025), together with an effective prompting strategy, will enable us to reveal whether LLMs exhibit the neglect-zero effect.

3 Method

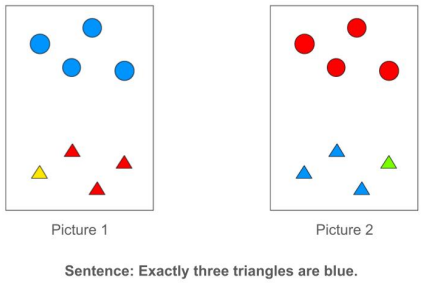
3.1 Task Setting

Before describing the experimental method in detail, we provide an overview of the task setting of Klochowicz et al. (2025). The experiment involves a picture-matching task, where the subject performs two kinds of inferences, one in the prime and the other in the target. The goal here is to test for the structural priming effect, thereby determining whether the two inferences are driven by a common mechanism. More specifically, the prime involves one of the three inferences $I_p \in \{\text{ESQ}, \text{DIS}, \text{UPP}\}$, while the inference in the target is fixed to ESQ. In a trial with the prime sentence, we "suppress" the I_p inference by forcing the subject to choose a picture that contradicts the conclusion of I_p but is true based on the literal meaning of the prime sentence. Then, if ESQ is also suppressed in a trial with the target sentence, this result supports the conclusion that I_p and ESQ are processed by a shared inference mechanism. In this way, we can use structural priming to investigate whether any two of ESQ, DIS, and UPP share a common mechanism.

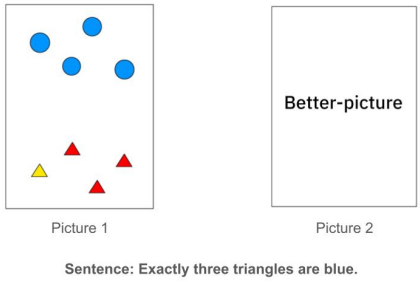
Having provided the general idea behind the experimental framework, we proceed to the details of how the experimental material is constructed.

3.1.1 The Better-Picture Paradigm

The paradigm of this experiment, which we refer to as the *better-picture* paradigm, involves a version of the picture-matching task. It is a two-alternative forced-choice task with two pictures (named "Picture 1" and "Picture 2") and a single sentence. A single execution of this task is referred to as a *trial*, and the trials are divided into two types based on the type of pictures. In the first type of trials, two open pictures and a sentence are presented, and the

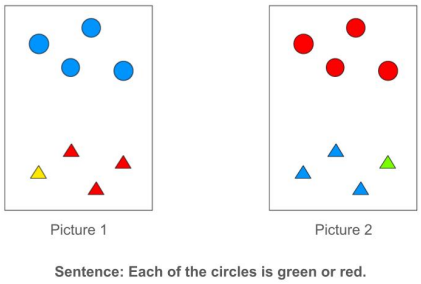


(a) An example of a trial with two open pictures. The correct answer in this trial is Picture 2.

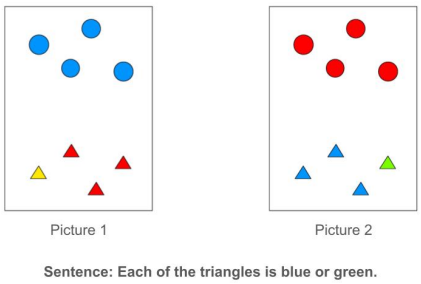


(b) An example of a trial with a better-picture. The correct answer in this trial is Picture 2 (the better-picture), since Picture 1 is incompatible with the sentence.

Figure 2: Examples of trials.



(a) An example of a critical-prime trial. The correct answer in this trial is Picture 2 (a zero-model). Note that Picture 1 is designed to be incompatible with the sentence.



(b) An example of a control-prime trial. The correct answer in this trial is Picture 2 (a non-zero-model).

Figure 3: Examples of a critical-prime trial and a control-prime trial (I_p = DIS).

subject answers which picture matches the sentence (see Figure 2a). In the second type of trials, an open picture and a covered picture are presented, and the covered picture is called the *better-picture* (see Figure 2b). When the subject addresses this type of trial, the better-picture has to be chosen if and only if the open picture does not match the sentence.

Next, we describe the content of an open picture. The open pictures feature two types of simple shapes, selected from the following: triangles, squares, crosses, hearts, and circles. One shape is in the upper half, and the other shape is in the lower half. In addition, the shapes are colored, and those in one half have a homogeneous color, and those in the other half have mixed colors. The colors used in this experiment are blue, brown, gray, orange, black, purple, green, pink, and yellow.

3.1.2 The Experimental Framework

In this experimental framework, there are three types of trials: *prime trials*, *target trials*, and *filler trials*. Prime trials and target trials correspond to the prime and the target, respectively (cf. Section 2.2). We group a sequence consisting of a prime trial and a target trial preceded by three filler trials as an *experimental item*, where the filler trials separate different experimental items, ensuring that the target trial in the former experimental item does not induce structural priming for the prime trial in the latter experimental item. In what follows, we explain the details of these three kinds of trials.

Prime trials, in which two open pictures are shown, are divided into two types based on whether the inference of interest is suppressed. One is the *critical-prime trial*, where Picture 1 does not match the sentence. Here, the subject is expected to choose Picture 2, which shows a zero-model, that is, the situation neglected in the inference I_p (see Figure 3a). The other is the *control-prime trial*, where Picture 1 does not match the sentence. Again, the subject is expected to choose Picture 2, which shows a non-zero-model of the sentence (see Figure 3b). In the prime trials, Picture 1 is designed to be clearly incompatible with the sentence. Since the task is a

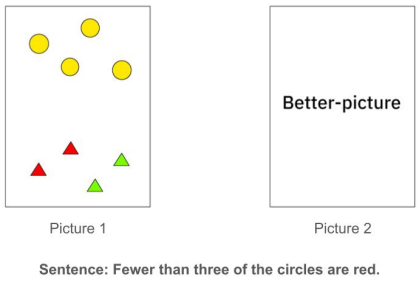


Figure 4: An example of a target trial. If priming occurs, Picture 1 should be selected, and if priming does not occur, Picture 2 should be selected.

two-alternative forced-choice, this design ensures that Picture 2 is inevitably selected. By comparing accuracy in target trials after critical-prime trials with that after control-prime trials, we can determine whether structural priming occurs.

In target trials, one of the pictures is a better-picture, and the sentence includes the phrase "fewer than three," which induces ESQ (see Figure 4). As the open picture shows a zero-model, it is expected that if structural priming is effective, the open picture will be chosen more often after the critical-prime trials than after the control-prime trials.

Finally, filler trials are designed so that the sentences in these trials do not induce the conversational implicatures of interest in this experiment (see Figure 2 for an example of this trial). In half of the prepared filler trials, Pictures 1 and 2 are open pictures, and the other half contain a better-picture. In addition, the correct answer in trials containing a better-picture is either an open picture or a better-picture, and the number of each case is the same.

Here are further details of the experimental setting. To prevent participants from predicting correct answers through memorization, the shapes and colors presented in the sentences and pictures are randomized for each trial. Furthermore, to counterbalance the potential effects arising from features such as the color and shape of the pictures, four variations of the pair of a prime trial and a target trial are prepared, and one of them is selected almost randomly.[1] These variations are classified based on two criteria: whether they correspond to a zero-model and whether they are pictures that match the sentence, and as a result, we obtain the following four variations: ⟨zero-model, answer⟩, ⟨zero-model, non-answer⟩, ⟨non-zero-model, answer⟩, ⟨non-zero-model, non-answer⟩.

In addition, before the experimental items, a short training consisting of eight filler trials is conducted to help the subject get accustomed to the task. After this training, 48 experimental items are presented to the subject.

3.1.3 Four Types of Sub-Experiments

To consider how differently structural priming occurs among the three inferences, Klochowicz et al. (2025) prepare four types of *sub-experiments*: *ESQ-sub-experiment*, *DIS-sub-experiment*, *UPP-sub-experiment*, and *BAS-sub-experiment*. These four sub-experiments have different types of prime trials, and the inference of interest is the same across all target trials. All graphical examples for this section are shown in Appendix A.

In ESQ-sub-experiments, the inference I_p is ESQ, and the sentence of the prime trial includes the superlative empty-set quantifier "at most two." In DIS-sub-experiments, the inference I_p is DIS, and the sentence of the prime trial includes "each." In UPP-sub-experiments, the inference I_p is UPP, and the sentence of the prime trial includes the scalar "some." As explained in Section 2.1, UPP is an inference triggered by scalar expressions and considered to be based on a principle distinct from the neglect-zero effect. Consequently, we expect this sub-experiment to yield results different from those observed in the ESQ and DIS sub-experiments.

In BAS-sub-experiments, the sentence of the prime trial is replaced by a filler trial, so it does not induce the inferences mentioned above. Therefore, this sub-experiment serves as the baseline for this experiment in terms of the choice rate of zero-models in target trials. In addition, by comparing this sub-experiment and the others, we can distinguish priming from general adaptation effects caused by repeated exposure to similar tasks in target trials.

3.2 The Evaluation Protocol for LLMs

Our evaluation protocol for LLMs follows the methodology of Tsvilodub et al. (2024), who have applied complex experimental designs for humans to LLMs. Specifically, we convert the visual contents of the pictures into natural language text, which we then combine with a short instruction and a sentence to form a single trial. In addition, we

[1] The choice is constrained to ensure an equal number of control-prime and critical-prime trials within a single sub-experiment.

randomize the order of the answer options ("First" and "Second").

Each sub-experiment is constructed by first presenting the task description as an instruction, followed by eight filler trials introduced as few-shot training, and finally arranging the trials as described in Section 3.1.2.

The prompt formats and examples are provided in Appendix B. This conversion is automatically performed by a Python program, utilizing the data provided by Klochowicz et al. (2025).

4 Experiments and Analysis

4.1 Settings

4.1.1 Models and Prompts

We evaluate five open LLMs: four instruction-tuned models from the Gemma-3 series (1B, 4B, 12B, and 27B models)[2] and Llama-4-Scout-17B-16E-Instruct.[3] Hereafter, we refer to Gemma-3-nB-it as "Gemma-3-nB" and Llama-4-Scout-17B-16E-Instruct as "Llama-4." In addition, we evaluate one closed LLM, GPT-5 nano.[4]

We treat each seed value assigned to the LLMs as an individual participant in human experiments (McCurdy et al., 2020). The total number of seeds is 80 per sub-experiment, for a total of 320 per model.

As mentioned in Section 3.2, we construct a dataset for our experiment with LLMs by using the method of Tsvilodub et al. (2024) for converting images into texts. The original dataset is provided by Klochowicz et al. (2025), and it contains 380 items for prime trials and target trials per sub-experiment, along with 250 items for filler trials. To ensure that the items presented to each seed are not identical, prompts are constructed by randomly sampling from this dataset.

For each model, 80 seeds are prepared for each of the four sub-experiments. The prompt for each seed consists of a few-shot training (consisting of eight filler trials) and 48 experimental items. For specific examples of the prompts and the detailed construction methodology, see Appendix B.

4.1.2 Metrics

In this experiment, we use the accuracy of LLMs' responses as a metric. Following Tsvilodub et al. (2024), the seed values that achieve an accuracy below 75% in filler trials are excluded. We also exclude experimental items if the answer in their prime trial is incorrect. We refer to the target trials that survive these exclusion steps as *valid target trials*.

For each sub-experiment, we calculate the accuracy of valid target trials under three conditions: following the critical-prime trials, following the control-prime trials, and following all prime trials. In the BAS-sub-experiment, as there is no distinction between control-prime and critical-prime trials, we only calculate accuracy following all prime trials. Based on these data, we conduct a statistical analysis using a *generalized linear mixed model* (GLMM) for each sub-experiment. See Appendix C for a detailed description of the GLMM.

In this experiment, a GLMM is expressed by the following equation:

$$\text{logit}(q_i) = \beta_i + \sum_j \beta_{ji} x_{ji} + \sum_k r_{ki}.$$

Here, q_i is the mean of the response variable, $\text{logit}(\cdot)$ is the logit function, β_i and β_{ji} are the parameters of this model, x_{ji} are the fixed effects, and r_j are the random effects. In our experiment, the response variable corresponds to the rate of choosing a zero-model in valid target trials. The fixed effects correspond to the effect of the critical-prime trials relative to the control-prime trials, which we call the *priming condition*, and the number indicating the order of the experimental items, which we call the *trial index*. The random effects correspond to the intercepts and slopes for each seed, as well as the intercepts for each experimental item.

Through this analysis, we investigate how each fixed effect influences the accuracy in each sub-experiment. Furthermore, by including the effect of each sub-experiment relative to the BAS-sub-experiment as fixed effects, we examine whether there are differences in accuracy between the BAS-sub-experiment and the other sub-experiments. For this comparison, we use the effect of the control-prime trials in each sub-experiment.

[2] https://huggingface.co/collections/google/gemma-3-release

[3] https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct

[4] https://developers.openai.com/api/docs/models/gpt-5-nano

	Gemma-3-27B					Llama-4				
	Filler	Prime			Target	Filler	Prime			Target
		Critical	Control	Overall			Critical	Control	Overall	
ESQ	74.9	99.9	95.4	97.6	97.1	91.3	96.0	99.4	97.7	74.1
DIS	71.1	98.2	97.5	97.8	95.6	91.6	92.3	95.6	94.0	72.3
UPP	72.8	100.0	91.0	95.5	96.2	89.8	98.3	84.4	91.6	29.8
BAS	76.0	-	-	93.8	97.2	90.7	-	-	97.2	49.5

	GPT-5 nano					Humans				
	Filler	Prime			Target	Filler	Prime			Target
		Critical	Control	Overall			Critical	Control	Overall	
ESQ	96.5	99.9	53.4	77.1	98.8	84.0	83.2	85.6	84.4	64.7
DIS	96.8	98.1	97.6	97.9	98.5	85.0	98.7	95.8	97.2	44.0
UPP	96.8	98.7	92.1	95.4	99.0	85.3	99.0	97.4	98.1	38.7
BAS	96.8	-	-	99.9	99.1	85.7	-	-	97.6	49.0

Table 1: The mean accuracy in filler trials and the rate of choosing a zero-model in prime and target trials in each sub-experiment of Gemma-3-27B, Llama-4, GPT-5 nano, and humans. Human data were sourced from Klochowicz et al. (2025).

4.2 Results and Analysis

Table 1 summarizes the mean accuracy across seeds for filler trials in each sub-experiment of Gemma-3-27B, Llama-4, GPT-5 nano, and humans, along with the rate of choosing a zero-model in prime trials (critical-prime only, control-prime only, and both) and target trials. The filler accuracy is calculated from data from all seeds, and all other values are calculated solely from the seeds retained after exclusion based on filler accuracy. Note that all values are rounded to one decimal place.

Regarding Gemma-3-1B, Gemma-3-4B, and Gemma-3-12B, all seeds were excluded because their accuracy on filler trials was under 75% across all seeds. Their results are shown in Appendix D. Regarding Gemma-3-27B, 192 seeds were excluded, consisting of 45 from the ESQ-sub-experiment, 57 from the DIS-sub-experiment, 50 from the UPP-sub-experiment, and 40 from the BAS-sub-experiment. In contrast, we used all seeds for Llama-4 and GPT-5 nano.

Note that there were rare cases where the model generated statements such as "Both, but I must choose one, {First/Second}." In such cases, the final choice mentioned was treated as the model's response. Furthermore, any other trials that did not yield a clear answer were treated as failed.

4.2.1 Statistical Analysis

In the following, we conduct a statistical analysis of the results for each model.

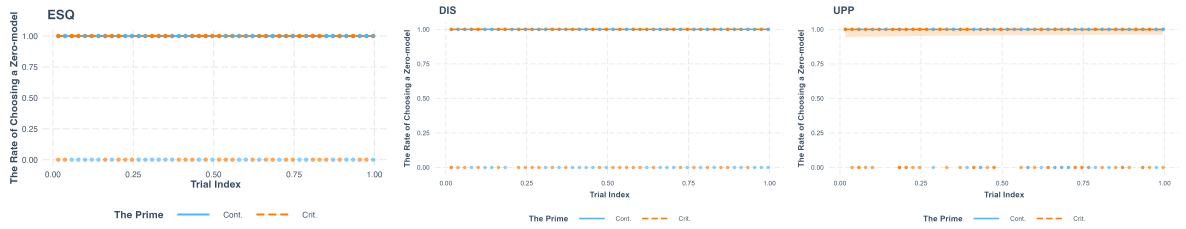
	BAS	Index	Prime	Index & Prime
ESQ	positive	positive	negative	positive
DIS	ns	positive	negative	positive
UPP	positive	ns	negative	ns
BAS	-	ns	-	-

Table 2: Summary of whether each fixed effect is significantly correlated with the observed data in the Gemma-3-27B experiment, and if so, whether the correlation is positive or negative. "ns" denotes non-significant.

Gemma-3-27B Figure 5 shows the regression curves derived from GLMM for each sub-experiment of Gemma-3-27B. Table 2 summarizes whether each fixed effect is significantly correlated with the observed data, and if so, whether the correlation is positive or negative.

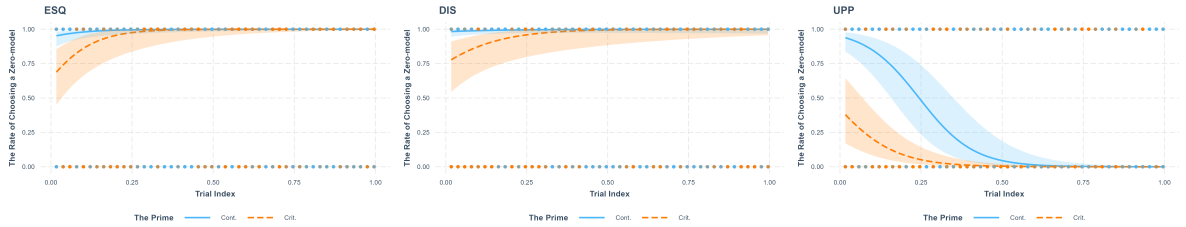
First, a general characteristic observed in all sub-experiments is that priming had a significant negative effect. That is, once it was forced to select a zero-model in the prime, Gemma-3-27B became less likely to select one in the target. This result indicates that Gemma-3-27B does not exhibit the neglect-zero effect.

However, since Table 2 shows that the model becomes significantly less likely to consider zero-models after initially taking them into account, it is evident that the model is sensitive to zero-models to some extent. These results suggest that semantic structural priming of zero-models in Gemma-3-27B manifests in a manner different from that in humans,



(a) ESQ-sub-experiment (b) DIS-sub-experiment (c) UPP-sub-experiment

Figure 5: The regression curves for GLMM in each sub-experiment of Gemma-3-27B. It should be noted that we determined the results of this analysis to be inappropriate.



(a) ESQ-sub-experiment (b) DIS-sub-experiment (c) UPP-sub-experiment

Figure 6: The regression curves for GLMM in each sub-experiment of Llama-4.

	BAS	Index	Prime	Index & Prime
ESQ	positive	positive	negative	positive
DIS	positive	ns	negative	positive
UPP	positive	negative	negative	positive
BAS	-	negative	-	-

Table 3: Summary of whether each fixed effect is significantly correlated with the observed data in the Llama-4 experiment, and if so, whether the correlation is positive or negative. "ns" denotes non-significant.

where zero-models become more likely to be chosen after considering zero-models. This point needs further detailed investigation.

In addition, because the accuracy in target trials is very high, it is possible that Gemma-3-27B rarely performs inference based on non-literal meanings and primarily processes sentences based on their literal meanings. It is worth noting that Yerukola et al. (2024) reported that LLMs struggle to generate responses based on non-literal meanings, which is consistent with our result here.

Llama-4 Figure 6 shows the regression curves derived from GLMM for each sub-experiment of Llama-4. Table 3 summarizes whether each fixed effect is significantly correlated with the observed data, and if so, whether the correlation is positive or negative.

As with Gemma-3-27B, priming had a significant negative effect in all sub-experiments. This result indicates that Llama-4 does not exhibit the neglect-zero effect.

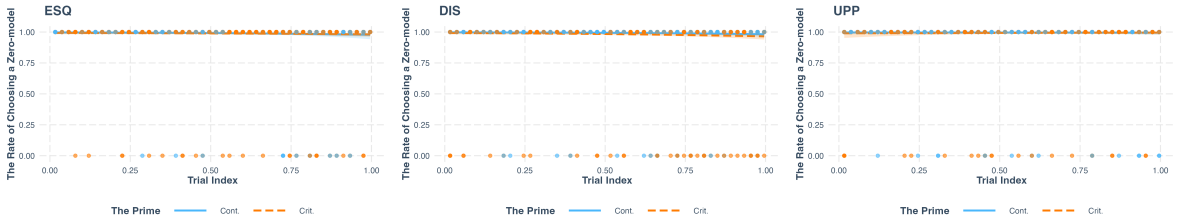
In addition, as indicated by the results in Figure 6, the behavior observed in UPP is apparently different from that in ESQ and DIS. This suggests that a different underlying mechanism is at play in UPP compared to ESQ and DIS.

Taken together, these findings imply that semantic structural priming in Llama-4 takes a form distinct from that in humans, and, furthermore, that the neglect-zero effect is unlikely to manifest.

GPT-5 nano Figure 7 shows the regression curves derived from GLMM for each sub-experiment of GPT-5 nano. Table 4 summarizes whether each fixed effect is significantly correlated with the observed data, and if so, whether the correlation is positive or negative.

First, priming had a non-significant effect in all sub-experiments. This result indicates that GPT-5 nano does not exhibit the neglect-zero effect and is not sensitive to zero-models.

In addition, as with Gemma-3-27B, due to the high accuracy across all sub-experiments, this model may place a greater emphasis on literal meaning during its inference process. On the other hand,



(a) ESQ-sub-experiment (b) DIS-sub-experiment (c) UPP-sub-experiment

Figure 7: The regression curves for GLMM in each sub-experiment of GPT-5 nano.

	BAS	Index	Prime	Index & Prime
ESQ	ns	ns	ns	ns
DIS	ns	negative	ns	ns
UPP	ns	ns	ns	ns
BAS	-	ns	-	-

Table 4: Summary of whether each fixed effect is significantly correlated with the observed data in the GPT-5 nano experiment, and if so, whether the correlation is positive or negative. "ns" denotes non-significant.

as shown in Table 1, the accuracy rate in the control-prime trials is lower than that in the critical-prime trials, which is particularly prominent in ESQ. Further investigation is necessary to elucidate the underlying causes of this phenomenon.

5 Conclusion

In this study, against the backdrop of growing interest in the similarities between the cognitive processes of LLMs and those of humans, we focused on a specific human cognitive bias known as the neglect-zero effect. To determine whether this effect is also observed in LLMs, we designed a framework for analyzing LLMs using structural priming by combining the methods of Klochowicz et al. (2025) and Tsvilodub et al. (2024), and conducted experiments on six models utilizing this framework.

Our experimental results suggest that these models do not exhibit the neglect-zero effect in their sentence processing, unlike humans. More specifically, the results from Gemma-3-27B and GPT-5 nano indicate that some models tend to primarily process the literal meaning instead of considering the non-literal meaning. On the other hand, the results from Gemma-3-27B and Llama-4 show that some models may be sensitive to zero-models in a non-human-like manner.

Given these findings, an important direction for future work would be to conduct experiments in other settings. For example, it would be valuable to conduct experiments on a larger number of models, or to use experimental settings capable of distinguishing between the manifestation of structural priming effects specific to LLMs and the disregard for non-literal meanings. Such studies will reveal whether a unified explanation for the neglect-zero effect in LLMs is possible.

Limitations

We observed a significant difference in the results between Gemma-3-27B/GPT-5 nano and Llama-4. In this study, we attributed the results to the possibility that Gemma-3-27B and GPT-5 nano place greater weight on literal meaning than non-literal meaning. However, to analyze this difference more precisely, it is necessary to conduct similar experiments on a larger number of models.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24H00809 and JST CREST Grant Number JPMJCR2565, Japan.

References

Maria Aloni. 2022. Logic and conversation: The case of free choice. *Semantics and Pragmatics*, 15(5):1–60.

Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Lewis Bott and Emmanuel Chemla. 2016. Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, 91:117–140.

Laurence Horn. 1984. Towards a new taxonomy for pragmatic inference: Q-based and R-based implicature. In Deborah Schiffrin, editor, *Meaning, Form and Use in Context*, pages 11–42.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. Association for Computational Linguistics.

Tomasz Klochowicz, Fabian Schlotterbeck, Sonia Ramotowska, Oliver Bott, and Maria Aloni. 2025. Neglect zero: Evidence from priming across constructions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, pages 5954–5960.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.

Marie-Christine Meyer and Roman Feiman. 2021. Priming reveals similarities and differences between three purported cases of implicature: Some, number and free choice disjunctions. *Journal of Memory and Language*, 120:104206.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, Ming Liu, Ziyuan Qin, Riyang Bao, Xinyuan Song, and Zekun Jiang. 2025. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *Preprint*, arXiv:2409.02387.

Martin Pickering and Victor Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427–459.

Polina Tsvilodub, Paul Marty, Sonia Ramotowska, Jacopo Romoli, and Michael Franke. 2024. Experimental pragmatics with machines: Testing LLM predictions for the inferences of plain and embedded disjunctions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, pages 3960–3967.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

A Examples of experimental items

Examples of experimental items for each sub-experiment are shown in Table 5 and Figure 8.

B Prompts

The prompt format provided to the LLM in a single trial is shown below.

> Currently the following pair of pictures is presented:
> {The explanation of Picture 1}
> {The explanation of Picture 2}
> The sentence is: {sentence}
> Which picture does match with the sentence in this situation? Here are your answer options:
> {option1}
> {option2}
> Your answer: I choose

A concrete example of it is shown below. This example uses the target row in Table 5 as the prompt.

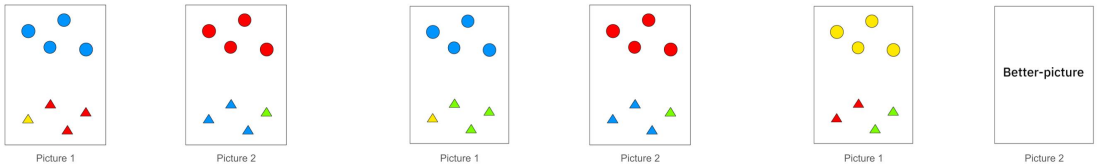
> Currently the following pair of pictures is presented:
> Picture 1 contains 4 yellow circles in the upper half, and 2 red triangles and 2 green triangles in the lower half.
> Picture 2 is the better-picture.
> The sentence is: Fewer than three of the circles are red.
> Which picture does match with the sentence in this situation? Here are your answer options:
> Second
> First
> Your answer: I choose

The prompt format provided to the LLM in each experiment is shown below.

> In the following, we will ask for your judgments about certain kinds of sentences in English.
> The sentences refers to pairs of pictures. The picture contains two types of geometrical shapes, one in the upper half and one in the lower half of the picture.
> One of these shapes in the picture were homogeneous with respect to their color, and the other set had mixed colors, containing one element with a different color.
> Only one of the pictures would match with the sentence in each trial, and your task is to choose that one.
> The covered picture, what we call ”better-picture”, is sometimes contained in the pairs.
>
> You will see many pairs, each of which will be accompanied by an sentence about the contents of the pictures.
> Your task is to decide which picture in a pair match this sentence.
>
> The better-picture should only be chosen if the open picture did not match the sentence.
> You will answer ’First’ if you consider the picture 1 match the sentence; otherwise you will answer ’Second’.
> Do not include any words or sentences other than ”First” or ”Second” in your answer.
>
> You will start with a short training to get you familiar with the response procedure.
> During this training, you will see examples of correct responses.
> ### Training 1
> {fewshot_1}
>
> ### Training 2
> {fewshot_2}
>
> ### Training 3
> {fewshot_3}
>
> ### Training 4
> {fewshot_4}
>
> ### Training 5
> {fewshot_5}
>
> ### Training 6
> {fewshot_6}
>
> ### Training 7
> {fewshot_7}
>
> ### Training 8

Sub-experiment	Type of trial	Example sentence	Pictures
ESQ	critical-prime	At most two of the circles are blue.	A
ESQ	control-prime	At most two of the triangles are green.	A*
DIS	critical-prime	Each of the circles is green or red.	A
DIS	control-prime	Each of the triangles is blue or green.	A
UPP	critical-prime	Some of the hearts are red.	A
UPP	control-prime	Some of the triangles are blue.	A
BAS	control-prime	Exactly three triangles are blue.	A
All	target	Fewer than three of the circles are red.	B

Table 5: Examples of experimental items in each sub-experiment. The alphabet in the Pictures column is based on Figure 8. This table is constructed based on the layout of Table 1 in Klochowicz et al. (2025).



(a) Example of A in Table 5. (b) Example of A* in Table 5. (c) Example of B in Table 5.

Figure 8: Illustration of A, A*, and B in Table 5.

{fewshot_8}

Your turn
As in the training, you will decide which pictures are appropriate to the sentence you see.
{48 experimental items}

Regarding the task dataset before the conversion, we utilize the experimental dataset in CSV format provided by Klochowicz et al. (2025) and annotate it with additional necessary data. Specifically, we add information about Picture 2 to the data for all sub-experiments.[5] For the filler trial data, we additionally append labels indicating which of Picture 1 or Picture 2 matches the sentence, and the validation of this annotation is performed by two annotators. Additionally, the filler trials originally contain sentences of the form "Half of the {shapes} are {color}," but because the phrase "Half of" may trigger scalar implicature, we replace them with "Exactly half of the {shapes} are {color }."

[5]Information about Picture 1 is already included.

C GLMM

Before explaining a GLMM, we first define several technical terms.

First, the observed data refer to the *response variable*, while the data considered to affect the variation in the response variable is termed the *explanatory variable*. The explanatory variables include the priming condition, the trial index, the intercepts and slopes for each seed, and the intercepts for each experimental item.

Next, the effects considered to affect the observed data are categorized into *fixed effects* and *random effects*. Fixed effects refer to the factors of primary interest in the analysis. Random effects refer to factors that are not the primary focus of the analysis but are expected to influence the observed data.

The GLMM equation used in this experiment is restated below.

$$\text{logit}(q_i) = \beta_i + \sum_j \beta_{ji} x_{ji} + \sum_k r_{ki}.$$

It is assumed that the response variable follows a binomial distribution, and that the random effects r_{kj} independently follow a normal distribution. The analysis is performed by estimating the model parameters using maximum likelihood estimation and conducting statistical tests.

	Gemma-3-1B	Gemma-3-4B	Gemma-3-12B
ESQ	23.8	43.7	59.3
DIS	24.2	40.2	59.2
UPP	25.9	40.3	62.0
BAS	25.9	49.7	62.0

Table 6: Summary of the mean accuracy in filler trials of Gemma-3-1B, Gemma-3-4B, and Gemma-3-12B. Their seeds were excluded because of their low accuracy in filler trials.

D Excluded Results

The mean accuracy in filler trials of Gemma-3-1B, Gemma-3-4B, and Gemma-3-12B is shown in Table 6.