There is No Spoon: Existential Presupposition in Large Language Models

Marie-Léontine Wörgötter[1,2*], Shikai Lai[3*], Sebastian Schuster[2,3]
[1]UniVie Doctoral School Computer Science, University of Vienna, Vienna, Austria,
[2]Faculty of Computer Science, University of Vienna, Vienna, Austria,
[3]Department of Linguistics, University College London, UK
marie-leontine.woergoetter@univie.ac.at, shikai.lai.24@ucl.ac.uk, sebastian.schuster@univie.ac.at

Abstract

Existential presupposition is a foundational component of meaning: it reflects implicit assumptions of existence that underlie interpretation, even when not explicitly stated. Sentences such as *Neo bends the spoon* presuppose that the entities referred to exist, independent of the truth-value of the sentence itself. Because this type of meaning is implied rather than explicitly asserted, it provides a diagnostic test of whether large language models (LLMs) display sensitivity to more abstract and less surface-driven layers of meaning. We adapt a natural language inference (NLI)–based probing setup, using a fine-tuned version of DeBERTa-v3-large as a baseline model and compare its behaviour to that of LLaMA-3.1-8B-Instruct and Gemma-3-12B-it under zero- and few-shot prompting, as well as to their fine-tuned base-variants. We find that while all models show sensitivity to existential presupposition across syntactic embeddings, determiner types and contextual cues, their behaviour differs markedly in strength and systematicity, with NLI-fine-tuned autoregressive models exhibiting the most coherent and stable projection patterns. They showed graded and theoretically aligned projection patterns, whereas instruction-tuned models remain largely prone to surface heuristics and prompt susceptibility. These results suggest that pre-trained LLMs exhibit sensitivity to existential presupposition but this behaviour surfaces only systematically when the models have learned the intricacies of the NLI task.

Keywords: existential presupposition, quantificational determiners, large language models, natural language inference

1. Introduction

A central property of natural language is its capacity to convey assumptions that are not explicitly asserted but are nonetheless taken for granted within discourse. *Existential presupposition* exemplifies this property: it encodes the background assumption that the referents of an expression exist in the world of interpretation (Beaver, 1997). For instance, the sentence *Neo bends the spoon* presupposes the existence of both a character named *Neo* and a spoon, independently of whether the statement itself is true or false. Crucially, the strength and persistence of presuppositions are not fixed but vary with the linguistic environment. When the sentence is embedded, for instance, in conditionals or questions, e.g. *If Neo bends the spoon, the Oracle will notice* or *Did Neo bend the spoon?*, the presupposition is typically weakened, even though not necessarily fully cancelled (Karttunen, 1971; Tonhauser et al., 2018). Adding quantifiers to the structure further modulates this behaviour: *Every spoon on the table is silver* strongly presupposes that spoons exist, whereas *Some spoons on the table are silver* carries a weakened existential commitment. Contextual framing can likewise change presupposition strength. For example, given the premise *The spoon might be bent by Neo*, adding the context *In the Matrix, there exists a spoon* supports the presupposition of existence, whereas *In the real world, there is no spoon* suppresses it, shifting how the presupposition in the premise is projected (Simons, 2001).

From the perspective of computational modelling, existential presupposition provides an opportunity to assess how large language models (LLMs) handle implicit information. Whereas explicit semantic relations such as entailment and contradiction can often be resolved from lexical or syntactic cues, presuppositional interpretation depends on whether models can recognise and sustain assumptions about the existence of entities and carry these assumptions through embedded or contextually modified structures. Evaluating LLMs on this phenomenon therefore tests whether they can maintain consistent representations of implied information rather than relying on surface patterns in word co-occurrence or syntax.

In this work, we adapt a natural language inference (NLI)–based framework, following related work such as IMPPRES (Jeretic et al., 2020), NOPE (Parrish et al., 2021) and PROPRES (Asami and Sugawara, 2023), to systematically test how models handle existential presupposition across structural and contextual variation. The central aim is to characterise whether, and how, model outputs align with established predictions about existential

*Equal contribution.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2048–2061
11–16 May 2026.

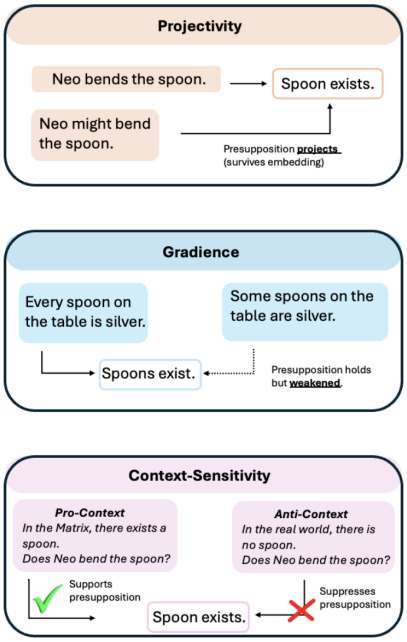

Figure 1: Schematic illustration of the three core dimensions of presupposition we examined. Presupposed content may persist under embedding (*projectivity*), vary in existential strength across embeddings and determiner types (*gradience*), and be modulated by supportive or suppressive discourse contexts (*context-sensitivity*).

presuppositions under the controlled NLI-probing setup. The analysis focuses on three core properties: *projectivity*, *gradience* and *context-sensitivity* (see Figure 1). Following prior work on masked language models (Sieker and Zarrieß, 2023; Asami and Sugawara, 2023), we use a DeBERTa-based model (He et al., 2021) as a baseline. However, given the more recent predominance of autoregressive models, we compare these baseline results to results of models from the LLaMA 3.1 (Grattafiori et al., 2024) and Gemma 3 (Team et al., 2025) families in different setups. Experiment 1 examines presuppositional behaviour across syntactic environments, determiner types and lexical plausibility. Experiment 2 further investigates context-sensitivity, assessing how supportive or suppressive discourse contexts influence existential inference.

We find that while all models exhibit partial sensitivity to existential presupposition and its modulation by structure and context, only those exposed to extensive fine-tuning display systematic, graded patterns that align with theoretical expectations. Since the NLI data that we used for fine-tuning, does not contain examples of this type of presuppositions, these results suggest that recent pretrained language models already exhibit sensitivity to existential presupposition, but that this behaviour becomes systematic and stable only once they have extensively learned the NLI task through fine-tuning.[1]

2. Related Work

Recent work has increasingly extended presupposition research into NLI, using diagnostic datasets to evaluate whether masked language models can recognise and reason about presupposed information. Several works have been central to this development. IMPPRES (Jeretic et al., 2020) introduced a controlled template-based approach for assessing projection behaviour in constructions such as clefts and possessive definites, showing that models like BERT capture some projectivity patterns but fail to generalise consistently across embedding environments. NOPE (Parrish et al., 2021) complemented this work with a corpus of naturally occurring presuppositions, evaluating model–human alignment. PROPRES (Asami and Sugawara, 2023) broadened the empirical coverage by incorporating a wider range of presupposition triggers (e.g. factives, implicatives and manner adverbs) and embedding them under negation, interrogative, conditional and modal environments on RoBERTa and DeBERTa.

More recently, Kabbara and Cheung (2022) evaluated BERT-large and RoBERTa-large models fine-tuned on MNLI on the presupposition parts of IMPPRES. By introducing contrastive perturbations, they demonstrated that models with otherwise strong benchmark performance often fail to generalise under minimally modified inputs on this task. Their findings suggest that these transformer-based masked LMs exploit surface-level lexical and structural cues rather than performing stable pragmatic reasoning.

Relatedly, Sieker and Zarrieß (2023) examined whether LMs obey the Maximize Presupposition! principle in determiner choice, which is a pragmatic constraint predicting preference for stronger presuppositional forms when contextually licensed. Their results showed that masked LMs such as BERT thereby largely rely on lexical cues and rarely exhibit the expected preference patterns, suggesting limited sensitivity.

These studies have remained confined to earlier masked LMs and do not extend to more recent and capable autoregressive LLMs. Moreover, none investigated existential presuppositions triggered by quantificational determiners. Existing evaluations have focused primarily on lexical triggers and general projection behaviour, without testing whether

[1] Project code: https://github.com/mariewoergoetter/existential_presupposition_llms

models distinguish between strong and weak quantifiers or how such sensitivity interacts with embedding and contextual variation.

We address this gap by probing large instruction- and fine-tuned models for their treatment of existential presupposition within a controlled NLI paradigm.

3. Background

Presupposition concerns information that is not explicitly asserted but implicitly assumed. Classical theories differ on whether presuppositions are conventional semantic content (Karttunen, 1971; Heim and Kratzer, 1998), constraints on discourse-update procedures (Beaver, 1997) or pragmatic inferences that arise from contextual reasoning (Simons et al., 2010). Despite these theoretical differences, it is agreed that presuppositions interact systematically with syntactic embedding, determiner type and discourse context and with that exhibiting different graded and context-sensitive patterns of projection.

As already mentioned in the introduction, there are three key characteristics to existential presuppositions:

- **Projectivity:** the extent to which presupposed content persists under embedding operators such as negation, interrogatives, conditionals, and modals (Karttunen, 1971; Tonhauser et al., 2018).
- **Gradience:** variation in the strength and cancellability of presuppositions across triggers and environments; strong quantifiers (e.g. *every*, *both*) project more robustly than weak ones (e.g. *some*, *no*) (Tonhauser et al., 2018; Tonhauser and Degen, 2020; de Marneffe et al., 2019).
- **Context-sensitivity:** the degree to which presuppositions can be strengthened, weakened or cancelled depending on discourse context (Simons, 2001; Simons et al., 2010; de Marneffe et al., 2019).

NLI Fine-Tuning and Transfer Effects. Recent research has shown that fine-tuning LLMs on NLI tasks can substantially improve their ability to perform reasoning beyond the NLI domain itself (Laurer et al., 2024; Ward et al., 2025). Because NLI supervision explicitly trains models to distinguish entailment, contradiction, and neutrality between premise–hypothesis pairs, it encourages the development of structured inferential representations that may generalise to other contexts (Ward et al., 2025). This process strengthens the model's capacity for resolving logical relations and applying inferential patterns to other reasoning tasks. In this sense, NLI fine-tuning acts as a form of scaffolding for latent competencies in the models. It is important to note, that in the training corpora we used for fine-tuning, no examples where present that explicitly instantiated existential presuppositions. Any observed presuppositional behaviour consequently in fine-tuned models cannot be attributed to direct exposure during training.

4. Experiment 1

In this experiment, we tested whether LLMs can infer the existential presuppositions of quantified expressions across four syntactic environments and between strong and weak determiners, and whether such inferences reflect structured projection patterns expected from theory.

4.1. Model Selection and Fine-Tuning

As a baseline, we used a model fine-tuned by Laurer et al. (2024) that is based on DeBERTa-v3-Large and has been fine-tuned on MultiNLI (Williams et al., 2018), FEVER-NLI (Thorne et al., 2018), Adversarial-NLI (ANLI) (Nie et al., 2020), LingNLI (Parrish et al., 2021), and WANLI (Liu et al., 2022).

To compare more recent generative models under similar inferential supervision, we evaluated LLaMA-3.1-8B-Instruct and Gemma-3-12B-it in zero- and few-shot NLI setups, as well as their pre-trained variants fine-tuned on MNLI, ANLI, and WANLI using parameter-efficient LoRA.

For the few-shot prompts, six NLI premise–hypothesis pairs from MNLI, ANLI, and WANLI were randomly sampled and balanced across labels (two entailments, two neutrals, and two contradictions) for each of five runs.

4.2. Pretest

Before testing for presuppositional inferences, it was necessary to establish that model behaviour in the NLI setup is not driven by surface-level heuristics. Prior studies have shown that pretrained models often rely on lexical overlap between premise and hypothesis (McCoy et al., 2019) or treat negation as a direct cue for contradiction (Gururangan et al., 2018). Such tendencies could confound any apparent sensitivity to presuppositional meaning. The pretest therefore served as a sanity check to verify that models distinguish entailment and contradiction based on semantic content rather than lexical cues.

4.2.1. Pretest Sentence Pair Construction

Each pretest item consists of a premise and a minimally altered hypothesis, either an affirmative rep-

etition or its negated counterpart (see Table 1). A model relying purely on lexical matching would label affirmative hypotheses as entailments, while one following a negation heuristic might label all negative hypotheses as contradictions.

To control for potential interactions between determiners and negation, all pretest sentences used the definite determiner *the*. Beyond this restriction, pretest items were systematically varied along the same factors used later for target items: syntactic environment (unembedded, conditional, modal, interrogative), noun category (real vs. fictional) and hypothesis polarity (positive vs. negative). This yielded 16 unique conditions, with 100 sentence pairs per condition (1,600 total).

4.2.2. Pretest Results and Implications

DeBERTa-v3-Large achieved near-ceiling performance in nearly all environments, with perfect accuracy (100%) on unembedded, modal, and interrogative items, and a modest drop to 69.8% in the conditional context. The small polarity asymmetries observed did not suggest reliance on negation heuristics.

In contrast, *LLaMA-3.1-8B-Instruct* and *Gemma-3-12B-it* in the zero-shot setting exhibited strong heuristic biases. Both models reached 100% accuracy for unembedded items but failed under embedding: LLaMA fell to 1.25% in conditionals, 6.00% in modals, and 4.50% in interrogatives; Gemma performed slightly better on conditionals (32.5%) but collapsed to 8.5% and 0.0% in modals and interrogatives. These near-uniform misclassifications of negated hypotheses as contradictions indicate a strong *negation = contradiction* heuristic.

Few-shot prompting improved results only slightly. For *LLaMA-3.1-8B*, mean accuracies rose to 14.92% (SD = 3.97) in conditionals, 17.42% (SD = 5.77) in modals, and 9.33% (SD = 5.70) in interrogatives, with ceiling performance on unembedded items (100%, SD = 0.00). Similarly, *Gemma-3-12B* exhibited 32.67% (SD = 14.52) in conditionals, 38.00% (SD = 1.09) in modals, and 9.33% (SD = 8.43) in interrogatives, again reaching 100% (SD = 0.00) in unembedded contexts.

Fine-tuned variants (*LLaMA-3.1-8B-pt* and *Gemma-3-12B-pt*) largely corrected misclassifications. *LLaMA-3.1-8B-pt* achieved 100% accuracy in conditionals and unembedded items, 95.0% in modals, and 62.4% in interrogatives; *Gemma-3-12B-pt* showed similar stabilisation with 100%, 96.8%, and 67.1% accuracies, respectively(see Figure 2).

Because zero- and few-shot configurations systematically misclassified negated hypotheses, all subsequent analyses for these approaches were restricted to positive-polarity hypotheses only.

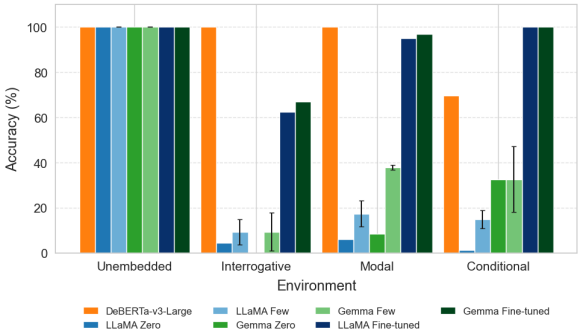

Figure 2: Mean accuracies across environments (unembedded, interrogative, modal and conditional) for all model variants. Error bars denote standard deviations for few-shot conditions.

4.3. Target Sentence Pairs

4.3.1. Target Sentence Pair Construction

Target items probed existential presupposition directly by pairing quantified premises with existential hypotheses. For example, a premise such as *All unicorns hit the ball* was paired with a hypothesis like *There was at least one unicorn* (entailment candidate) or *There was no unicorn* (contradiction candidate). The design fully crossed four linguistic dimensions, ten determiners, two noun categories, and negative and positive hypothesis polarity.[2]

The experimental design included four linguistic dimensions: determiner type, noun category, syntactic environment, and hypothesis polarity. Table 2 summarises these dimensions and their levels.

This design produced 32 unique experimental conditions ($2 \times 2 \times 4 \times 2$). Each determiner appeared in 100 sentence pairs for all environment–noun–polarity combinations, yielding 16,000 target pairs (32,000 individual sentences).

Expected theoretical patterns. Although no gold labels were assigned to the target sentence pairs, as linguistic theory provides no categorical consensus on existential presupposition judgements, all theories agree on severak generalisations. If models adequately process sentences with

[2]We deliberately exclude the negation environment as a fifth embedding type, following theoretical accounts that negation introduces scope ambiguity (Ladusaw, 1979; Heim and Kratzer, 1998), particularly in combination with quantifiers. Sentences like *All unicorns did not hit the ball* may be interpreted as either *Not all unicorns hit the ball* or *It was all unicorns that did not hit the ball*. Such ambiguities, together with focus-sensitive pragmatic effects (Horn, 1989), would obscure measurement of presupposition projection. Nevertheless, because the determiner *no* inherently encodes negation, its behaviour may serve as a proxy for testing how models handle negation.

Environment	Premise	Hypothesis	Label
Unembedded	The unicorns hit the ball.	The unicorns (did not) hit the ball.	E(C)
Interrogative	Did the unicorns hit the ball?	The unicorns (did not) hit the ball.	N(N)
Conditional	If the unicorns had hit the ball, we would have begun.	The unicorns (did not) hit the ball.	C(E)
Modal	The unicorns might hit the ball.	The unicorns (did not) hit the ball.	N(N)

Table 1: Examples of pretest sentence pairs. *E* = Entailment, *C* = Contradiction, and *N* = Neutral.

Dimension	Level	Examples
Determiner type	Strong	*every, most, the, both, all*
	Weak	*some, no, many, two, few*
Noun category	Real	*e.g. student, teacher, doctor, actor*
	Fictional	*e.g. unicorn, dragon, vampire, mermaid*
Syntactic environment	Unembedded	*The unicorn hit the ball.*
	Interrogative	*Did the unicorn hit the ball?*
	Conditional	*If the unicorn hit the ball, we would have started.*
	Modal	*The unicorn might hit the ball.*
Hypothesis polarity	Positive	*There was at least one unicorn.*
	Negative	*There was no unicorn.*

Table 2: Overview of dimensions used in target sentence pair construction.

presuppositions, we therefore expect the following general observations:

- **Strong determiners** are expected to show high entailment rates for existential hypotheses across all environments. Their presuppositions should remain relatively stable, though weaker under embeddings. This is because they typically imply that the set of entities they refer to is non-empty, leading to necessary projection of the presupposition.
- **Weak determiners** are expected to show lower entailment rates overall and even stronger reduction in entailment under embedding, as they do not necessarily imply that the set they define contains elements.
- **The determiner *no*** is expected to show very low or near-zero entailment and high contradiction rates. It introduces a negative statement over the referent set, thereby assuming its emptiness.
- **Real nouns** are expected to yield stronger and more consistent presupposition projections than **fictional** referents. This is because references to real entities are more likely to align with general world knowledge and are therefore more readily interpreted as semantically plausible. Since existential presupposition depends on assumptions about what exists within an assumed world, they are more prone to be maintained for real than for fictional referents.

4.4. Results and Discussion

Across models, existential presuppositions were projected with varying strength and stability across syntactic environments (see Figure 3) and determiner types. While all architectures exhibited some degree of existential inference, their behaviours diverged markedly in consistency, sensitivity to embedding and determiner type, and degree of overprojection.

DeBERTa-v3-Large showed a clear and theoretically interpretable gradient across environments. Entailment probabilities were highest in unembedded (86.6%) and conditional (87.1%) sentences and decreased in modal (74.8%) and interrogative (72.2%) contexts, with corresponding increases in neutral and contradictory responses. This pattern aligns closely with the expected projectivity hierarchy and indicates that DeBERTa adjusts existential inference systematically in response to embedding. DeBERTa also showed clear differentiation between determiners: strong quantifiers maintained near-ceiling entailment, whereas weak determiners exhibited reduced projection, particularly under modal and interrogative embeddings. The determiner *no* produced the lowest entailment and highest contradiction rates (up to 91.5%), although surprisingly produced full entailment in conditionals. A possible reason for this anomaly might be that DeBERTa treat hypothetical antecedents as neutralising the negative semantics of *no*.

In contrast, LLaMA-3.1-8B-Instruct exhibited more variable projection in zero-shot setting. En-

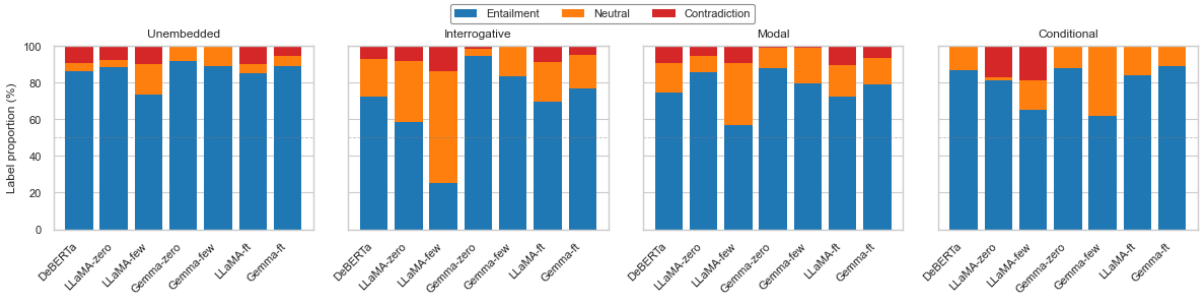

Figure 3: Label distributions across environments for positive hypotheses: Stacked bars show proportions of entailment (blue), neutral (orange), and contradiction (red) responses for each model.

tailment remained high in unembedded (88.6%), modal (85.5%) and conditionals (81.4%) but dropped markedly in interrogatives (58.8%). Conditionals sentences showed relative intermediate entailment (81.4%). The determiner *no* produced sharp drops in entailment and elevated contradiction rates (up to 84%). Overall, zero-shot LLaMA displayed some structural sensitivity but lacked consistent gradience, especially across determiners. Gemma-3-12B-it in zero-shot setting showed general over-projection. Entailment remained uniformly high across all environments (87.9% in conditionals, 94.6% in interrogatives, 87.8% in modals, and 91.7% in unembedded) indicating a general insensitivity to embedding. Only the determiner *no* broke this pattern, yielding strong contradictions (up to 95%). This suggests that Gemma struggled to adjust existential import in response to syntactic environment and quantifier type.

When exposed to few-shot examples, both LLaMA and Gemma began to show clearer gradience, with entailment weakening under embeddings. For LLaMA, mean entailment dropped to ≈70% in embedded clauses with high standard deviation across runs (up to 15 pp), while Gemma showed a similar but even less stable patterns, with standard deviations up to 20 percentage points. On a determiner-level, the few-shot settings exhibited varied and non-gradient behaviour. These results indicate that few-shot prompting introduced some differentiation tendencies, but the observed effects were unstable and highly prompt-dependent. Both models captured broad contrasts between strong and weak determiners, yet the relative hierarchy among determiners remained unsystematic and highly variable across runs.

Fine-tuning produced the most coherent and theoretically consistent projection profiles. Both LLaMA-3.1-8B-pt (see Figure 4) and Gemma-3-12B-pt displayed graded attenuation of entailment across embeddings, closely matching the hierarchy observed in DeBERTa. Entailment was strongest for unembedded and conditional sentences and reduced for modal and interrogative ones. Fine-tuning also led to determiner-specific behaviour: strong quantifiers consistently yielded high entailment, while weak ones showed systematic reductions. The determiner *no* reliably suppressed existential inference, showing minimal entailment (below 6%) and high contradiction rates (above 85%). Compared to LLaMA, Gemma exhibited a flatter gradient across both embeddings and determiners, suggesting a more conservative but less differentiated inferential profile. This suggests that explicit NLI supervision does not create presuppositional understanding from scratch but consolidates pre-existing, unsystematic sensitivities into stable, structured inferential behaviour aligned with theoretical predictions. As discussed earlier, our training data for fine-tuning did not contain examples of existential presupposition, so any systematic sensitivities towards it emerges from transfer-based adaptation.

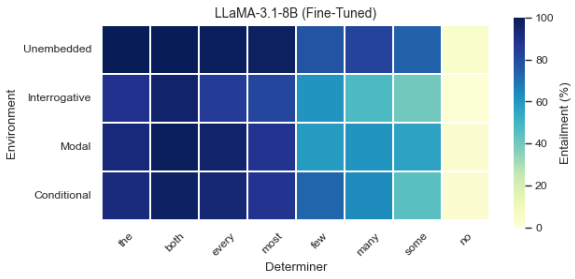

Figure 4: Entailment projection strength (0-100%) by determiner and environment for LLaMA-3.1-8B-pt fine-tuned.

Noun Plausibility (Real vs. Fictional Referents). The plausibility of the restrictor noun modulated existential projection, pointing at how world knowledge may interact with presuppositional inference. DeBERTa-v3-Large showed slight sensitivity to noun plausibility: entailment averaged 82.8% for real nouns vs. 77.6% for fictional ones, with slightly elevated neutral responses for the latter, indicating that DeBERTa downweights existential import when referents are implausible. LLaMA-3.1-8B-

Instruct exhibited a similar effect in the zero-shot setting (81.4% vs.75.8%), which widened (67.1% vs.43.1%) in few-shot setting, accompanied by higher neutral responses for fictional nouns. Variance between runs was ≈ 7 pp. Gemma-3-12B-it in zero-shot mode only showed small differentiation (92.3% vs.88.8%), similar in few-shot prompting (79.8% vs.77.3% with run-to-run variance ≈ 5 pp). Fine-tuned LLaMA-3.1-8B-pt exhibited a more pronounced gap between real and fictional referents (82.1% vs.68.5%), while Gemma-3-12B-pt showed a milder effect (86.5% vs.80.8%).

5. Experiment 2

The previous experiment did not explicitly account for the context-sensitivity of presupposition: all sentences were presented without any supporting or denying discourse context. It is possible that, during training, language models have learned to interpret *Det + NP + VP* constructions under a default assumption of referent existence. Providing explicit contextual cues that either support or contradict this assumption allows us to test whether models can modulate existential inference in response to pragmatic information.

Furthermore, presuppositional gradience implies that distinctions between strong and weak determiners, or between supportive and neutral contexts, may not always produce categorical shifts in prediction. Even subtle changes in entailment probability could therefore reveal meaningful sensitivity to presuppositional strength. To investigate these questions, we conducted a follow-up experiment focusing on fictional referents, designed to probe whether models adjust existential inference when provided with explicit contextual cues.

5.1. Sentence Pairs Construction

Context manipulation We introduced a three-level manipulation of preceding context to examine whether contextual information affects the model's tendency to project existential presuppositions:

- **Pro-presuppositional context:** supports or reinforces the existence of the fictional entity (e.g. *There is a world where unicorns were real.*)
- **Anti-presuppositional context:** denies or weakens existential assumptions (e.g. *There is a world where unicorns were fictional.*)
- **Neutral context:** no preceding context; serves as the baseline for default presuppositional projection.

Basic Premise and Hypothesis Each item consisted of a premise–hypothesis pair formatted for NLI evaluation. Unlike the previous experiment, which included hypothesis polarity manipulation, here all hypotheses affirmed the existence of the referent noun phrase (e.g. *There was at least one unicorn*).

Overall design All items followed the structure:

`[Optional Context] + [Premise] → Hypothesis`

The resulting design comprised 120 experimental conditions (10 determiners × 4 environments × 3 contexts), with 100 sentence pairs per condition (12,000 total). [3]

5.2. Results and Discussion

DeBERTa-v3-Large displayed clear variation across contexts and environments (see Figure 5). Entailment dropped sharply in anti-contexts, especially in unembedded and modal environments (0.09–0.03), though was partially mitigated conditionals and interrogatives, suggesting that contextual and structural cues interact. Pro-contexts slightly increased entailment relative to the neutral condition. Which also showed in the decrease of pro-context facilitation in interrogatives, leading to a reversed effect and suggesting that DeBERTa downweights supportive cues when projection is already weakened by embedding. Overall, variation was greater in embedded than in unembedded settings.

LLaMA-3.1-8B-Instruct exhibited a qualitatively similar hierarchy but with higher variability across contexts and prompt settings. In the zero-shot configuration, anti-contexts caused a large drop in entailment (0.04–0.19), while pro-contexts remained high (around 0.85–0.88) and on par with neutral context. Environment effects were fairly pronounced but highly variable across runs in few-shot setting.

Gemma-3-12B-it in zero-shot setting maintained similarly high entailment values for both neutral and pro-contexts (≈0.90–1.00). Anti-contexts produced some reduction (0.30–0.77 across environments). In few-shot prompting, behaviour fluctuated, with

[3] For each sentence pair, the predicted probability of the entailment label was extracted and used as a continuous score in the analyses. This value serves only to compare relative differences across syntactic environments and discourse contexts and is not interpreted as the model's calibrated confidence in the hypothesis. Following Guo et al. (2017) and Desai and Durrett (2020), these probabilities are treated as uncalibrated model-internal preferences for the entailment label under the probing setup.

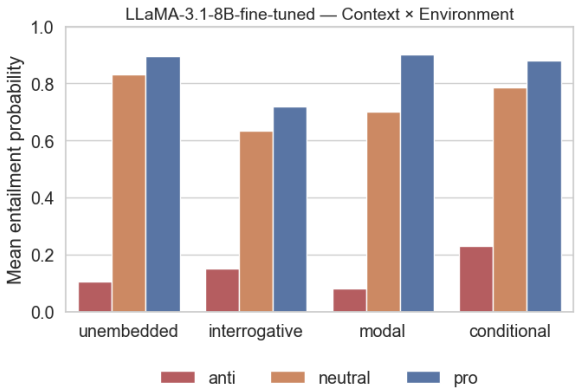

Figure 5: Mean entailment probabilities for fine-tuned LLaMA-3.1-8B-pt across contexts (pro, neutral, anti) and syntactic environments.

anti-contexts between 0.15 and 0.36 and neutral contexts between 0.45 and 0.90.

Fine-tuned models exhibited more consistent patterns across environments. LLaMA-3.1-8B-pt showed distinct and consistent lower entailment in anti-contexts (0.10–0.23) and slightly higher pro-contexts than neutral. Suppression was strongest in unembedded and modal sentences and weaker in conditionals. Gemma-3-12B-pt followed a similar pattern, with anti-contexts between 0.15 and 0.34, though pro-contexts exhibiting no significant entailment support.

Overall, the results show that anti-presuppositional contexts consistently lowered entailment probabilities across models, whereas pro-contexts produced smaller or no increases relative to neutral baselines. Embedding environments modulated these effects: suppression was generally strongest in unembedded and modal sentences and weaker in conditionals and interrogatives.

6. Conclusion

Across both experiments, the results converge on a coherent picture of how LLMs in different setups process existential presuppositions. While all models displayed at least partial sensitivity to presupposed content, only those trained on extensive NLI data exhibited systematic and graded projection patterns.

Experiment 1 demonstrated that fine-tuned models reproduce the expected hierarchy of presupposition projection across syntactic embeddings. Strong determiners consistently elicited robust existential entailment that weakened in modal and interrogative contexts, while weak or negative quantifiers showed attenuated projection. Models without fine-tuning tended to over-project existence or vary idiosyncratically with prompt formulation, suggesting reliance on surface heuristics. Noun plausibility effects further revealed that fine-tuning enhances the integration of world knowledge with semantic inference, reducing overgeneralised existential assumptions.

Experiment 2 provided additional supportive or suppressive contextual cues. Anti-presuppositional contexts thereby systematically reduced entailment probabilities, while supportive contexts induced only minor or no presupposition projection. Importantly, syntactic embedding interacted with discourse effects: context suppression was strongest in unembedded clauses and partially neutralised under conditionals and interrogatives, indicating that models can coordinate discourse- and structure-based constraints in a graded fashion. Per-determiner analyses further confirmed determiner-specific modulation, with every, both, and the showing stable projection even under contextual suppression, while some, many, and few exhibited high pragmatic susceptibility.

Taken together, these findings suggest that pre-trained language models already display sensitivity to existential presupposition and its modulation by structural and contextual factors. However, this sensitivity becomes systematic, graded and theoretically aligned only once models have learned the inferential structure of the NLI task through extensive fine-tuning. Fine-tuning therefore does not introduce presuppositional meaning, but rather consolidates and stabilises latent inferential capacities that are already present in pre-trained models. In particular, newer NLI-fine-tuned autoregressive LMs demonstrate the most coherent projection behaviour across dimensions.

Furthermore it is worth situating these results in a broader perspective on discourse-level semantic competence in LLMs. The here examined existential presupposition task relates closely to *entity tracking* - the ability to maintain or suppress discourse referents as contexts change, which, as Schuster and Linzen (2022) have shown, remains only partially systematic even in large models. Existential presuppositions, however, pose an even greater challenge, as they concern entities whose existence is not explicitly asserted, requiring the model to infer and sustain implicit assumptions. At the same time, the here presented setups benefit from more extensive supervision, making it difficult to conclude that the task is resolved even in most recent models.

7. Limitations

One limitation of our study is that the analysis did not include human data as a comparative baseline, which constrains the interpretability of model performance relative to human intuitions about pre-

supposition projection. While this may limit very fine-grained comparisons, we focused in this work on patterns that all linguistic theories agree on and therefore we would expect that these patterns would also be confirmed in a human experiment.

Furthermore, the experimental design and dataset were deliberately constrained to ensure interpretability and control, yet this restriction limits ecological validity. The constructed probing sentences may not fully capture the variability and complexity of presuppositions as they occur in natural discourse. Extending this approach to corpus-based or conversational data could reveal whether the projection patterns identified here generalise into real-world language use. On the other hand, our controlled setup makes it possible to both verify that no examples of similar structure exist in the fine-tuning data and by not relying on corpora, we also make it less likely that similar examples appear in the pre-training data of the models, thus avoiding contamination issues.

Finally, explicit negation was excluded from the experimental environments to avoid scope ambiguity. Nevertheless, the distinct behaviour of the determiner *no* and the models' graded responses to anti-presuppositional contexts suggest some latent sensitivity to polarity cues. Future research could look at these interactions more directly, introducing negation as its own embedding category.

8. Acknowledgements

This research has been funded by the Vienna Science and Technology Fund by the Vienna Science and Technology Fund (WWTF) [10.47379/VRG23007] "Understanding Language in Context."

9. Bibliographical References

Daiki Asami and Saku Sugawara. 2023. Propres: Investigating the projectivity of presupposition with various triggers and environments. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 122–137, Singapore. Association for Computational Linguistics.

David I. Beaver. 1997. Presupposition. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 939–1008. MIT Press and North-Holland, Cambridge, MA and Amsterdam.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302. Online, November 2020.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 70:1321–1330. In D. Precup and Y. W. Teh (Eds.), *Proceedings of Machine Learning Research*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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

Laurence R. Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

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.

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based nli models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785.

Lauri Karttunen. 1971. Some observations on factivity. *Papers in Linguistics*, 4:55–69.

William A. Ladusaw. 1979. *Polarity Sensitivity as Inherent Scope Relations*. Ph.D. thesis, University of Texas at Austin.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*. Originally published in the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), pages 4885–4901.

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.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. *Transactions of the Association for Computational Linguistics*, 10:1204–1220.

Judith Sieker and Sina Zarrieß. 2023. When your language model cannot even do determiners right: Probing for anti-presuppositions and the maximize presupposition! principle. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 180–198, Singapore. Association for Computational Linguistics.

Mandy Simons. 2001. On the conversational basis of some presuppositions. In *Proceedings of Semantics and Linguistic Theory (SALT)*, volume 11, pages 431–448.

Mandy Simons, Judith Tonhauser, David I. Beaver, and Craige Roberts. 2010. What projects and why. In *Proceedings of Semantics and Linguistic Theory (SALT)*, volume 20, pages 309–327.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: A large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Judith Tonhauser, David I. Beaver, and Judith Degen. 2018. How projective is projective content? gradience in projectivity and at-issueness. *Journal of Semantics*, 35(3):495–542.

Judith Tonhauser and Judith Degen. 2020. Which predicates are factive? an empirical investigation. In *Proceedings of Sinn und Bedeutung*.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. 2025. Reasoning-finetuning repurposes latent representations in base models. *ICML Workshop on Actionable Interpretability*, pages 1–6. ArXiv:2507.12638 [cs.LG].

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1804.07461*. Originally published in the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), pages 1112–1122.

A. Templates and Prompts

Tables 3 and 4 present the full set of sentence templates and representative examples used to generate the premise–hypothesis pairs in the pretest and target conditions, respectively. The templates were systematically varied across four syntactic environments: unembedded, interrogative, conditional, and modal as well as controlled for morphosyntactic features such as determiner type, noun plurality, and verb form. In the target set, the determiner position was systematically manipulated, comprising ten determiners: five strong (*every, most, the, both, all*) and five weak (*some, no, many, two, few*). In contrast, the pretest set used only the definite determiner *the*, in order to isolate the effects of lexical overlap and hypothesis polarity without determiner variation.

Zero-shot Prompt Template

```
[
  {
    "role": "system",
    "content": "You are a model
    performing a natural language
    inference task. There are two texts
    in English: a premise and a
    hypothesis. The NLI label should be
    'entailment' if whenever the premise
     is true, the hypothesis is also
    true. The NLI label should be '
    contradiction' if whenever the
    premise is true, the hypothesis is
    not true. The NLI label should be '
    neutral' if neither of the other
    cases holds. Respond only with one
    word."
  },
  {
    "role": "user",
    "content": "Premise: <PREMISE>\
    nHypothesis: <HYPOTHESIS>\n\nWhat is
     the NLI label for this pair of
    texts?"
  }
]
```

Few-shot Prompt Template

```
[
  {
    "role": "system",
    "content": "You are a model
    performing a natural language
    inference task. There are two texts
    in English: a premise and a
    hypothesis. The NLI label should be
    'entailment' if whenever the premise
     is true, the hypothesis is also
    true. The NLI label should be '
    contradiction' if whenever the
    premise is true, the hypothesis is
    not true. The NLI label should be '
    neutral' if neither of the other
    cases holds. Respond only with one
    word."
  },
  {
    "role": "user",
    "content": "Premise: <PREMISE>\
    nHypothesis: <HYPOTHESIS>\nWhat is
    the NLI label for this pair of texts
    ?"
  },
  {
    "role": "assistant",
    "content": "entailment"
  },
  {
    "role": "user",
    "content": "Premise: <PREMISE2>\
    nHypothesis: <HYPOTHESIS2>\nWhat is
    the NLI label for this pair of texts
    ?"
  },
  {
    "role": "assistant",
    "content": "contradiction"
  },
  {
    "role": "user",
    "content": "Premise: <
    PREMISE_TO_TEST>\nHypothesis: <
    HYPOTHESIS_TO_TEST>\nWhat is the NLI
     label for this pair of texts?"
  }
]
```

B. Fine-Tuning Details

Both LLaMA-3.1-8B and Gemma-3-12B-pt were fine-tuned using parameter-efficient LoRA adaptation with a rank of 16, scaling factor $\alpha = 32$, dropout 0.05, and two training epochs. The maximum sequence length was set to 256. For LLaMA-3.1-8B, the learning rate was 2×10^{-4} and the batch size 16, while for Gemma-3-12B-pt, a slightly lower learning rate of 1.5×10^{-4} and smaller batch size of 12 were used. Validation accuracies for the best-performing checkpoints on held-out sets reached 91.3% and 93.4%, respectively.

Environment	Templates		Examples	
	Premise	**Hypothesis**	**Premise**	**Hypothesis**
Unembedded	The NP VP.	Affirmative: The NP VP.	The unicorns hit the ball	Affirmative: The unicorns hit the ball.
Interrogative	Did the NP VP?	Negative: The NP did not VP.	Did the unicorns hit the ball?	Negative: The unicorns did not hit the ball.
Conditional	If the NP had VP, we would have begun.		If the unicorns hit the ball, we would have begun.	
Modal	The NP might VP.		The unicorns might hit the ball.	

Table 3: Sentence templates and examples for pretest sentence pairs.

Environment	Templates		Examples	
	Premise	**Hypothesis**	**Premise**	**Hypothesis**
Unembedded	Det NP VP.	Positive: There was at least one NP.	Every unicorn hit the ball	Positive: There was at least one unicorn.
Interrogative	Did Det NP VP?	Negative: There was no NP.	Did every unicorn hit the ball?	Negative: There was no unicorn.
Conditional	If Det NP had VP, we would have begun.		If every unicorn hit the ball, we would have begun.	
Modal	Det NP might VP.		Every unicorn might hit the ball.	

Table 4: Sentence templates and examples for target sentence pairs.

C. Supplementary Results

Figures 6–12 provide the full set of determiner-level heatmaps, illustrating entailment projection strength (0–100%) across quantificational determiners and syntactic environments, per model approach.

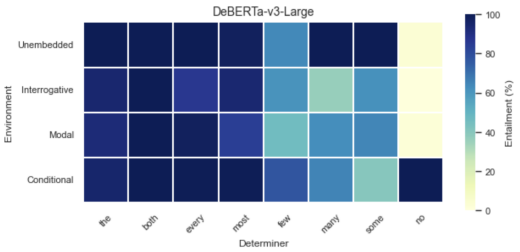

Figure 6: DeBERTa-v3-Large

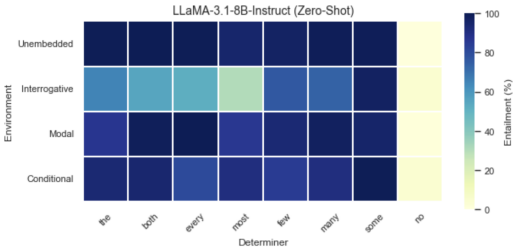

Figure 7: LLaMA-3.1-8B (zero-shot)

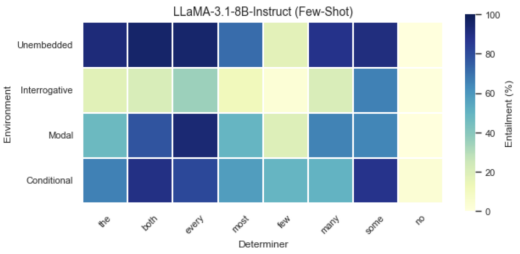

Figure 8: LLaMA-3.1-8B (few-shot)

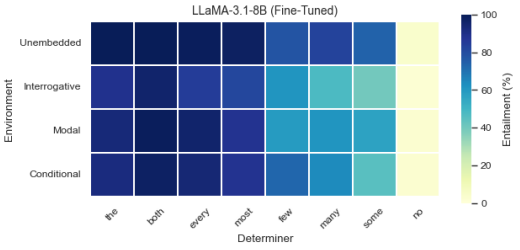

Figure 9: LLaMA-3.1-8B-pt (fine-tuned)

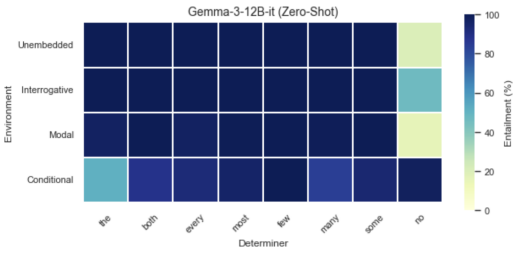

Figure 10: Gemma-3-12B-it (zero-shot)

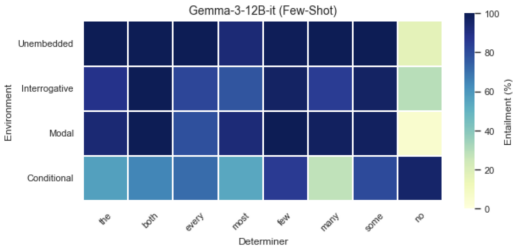

Figure 11: Gemma-3-12B-it (few-shot)

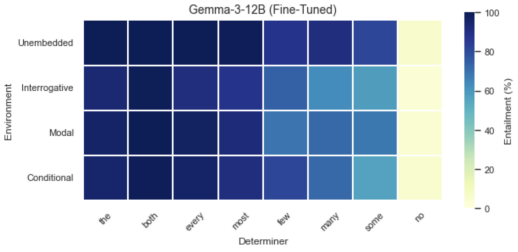

Figure 12: Gemma-3-12B-pt (fine-tuned)

The following figures 13–19 depict mean entailment probabilities across supportive (*pro*), neutral, and suppressive (*anti*) discourse contexts for each syntactic environment, per model approach. The error bars for the few-shot approaches indicate minimum to maximum range across runs.

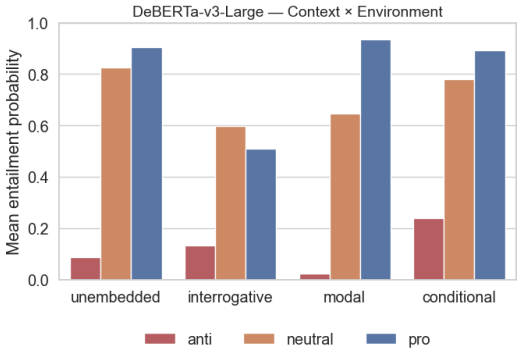

Figure 13: DeBERTa-v3-Large

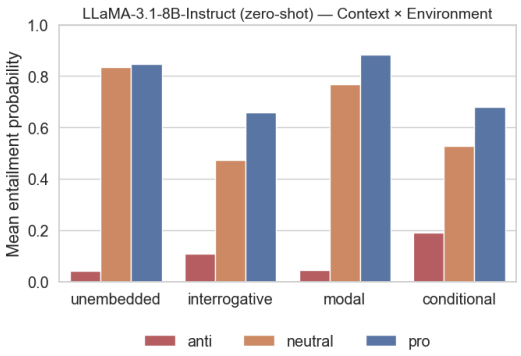

Figure 14: LLaMA-3.1-8B (zero-shot)

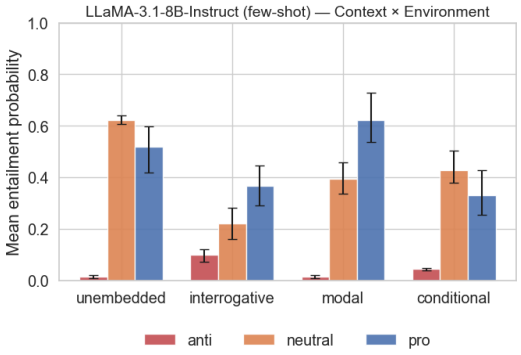

Figure 15: LLaMA-3.1-8B (few-shot)

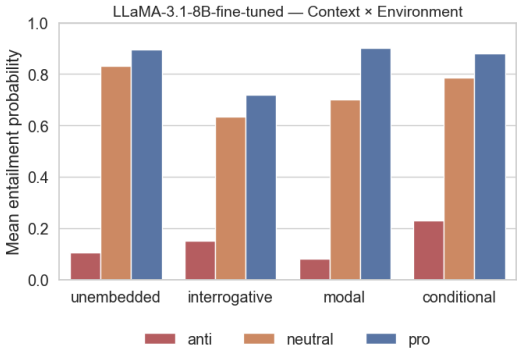

Figure 16: LLaMA-3.1-8B-pt (fine-tuned)

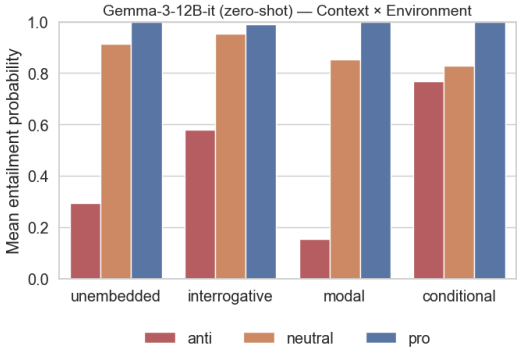

Figure 17: Gemma-3-12B-it (zero-shot)

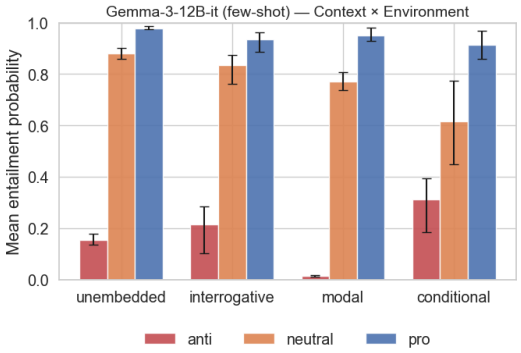

Figure 18: Gemma-3-12B-it (few-shot)

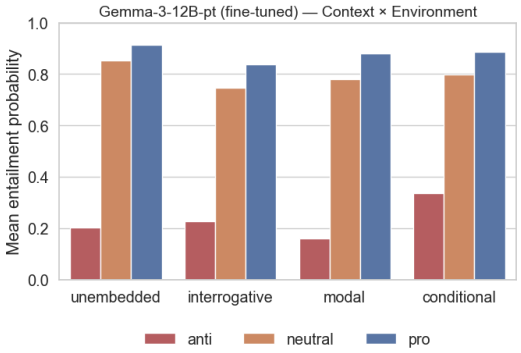

Figure 19: Gemma-3-12B-pt (fine-tuned)