Do Language Models Know Theo Has a Wife? Investigating the Proviso Problem

Anonymous IWCS submission

Abstract

We investigate the proviso problem, an unresolved issue in formal semantics concerning presupposition projection in conditionals, using four language models (RoBERTa, DeBERTa, LLaMA, and Gemma) and AI explainability methods. Our findings show that these models process presuppositions differently from humans. The results offer insights for evaluating and refining existing semantic theories.

1 Introduction

The proviso problem (Geurts, 1996) is an ongoing challenge in formal semantics, particularly for satisfaction-based theories (Heim, 1983; Beaver, 1997; Mayr and Romoli, 2016; Mandelkern, 2016). According to dynamic semantics and related accounts, in conditionals of the form *If A,* B_p, where p is the direct presupposition of the consequent B, the presupposition of the sentence is predicted to project conditionally. For example, in the sentence *"If Theo hates sonnets, so does his wife"*, the predicted presupposition is *"If Theo hates sonnets, then Theo has a wife"*. However, in practice, speakers typically accommodate a stronger, unconditional presupposition: *"Theo has a wife"*. The discrepancy is often explained by presupposition accommodation, where the listener updates the context to satisfy the sentence's projected semantic presupposition (Lewis, 1979; Von Fintel, 2008; Singh, 2020). This shows a gap between theoretical predictions and intuitive judgments. This gap extends beyond conditionals to structures with disjunctions, conjunctions, and attitude verbs, making the proviso problem a core problem in the study of presupposition.

Despite its importance in semantic theory, no prior work has empirically investigated how language models handle presupposition projection in conditionals. We address this gap by reformulating the proviso problem as a Natural Language Inference (NLI) task, an established natural language processing framework for modeling inferential relations between a premise and a hypothesis using Entailment, Neutral, and Contradiction labels (Bowman et al., 2015; Williams et al., 2018). We apply recent explainability techniques, such as attribution-based saliency and attention analysis (Sundararajan et al., 2017; Atanasova et al., 2020), to better understand how models process context-sensitive meaning and which lexical or structural features influence their predictions. In this paper, we specifically ask whether language models project presuppositions in conditionals as predicted by existing semantic theories or whether they do so in line with human judgment.

2 Related Work

Most existing NLI datasets (Bowman et al., 2015; Williams et al., 2018) focus on classical inference and overlook pragmatic phenomena such as presupposition. Datasets such as IMPPRES (Jeretic et al., 2020) and NOPE (Parrish et al., 2021) include presuppositional content, but either rely on simple conditional forms or lack coverage of structurally complex cases like embedded conditionals. A recent dataset, CONFER (Azin et al., 2025), has introduced an NLI benchmark for testing presuppositional reasoning in conditionals; however, it does not address structural variations relevant to projection.

3 Dataset Construction

We build on the CONFER dataset (Azin et al., 2025), which contains sentence pairs with five types of conditional constructions. Each conditional takes the form *If A,* B_p (see section 1). These conditionals are paired with a hypothesis p. Each conditional type encodes a specific logical relation-

ship between *A* and *p* (see Appendix for examples). For this study, we selected types 4 and 5 conditionals, in which antecedent *A* and presupposition *p* are logically independent and their NLI label is Entailment (E). These types include the presupposition triggers *again* and possessive constructions.[1] We began with 900 unique conditionals and then extended the dataset, constructing four subsets to evaluate neural models' sensitivity to structural and contextual variation based on existing theories related to the proviso problem. In total, our dataset includes approximately 8,500 sentence pairs, with individual subsets ranging from about 900 to 3,600 examples.[2] Below, we briefly describe the structure of each subset.

Subset 1: Original Sentences

This subset used the original NLI pairs from CONFER as a baseline to evaluate models' performance.

Subset 2: Structural Variation

To test how structural manipulation affects projection, we created three modified versions of each original sentence: (1) Conjunction: adding a conjunct to the antecedent (e.g., *If A and B, then C*); (2) Disjunction: rephrasing the conditional as *Either not-A or B*; and (3) Belief Embedding: embedding the consequent under an attitude verb (e.g., *X believes that B*).

Subset 3: Trigger–Hypothesis Relatedness

In this subset, we modified the trigger phrase by manipulating its semantic similarity to the hypothesis. Using WordNet (Miller, 1995) and ConceptNet (Speer et al., 2017) relations (e.g., `is-a`, `part-of`, `hyponymy`), we generated *related*, *somewhat related*, and *unrelated* variants by substituting key lexical items. For example, in the original premise "If Randolf is a carpenter, he'll use his beading tools for designing" with the hypothesis "Randolf has beading tools," we replaced "his beading tools" in the premise with "his round nose pliers" to create a "somewhat related" to the hypothesis version. The gold labels were updated from E to Neutral (N), as the relationship no longer clearly supports entailment.

[1]Throughout this paper, we refer to these two types as type 4 and type 5. type 5 sentences containing the trigger *again* are labeled as type 5_a, and those with possessive constructions are labeled as type 5_p.

[2]The dataset and detailed results of the experiments will be made publicly available upon acceptance.

Subset 4: Context–Trigger Relatedness

Subset 4 tested how premise-level modifications, either in the antecedent or the consequent, affect the models' behavior. Each sentence included two key phrases: K1 (the presupposition trigger) and K2 (an added or modified contextual phrase, such as an unrelated event). We created two variants: one where the antecedent and consequent events are logically related (e.g., *If Sarah attends a movie festival, she'll never watch Star Wars again*) and one where they are unrelated (e.g., *If Sarah attends the conference, she'll never watch Star Wars again*).

4 Experiments

Our experimental setup combines classification accuracy with explainability analyses using saliency methods. We compare outputs from four language models against both human and theoretical expectations. Human labels come from CONFER, while theoretical labels are based on the existing theories of presupposition projection in conditionals discussed in section 1. Table 1 outlines the metrics we used to evaluate the models on each subset.

Metric	Description
Accuracy	Percentage of classifications that are correct relative to human-annotated and theory-predicted NLI labels.
(IG) Integrated Gradients	Measures the overall attribution strength assigned to input tokens (Sundararajan et al., 2017).
Trigger IG Ratio	Ratio of influence assigned to the presupposition trigger word (as measured by IG) to the average influence of all words in the sentence.
IG by POS	Aggregates IG scores by part-of-speech.
Keyword Percentile Score	Ranks the salience of the presupposition-bearing phrase relative to all tokens in the input.
K1→K2 Attention	Average attention between key phrases 1 and 2, normalized by average attention.
K2→Special Token Attention	Attention from key phrase 2 to the special token (e.g., [SEP], <bos>, [CLS]), normalized by average special token attention.
T-test p-values	Results of statistical tests comparing IG, Keyword Percentile, and Attention scores between subsets.

Table 1: Metrics used for model evaluation.

Models and Setup

We used four language models: RoBERTa-large-MNLI, DeBERTa-large-MNLI, Llama-3.2-1B, and

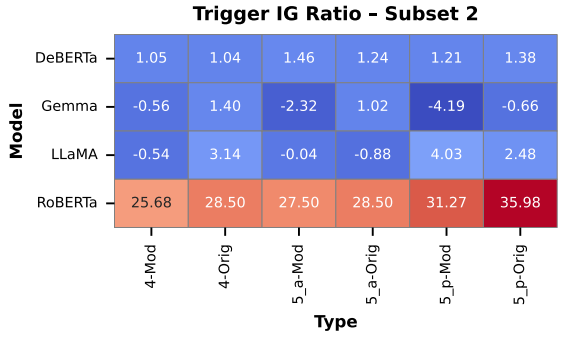

Figure 1: Trigger IG Ratios for subset 2 across four models. "Org" refers to the original conditionals and "Mod" to the structurally modified versions. 5_a and 5_p correspond to *again* and possessive triggers in type 5 sentences, respectively.

Gemma-3-1B. All models were fine-tuned on the CONFER dataset. Additionally, RoBERTa and DeBERTa had been fine-tuned on the MultiNLI dataset (Williams et al., 2018).

Evaluation without Fine-tuning: All Subsets

In this experiment, we evaluated RoBERTa and DeBERTa on the original conditional sentences (subset 1) without any structural modifications, as well as all modified examples created in subsets 2-4. The models were not fine-tuned on CONFER for this experiment. Gemma and LLaMA were evaluated through zero-shot prompting asking to predict the NLI label. The accuracy of each model with respect to both the gold (human) labels and the theoretical labels was computed.[3]

Structural Variation: Subset 2

Subset 2 tests how structural changes such as conjunctions, disjunctions, and belief embeddings affect presupposition projection. In this experiment, we evaluate classification accuracy, the Trigger IG Ratio, IG grouped by POS, and Keyword Percentile Scores.

Trigger–Hypothesis Relatedness: Subset 3

Subset 3 evaluates how semantic relatedness between the trigger phrase and the hypothesis affects the models' behaviour. We calculated accuracy, the Trigger IG Ratio, IG by POS, and Keyword Percentile Scores to examine whether substituted phrases, less related to the hypothesis, reduce trigger saliency and weaken presupposition projection.

[3]This dual evaluation, against both gold and theoretical labels, was applied consistently across all subsets.

Relatedness to Premise: Subset 4

Subset 4 examines how modifying the premise, either in the antecedent or consequent, affects models' behavior, focusing on two key phrases: K1 and K2. We compared classification accuracy and attention-based metrics (e.g., K1–K2 attention, special token attention) across related and unrelated variants. The aim is to show how context modifications influence the salience of presuppositional content and whether models maintain focus on the intended trigger.

5 Results and Analysis

Zero-shot Evaluation on Presupposition Triggers

We first analyzed the models' zero-shot behavior on unmodified conditionals. Generally, both RoBERTa and DeBERTa achieved high accuracy (>97%) against the NLI gold labels, and 0% accuracy with respect to theoretical labels. Llama only achieved an accuracy between 80% and 83%. On modified examples from subsets 2 and 4, the models also matched the human gold labels except for subset 3 where models only correctly predicted the labels as neutral 20% to 52% of the time. Gemma performed poorly with less than 47% of examples predicted correctly across all types.

Influence of Structural Modifiers

In evaluating the fine-tuned models on subset 2 examples, RoBERTa achieved high accuracy (0.99–1.0) and strong trigger IG ratios, but its saliency on presupposition triggers declined when structures were modified (Figure 1). Both keyword percentile and POS-based IG dropped (e.g., the noun phrase (NN) in possessive constructions). DeBERTa also maintained high accuracy in line with human labels but showed lower and more evenly distributed IG, indicating reduced sensitivity to structural changes. LLaMA and Gemma showed greater variability in trigger IG, including negative values, and higher misclassification rates (6–32%), suggesting weaker reliance on trigger phrases.

Lexical Relatedness and Heuristic

In the experiment on subset 3, replacing the trigger phrase resulted in almost all examples misclassified (e.g., accuracy <4% for type 5_A across all models), with models continuing to predict E despite there no longer being a relationship between the trigger phrase and the hypothesis. RoBERTa

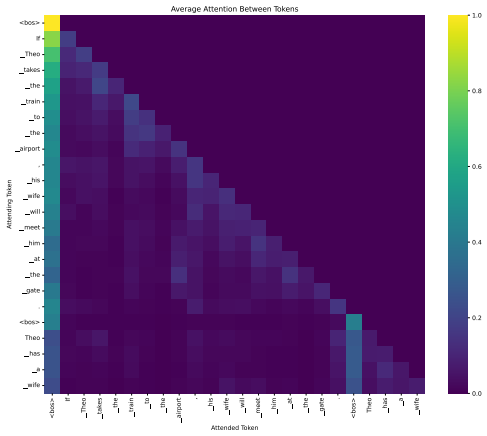

Figure 2: Average attention between tokens across all layers and heads in Gemma, for the sentence *If Theo's train is delayed, his wife will wait for him at the station*, and its hypothesis *Theo has a wife*. Brighter cells indicate higher attention values. First column corresponds to the special token attention.

showed strong reliance on triggers in related contexts, with notable drops in both attribution and accuracy as relatedness decreased. DeBERTa also distinguished between levels of relatedness but consistently showed lower trigger IG values. In contrast, LLaMA and Gemma were less consistent: LLaMA occasionally showed increased trigger IG in less related contexts (e.g., IG = 1.64 in type 4 Unrelated), while Gemma's attributions were often low or even negative across all conditions. (e.g., IG = –0.58 in type 5 Related).

Structural Change and Attention Drift

In these experiments, all models maintained accuracy near 100% , in line with human labels, on the modified type 4_p and type 5_p examples. However, they all misclassified most of the type 5_a examples (6% - 47% accuracy).

RoBERTa showed stable K1→K2 attention between related and unrelated cases for type 4_P (e.g., ∼1.7–1.8), but a significant difference observed in type 5 examples (t-test, p = 0 for type 5_a). On the other hand, DeBERTa showed stronger K1→K2 attention overall, especially in related contexts.

LLaMA and Gemma showed high K1→K2 attention (e.g., 3.95 in Related, 5.00 in Unrelated for type 5_a in Gemma). LlaMa also showed high attention from K2 to the special token (e.g., >1.1× average). Higher attention values are also observed on the original examples compared to the modified examples, especially in type 5_a examples, where this difference is significant across all models (and where models also misclassify the most). Figure 2 illustrates the average attention between tokens in Gemma, and Figure 3 shows an example of IG in type 5_p.

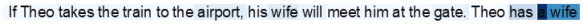

Figure 3: Visualization of the IG with respect to each input token where the darker the word, the higher its IG value.

What This Means for the Proviso Debate

Our results show that presupposition projection is fragile in language models. Attention towards the model's classification tokens drifts due to distracting phrases, and attribution to implied meaning weakens even with minor changes. This highlights that language models still struggle with pragmatic aspects of meaning, such as presupposition.

Regarding the proviso problem, classic theories treat projection as binary (Karttunen, 1973), while recent works argue that it depends on independence between antecedent and presupposition (Mandelkern, 2016). Our findings support the latter view: when A and *p* are logically independent, models show graded and context-sensitive behavior, influenced by how attention flows through the sentence. Although models often fail on nuanced semantic tasks, these failures provide diagnostic insights into how attention and attribution patterns reflect underlying generalizations about meaning in natural language.

6 Conclusion and Future Work

Based on the results, models are mostly in agreement with humans as opposed to theoretical-based NLI labels. They reliably identified trigger phrases even in modified structures, as evidenced by high accuracy and trigger IG ratios above 1 in subset 2. However, subset 3 modifications indicate that the models can easily be fooled by replacing the trigger phrase. Findings on subset 4 also show that even on fine-tuned models, phrases that are irrelevant to the NLI relationship influence attention-based computations and can end up producing significant misclassifications.

We suggest that future research extend presupposition analysis to more complex sentence structures. Parallel human experiments are also necessary, as many longstanding semantic theories remain untested empirically.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Tara Azin, Daniel Dumitrescu, Diana Inkpen, and Raj Singh. 2025. Let's confer: A dataset for evaluating natural language inference models on conditional inference and presupposition. In *Proceedings of the Canadian Conference on Artificial Intelligence (CAIAC)*. Retrieved from https://caiac.pubpub.org/pub/keh8ij01.

David Ian Beaver. 1997. Presupposition. In Johan Van Benthem and Alice Ter Meulen, editors, *Handbook of Logic and Language*, chapter 17, pages 939–1008. North-Holland.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Bart Geurts. 1996. Local satisfaction guaranteed: A presupposition theory and its problems. *Linguistics and Philosophy*, 19:259–294.

Irene Heim. 1983. On the projection problem for presuppositions. In Paul Portner and Barbara H. Partee, editors, *Formal Semantics: The Essential Readings*, pages 249–260. Blackwell.

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.

Lauri Karttunen. 1973. Presuppositions of compound sentences. *Linguistic Inquiry*, 4(2):169–193.

David Lewis. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1):339–359.

Matthew Mandelkern. 2016. Dissatisfaction theory. In *Proceedings of the 26th Semantics and Linguistic Theory (SALT)*, pages 391–416.

Clemens Mayr and Jacopo Romoli. 2016. Satisfied or exhaustified: an ambiguity account of the proviso problem. In *Semantics and Linguistic Theory*, pages 892–912.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Alyssa Parrish, Sebastian Schuster, Alex Warstadt, et al. 2021. Nope: A corpus of naturally-occurring presuppositions in english. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366.

R. Singh. 2020. Matrix and embedded presuppositions. In *The Wiley Blackwell Companion to Semantics*, pages 1–42. Wiley-Blackwell.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 4444–4451. AAAI Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Kai Von Fintel. 2008. What is presupposition accommodation, again? *Philosophical Perspectives*, 22(1):137–170.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.