Presupposition and Reasoning in Conditionals: A Theory-Based Study of Humans and LLMs

Tara Azin[1]*, **Yongan Yu**[2,3], **Raj Singh**[1], **Olessia Jouravlev**[1]

[1]Department of Cognitive Science, Carleton University
[2]School of Computer Science, McGill University
[3]Mila – Quebec AI Institute

Abstract

Presupposition projection in conditionals is central to theories of meaning and pragmatics, yet it remains largely unevaluated in large language models. We address this gap through a parallel behavioral study comparing human judgments and LLM predictions on a normed dataset of conditional sentences that controls the relation between the antecedent and the projected presupposition. We collect likelihood ratings from 120 participants and four LLMs under matched contextual conditions. Results show that humans integrate probabilistic and pragmatic cues in their judgment, whereas LLMs show variable alignment with human patterns. Using a linguistically motivated checklist within an LLM-as-a-Judge framework, we further evaluate model reasoning. We observe models that best match human ratings often lack coherent pragmatic reasoning, while models with stronger reasoning produce less human-like judgments. These findings suggest that LLMs' performance on such tasks may result from surface pattern matching rather than pragmatic competence. Our findings highlight the importance of benchmarks grounded in linguistic theory for comparing humans and models.

1 Introduction

Understanding how implicit meaning is inferred in context has remained a complicated challenge for theories of natural language interpretation. A classic puzzle is the proviso problem (Geurts, 1996) in presupposition projection, which concerns how presuppositions triggered in the consequent of conditionals are interpreted. In some cases, the presupposition is clearly understood as unconditional (e.g., *If John flies to London, his sister will pick him up* presupposes that John has a sister). In other cases, however, the presupposition may be understood either unconditionally or conditional on the antecedent (e.g., *If John is a scuba diver, he will bring his wetsuit*). According to satisfaction theories of presupposition projection (Heim, 1983; Beaver, 2001; Schlenker, 2008), a sentence of the form *If* A, B_p, where B contains a presupposition trigger, and p is the presupposition of B, projects the conditional presupposition $A \rightarrow p$. In cases where accommodation is required[1] (Lewis, 1979; von Fintel, 2008), listeners may either accommodate $A \rightarrow p$ or strengthen it to p itself (Singh, 2007, 2020). The proviso problem thus concerns how this accommodation decision is made and what role plausibility and context-sensitive reasoning play in determining whether p is interpreted conditionally or unconditionally (Mandelkern, 2016; Mandelkern and Rothschild, 2019). Despite decades of theoretical work, no consensus has been reached on this problem.

Recent work evaluates the linguistic capabilities of large language models (LLMs) (Hale and Stanojević, 2024; Søgaard, 2025), but existing studies suggest that they often struggle with semantic and pragmatic reasoning tasks compared to human judgments (Sravanthi et al., 2024). From a linguistic perspective, as noted above, presupposition reasoning reflects the interaction of semantics, pragmatics, discourse context, and probabilistic reasoning, and is sensitive to relevance and world knowledge (Mandelkern, 2016; Domaneschi et al., 2016). However, these interactions have not yet been examined in controlled human–LLM comparisons.

In this paper, we present a comparison of human and LLM presupposition judgments based on the proviso problem in conditional sentences. We investigate how probabilistic relevance between the antecedent and the presupposition affects projection patterns in conditionals of the form *If* A, B_p, where B contains a possessive trigger and p is the

*Corresponding author: taraazin@cmail.carleton.ca

[1]Accommodation refers to the pragmatic process by which listeners update the common ground to satisfy a presupposition when it is not already entailed.

Proceedings of the 30th Conference on Computational Natural Language Learning, pages 452–470

presupposition of B. We construct a dataset of 90 sentences based on 30 base propositions, each instantiated in three variants that manipulate the logical and probabilistic relationship between A and p. We collect likelihood ratings from 120 human participants and four LLMs on a 0–7 Likert scale (0 = very unlikely, 7 = very likely), with and without minimal contextual information. This design allows us to examine how presupposition judgments vary with A–p relevance and how they are influenced by contextual information. The design is in line with current works using parallel human-LLM comparisons to assess linguistic competence (Qiu et al., 2024).

In addition, we conduct an LLM-as-a-judge experiment to analyze the reasoning underlying models' predictions. Using a theory-informed checklist that is grounded in formal semantics and pragmatic principles, a judge model evaluates whether explanations generated by the models reflect valid inferential patterns. Human experts are involved in both designing the checklist and in the meta evaluation stage. We argue that this approach moves beyond evaluation that is solely based on accuracy metrics toward a more interpretable analysis of model behavior, and is better suited to semantic and pragmatic tasks such as presupposition handling.

Our study addresses the following research questions: (i) How do human presupposition judgments in conditional sentences vary with antecedent-presupposition relevance? (ii) How closely do LLM judgments align with human patterns across contextual conditions? (iii) How does minimal discourse context influence interpretation for humans and models? (iv) To what extent do LLM explanations represent theoretically grounded reasoning about presupposition projection?[2]

2 Related Work

2.1 Human-Machine Presupposition Studies

Research on presuppositional reasoning in language models is relatively recent. Prior work has examined conditional inference and presupposition judgment using prompt-based and NLI-style evaluations (Holliday et al., 2024; Atwell et al., 2025). Earlier studies evaluated models on datasets such as IMPPRES, NOPE, and PROPRES (Jeretic et al., 2020; Parrish et al., 2021; Asami and Sugawara, 2023). These datasets mainly focus on entailment reasoning with simple conditional structures and do not address pragmatic factors that influence presupposition projection. As a result, they offer limited coverage of presupposition projection in complex conditionals. Overall, existing work primarily adopts classification-based evaluation and rarely conducts controlled behavioral comparisons with human judgments. Our study addresses this gap through a theory-driven human-LLM comparison of presupposition projection in conditionals.

2.2 Pragmatic Evaluation of LLMs

Recent work has examined whether LLMs exhibit pragmatic competence in areas such as implicature, presupposition, and reference (Jeretic et al., 2020; Kabbara and Cheung, 2022). The Pragmatics Understanding Benchmark (Sravanthi et al., 2024) provides the only large-scale theory-based resource with parallel human-LLM evaluation across multiple pragmatic phenomena. Azin et al. (2025) introduce CONFER, an NLI benchmark for this phenomenon, showing that models fail to generalize presuppositional reasoning to complex conditional structures. They further probe this behavior through explainability analyses, finding that models broadly align with human judgments on the proviso problem but rely on shallow pattern matching rather than pragmatic reasoning (Azin et al., 2026). Our work builds on this line of research by focusing specifically on presupposition in conditional contexts and by providing a behavioral comparison based on pragmatic theories.

2.3 LLM-as-a-Judge

Traditional automatic metrics such as BLEU and ROUGE are limited in evaluating semantic and pragmatic reasoning (Papineni et al., 2002; Lin, 2004). Recent work has therefore explored more interpretability-oriented LLM-as-a-Judge approaches (Zheng et al., 2023). Lee et al. (2025b) propose a checklist-based framework that decomposes evaluation criteria into interpretable binary judgments. Building on this approach, we adopt and extend a theory-informed checklist tailored to presupposition inference.

3 Methodology

Our methodology consists of four stages: (1) a norming study to construct a controlled dataset, (2) collection of parallel presupposition judgments and reasoning from human participants and LLMs, (3)

[2]Codes and dataset are available at https://github.com/proviso-bench/Presupposition-and-Reasoning-in-Conditionals

development of a checklist for evaluating pragmatic reasoning, and (4) automated evaluation using an LLM-as-a-Judge framework, followed by human validation.

3.1 Data Construction and Norming

Our dataset design is motivated by probabilistic accounts of presupposition accommodation in conditionals. Based on existing theories, when interpreting the presupposition (p) of a conditional sentence of the form *If* A, B_p, where p is the direct presupposition of the consequent B, listeners compare how likely p is in general given background context c[3], $Pr(p \mid c)$, with how likely p is under the assumption that A holds, $Pr(p \mid A, c)$ (Carnap, 1950; Strawson, 1950; Stalnaker, 1973, 1998). In simple words, listeners consider whether assuming that A holds makes p much more likely to be true. If it does, they tend to interpret p as holding only under the condition A; if it does not, they tend to interpret p as generally true. For example, in *If John is a scuba diver, then he will bring his wetsuit*, being a scuba diver (A) makes having a wetsuit (p) much more likely, whereas in *If John flies to London, then his sister will pick him up*, flying to London (A) does not affect whether John has a sister (p). To operationalize this idea, we conduct a norming study to quantify and validate the probabilistic relevance between antecedents (A) and presuppositions (p). This allows us to obtain graded levels of A-p relevance, which are used to construct the main study items for presupposition judgment.

We construct 30 base propositions corresponding to presupposed content (e.g., having a guitar, having an apron, having a boat). These propositions span high-probability ownerships (e.g., having a smartphone), moderate-probability cases (e.g., having siblings), and low-probability cases (e.g., having a boat). We also include neutral contextual constraints (e.g., someone being at a gym or at an airport) to restrict the possible world without directly affecting p.

For each proposition, we design four norming conditions: a baseline condition measuring $Pr(p \mid c)$ and three conditional criteria measuring $Pr(p \mid A, c)$ with high, mid, and low A–p relevance antecedents. For example, for *having a guitar* in the context *someone at a gym*, we ask about the likelihood that the person (i) has a guitar (baseline), (ii) is a musician and has a guitar (high relevance), (iii) likes music and has a guitar (mid relevance), and (iv) speaks French and has a guitar (low relevance). Participants (thirty native English speakers residing in the US, Canada, and the UK, recruited via Prolific[4]) rated 120 items on a 1–7 Likert scale by judging how likely each statement was to be true, with items presented in randomized order.

The norming results show clear statistical separation between low, mid, and high probability conditions, confirming our initial predictions. We therefore excluded the baseline items and retained the 90 conditional items for the main study. Each main study item followed the form *If* A, B_p, with antecedents selected to induce relevant, somewhat relevant, or irrelevant A-p relations, and with target propositions representing low-, mid-, and high-probability categories derived from the norming data (e.g., having a watch, having a smartphone).

For example, in *If Daniel drinks tea, he will make tea for his sister*, the antecedent A (*drinks tea*) is semantically unrelated to the presupposition p (*Daniel has a sister*), yielding an irrelevant A-p relation, and the proposition *having a sister* corresponds to a mid-probability category based on norming results (see Appendix B for more examples).

3.2 Human and LLM Evaluation

Human Participants We recruit 128 native English speakers via Prolific. All participants report being born and raised speaking English and having no history of neurological or cognitive conditions affecting language comprehension. Two attention-check questions are included, and participants who fail either check are excluded. 120 participants were retained after applying the exclusion criteria (see Appendix A for details).

Both human participants and LLMs receive identical instructions. The full instructions and prompts are reported in Appendix B and Appendix F.

Experimental Procedure Participants complete a presupposition judgment task consisting of the same 90 conditional sentences of the form *If* A, B_p. They are instructed to assume that the speaker is honest, reliable, and cooperative, and that relevant background assumptions are shared. For each item,

[3] By background context, we refer to shared world knowledge and situational assumptions available to interlocutors, such as general facts about people, places, and everyday situations.

[4] https://www.prolific.com

Dimension	Theory Base	Examples of What It Tests
Accuracy	Discourse Representation Theory (DRT)	Trigger and anaphora detection
Context	Dynamic Semantics	Robustness to context effects
Pragmatic	Gricean Pragmatics	Relevance and common ground reasoning
Presupposition Handling	Projection and Accommodation Theory	Projection and accommodation control
Coherence	Argumentation Theory	Consistency between reasoning and judgments

Table 1: Checklist dimensions and representative evaluation criteria used by the LLM-as-a-judge model to assess LLM reasoning steps against semantic and pragmatic theories in conditional presupposition tasks.

participants rate how likely the presupposition p is to be true on a 0–7 Likert scale (0 = very unlikely, 7 = very likely). Examples of this task are provided in Appendix B.2, Table 5.

The experiment follows a between-participants design with two conditions: *without-context* and *with-context*. All participants in both conditions answered the same 90 items, however, in the with-context condition, each item additionally included a brief identifying background description (e.g., the person being discussed is from Toronto). Providing minimal identifying context allows us to show how background information modulates perceived evidential reliability. This is consistent with prior work showing that contextual cues influence how speakers evaluate the strength of testimonial and inferential evidence (Lesage et al., 2015). The rationale for this manipulation is based on psycholinguistic work on context-driven interpretation (Crain and Steedman, 1985). In the absence of any contextual framing, participants must implicitly decide which situation or possible world they are meant to reason about, leaving many parameters of the interpretive problem underspecified. Even a minimal context line, such as identifying where an individual is from, serves two related functions. First, it fixes parameters that would otherwise remain open (e.g., the individual's background properties), and it increases the salience of those parameters, potentially triggering relevance-based reasoning about whether the contextual information bears on the target presupposition (Sperber and Wilson, 1986). These are genuine contextual changes, not trivial additions, and our design allows us to examine how much such changes actually shift presupposition judgments in both humans and LLMs.

To reduce fatigue in the longer with-context condition, two mandatory 8-second pauses were inserted near the beginning and end of the task. All items were presented in randomized order, and participants completed the task in a single session.

LLM Setup and Prompting We evaluate four LLMs: GPT-5, Gemini-2.5-flash, Llama-3.1-8B-Instruct, and Qwen2.5-7B-Instruct. Model versions, access methods, and generation parameters are reported in Appendix D, and the full decoding configuration in Appendix E. Each model receives the same instructions and stimuli as human participants. All models are instruction-tuned versions, denoted as "IT". Details are in Appendix F.

Each model is presented with the same 90 items under both experimental conditions. Models produce (i) a likelihood judgment on the same 0–7 Likert scale and (ii) an output of CoT reasoning steps. This yields parallel human and model judgments and corresponding reasoning traces.

3.3 LLM-as-a-Judge Framework

Because presupposition inference relies on pragmatic reasoning, evaluating final judgments, made by LLMs, alone is insufficient. We therefore adopt an LLM-as-a-Judge framework inspired by Lee et al. (2025b) and Yu et al. (2025), which enables structured assessment of reasoning quality of the models using theory-driven criteria.

Checklist Design We design a checklist that decomposes pragmatic competence into explicit yes/no questions. Seed questions are created by a linguistics expert based on theories of presupposition accommodation, conditional semantics, and pragmatic reasoning, as well as coherence and consistency criteria (Wang et al., 2020; Fabbri et al., 2021). These questions are grouped into four dimensions: logical accuracy, presupposition handling, pragmatic criteria, and coherence. For the with-context condition, we add a fifth dimension that addresses context integration.

Each dimension is further divided into subdimensions targeting specific reasoning behaviors, based on established work in discourse and dynamic semantics, presupposition projection, and pragmatic reasoning (Grice, 1975; Heim, 1983; Kamp and Reyle, 1993; Beaver, 2001). Accuracy

evaluates whether the conditional structure is correctly interpreted, presupposition handling assesses identification of presupposition triggers and the distinction between presupposition and entailment, pragmatic criteria examine cooperativity, relevance, and common ground, and coherence evaluates consistency between numerical judgments and stated reasoning. Table 1 summarizes these dimensions, their theoretical bases, and representative evaluation criteria.

To expand coverage, we use `Llama-3.1-8B-Instruct` to diversify and augment the seed questions and then further refine them through manual elaboration. The same model is used to filter redundant items. All questions are binary, with "yes" indicating alignment with theoretical expectations. The final checklist is reviewed by a linguist to ensure theoretical validity and non-redundancy.

The resulting checklist contains 59 questions for the with-context condition (5 dimensions) and 52 questions for the without-context condition (4 dimensions). A sample is provided in the Appendix G.[5] Thus, we define the total criteria space as:

$$N = \underbrace{|M|}_{\text{Models}} \times \left(\underbrace{|I_{\text{wc}}| \times |\mathcal{K}_{\text{wc}}|}_{\text{With-Context}} + \underbrace{|I_{\text{nc}}| \times |\mathcal{K}_{\text{nc}}|}_{\text{Without-Context}} \right)$$

where $|M|$ is the number of models, $|I_{\text{wc}}|$ and $|I_{\text{nc}}|$ are the numbers of items in the with-context and without-context conditions, respectively, and $|\mathcal{K}_{\text{wc}}|$ and $|\mathcal{K}_{\text{nc}}|$ are the corresponding numbers of checklist questions.

Judge Model Configuration Given the scale of the model checkpoints and the complexity of pragmatic reasoning tasks, we employ an LLM-as-a-Judge framework to enable scalable and consistent evaluation of model reasoning. The judge evaluates whether each reasoning trace satisfies the pragmatic constraints specified by the checklist.

We use `claude-haiku-4` as the judge model, based on preliminary experiments indicating strong performance on the recent fact-checking benchmark (Theologitis et al., 2026), while also mitigating potential bias arising from shared training corpora within the same provider. For each evaluation instance, the judge receives an input tuple:

$$\mathcal{I} = (c, s, \tau, r_{\text{CoT}})$$

[5]The full checklist is available in our GitHub repository.

where c is the discourse context (or $\emptyset$ in the without-context condition), s is the conditional sentence, τ is the target presupposition, and r_{CoT} is the reasoning trace generated by the evaluated model.

Evaluation Procedure A prompt template P presents (c, s) and instructs models to act as pragmatic listeners (see Appendix F for the prompts). Each model first produces an explicit reasoning trace and then provides a final likelihood judgment. Finally, we deploy the judge model configured in the previous section. For a given model response r, the judge iterates through the condition-specific checklist $\mathcal{K}$. For each criterion $\kappa_j \in \mathcal{K}$, the judge model J produces a binary decision $J\big((c, s, \tau, r), \kappa_j\big)$, where 1 indicates that the response satisfies the criterion and 0 otherwise. The final score (S) for that response is computed as the average compliance across all criteria:

$$S = \frac{1}{|\mathcal{K}|} \sum_{j=1}^{|\mathcal{K}|} J\big((c, s, \tau, r), \kappa_j\big), J(\cdot) \in \{0, 1\}$$

This procedure yields both aggregate performance metrics and granular, dimension-level diagnostics. To assess the reliability of this automated evaluation, we conduct a human validation study on a subset of the judged outputs below.

Human Validation Two PhD students in linguistics independently evaluated a randomly sampled 5% of the outputs (36 items, 1,992 binary judgments), drawn with equal representation across models and conditions. Inter-annotator agreement was 89%, (exact match accuracy) and agreement with the judge model was 79.46%.

4 Experiments and Results

4.1 Norming Results

The norming study results closely matched our intended low, mid, and high probability conditions. Mean ratings increased monotonically across relevance levels, with low probability items receiving lower ratings (M = 2.76), mid probability items intermediate ratings (M = 4.38), and high probability items the highest ratings (M = 5.52).

We analyzed the results using a linear mixed-effects model (Baayen et al., 2008) with expected probability level (low, mid, high) as a fixed effect and random intercepts for participants and items. This approach allows us to account for repeated

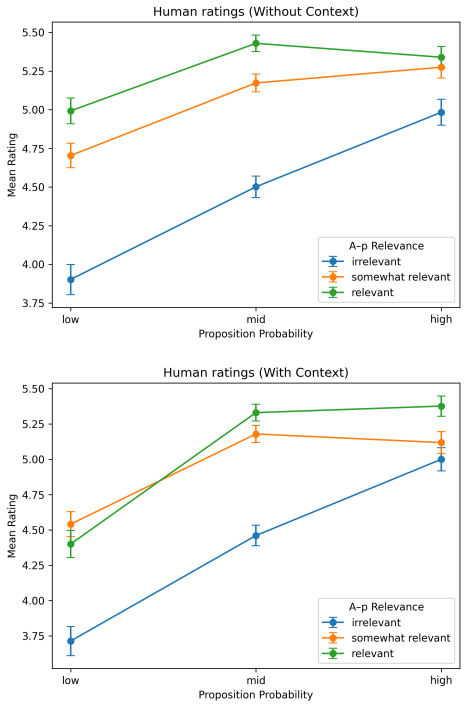

Figure 1: Human mean Likert scores in the main presupposition judgment experiment, across participants and items as a function of proposition probability (low, mid, high) and antecedent–presupposition ($A - p$) relevance (relevant, somewhat relevant, irrelevant). Top: without-context condition; bottom: with-context condition. Error bars indicate standard errors.

		Probability Level (0–7)		
Model	**Type**	**Rel.**	**s/w Rel.**	**Irrel.**
Closed Source				
Gemini-2.5-flash	w Context	6.43	6.57	6.87
	w/o Context	6.60	6.53	6.93
GPT-5	w Context	5.43	5.87	6.17
	w/o Context	5.83	5.93	6.53
Open Source				
Llama3.1-8B-IT	w Context	5.07	4.70	4.30
	w/o Context	5.60	4.83	5.17
Qwen2.5-7B-IT	w Context	6.00	5.80	5.73
	w/o Context	6.37	6.20	6.03

Table 2: Mean likelihood ratings (0–7 scale) produced by each model across A-p relevance (relevant, somewhat relevant, irrelevant) under with-context and without-context prompting conditions. Shaded rows indicate the with-context condition. IT abbreviates Instruct.

measurements as well as variability across participants and items, making it appropriate for our experimental design. The model showed a significant effect of probability level, with ratings increasing from low to mid probability ($\beta = 1.62$, $p < .001$) and from low to high probability ($\beta = 2.75$, $p < .001$). These results confirm that the norming study can successfully be used as the basis for graded A-p probability, as intended.

4.2 Human Performance

A-p Relevance and Probability We analyzed human judgments using linear mixed-effects models with fixed effects of proposition probability (low, mid, high), A-p relevance (relevant, somewhat relevant, irrelevant), and their interaction, with random intercepts for participants. Proposition probability was treated as a three-level categorical factor rather than a continuous predictor to preserve alignment with the experimental design, in which items were constructed to instantiate discrete probability levels, and to facilitate interpretable comparison across conditions. Full model output, including the model formula and fit statistics, is reported in Appendix C (Table 6). Separate models were fitted for the without-context and with-context conditions. Figure 1 summarizes the resulting response patterns. In the without-context condition, items with high proposition probability and high A-p relevance (e.g., having a smartphone in *If Alex is a college student, he will use his smartphone to take notes*) received the highest ratings (M $\approx$ 5.34). Low-probability items (e.g., *having a Rolex*) were rated significantly lower, while mid-probability items (e.g., *having a brother*) were not reliably distinguished from high-probability ones. Similarly, items with irrelevant antecedents (e.g., *liking coffee* (A) and *having a wetsuit* (p)) received substantially lower ratings, whereas somewhat relevant items (e.g., *liking swimming* and *having a wetsuit*) were treated similarly to highly relevant ones. A significant interaction showed that low and mid probability led to especially strong penalties when A and p were irrelevant (low $\times$ irrelevant: $\beta = -0.73$, $z = -5.47$, $p < .001$; mid $\times$ irrelevant: $\beta = -0.57$, $z = -4.61$, $p < .001$; see Table 6). This indicates that participants combined probability and relevance in a non-additive manner, consistent with probabilistic accounts that predict interactive effects between antecedent relevance and prior likelihood. Notably, when A and p were irrelevant, ratings tracked proposition probability closely, suggesting that participants fell back on prior likelihood in the absence of a meaningful antecedent-presupposition relation. In contrast, even partial $A - p$ relevance produced a marked boost in ratings, indicating that the relevance cue

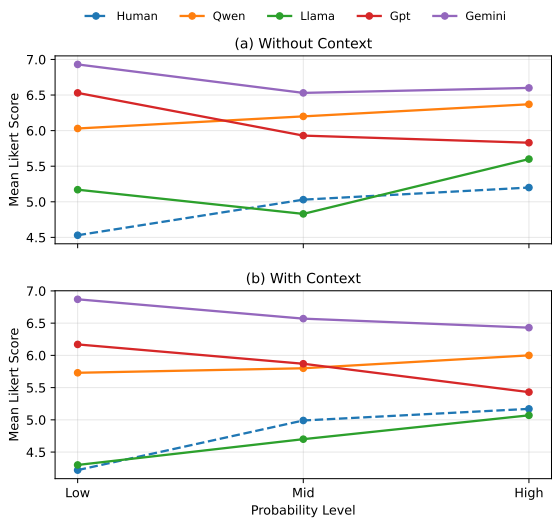

Figure 2: Mean Likert scores for human participants and each LLM in the main presupposition judgment experiment, as a function of proposition probability (low, mid, high) and $A - p$ relevance, under without-context (top) and with-context (bottom) conditions.

acts less as a graded scalar and more as a threshold.

In the with-context condition, baseline ratings for highly probable and highly relevant items remained similar (M $\approx$ 5.38), indicating no overall shift in response levels. However, participants used both cues more consistently. Low probability propositions showed a stronger decrease in ratings, and both somewhat relevant and irrelevant items were rated significantly lower than highly relevant ones. This suggests that contextual information encouraged more graded integration of probability and relevance.

Context vs. No-Context Effects Across both conditions, baseline judgments for highly probable and highly relevant items are nearly identical, suggesting that minimal context does not affect default interpretations in clear cases. However, context modulates how participants use probability and relevance cues. Without context, participants primarily distinguish between relevant and irrelevant items, with irrelevant A–p relations acting as a strong gating factor and probability playing a relatively modest role. In contrast, with context, participants show a more differentiated use of both cues, consistent with the interaction effects observed in the mixed-effects analysis. For illustration, consider the item *If Jack is a scuba diver, he will bring his wetsuit*, which exemplifies the overall pattern visible in Figure 1. Statistical conclusions are based on the full model reported above.

4.3 Human-LLM Comparison

We assessed the alignment between human and model judgments using Spearman rank correlations and mean absolute error (MAE). Table 2 summarizes the average model ratings across A-p relevance and context conditions. Figure 2 provides a visual comparison of human and model responses across probability levels under both contextual conditions. At the item level, models differed substantially in their degree of alignment with human judgments. Qwen-2.5-7B-IT showed the strongest and most reliable correlations in both conditions (without context: $\rho = 0.25$, $p = .016$; with context: $\rho = 0.38$, $p < .001$), very close to the human's ranking of items. Llama3.1-8B-Instruct also showed consistent moderate alignment (without context: $\rho = 0.21$, $p = .047$; with context: $\rho = 0.30$, $p = .004$). In contrast, Gemini-2.5-flash showed meaningful alignment only when context was provided ($\rho = 0.26$, $p = .013$). GPT-5 did not show significant overall correlations in either condition.

These patterns were mirrored in the MAE results (Figure 3). Llama3.1-8B-Instruct achieved the lowest overall error (MAE = 1.14), followed by Qwen2.5-7B-IT (1.32) and GPT-5 (1.34), whereas Gemini-2.5-flash showed the largest deviation from human judgments (1.90). For GPT-5 and Qwen2.5-7B-IT, MAE was slightly lower in the with-context condition, and Llama3.1-8B-Instruct and Gemini-2.5-flash showed comparable error levels across conditions. Overall, Qwen2.5-7B-IT and Llama3.1-8B-Instruct demonstrated the strongest and most reliable correspondence with human presupposition judgments based on Likert scale. Gemini-2.5-flash's performance depended on contextual information, and GPT-5 showed limited alignment across conditions.

To examine whether these quantitative patterns are reflected in the models' reasoning, we next used the LLM-as-a-Judge framework to evaluate whether their reasoning steps are consistent with the scores they assigned on the Likert scale.

4.4 Reasoning Analysis

To examine whether model explanations align with their assigned Likert scores, we analyze checklist-based evaluation results produced by the LLM-as-a-Judge framework. The breakdown of reasoning scores is presented in Table 3.

	Proprietary Models				Open-source Models			
	Gemini-2.5-flash		GPT-5		Llama3.1-8B-IT		Qwen2.5-7B-IT	
Category	**With Ctx (%)**	**Δ**	**With Ctx (%)**	**Δ**	**With Ctx (%)**	**Δ**	**With Ctx (%)**	**Δ**
Accuracy	61.39	↓ 6.39	60.74	↓ 2.50	16.67	↓ 13.98	22.96	↓ 13.80
Coherence	69.41	↑ 5.30	75.11	↑ 4.56	62.81	↑ 6.15	58.74	↑ 5.63
Pragmatic	82.84	↓ 5.63	78.89	↓ 3.19	54.94	↑ 5.63	56.42	↓ 3.86
Presupposition	61.11	↓ 12.93	71.02	↑ 0.51	49.35	↓ 8.32	42.04	↓ 15.44
Context Util.	30.10	–	28.18	–	14.75	–	12.42	–
TOTAL	**60.81**	**↓ 11.79**	**63.18**	**↓ 7.47**	**40.53**	**↓ 7.36**	**39.08**	**↓ 11.82**

Table 3: The Δ columns indicate the performance gap relative to the *Without Context* condition (Green ↑ indicates *With Context* scored higher; Red ↓ indicates *With Context* scored lower). IT abbreviates Instruct.

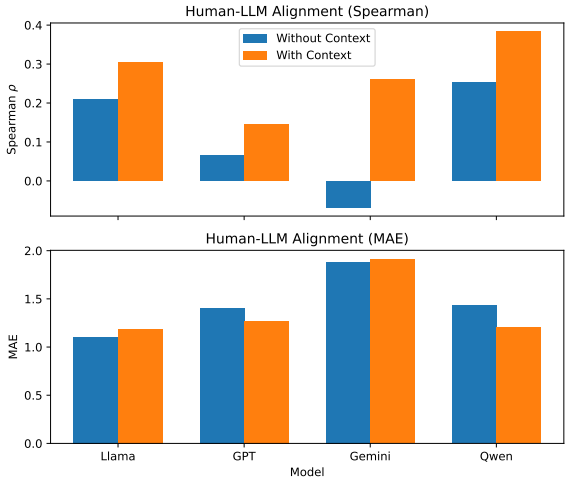

Figure 3: Human-LLM alignment measured using Spearman's rank correlation (ρ) and mean absolute error (MAE) under with-context and without-context conditions. Higher Spearman values and lower MAE indicate closer alignment between model predictions and human mean Likert scores.

Stronger Theoretical Alignment in Larger Instruction-Tuned Models Across evaluation dimensions, larger and more heavily instruction-tuned models in our comparison (Gemini-2.5-flash and GPT-5) achieve higher checklist compliance than smaller open-source models. In the with-context condition, proprietary models achieve total scores above 60%, while open-source models remain near 40%. This gap is particularly pronounced in the Accuracy and Presupposition dimensions.

These results align with scaling-law observations (Ruan et al., 2024) and the effects of instruction tuning (Jiang et al., 2024). Larger instruction-tuned models appear better at maintaining structured reasoning and tracking presuppositional dependencies across multi-step explanations. Nevertheless, even these models exhibit limitations in graded presupposition reasoning. For instance, GPT-5 assigns the same Likert rating (6) to both *If Jack is a scuba diver, he'll bring his wetsuit* and *If Sara likes coffee, she'll bring her wetsuit*. In its explanation, the model emphasizes the possessive trigger ("his wetsuit") and general ownership knowledge without evaluating whether the antecedent increases $Pr(p \mid A, c)$ relative to $Pr(p \mid c)$. This pattern suggests reliance on lexical cues rather than genuine sensitivity to presupposition relevance.

Context Improves Checklist Performance Providing minimal identifying context generally increases checklist compliance. In particular, Llama-3.1-8B-IT and Qwen2.5-7B-IT show notable gains in the Accuracy and Presupposition dimensions under the with-context condition. Proprietary models also benefit in some dimensions, although improvements are more modest.

This trend aligns with prior work (Lee et al., 2025a), which finds that abstract context and world knowledge do not shift baseline interpretations but increase sensitivity to probabilistic structure. For models, however, context often introduces reasoning noise, as attempts to incorporate contextual details frequently fail to align consistently with conditional semantics, leading to reduced coherence compliance.

Dissociation Between Theoretical Reasoning and Probabilistic Presupposition We observe a clear dissociation between behavioral alignment (Likert ratings) and explicit reasoning quality (checklist compliance). As shown in Table 2 and Figure 3, Qwen2.5-7B-IT is most closely aligned with human Likert ratings across all conditions. However, the reasoning results in Table 3 reveal a paradox: Qwen2.5-7B-IT shows the lowest average compliance with our theory-informed checklist.

This pattern suggests that smaller models ap-

proximate human-like response distributions by relying on surface-level lexical associations and distributional world knowledge, without consistently implementing the formal pragmatic constraints underlying the proviso problem. In contrast, proprietary models generate more structured and theory-consistent reasoning steps, yet their final Likert ratings deviate more from human judgments. We interpret this dissociation as suggesting that behavioral similarity does not necessarily reflect stable pragmatic competence. High alignment in ratings may arise from probabilistic pattern matching over training data, rather than explicit modeling of relevance relations between antecedent and presupposition. At the same time, we acknowledge that checklist compliance may partly reflect verbalization quality rather than reasoning depth, given evidence that chain-of-thought traces can function as post-hoc rationalizations in instruction-tuned models (Turpin et al., 2023). The dissociation is therefore consistent with either a genuine competence gap in smaller models or a verbalization advantage in larger ones, and disentangling these interpretations is an important direction for future work.

5 Conclusion

This study presents an empirical comparison of human and LLM presupposition judgments in conditional sentences grounded in pragmatic theory. Using a norming study and parallel behavioral experiments, we show that human presupposition judgments are guided by intra-sentential relationships (e.g., between antecedent and presupposition), proposition probability, and contextual information. In contrast, LLMs show varying degrees of alignment, with Qwen2.5-7B-IT closest to human patterns, but generally struggle with graded presupposition accommodation. Our LLM-as-a-Judge analysis reveals a dissociation between behavioral alignment and reasoning quality. Models that best match human ratings often fail to satisfy theory-derived pragmatic criteria, while models that better meet theoretical standards produce less human-like judgments. This suggests that current performance may reflect surface-level pattern matching rather than stable pragmatic understanding.

Future work should extend this framework to additional presupposition triggers and more complex constructions. Theory-driven evaluation remains crucial for characterizing semantic and pragmatic reasoning in language models and improving human–model alignment.

Limitations

Our study focuses on conditional sentences of a specific structural form (*If* A, B_p) and investigates presupposition projection through possessive pronoun triggers (e.g., *his*, *her*, *their*). We focus on possessive pronouns because they introduce simple existential presuppositions that are well studied in the projection literature (e.g. Karttunen 1973, Heim 1983) and allow controlled testing of the A-p relevance without additional lexical confounding. As a result, our findings may not generalize to other types of conditionals, presupposition triggers (e.g., factives, change-of-state verbs), or more complex embedding environments.

Moreover, our human data is collected from English speaking participants in a limited set of countries, which may restrict cross-linguistic and cross-cultural generalizability. Although we validated a subset of the LLM-as-a-Judge outputs with human annotators due to practical cost considerations, future work could expand this validation to additional samples to further support and refine the evaluation framework.

Finally, the LLM-as-a-Judge checklist operationalizes pragmatic competence through satisfaction-theoretic principles, which may not fully capture the heuristic and probabilistic nature of human presupposition accommodation. The observed dissociation between behavioral alignment and checklist compliance is therefore potentially ambiguous: it may reflect a genuine competence gap in models, or it may reflect a mismatch between the formal criteria encoded in the checklist and the cognitive processes that actually drive human judgments. Future work could complement theory-driven evaluation with process-level measures to disentangle these interpretations.

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.

Katherine Atwell, Mandy Simons, and Malihe Alikhani. 2025. Measuring bias and agreement in large language model presupposition judgments. In *Find-*

ings of the Association for Computational Linguistics: ACL 2025, pages 2096–2107, Vienna, Austria. 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. *arXiv preprint arXiv:2506.06133*.

Tara Azin, Daniel Dumitrescu, Diana Inkpen, and Raj Singh. 2026. Do language models know theo has a wife? investigating the proviso problem. *arXiv preprint arXiv:2603.08358*.

R. Harald Baayen, Douglas J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

David I. Beaver. 2001. *Presupposition and Assertion in Dynamic Semantics*. CSLI Publications, Stanford, CA.

Rudolf Carnap. 1950. *Logical Foundations of Probability*. University of Chicago Press, Chicago, IL.

Stephen Crain and Mark Steedman. 1985. *On not being led up the garden path: the use of context by the psychological syntax processor*, page 320–358. Studies in Natural Language Processing. Cambridge University Press.

Filippo Domaneschi, Elena Carrea, Carlo Penco, and Alberto Greco. 2016. Selecting presuppositions in conditional clauses. results from a psycholinguistic experiment. *Frontiers in Psychology*, 6.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

Herbert P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*, pages 41–58. Brill.

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

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.

Wesley H. Holliday, Matthew Mandelkern, and Cedegao E. Zhang. 2024. Conditional and modal reasoning in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3800–3821, Miami, Florida, USA. Association for Computational Linguistics.

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.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. 2024. Instruction-tuned language models are better knowledge learners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5421–5434.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model-Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht.

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

Seungpil Lee, Woochang Sim, Donghyeon Shin, Wongyu Seo, Jiwon Park, Seokki Lee, Sanha Hwang, Sejin Kim, and Sundong Kim. 2025a. Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–52.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025b. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15771–15798, Suzhou, China. Association for Computational Linguistics.

Clair Lesage, Nalini Ramlakhan, Ida Toivonen, and Chris Wildman. 2015. The reliability of testimony and perception: Connecting epistemology and linguistic evidentiality. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, Pasadena, CA, USA.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Matthew Mandelkern and Daniel Rothschild. 2019. Independence day? *Journal of Semantics*, 36(2):193–210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2024. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 189–198, Bangkok, Thailand. Association for Computational Linguistics.

Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. 2024. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892.

Philippe Schlenker. 2008. Presupposition projection: Explanatory strategies. *Theoretical Linguistics*, 34(3):287–316.

Raj Singh. 2007. Formal alternatives as a solution to the proviso problem. In *Semantics and Linguistic Theory (SALT) 17*.

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

Anders Søgaard. 2025. Do language models have semantics? on the five standard positions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25910–25922, Vienna, Austria. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Blackwell.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Robert Stalnaker. 1973. Presuppositions. *Journal of Philosophical Logic*, 2(4):447–457.

Robert Stalnaker. 1998. On the representation of context. *Journal of Logic, Language and Information*, 7(1):3–19.

Peter F. Strawson. 1950. On referring. *Mind*, 59(235):320–344.

Michael Theologitis, Preetam Prabhu Srikar Dammu, Chirag Shah, and Dan Suciu. 2026. Claimdb: A fact verification benchmark over large structured data. *arXiv preprint arXiv:2601.14698*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc.

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

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Yongan Yu, Mengqian Wu, Yiran Lin, and Nikki G Lobczowski. 2025. Think: Can large language models think-aloud? *arXiv preprint arXiv:2505.20184*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Human Study Materials

A.1 Norming

The norming study included 30 native English speakers (18 men, 12 women), aged between 25 and over 65. Participants were primarily based in the United States ($n = 22$), with other participants from the United Kingdom ($n = 4$) and Canada

($n = 4$). Educational backgrounds ranged from high school to graduate degrees.

Participants were recruited through an online platform (Prolific) and were required to be native or near-native English speakers who were raised speaking English from birth. Eligibility criteria included current residence in the United States, the United Kingdom, or Canada, age between 18 and 99, and a high prior approval rate on Prolific (95–100%). Participants were screened for language background, nationality, and country of birth to ensure alignment with the target population. Individuals reporting a history of neurological, cognitive, or language-related conditions were excluded. All participants completed the study in a single session and received monetary compensation at an hourly rate of £12. All participants provided informed consent electronically, on the platform, prior to participation and were informed that their anonymized responses would be used for research purposes. The same recruitment requirements, screening criteria, and consent procedures are applied in the main experiment. The study protocol for both norming and presupposition judgment task is approved by the institutional Research Ethics Board.

In this task, participants rated 120 short sentences on a 0–7 Likert scale (0 = very unlikely, 7 = very likely) indicating how likely each statement was to be true in the real world. Two attention checks were included. The sentences varied in expected likelihood. For example, a baseline item asked, *How likely is it that someone chosen at random from your hometown has a stamp collection?*, High-probability items added a strongly relevant modifier (e.g., "someone who is interested in postal history"), mid-probability items included a moderately relevant modifier (e.g., "a retired postman"), and low-probability items included an unrelated modifier (e.g., "someone who is vegetarian"). The expected completion time was approximately 35–40 minutes the task was designed to be finished in one sitting (median completion time: 26.8 minutes).

A.2 Main Presupposition Judgment Study

The main study consisted of two independent groups corresponding to the with-context and without-context conditions. Different participants were recruited for each condition to avoid carryover effects. In total, data from 128 participants were collected. After excluding eight participants who failed attention checks, data from 120 participants (60 per condition) were retained for analysis. The final sample consisted of 55 women, 64 men, and one non-binary participant, residing in the United States ($n = 73$), the United Kingdom ($n = 29$), and Canada ($n = 18$), aged between 25 and over 65. Educational backgrounds ranged from high school to graduate degrees.

Participants completed 90 items, each consisting of a conditional statement and a corresponding target statement, and rated the likelihood of the target statement on a 0–7 Likert scale. In the with-context condition, participants were additionally provided with brief identifying background information about the individual described in each item.

The median completion time was 23.5 minutes for the with-context condition and 19.9 minutes for the without-context condition.

Participants were compensated at an hourly rate of approximately £10. To reduce fatigue and cognitive load, two mandatory 8-second pauses were included in the with-context study.

Since human participants and LLMs received largely identical instructions, except that LLMs were additionally asked to provide step-by-step reasoning, we do not reproduce them here. The full instructions and prompts are reported in the following section of the Appendix.

B Example Stimuli

This section presents representative examples of stimuli used in the norming study and the main presupposition judgment experiment. All stimuli were presented in randomized order during the experiments, and probability and A-p relevance labels are shown here for explanatory purposes only.

B.1 Norming Study Items

The norming study was conducted to elicit high-, mid-, and low-probability propositions and to confirm, through human judgments, that these propositions exhibit the intended probability levels in real-world contexts. Table 4 presents representative examples from the norming study. Each item consists of a baseline condition and three conditional variants corresponding to high, mid, and low probability levels. The resulting ratings were used to construct the stimulus set for the main presupposition judgment experiment.

Condition	Item
Low-Probability Example: Stamp Collection	
Scenario	Someone from your hometown / having a stamp collection
Baseline	How likely is it that someone chosen at random from your hometown has a stamp collection?
High	How likely is it that someone chosen at random from your hometown who is interested in postal history has a stamp collection?
Mid	How likely is it that someone chosen at random from your hometown who is a retired postman has a stamp collection?
Low	How likely is it that someone chosen at random from your hometown who is vegetarian has a stamp collection?
Mid-Probability Example: Having Grandchildren	
Scenario	Someone from your neighborhood / having grandchildren
Baseline	How likely is it that someone chosen at random from your neighborhood has grandchildren?
High	How likely is it that someone chosen at random from your neighborhood who has adult children has grandchildren?
Mid	How likely is it that someone chosen at random from your neighborhood who is over 60 has grandchildren?
Low	How likely is it that someone chosen at random from your neighborhood who is left-handed has grandchildren?
High-Probability Example: Credit Cards	
Scenario	Someone at a coffee shop / having credit cards
Baseline	How likely is it that someone chosen at random at a coffee shop has credit cards?
High	How likely is it that someone chosen at random at a coffee shop who travels frequently has credit cards?
Mid	How likely is it that someone chosen at random at a coffee shop who works full-time has credit cards?
Low	How likely is it that someone chosen at random at a coffee shop who drinks tea has credit cards?

Table 4: Representative examples from the norming study showing baseline and high-, mid-, and low-relevance conditions for the propositions.

B.2 Main Presupposition Judgment Items

The main presupposition judgment task is designed to investigate how participants integrate antecedent-presupposition relevance and contextual information when interpreting conditional sentences. Each item consists of a background description, a conditional statement, and a target presupposition, and is presented under either with-context or without-context conditions. Table 5 presents representative examples of the experimental stimuli.

C Linear Mixed-Effects Model Results

Table 6 reports the full fixed-effects output for the linear mixed-effects models fitted to human presupposition judgments in the with-context and without-context conditions, respectively. Both models were specified as follows:

$$\text{rating} \sim \text{probability} \times \text{relevance} + (1 \mid \text{participant})$$

In this specification, probability (low, mid, high) and $A - p$ relevance (irrelevant, somewhat relevant, relevant) were treatment-coded factors, with high probability and relevant as the reference levels. Random intercepts were included for participants. Models were fit by restricted maximum likelihood (REML) using the `MixedLM` implementation in `statsmodels`. The without-context model included 5,400 observations from 60 participants (log-likelihood $= -10018.32$, residual scale $= 2.29$). The with-context model included 5,400 observations from 60 participants (log-likelihood $= -10240.25$, residual scale $= 2.48$).

D Model Sources and Computational Cost

The models evaluated in this paper are obtained from the following sources:

1. **GPT-5** is provided by OpenAI. The corresponding API documentation is available at https://platform.openai.com/docs/models.

2. **Gemini-2.5-flash** is provided by Google

Background	Statement 1 (Conditional)	Target
having an Instagram account (High-Probability Example)		
Alex works at the airport in your town.	**Rel:** If Alex uses social media, he will post a photo to his Instagram account.	Alex has an Instagram account.
Maya works at the airport in your town.	**S-Rel:** If Maya is over 50, she will check her Instagram account.	Maya has an Instagram account.
Daniel works at the airport in your town.	**Irrel:** If Daniel wears blue, he will log out of his Instagram account.	Daniel has an Instagram account.
having a brother (Mid-Probability Example)		
John works out at the downtown gym.	**Rel:** If John has siblings, he will call his brother on the weekend.	John has a brother.
Nadine works out at the downtown gym.	**S-Rel:** If Nadine has a large family, she will invite her brother.	Nadine has a brother.
Nick works out at the downtown gym.	**Irrel:** If Nick is left-handed, he will teach his brother mathematics.	Nick has a brother.
Having a Wetsuit (Low-Probability Example)		
Jack is from Ottawa.	**Rel:** If Jack is a scuba diver, he will bring his wetsuit.	Jack has a wetsuit.
Tim is from Ottawa.	**S-Rel:** If Tim likes swimming, he will pack his wetsuit.	Tim has a wetsuit.
Sara is from Ottawa.	**Irrel:** If Sara likes coffee, she will forget her wetsuit.	Sara has a wetsuit.

Table 5: Representative examples from the main presupposition judgment task. Each item consists of a background description, a conditional statement, and a target presupposition. Background information was provided only in the with-context condition and omitted in the without-context condition. Probability-level labels were used for analysis and were not shown to participants. The labels Rel, S-Rel, and Irrel indicate levels of antecedent–presupposition relevance (relevant, somewhat relevant, and irrelevant) and were not shown to the participants.

Gemini, with API documentation available at https://ai.google.dev/gemini-api/docs.

3. **Claude-haiku-4** is provided by Anthropic. The corresponding API documentation is available at https://platform.claude.com/docs/en/intro.

4. **Qwen2.5-7B-Instruct**[6] and **Llama3.1-8B-Instruct**[7] are open-source base model weights obtained from Hugging Face (https://huggingface.co/).

For large proprietary models (e.g., GPT-5 and Gemini-2.5-Flash), a single run on 360 samples costs approximately $14 CAD for explanation generation. For the judge model, Claude-Haiku-4, evaluating approximately 40,000 checklist items costs around $55 CAD. All open-source model evaluations are conducted on a system equipped with two NVIDIA RTX 4090 GPUs (32 GB memory each). Overall, the modest computational requirements demonstrate that the proposed evaluation protocol is accessible to researchers with limited computational resources, while still enabling a comprehensive assessment of state-of-the-art models.

[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

E Inference and Decoding Configuration

All four evaluated LLMs produce a single response per item; no self-consistency or re-sampling is used.

Open-source Models (Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct). Run locally with HuggingFace `transformers` in `bfloat16` via `AutoModelForCausalLM.generate` (`do_sample=True`, `temperature`= 0.7, `top_p`= 0.9, `max_new_tokens`= 1024; `top_k` and `repetition_penalty` left at library defaults of 50 and 1.0). Prompts use the tokenizer's chat template, are truncated to 4,096 input tokens, and generation stops on the model's default `eos_token_id`; no custom stop strings are set.

Closed-source Models (GPT-5, Gemini-2.5-Flash). Accessed through the official OpenAI and Google Generative APIs (`chat.completions.create` and

(A) With-context condition

Predictor	β	SE	z	p	95% CI
Intercept	5.377	0.159	33.82	$< .001$	$[5.065, 5.689]$
Probability: low	-0.977	0.099	-9.89	$< .001$	$[-1.171, -0.783]$
Probability: mid	-0.046	0.091	-0.51	.612	$[-0.225, 0.133]$
Relevance: irrelevant	-0.377	0.102	-3.71	$< .001$	$[-0.576, -0.178]$
Relevance: somewhat relevant	-0.258	0.102	-2.54	.011	$[-0.458, -0.059]$
Prob.: low × Rel.: irrelevant	-0.310	0.140	-2.22	.027	$[-0.584, -0.036]$
Prob.: mid × Rel.: irrelevant	-0.493	0.129	-3.82	$< .001$	$[-0.747, -0.240]$
Prob.: low × Rel.: somewhat relevant	0.399	0.140	2.86	.004	$[0.125, 0.673]$
Prob.: mid × Rel.: somewhat relevant	0.107	0.129	0.83	.407	$[-0.146, 0.360]$
Random intercept variance (Participant)	1.206	0.145		—	

(B) Without-context condition

Predictor	β	SE	z	p	95% CI
Intercept	5.340	0.141	37.99	$< .001$	$[5.064, 5.615]$
Probability: low	-0.347	0.095	-3.66	$< .001$	$[-0.533, -0.161]$
Probability: mid	0.090	0.088	1.02	.306	$[-0.082, 0.262]$
Relevance: irrelevant	-0.356	0.098	-3.65	$< .001$	$[-0.548, -0.165]$
Relevance: somewhat relevant	-0.065	0.098	-0.66	.509	$[-0.256, 0.127]$
Prob.: low × Rel.: irrelevant	-0.734	0.134	-5.47	$< .001$	$[-0.998, -0.471]$
Prob.: mid × Rel.: irrelevant	-0.572	0.124	-4.61	$< .001$	$[-0.815, -0.329]$
Prob.: low × Rel.: somewhat relevant	-0.224	0.134	-1.67	.095	$[-0.487, 0.039]$
Prob.: mid × Rel.: somewhat relevant	-0.192	0.124	-1.55	.122	$[-0.435, 0.052]$
Random intercept variance (Participant)	0.899	0.113		—	

Table 6: Linear mixed-effects model results for human presupposition judgments. (A) With-context condition. (B) Without-context condition. Reference levels: Probability = high, Relevance = relevant. Models were fit using REML; each includes 5,400 observations from 60 participants.

`GenerativeModel.generate_content`, respectively). `temperature` and `max_(output_)tokens` are set with the same parameter as open-source LMs, matching the condition a typical downstream user encounters.

Judge Model (Claude-Haiku-4-5). Called via the Anthropic APIs (`messages.create`) with `max_tokens`$= 100$ to accommodate the binary True/False; all other sampling parameters use Anthropic defaults. We release all raw generations and judge decisions alongside the code.

F Prompt Design

This section documents the prompt templates used for eliciting likelihood judgments and reasoning from LLMs, as well as the prompts used in the LLM-as-a-Judge evaluation. All task prompts were designed to closely parallel the instructions provided to human participants, with the additional requirement that models produce explicit step-by-step reasoning prior to reporting a final numerical rating.

Figure 4 presents the system prompts used in the without-context and with-context conditions of the main presupposition judgment task. In both conditions, models were instructed to interpret the speaker as honest and cooperative, to consider relevant background knowledge, and to provide a likelihood rating on a 0–7 Likert scale. The with-context prompt also included identifying background information, parallel to the corresponding human condition.

We use a separate set of prompts for the LLM-as-a-Judge framework. Figure 5 shows the evaluation prompts used in the without-context and with-context conditions. These prompts present the judge model with the original stimulus, the evaluated model's response, and a checklist question, and require a binary judgment indicating whether the response satisfies the specified criterion.

All prompt templates were held constant across models and experimental conditions, except for condition-specific contextual information.

G LLM-as-a-Judge Checklist

This section presents representative examples from the theory-informed checklist used in the LLM-as-a-Judge evaluation. The checklist consists of binary (yes/no) questions designed to assess whether model-generated reasoning aligns with established principles from formal semantics and pragmatics. The questions are organized into multiple dimensions and sub-dimensions targeting distinct aspects of conditional interpretation and presupposition handling. Table 7 provides a sample of the evaluation criteria.

Dimension	Sub-Dimension	Sample Question
Accuracy	Trigger Identification	Does the model correctly identify the presupposition trigger in the consequent?
	Anaphora Resolution	Does the model distinguish possessive triggers from non-trigger pronouns?
	Conditional Scope	Does the model correctly identify the scope of the conditional over the consequent?
Context	Contextual Influence	Does the model avoid using identifying context in its judgment?
	Presupposition–Context Distinction	Does the model distinguish between the presupposition and other contextual features?
Pragmatic	Speaker Cooperativity	Does the model consider the speaker's communicative intentions?
	Common Ground	Does the model use shared background knowledge in interpretation?
	Relevance	Does the model assess relevance between antecedent and presupposition?
Presupposition Handling	Projection	Does the model correctly project triggers under conditional embedding?
	Inference Validity	Does the model account for the fact that the inference may be invalid if the presupposition is false?
Coherence	Reasoning Alignment	Does the numerical rating align with the reasoning process?
	Evidence Use	Does the model justify its judgment using available evidence?
	Transparency	Is the reasoning process clearly articulated and interpretable?

Table 7: Representative sample of checklist questions indicating major dimensions and sub-dimensions used in the evaluation.

Without-Context Condition Prompt

```
In each item, you will be given a statement made by a speaker about a person.
Assume that the speaker is honest, reliable, and helpful, and that you share
relevant background assumptions with them.

After reading the statement, you will be asked to judge how likely a second
statement is to be true about that person.

Please respond using a scale from 0 (very unlikely) to 7 (very likely).

Please think step by step before giving your final rating. Consider:
1. What does the speaker's conditional statement imply about the subject?
2. What background assumptions or world knowledge are relevant?
3. How does the conditional relationship affect the likelihood of the target statement?

After your reasoning, provide your final answer in the format:
"Final Rating: [0--7]"

---

Speaker's statement: {statement_1}

Question: How likely is it that the following statement is true?

Statement: {statement_2}
```

With-Context Condition Prompt

```
In this task, imagine you are participating in a get-to-know-people activity.
Suppose that there are a bunch of cards with names on them, along with a small
identifying detail about the person (e.g., where the person is from). This
information is provided only to identify which person is being talked about.

Suppose that someone is tasked with selecting a card at random, and then --
depending on who the card identifies -- they have to say something they know
about that person. Assume that the speaker is honest, reliable, and helpful,
and that you share relevant background assumptions with them.

You will then be asked to judge how likely a second statement is to be true on
the assumption that what the speaker said is true. Please respond using a scale
from 0 (very unlikely) to 7 (very likely).

Please think step by step before giving your final rating. Consider:
1. What does the speaker's conditional statement imply about the subject?
2. What background assumptions or world knowledge are relevant?
3. How does the conditional relationship affect the likelihood of the target statement?

After your reasoning, provide your final answer in the format:
"Final Rating: [0--7]"

---

background: {background}

Speaker's statement: {statement_1}

Question: How likely is it that the following statement is true?

Statement: {statement_2}
```

Figure 4: System prompts used for likelihood judgment in the without-context and with-context conditions. The instruction to provide step-by-step reasoning was used only in the LLM study and was not included in the human evaluation.

LLM-as-a-Judge Prompt (Without-Context Condition)

```
You are an expert evaluator assessing how well a model reasons about the
likelihood that a presupposition (the target statement) in a conditional
of the form If A, then B, where B contains a presupposition trigger, is true,
using a 0--7 Likert scale.

Given the following information:
- Title: {title}
- Statement 1 (Conditional): {statement_1}
- Statement 2 (Target): {statement_2}
- Model Response: {response}

Evaluate the model's response against the following checklist question:
{checklist_question}

Based on your evaluation, respond with only "True" if the model's response
satisfies the checklist criterion, or "False" if it does not.

Your answer (True/False):
```

LLM-as-a-Judge Prompt (With-Context Condition)

```
You are an expert evaluator assessing how well a model reasons about the
likelihood that a presupposition (the target statement) in a conditional
of the form If A, then B, where B contains a presupposition trigger, is true,
using a 0--7 Likert scale.

Given the following information:
- Title: {title}
- Background: {background}
- Statement 1 (Conditional): {statement_1}
- Statement 2 (Target): {statement_2}
- Model Response: {response}

Evaluate the model's response against the following checklist question:
{checklist_question}

Based on your evaluation, respond with only "True" if the model's response
satisfies the checklist criterion, or "False" if it does not.
# DO NOT OUTPUT Judging response, only TRUE/FALSE.

Your answer (True/False):
```

Figure 5: System prompts used by the LLM-as-a-judge model for checklist-based evaluation in the without-context and with-context conditions.