

If Probable, Then Acceptable? Understanding Conditional Acceptability Judgments in Large Language Models

Jasmin Orth*¹ Philipp Mondorf*^{1, 2} Barbara Plank^{1, 2}

¹MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

²Munich Center for Machine Learning (MCML), Munich, Germany

jasmin.orth@campus.lmu.de, {p.mondorf, b.plank}@lmu.de

Abstract

Conditional acceptability refers to how plausible a conditional statement is perceived to be. It plays an important role in communication and reasoning, as it influences how individuals interpret implications, assess arguments, and make decisions based on hypothetical scenarios. When humans evaluate how acceptable a conditional “If A, then B” is, their judgments are influenced by two main factors: the *conditional probability* of B given A, and the *semantic relevance* of the antecedent A given the consequent B (i.e., whether A meaningfully supports B). While prior work has examined how large language models (LLMs) draw inferences about conditional statements, it remains unclear how these models judge the *acceptability* of such statements. To address this gap, we present a comprehensive study of LLMs’ conditional acceptability judgments across different model families, sizes, and prompting strategies. Using linear mixed-effects models and ANOVA tests, we find that models are sensitive to both conditional probability and semantic relevance—though to varying degrees depending on architecture and prompting style. A comparison with human data reveals that while LLMs incorporate probabilistic and semantic cues, they do so less consistently than humans. Notably, larger models do not necessarily align more closely with human judgments.

1 Introduction

Conditional statements of the form “If A, then B” are fundamental to how humans make predictions, draw conclusions, and offer explanations (Byrne, 2007; Evans and Over, 2004; Douven, 2015). As such, they play a central role in human communication, reasoning, and decision-making. Understanding the conditions under which humans judge conditional statements as *acceptable* or *plausible* sheds light on the principles that underlie everyday

*Equal contribution.

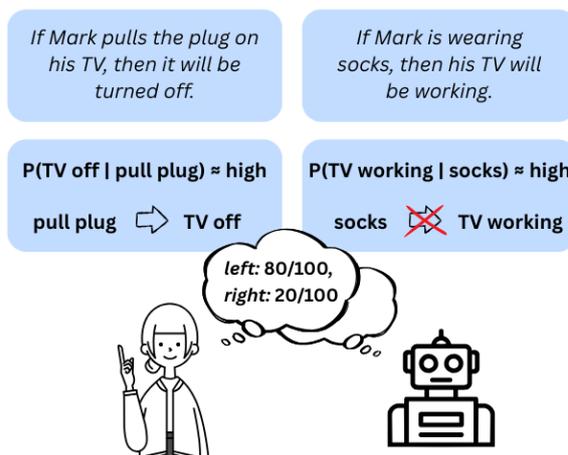


Figure 1: Illustration of two conditionals with equally high conditional probabilities but differing evidential relevance. While both are probable, only the left one encodes a plausible causal or evidential link between antecedent and consequent, appearing more acceptable.

language use, from casual conversation to scientific explanations. Consider the following example, drawn from Skovgaard-Olsen et al. (2016):

- (i) *If Mark pulls the plug on his TV, then it will be turned off.*
- (ii) *If Mark is wearing socks, then his TV will be working.*

While both sentences are grammatically sound, the second appears less acceptable: it is unlikely that Mark wearing socks has any effect on whether his TV is working. This notion of *acceptability*, or how appropriate a conditional statement seems in a given context, is often judged intuitively by humans (Douven and Verbrugge, 2012; Skovgaard-Olsen et al., 2016). Two main hypotheses aim to explain such intuitions. The *conditional probability hypothesis* holds that acceptability depends primarily on the *conditional probability* of B given A (Evans et al., 2003; Over et al., 2007). By this account, both conditionals above might be judged acceptable, as the probability of the TV being turned

off given that the plug has been pulled, and the probability of the TV working given that Mark is wearing socks, are both high. In contrast, the *evidentiality hypothesis* emphasizes an additional influence of *relevance* or *evidential connection* encoded within the statement (Douven and Verbrugge, 2012; Skovgaard-Olsen et al., 2016). From this perspective, (i) is acceptable because pulling the plug is causally connected to the TV turning off, whereas (ii) is not, because wearing socks lacks such a connection. LLMs have demonstrated notable capabilities across diverse linguistic and cognitive tasks, including general reasoning (Hao et al., 2024; Mondorf and Plank, 2024a), linguistic inference (Clark et al., 2020), and causal reasoning (Dettki et al., 2025; Yu et al., 2024). However, despite increasing work on LLM reasoning, research on conditional *acceptability* judgments in LLMs remains virtually nonexistent. Most existing studies on LLMs and conditional statements focus on their causal, abductive, or inferential capabilities (Dettki et al., 2025; Holliday et al., 2024; Liu et al., 2023; Mondorf and Plank, 2024b), rather than on whether conditionals are judged contextually *acceptable*.

We address this gap by conducting a targeted investigation of conditional acceptability judgments in LLMs. Building on prior work from cognitive psychology (Skovgaard-Olsen et al., 2016), we assess the acceptability judgments of two model families: Llama 3.1 (8B and 70B) and Qwen 2.5 (7B and 72B), and examine how their judgments vary as a function of both the conditional probability and the evidential connection encoded within a statement. Our findings reveal that:

- Acceptability judgments of models are influenced by both conditional probability and semantic relation, though the strength and consistency of this effect vary substantially by model family, size, and prompting technique.
- Comparison with human data suggests that while the overall trends in LLM judgments broadly resemble those of humans, LLMs show greater variability across prompts and model sizes and a less *systematic* integration of probabilistic and semantic cues.
- Larger model size does not necessarily lead to more alignment with human trends, and while few-shot prompting can enhance LLMs’ sensitivity to semantic relevance, it may introduce bias toward specific semantic relations.

2 Background

2.1 Conditional Acceptability

The key concept of this study is *conditional acceptability*: the degree to which a conditional statement is perceived to be appropriate, plausible, or natural in context (Skovgaard-Olsen et al., 2016). While often associated with *conditional probability* $P(B | A)$, such as proposed by the *conditional probability hypothesis* (Evans et al., 2003; Oaksford and Chater, 2003; Over et al., 2007), the two can diverge significantly in practice. Take, for instance, the statement from the introduction: *If Mark is wearing socks (A), then his TV will be working (B)*. This may be judged unacceptable or odd, even when both $P(A)$ and $P(B)$ are high, and even when $P(B | A)$ is high. The issue lies not in probability, but in the perceived irrelevance of A to B . The statement violates intuitive notions of coherence or causality: wearing socks does not cause the TV to function, nor would not wearing them affect the outcome. As such, the distinction reflects the broader gap between *correlation* and *causation* (Barrowman, 2014). This disconnect has prompted extensive psychological research on how humans evaluate conditional statements (e.g., Douven, 2015; Evans and Over, 2004; Hardman and Macchi, 2003), suggesting that individuals often rely on a combination of semantic, contextual, and probabilistic cues when assessing conditionals (Berto and Özgün, 2021; Douven and Verbrugge, 2012; Over et al., 2007). Specifically, the *evidentiality hypothesis* argues that the semantic relation between antecedent A and consequent B —whether A supports, contradicts, or is irrelevant to B —plays a crucial role in human acceptability judgments (Douven and Verbrugge, 2012; Skovgaard-Olsen et al., 2016; Berto and Özgün, 2021). Returning to the earlier example, *If Mark is wearing socks, then his TV will be working* may be judged unacceptable not because $P(B | A)$ is low, but because there is no plausible explanatory or causal (*supporting*) link between the two clauses. In particular, Skovgaard-Olsen et al. (2016) find that humans generally judge conditionals with a *supporting* relation more acceptable than those with a *contradicting* or *irrelevant* relation.

Although studies have investigated how *humans* judge the acceptability of conditional statements, it remains unclear how and under what circumstances LLMs judge conditionals as acceptable.

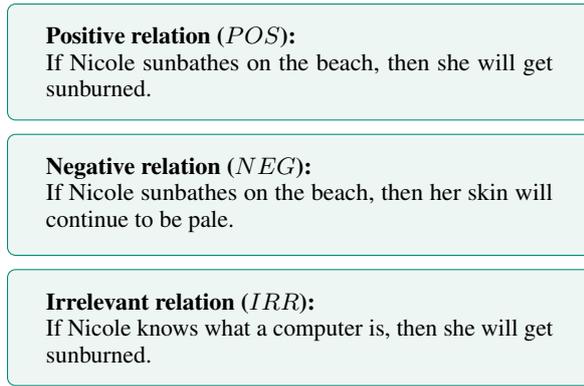


Figure 2: Example of three conditional statements representing different relations.

2.2 Dataset

To investigate how LLMs assess the acceptability of conditional statements, we draw on a dataset from Skovgaard-Olsen et al. (2016). Their study aims to understand how humans judge conditional acceptability using 144 conditional statements embedded in 12 short, realistic contexts.¹ Each context is tied to a unique everyday scenario, such as going on a date or attending a meeting.

To enable systematic analyses, the 12 conditionals that share the same context are organized along two dimensions. First, each conditional can be classified into one of three *relation types*, indicating the semantic connection between antecedent A and consequent B : a **positive** relation (*POS*) means A supports B , a **negative** relation (*NEG*) denotes that A undermines or contradicts B , and an **irrelevant** relation (*IRR*) means A is unrelated to B (see Figure 2). Second, the conditionals vary in their *prior probability*: each scenario includes four levels based on combinations of high or low probability for both antecedent $P(A)$ and consequent $P(B)$ (e.g., high-high, high-low, low-high, low-low; see Appendix A.1 for details). This yields 12 unique configurations per context (3 relation types \times 4 prior probabilities, as shown in Table 5 in the Appendix), resulting in 144 conditionals in total. The following example illustrates how relation types can be defined within a given context:

Nicole’s skin is pale and she wants a suntan with a mild brown color. It’s summer, the weather is nice, and there is a beach nearby. However, Nicole’s skin is sensitive to the sun, and she is bad at controlling the amount of her exposure to the sun.

¹More information about the dataset and demographics can be found in Appendix A.1.

As shown in Figure 2, a **positive** relation is illustrated by the conditional statement: *If Nicole sunbathes on the beach, then she will get sunburned*. Here, under the contextual constraint of her having sensitive skin, *Nicole sunbathing* actively causes her to *get sunburned*. In the **negative** example, the antecedent *Nicole sunbathing on the beach* contradicts her *staying pale*, while the **irrelevant** example pairs two unrelated ideas: *knowing what a computer is* has no causal connection to *getting sunburned*. These structured scenarios are designed to help participants apply real-world knowledge while simultaneously constraining interpretation, anchoring relations and probability judgments to a shared context. In the present study, we use this dataset to evaluate how LLMs judge the acceptability of conditionals and whether they exhibit similar patterns as humans. By prompting models with these same context-rich scenarios and conditionals, we can systematically assess their responses across relation types and probability levels.

3 Method

3.1 Experimental Setup

Tasks We follow the approach of Skovgaard-Olsen et al. (2016) and present the 144 conditional statements described in Section 2.2 to LLMs in three distinct judgment tasks: (i) estimating the conditional probability $P(B | A)$ of the consequent given the antecedent, (ii.a) predicting the overall probability of the conditional $P(\text{“If } A, \text{ then } B\text{”})$, and (ii.b) judging the *acceptability* of the conditional statement *“If A , then B ”*. While tasks (ii.a) and (ii.b) present the LLMs with full conditional statements (e.g., *“If Mark is wearing socks, then his TV will be working”*), the conditional probability task (i) presents the statement in a split format, consisting of a supposition such as *“Suppose Mark is wearing socks”* and an outcome, such as *“His TV will be working”*, which needs to be judged (see Figure 3 lower bounding box).

To analyze whether LLMs’ acceptability judgments align with a probabilistic or relevance-based interpretation, these tasks are designed to separate probabilistic reasoning from pragmatic or communicative considerations. The conditional probability task described above provides a reference point reflecting pure probabilistic reasoning: if large language models rely solely on conditional probability when judging conditional statements,

<p>Statement: If Mark is wearing socks, then his TV will be working.</p>
<p>Supposition: Suppose Mark is wearing socks.</p> <p>Outcome: His TV will be working.</p>

Figure 3: Example of a conditional presented in full form (top) and its corresponding split version (bottom), used to elicit different types of model judgments.

their *if-probability*² and acceptability ratings would be expected to closely match their corresponding $P(B | A)$ values. The split between acceptability and if-probability judgments further allows to determine whether assessments of conditional statements differ depending on the type of judgment. All tasks instruct the model to respond with a single number between 0 – 100, indicating the degree of probability or acceptability. Full prompt templates are provided in Figures 8 to 10 in Appendix A.3.

Post-Processing To robustly extract LLM judgments, we use a regular expression to match numbers between 0 and 100 following either the token “answer” or “assistant”. If no valid number in the expected range is found, the response is recorded as missing (*NaN*). Notably, across all LLMs and prompting techniques, we report no missing responses. For subsequent analyses, all numerical ratings are mean-centered, resulting in a scale from -0.5 to 0.5 ; all reported values reflect this transformation. p -values (statistical significance: $p < 0.05$) are reported to three decimal places, except when highly significant, in which case thresholds are used ($p < 0.001$, $p < 0.0001$).

3.2 Prompts & Models

We aim to examine the impact of model family, model size, and prompting technique on the conditional acceptability judgments of LLMs. Therefore, we assess a total of four language models: Llama-3.1-8B-Instruct,³ Llama-3.1-70B-Instruct,³ Qwen2.5-7B-Instruct,⁴ and Qwen 2.5-72B-Instruct,⁴ using publicly accessible weights obtained from the Hugging Face platform.

²To make the distinction between the conditional probability $P(B | A)$ and the probability of the full conditional statement as clear as possible, we denote the probability of the full conditional statement the *if-probability* in this study.

³<https://huggingface.co/meta-llama/models>

⁴<https://huggingface.co/Qwen/collections>

	Cond. Prob.	If-Prob.	Acceptability
Vanilla			
Llama 8B	0.7685	0.7724	0.7529
Llama 70B	0.9468	0.9344	0.9201
Qwen 7B	0.9768	0.9716	0.9450
Qwen 72B	0.9693	0.9673	0.9784
Few-shot			
Llama 8B	0.7930	0.7538	0.7596
Llama 70B	0.9438	0.9556	0.8820
Qwen 7B	0.9542	0.9556	0.9596
Qwen 72B	0.9559	0.9437	0.9768
CoT			
Llama 70B	0.9014	0.8497	0.8757
Qwen 72B	0.9331	0.9127	0.9661

Table 1: Intraclass correlation scores across five model runs for each judgment type, model, and prompt type.

All models are prompted using a vanilla (zero-shot) and three-shot format, while the two larger models (Llama 70B and Qwen 72B) are also tested with chain-of-thought (CoT) prompting (Wei et al., 2022). For further details on models and prompts, please refer to Appendix A.2. Our code is publicly available at: <https://github.com/mainlp/conditional-acceptability>.

Consistency of Model Responses We generate model responses via nucleus sampling (top-p = 1.0) and a temperature of $T = 1.0$. As this decoding procedure is non-deterministic, we ensure the consistency of the models’ numerical ratings through repeated trials and a post-hoc correlation analysis. Each model is prompted five times per data sample to account for potential instabilities in ratings. By computing the intraclass correlation of these five ratings per sample, we further evaluate the reliability of the models’ judgments. As shown in Table 1, all models demonstrate good-to-excellent consistency across all judgment types and prompting strategies.

3.3 Statistical Methods

We follow the core analytic approach of Skovgaard-Olsen et al. (2016), applying *linear mixed-effects models* to investigate how conditional probability and relation type affect acceptability judgments of conditional statements.

Linear Mixed-Effects Models Mixed-effects models allow for the analysis of how conditional probability and relation type influence conditional acceptability judgments while accounting for addi-

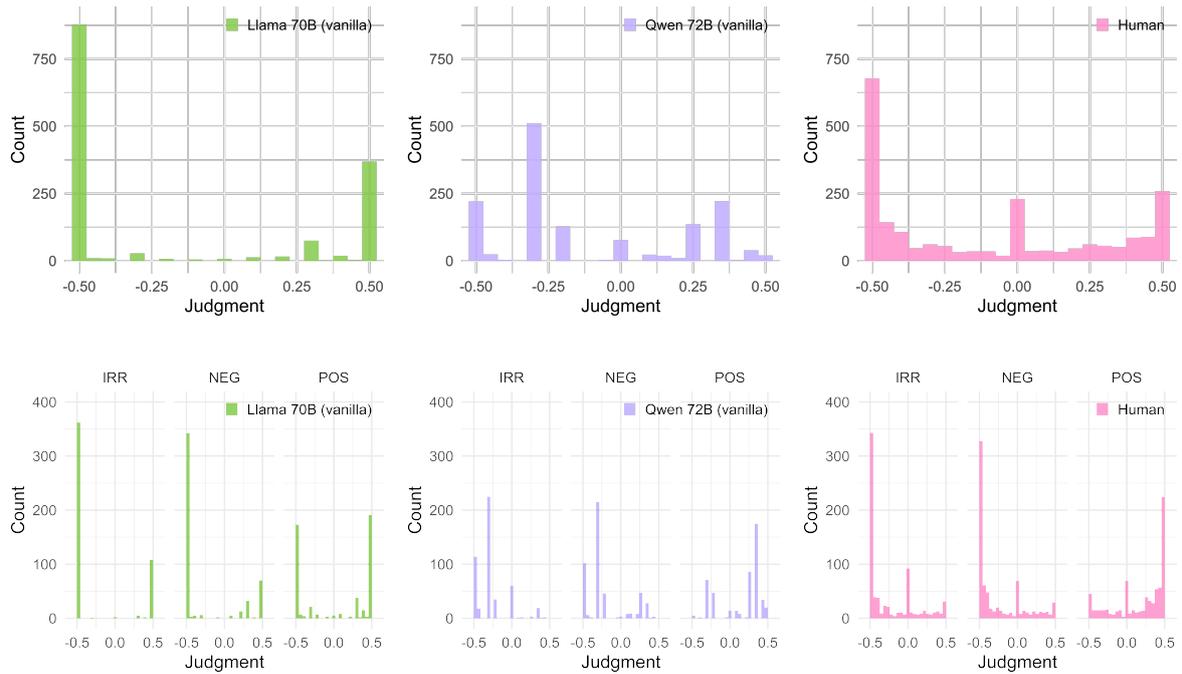


Figure 4: Distribution of center-scaled judgments across humans and LLMs. Top row: overall distributions. Bottom row: distributions divided by relation type. Left: Llama 70B (vanilla), center: Qwen 72B (vanilla), right: human data obtained from the study by Skovgaard-Olsen et al. (2016).

tional sources of variation. In particular, different *scenarios, participants*,⁵ and *LLM sampling iterations* introduce underlying variation that might otherwise obscure meaningful effects. Linear mixed-effects models account for this structure by incorporating both fixed effects (the predictors of interest) and random effects (uncontrolled variation across groups) (Brown, 2021). For further information on model fitting, please refer to Appendix A.5.

Additional Measures We report Type III ANOVA, fixed effect estimates, slopes, and estimated marginal means (EMMs) to assess the impact of predictors, such as conditional probability and relation types, on judgments of conditional statements. ANOVA tests show whether mean differences are statistically significant across conditions (e.g., relation types) (Brown, 2021; Shaw and Mitchell-Olds, 1993; Tabachnick and Fidell, 2007), while EMMs represent the model-adjusted average outcomes for each condition, allowing for comparisons between groups while controlling for other variables (Searle et al., 1980; CRAN, 2025a,b).

⁵While we do not collect human data ourselves, we use human data from Skovgaard-Olsen et al. (2016) to compare LLM trends.

4 Results

Figure 4 shows the distribution of vanilla LLM judgments (left: Llama 70B, center: Qwen 72B). Responses cluster around specific values, leading to a scattered overall distribution. This pattern is even more pronounced when split by relation type (bottom row); for example, Llama 70B gives almost exclusively low ratings for negative and irrelevant conditions. This supports prior findings that models prefer specific tokens or rating values in their judgments (Stureborg et al., 2024; Zheng et al., 2023).

Impact of Relation Type We find that all LLMs exhibit sensitivity to relevance when judging conditional statements. As shown in Figure 5, mean judgments differ systematically across relation types: conditionals with a positive (supporting) relation receive more favorable ratings than those with a negative (contradicting) or irrelevant (unrelated) relation. The highest positive mean is observed for Qwen 72B (vanilla) at 0.16, while Llama 8B (vanilla) shows the lowest, barely exceeding the midpoint, indicating that this model rates positively related statements less favorably than the others. Notably, Llama 70B presents the greatest standard deviation of all vanilla models (see Figure 5), which might be attributable to its spiked data

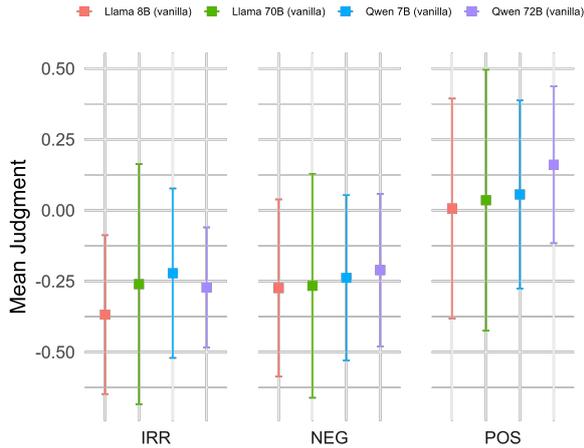


Figure 5: Mean and standard deviation across relevance conditions and vanilla models (red: Llama 8B, green: Llama 70B, blue: Qwen 7B, purple: Qwen 72B).

distribution (mainly concentrating on either very high or very low values, see Figure 4). Exact values for all models and prompting techniques can be found in Table 8 in the Appendix. ANOVA results confirm a highly significant main effect of relation type for all models except Llama 70B (vanilla) ($p = 0.060$; see Table 10 in Appendix A.6). Fixed effects estimates (Table 2) further highlight these differences: for instance, Qwen 72B (vanilla) rates negatively related conditionals 0.20 higher, and positively related ones 0.30 higher, than irrelevant ones. In contrast, Llama 70B (vanilla) shows minimal differentiation between relation types (0.07 for NEG, 0.01 for POS), suggesting lower sensitivity to relevance distinctions. In summary, while most models exhibit significant sensitivity to semantic relation type, Llama 70B (vanilla) appears to do so less than other models.

Impact of Conditional Probability For all LLMs, judgments of conditional statements are strongly influenced by the underlying conditional probability $P(B | A)$. ANOVA results (Tables 10 to 18 in the Appendix) confirm a highly significant main effect of conditional probability across all models ($p < 0.0001$). As shown in Figure 6, judgments increase with rising conditional probability, a pattern further supported by fixed effects estimates in Table 2, which quantify the effect of $P(B | A)$ on conditional statement judgments across models. Llama 70B (vanilla) shows the strongest sensitivity, with a slope of 0.65 per unit increase in $P(B | A)$, suggesting that, with the absence of relevance influence, it is particularly sensitive to conditional

	$P(B A)$	NEG	POS	Metric (P)
Vanilla				
Llama 8B	0.35	0.10	0.30	-0.06
Llama 70B	0.65	0.07	0.01	0.05
Qwen 7B	0.38	0.06	0.26	0.05
Qwen 72B	0.21	0.20	0.30	-0.03
Few-shot				
Llama 8B	0.29	0.23	0.33	0.12
Llama 70B	0.51	0.26	0.19	0.08
Qwen 7B	0.49	0.13	0.23	0.05
Qwen 72B	0.30	0.23	0.28	0.09
CoT				
Llama 70B	0.51	0.30	0.31	0.10
Qwen 72B	0.19	0.27	0.35	0.07

Table 2: Fixed effect estimates for LLMs predicting probability and acceptability judgments. Effects are shown as contrasts relative to the reference levels: relation = irrelevant, metric = acceptability.

probability. Out of the vanilla models, Qwen 72B shows the weakest effect (0.21).

How Conditional Probability and Relation Types Interact For all models except Llama 70B (vanilla), relation type significantly modulates the effect of conditional probability, as indicated by highly significant interaction effects in the ANOVA results ($p < 0.001$ for the vanilla models; see Tables 9 to 12 in the Appendix). Table 3 presents pairwise comparisons of interaction slopes, which indicate how much more strongly judgments increase with conditional probability for one relation type compared to another. These comparisons reveal dis-

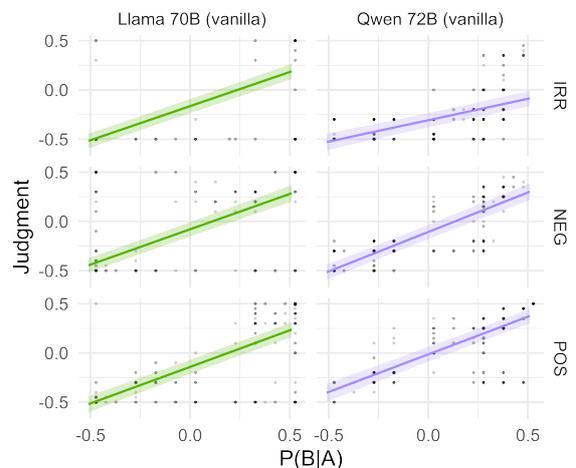


Figure 6: Scatterplots of data points and trend lines for Llama 70B (vanilla) and Qwen 72B (vanilla), divided by relation types. Plots for all vanilla models can be found in Figure 7 in the Appendix.

	IRR × NEG		IRR × POS		NEG × POS	
	Estimate	p	Estimate	p	Estimate	p
Llama 8B (vanilla)	-0.10	0.064	-0.28	< 0.0001	-0.18	0.001
Llama 70B (vanilla)	-0.02	1.000	-0.06	0.604	-0.03	1.000
Qwen 7B (vanilla)	-0.13	0.002	-0.22	< 0.0001	-0.09	0.029
Qwen 72B (vanilla)	-0.38	< 0.0001	-0.35	< 0.0001	0.03	0.420
Llama 8B (few-shot)	-0.21	< 0.001	-0.18	0.001	0.03	0.573
Llama 70B (few-shot)	-0.37	< 0.0001	-0.17	< 0.001	0.20	< 0.001
Qwen 7B (few-shot)	-0.25	< 0.0001	-0.12	0.017	0.13	0.017
Qwen 72B (few-shot)	-0.35	< 0.0001	-0.28	< 0.0001	0.07	0.077
Llama 70B (CoT)	-0.31	< 0.0001	-0.09	0.034	0.22	< 0.0001
Qwen 72B (CoT)	-0.51	< 0.0001	-0.44	< 0.0001	0.08	0.061

Table 3: Pairwise comparisons of interaction slopes for LLMs. SE: 0.04 – 0.05. Statistical significance at $p < 0.05$.

tinct patterns across models. Llama 70B (vanilla) shows no significant slope differences (IRR × POS: $p = 0.604$; others: $p = 1.000$), reinforcing its insensitivity to relevance distinctions. Among the remaining models, two distinct trends emerge. In the smaller models (Llama 8B and Qwen 7B, both vanilla), positive relations yield the steepest slopes (e.g., Llama 8B slope differences: IRR × POS: -0.28 ; NEG × POS: -0.18 ; $p \leq 0.001$ respectively), suggesting a stronger probability effect for supportive conditionals. By contrast, larger models and all few-shot or CoT variants exhibit steeper slopes for negatively related conditionals than for positive ones (e.g., Llama 70B (few-shot): 0.20, $p < 0.001$) or, as in Qwen 72B (vanilla), show no significant difference between the two (0.03, $p = 0.420$). Not only does relevance modulate the effect of conditional probability, but the reverse also seems true: the influence of relation type depends on the underlying probability of the conditional statement. Pairwise comparisons of estimated marginal means at low, medium, and high probability levels (see Table 20 in the Appendix) reveal that in both Qwen models (7B and 72B, vanilla), all relation types differ significantly; except at low probability, where the irrelevant and negative conditions do not differ (e.g., Qwen 72B: $p = 0.232$). A similar pattern appears in Llama 8B (vanilla), which shows no significant difference between negative and positive relations at low probability ($p = 0.148$). These results suggest that while these LLMs reliably distinguish relation types at medium and high conditional probabilities, their sensitivity decreases at low probabilities; perhaps because relevance becomes harder to evaluate when the conditional itself seems implausible.

Impact of Judgment Type Up to this point, we have analyzed the if-probability and acceptability judgments of LLMs together. However, ANOVA results (see Tables 10 to 18 in the Appendix) suggest that some LLMs may differentiate between these two metrics. Only Llama 70B (vanilla) and Qwen 72B (vanilla) show no significant main effect of judgment type, though both approach the threshold of significance (Llama $p = 0.058$ Table 10; Qwen $p = 0.068$ Table 12). Fixed effects estimates (using acceptability as the reference level, see Table 2) reveal no consistent trend among the vanilla models: Qwen 72B and Llama 8B give slightly lower ratings to if-probability (e.g., Qwen: -0.03), while Qwen 7B and Llama 70B show the opposite pattern (both: 0.05). In contrast, all few-shot and CoT variants assign higher ratings to if-probability than to acceptability (see Table 2). This effect is most pronounced in Llama 8B (few-shot: 0.12) and smallest in Qwen 7B (few-shot: 0.05). These results suggest that prompting examples consistently shift model ratings upward for probability judgments relative to acceptability. However, for zero-shot, this distinction appears less stable and may depend on model size or architecture. Larger models like Llama 70B and Qwen 72B appear less influenced by judgment type overall, whereas prompting reintroduces or amplifies this sensitivity.

Impact of Prompting Strategy All LLMs appear highly susceptible to few-shot and CoT prompting, though the direction of these effects varies across models. While few-shot prompting aids Llama 70B in incorporating relation types in its judgments (ANOVA: $p = 0.060$ to < 0.0001 , see Tables 10 and 14 in the Appendix), it seems to introduce a negativity bias in other models: As can be seen in Table 3, all few-shot variants pro-

duce steeper interaction slopes for negatively related conditionals than for positively related ones (e.g., Llama 70B (few-shot): 0.20, $p < 0.001$), or, as in Llama 8B (few-shot), eliminate the distinction entirely: $p = 0.573$. Prompting effects are also evident in the pairwise comparisons of estimated marginal means across conditional probability levels (see Table 20 in the Appendix). In the vanilla Qwen models, relation type distinctions collapse only at low probability between irrelevant and negative conditionals. However, the few-shot and CoT variants additionally produce insignificant differences between negative and positive conditionals at *high* conditional probability (e.g., Qwen 72B CoT $p = 1.000$), suggesting that the prompting examples may, in some cases, reduce sensitivity to fine-grained relevance contrasts. Contrary to expectations, prompting examples do not appear to increase alignment with human behavior, which typically shows great sensitivity to relevance cues. Notably, the Llama models do not exhibit this degradation under prompting, except for Llama 70B CoT, which converges negative and positive at *medium* conditional probability ($p = 0.459$).

Impact of Model Size Intriguingly, smaller models show a more consistent influence of relation type than their larger counterparts. For instance, as shown in Table 3, Llama 8B (vanilla) exhibits significant slope differences between relation types and strong fixed effects estimates, unlike Llama 70B (vanilla), which shows virtually no influence of relevance. Similarly, Qwen 7B (vanilla) does not display the reduced contrast between positive and negative slopes observed in Qwen 72B (vanilla). This pattern suggests that smaller models may more consistently integrate relevance cues, especially for supporting conditionals. Conversely, larger models tend to show reduced variation across the type of judgment (if-probability vs. acceptability), indicating a different kind of internal consistency.

4.1 Comparison Humans vs. LLMs

Data Distribution Overall, the rating patterns of humans and LLMs show broad similarities, yet systematic differences emerge. As shown in Figure 4, the general distribution of human ratings appears continuous, whereas LLMs tend to concentrate responses around specific values. Broken down by relation type (bottom row), both humans and LLMs generally rate positively related conditionals more favorably, and irrelevant and negative condition-

als less so. However, human ratings show tighter clustering: irrelevant and negative statements consistently fall at the lower end of the scale, while positive ones cluster at the higher end. In contrast, LLMs exhibit greater dispersion, with outlier judgments appearing across the scale. For example, Llama 70B (vanilla) assigns both very high and very low ratings to positively related conditionals.

Systematic Divergences Despite these differences, the overall mean ratings by relation type are similar. Human means for the irrelevant and negative conditions (IRR: -0.25 , NEG: -0.26) closely match the ones of all LLMs, which range from -0.37 to -0.20 for IRR and -0.27 to -0.19 for NEG (see Table 8 in the Appendix). The biggest difference occurs in the positive condition (humans: 0.19), where most LLMs fall short. Both Llama models and the smaller Qwen variant barely exceed the midpoint (Llama 8B: 0.01, 70B: 0.04; Qwen: 0.06). Qwen 72B is the only LLM that approximates this high mean across prompting techniques (vanilla: 0.16, few-shot: 0.15, CoT: 0.14). In terms of the influence of predictors, both humans and LLMs show strong effects of conditional probability and relation type. For humans, ANOVA results (see Table 19 in the Appendix) indicate significant main effects of probability and relevance, as well as their interaction ($p < 0.0001$), but no effect of judgment metric. This stands in contrast to LLMs, where the smaller models and all few-shot and CoT models consistently show modest but significant effects of the judgment type. In terms of slope comparisons, human pairwise interactions across relation types are highly significant ($p < 0.0001$), indicating robust modulation by relevance. This pattern most closely resembles Qwen 7B (vanilla), Llama 70B (few-shot), and Qwen 7B (few-shot), although the LLMs generally produce weaker effects and differ in direction: humans show steeper slopes for positively related conditionals, whereas LLMs sometimes favor negative ones. Finally, estimated marginal means comparisons show that humans distinguish between relation types at all levels of conditional probability; except between irrelevant and negative conditionals at low probability ($p < 0.0001$ for all other contrasts, see Table 20). This pattern is mirrored in the Qwen models and partially in the Llama models, which also exhibit collapsed distinctions at low probability levels.

In sum, while all LLMs except Llama 70B (vanilla) incorporate both conditional probability

	Sentence Prob.		Perplexity	
	r	ρ	r	ρ
Cond. Prob.				
Llama 8B	0.15	0.25	-0.28	-0.32
Llama 70B	0.16	0.26	-0.28	-0.38
Qwen 7B	0.08	0.22	-0.17	-0.27
Qwen 72B	0.16	0.33	-0.21	-0.36
If-Prob.				
Llama 8B	0.19	0.25	-0.32	-0.32
Llama 70B	0.11	0.31	-0.23	-0.31
Qwen 7B	0.07	0.20	-0.08	-0.27
Qwen 72B	0.14	0.19	-0.19	-0.31

Table 4: Pearson (r) and Spearman correlation (ρ) of sentence probability and perplexity with probability judgments obtained through direct elicitation.

$P(B | A)$ and semantic relevance in their judgments of conditionals, they do not exhibit the systematic integration of these factors observed in humans. In particular, the consistently elevated ratings and steepest interaction slopes for positively related conditionals appear to reflect a uniquely human sensitivity that LLMs can only approximate.

4.2 Eliciting Probability Judgments

In our study, we directly prompt models to judge the probability or acceptability of a conditional statement. While this method is well-established and has been used in existing studies (Tian et al., 2023; Yang et al., 2025), there also exist other methods for eliciting ratings from LLMs.⁶ For instance, we can compute a model’s sentence probability or its perplexity as an estimate of the probability it assigns to a given statement.

To compare sentence probabilities with our direct elicitation approach, we compute the model-assigned probabilities of both the conditional probability $P(B | A)$ and the if-probability $P(\text{“If } A, \text{ then } B\text{”})$. In particular, for conditional probability, we use the following prompt M :

Context: <context>
Given: <A>
Consequence:

and compute $P(B | M)$ using the probabilities the model assigns to each token in B . Similarly, for if-probability, we compute $P(\text{If } A, \text{ then } B | M)$

⁶A discussion of direct elicitation and alternative methods is provided in Appendix A.4.

using the prompt M :

Context: <context>
Statement:

Because sentence probability strongly depends on sentence length, we additionally compute the model’s perplexity over the statements. To study how well these probability-based judgments correlate with those obtained via direct elicitation, we compute both Pearson and Spearman correlations for each metric. As shown in Table 4, we observe a weak positive correlation for sentence probability and a weak-to-moderate negative correlation for perplexity scores, underlining that these metrics are related but ultimately capture different aspects.

5 Related Work

Conditional Reasoning in LLMs Several recent studies examine LLMs’ abilities in conditional and causal reasoning, though without addressing the acceptability of statements themselves. Liu et al. (2023) find that Code-LLMs outperform text-only models on abductive and counterfactual tasks when prompted in code, suggesting improved causal sensitivity. Dettki et al. (2025) show that LLMs draw on domain knowledge in causal inference rather than relying purely on abstract rules. Other work focuses on logical inference: Holliday et al. (2024) report frequent errors in detecting invalid inferences, though CoT prompting improves performance. Debray (2023) find that models like OPT show limited validity assessment but some semantic sensitivity when compared to human judgments.

6 Conclusion

In this paper, we investigate how LLMs assess conditional statements, a task that has long been debated in the context of human reasoning, but remains underexplored for LLMs. Specifically, we examine whether LLMs rely on conditional probability, relation type, or a combination of both when judging the acceptability of conditionals. Our findings suggest that both factors contribute to LLM judgments, although the degree of this influence varies substantially across models and prompting strategies. Notably, while LLMs and humans perform generally similar, key distinctions emerge, suggesting that LLMs may not *consistently* integrate both semantic relation and probabilistic strength in the systematic way that humans do.

7 Limitations

While our work sheds light on LLM capabilities in assessing conditional statements, several limitations warrant consideration and future investigation.

Human Data As discussed in Appendix A.1, the human judgment data from Skovgaard-Olsen et al. (2016) is collected via crowdsourcing. Although quality controls are implemented (such as restricting participation to native English speakers and setting time thresholds), these precautions are introduced only after data collection had already taken place. Consequently, the dataset features uneven response counts per conditional statement (e.g., some statements rated by ten participants, others by only three), introducing variability and inconsistency. Future work could benefit from eliciting human judgments in more controlled environments, such as laboratory or classroom settings, ideally synchronized with LLM data collection to enable direct comparisons under matched conditions.

Prompts and Prompting As noted in Section 4, the few-shot and CoT prompting conditions appear to bias model sensitivity towards negatively related conditionals. Since this effect occurred in all few-shot and CoT variants, the random selection of prompt examples (see Appendix A.3) may have inadvertently biased model behavior. Given the well-documented prompt sensitivity of LLMs (Errica et al., 2024; Loya et al., 2023; Sclar et al., 2023), future work should take greater care in constructing prompt sets. One possible improvement could be to include a more balanced set of few-shot examples, e.g., one for each combination of relation type and conditional probability level (resulting in nine examples total). This might help ensure that models are not inadvertently primed to favor particular judgment patterns.

Limits of Relying Solely on Ratings A key limitation of this study is the reliance on numerical ratings alone. While this approach enables systematic comparisons, it may not fully capture the reasoning processes that underlie those judgments. Conditional evaluation is a complex task for both humans and LLMs. That some aspects elicit greater concordance across models and participants than others suggests that numerical ratings may obscure important distinctions in underlying reasoning. Future work may benefit from incorporating richer diagnostic measures. A more informative approach

could involve qualitative analysis of both LLM and human reasoning. Such methods would offer deeper insight into the reasoning strategies underlying judgment patterns and help determine whether similar ratings reflect genuinely similar reasoning or merely coincidental alignment. This, in turn, would enable more nuanced evaluations of LLM-human alignment in conditional reasoning.

Acknowledgments

We would like to thank Alessandro Bogani for his initial input to the project and for his help in identifying relevant literature. Our appreciation extends to the anonymous reviewers for their valuable comments and suggestions. Lastly, we acknowledge the support for BP through the ERC Consolidator Grant DIALECT 101043235.

References

- Nick Barrowman. 2014. [Correlation, causation, and confusion](#). *The New Atlantis*, pages 23–44.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. [Fitting linear mixed-effects models using lme4](#). *Journal of statistical software*, 67:1–48.
- Francesco Berto and Aybüke Özgün. 2021. [Indicative conditionals: Probabilities and relevance](#). *Philosophical Studies*, 178(11):3697–3730.
- Violet A Brown. 2021. [An introduction to linear mixed-effects modeling in r](#). *Advances in Methods and Practices in Psychological Science*, 4(1).
- Ruth M. J. Byrne. 2007. *The Rational Imagination: How People Create Alternatives to Reality*. MIT Press.
- Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. [Transformers as soft reasoners over language](#). *arXiv preprint arXiv:2002.05867*.
- CRAN. 2025a. [Package ‘emmeans’](#). Accessed: 2025-05-30.
- CRAN. 2025b. [Quick start guide for emmeans](#). Accessed: 2025-05-30.
- Samuel Debray. 2023. [Conditional Reasoning in Large Language Models](#). Ph.D. thesis, Universiteit van Amsterdam.
- Hanna M Dettki, Brenden M Lake, Charley M Wu, and Bob Rehder. 2025. [Do large language models reason causally like us? even better?](#) *arXiv preprint arXiv:2502.10215*.
- Igor Douven. 2015. *The epistemology of indicative conditionals: Formal and empirical approaches*. Cambridge University Press.

- Igor Douven and Sara Verbrugge. 2012. [Indicatives, concessives, and evidential support](#). *Thinking & Reasoning*, 18(4):480–499.
- Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. [What did i do wrong? quantifying llms’ sensitivity and consistency to prompt engineering](#). *arXiv preprint arXiv:2406.12334*.
- Jonathan St B T Evans, Simon J Handley, and David E Over. 2003. [Conditionals and conditional probability](#). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(2):321–335.
- Jonathan St B T Evans and David E Over. 2004. *If: Supposition, Pragmatics, and Dual Processes*. Oxford University Press.
- Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, and 1 others. 2024. [Llm reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models](#). *arXiv preprint arXiv:2404.05221*.
- David Hardman and Laura Macchi. 2003. *Thinking: psychological perspectives on reasoning, judgment and decision making*. Wiley Online Library.
- Wesley H Holliday, Matthew Mandelkern, and Cedegao E Zhang. 2024. [Conditional and modal reasoning in large language models](#). *arXiv preprint arXiv:2401.17169*.
- Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. [The magic of if: Investigating causal reasoning abilities in large language models of code](#). *arXiv preprint arXiv:2305.19213*.
- Manikanta Loya, Divya Anand Sinha, and Richard Futrell. 2023. [Exploring the sensitivity of llms’ decision-making capabilities: Insights from prompt variation and hyperparameters](#). *arXiv preprint arXiv:2312.17476*.
- Philipp Mondorf and Barbara Plank. 2024a. [Beyond accuracy: Evaluating the reasoning behavior of large language models - a survey](#). In *First Conference on Language Modeling*.
- Philipp Mondorf and Barbara Plank. 2024b. [Comparing inferential strategies of humans and large language models in deductive reasoning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9402, Bangkok, Thailand. Association for Computational Linguistics.
- Mike Oaksford and Nick Chater. 2003. [Conditional probability and the cognitive science of conditional reasoning](#). *Mind & Language*, 18(4):359–379.
- David E Over, Constantinos Hadjichristidis, Jonathan St B T Evans, Simon J Handley, and Steven A Sloman. 2007. [The probability of causal conditionals](#). *Cognitive psychology*, 54(1):62–97.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. [Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting](#). *arXiv preprint arXiv:2310.11324*.
- SR Searle, FM Speed, and GA Milliken. 1980. [Population marginal means in the linear model: An alternative to least squares means](#). *American Statistician*, pages 216–221.
- Ruth G Shaw and Thomas Mitchell-Olds. 1993. [Anova for unbalanced data: An overview](#). *Ecology*, 74(6):1638–1645.
- Niels Skovgaard-Olsen, Henrik Singmann, and Karl Christoph Klauer. 2016. [The relevance effect and conditionals](#). *Cognition*, 150:26–36.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. [Large language models are inconsistent and biased evaluators](#). *arXiv preprint arXiv:2405.01724*.
- Barbara G Tabachnick and Linda S Fidell. 2007. *Experimental designs using ANOVA*, volume 724. Thomson/Brooks/Cole Belmont, CA.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Shravan Vasishth and Michael Broe. 2010. *The foundations of statistics: A simulation-based approach*. Springer Science & Business Media.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2025. [On verbalized confidence scores for LLMs](#). In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.
- Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Qingzhen Liu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2024. [Improving causal reasoning in large language models: A survey](#). *arXiv preprint arXiv:2410.16676*.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. [Large language models are not robust multiple choice selectors](#). *arXiv preprint arXiv:2309.03882*.

Relation / Probability	High-High	High-Low	Low-High	Low-Low
Positive (POS)	POS × High-High	POS × High-Low	POS × Low-High	POS × Low-Low
Negative (NEG)	NEG × High-High	NEG × High-Low	NEG × Low-High	NEG × Low-Low
Irrelevant (IRR)	IRR × High-High	IRR × High-Low	IRR × Low-High	IRR × Low-Low

Table 5: Design matrix showing all 12 conditional statement configurations for a single scenario, based on 3 relation types (positive, negative, irrelevant) and 4 prior probability levels (high-high, high-low, low-high, low-low).

A Appendix

A.1 Prior Human Study

Dataset In the dataset, as shown in Table 5, each conditional varies along two dimensions: The first dimension is prior probability assignment, where antecedent and consequent are each classified by Skovgaard-Olsen et al. (2016) as either high or low probability based on contextual cues provided in the scenario. This yields four prior configurations (*high-high*, *high-low*, *low-high*, *low-low*). According to Skovgaard-Olsen et al. (2016), this manipulation aims not to test probability effects directly, but to increase the spread of conditional probability judgments across statements.

The second dimension is the relation type, indicating the type of connection between antecedent *A* and consequent *B*. A positive (*POS*) relation means *A* supports *B*, a negative (*NEG*) relation means *A* undermines *B*, and an irrelevant (*IRR*) relation means *A* is unrelated to *B*.

In total, this then results in 12 statements per scenario, each representing a unique combination of prior probability and relation type.

Data Collection Skovgaard-Olsen et al. (2016) use this experimental setup to investigate how humans assess the acceptability of conditional statements *If A, then B*. Their dataset and supplementary materials are openly available on OSF. All ratings are collected via a crowd-sourcing platform (*CrowdFlower*), with participants from the United States, United Kingdom, and Australia. With this setup, the study shows that human judgments are sensitive to the semantic relation between *A* and *B* (positive, negative, or irrelevant) as well as the conditional probability of conditional statements.

While appropriate measures are implemented to ensure data quality, such as restricting participation to native English speakers and setting minimum and maximum completion times, these precautions are introduced only after data collection had already taken place. As a result, the dataset contains variable numbers of responses per conditional state-

ment (e.g., some statements are rated by ten participants, others by only three), which introduces inconsistency.

A.2 Models and Prompts

Language Models This study employs four publicly available large language models of comparable size (7 – 8 and 70 – 72 billion parameters): Llama 3.1 (8B and 70B) and Qwen 2.5 (7B and 72B). All models are instruction-tuned variants designed to follow user prompts and generate coherent natural language responses.

A summary of key model specifications is provided in Table 6. The Llama models, released in July 2024, consist of either 8.03 billion or 70.6 billion parameters, and the Qwen models, released in September 2024, contain 7.61 billion or 72.7 billion parameters. All models accommodate context windows of up to 128,000 tokens.

Prompts While the vanilla prompts follow a zero-shot format, providing only the task instruction, the scenario context, and the input item, the few-shot variants additionally include three example judgments based on a shared narrative context (see Figure 11). These examples span a range of output values (high, medium, low) to illustrate the rating scale. The same context and examples are used across all judgment types, with minor numerical variation to avoid repetition. Relation types are omitted to avoid bias. All prompts are constructed using each model’s respective chat formatting template. Instruction-tuned LLMs typically expect input messages to be wrapped in a specific format. These templates are applied using each model’s tokenizer. Where applicable, the model’s default sys-

Model ID	Parameters	Context Length
Llama-3.1-8B-it	8.03B	128K
Llama-3.1-70B-it	70.6B	128K
Qwen2.5-7B-it	7.61B	128K
Qwen2.5-72B-it	72.7B	128K

Table 6: Overview of selected large language models used in this study.

tem prompt is used. Model outputs are generated using the default decoding parameters provided by the Hugging Face Transformers library, with nucleus sampling ($top_p = 1.0$) and temperature ($T = 1.0$).

Outputs Each model judgment is expected to be a numerical value between 0 (completely improbable or unacceptable) and 100 (highly probable or acceptable). To elicit consistent responses, all prompts explicitly instruct models to return a single numerical value using both a textual instruction (Respond only with this number and nothing else.) and a predefined output format (Answer: <number>).

A.3 Prompt Templates

Vanilla Figure 8 shows the vanilla prompt used to elicit *conditional probability* estimates. Similar to the original human study (Skovgaard-Olsen et al., 2016), the model is given a context, a suppositional assumption (e.g., “Suppose Nicole sunbathes on the beach.”), and a target sentence (e.g., “She will get sunburned.”). It is then asked how probable the target sentence is, given the assumption.

Figure 9 shows the vanilla prompt used to elicit *if-probability* judgments. In this setup, the model is shown a context and a conditional statement (e.g., “If Nicole sunbathes on the beach, then she will get sunburned.”) and asked how probable the conditional is.

Figure 10 presents the vanilla prompt for *acceptability* ratings. The model receives a context and a conditional statement and is asked to assess the acceptability of the statement based on general knowledge or contextual plausibility. As Skovgaard-Olsen et al. (2016) include an instruction paragraph on what exactly is meant by acceptability in their task, a summarized version of this is also included in the LLM prompts for acceptability:

If a statement is considered "acceptable", it means that the information expressed by it is reasonable given some context or prior knowledge about the world.

Few-shot and CoT While the vanilla prompts follow a zero-shot format, providing only the task instruction, the scenario context, and the input item, the few-shot variants expand on this by including three illustrative examples after the instruction. As shown in Figure 11, these examples are embedded within a single, shared context and represent

a range of output values (one high, one moderate, and one low) to exemplify the scale of possible responses.

The example context is manually constructed to closely resemble the scenario style used in the original study: it establishes that a character, "Luisa", wants to pick up some groceries after work, and then adds constraints on the situation by stating that the supermarket is small and might not carry all the necessary items. The example inputs (e.g., *If Luisa drives to a bigger supermarket, she will find all the ingredients she needs*) are randomly selected to represent diverse output levels without biasing towards specific relation types. Moreover, relation types are deliberately omitted from the prompt to prevent influencing the model’s judgments. Those examples, designed to illustrate a clear probability gradient within the example scenario, remain identical across all prompts. Since the example output values are primarily intended as proxies to indicate a realistic range of responses, minor variations are introduced between judgment types (e.g., 92 → 94 → 95 for the high example) to avoid repetition. Additionally, as shown in Figure 12, the CoT prompts also include step-by-step reasoning that exemplifies how to arrive at the final judgment score.

A.4 Discussion of Elicitation Methods

While we use direct elicitation to retrieve (conditional) probability and acceptability ratings (cf. Sections 3.1 and A.2), a well-established approach (e.g., Tian et al., 2023; Yang et al., 2025), other methods exist for obtaining a model’s probability estimate for a given statement. As an alternative to direct elicitation, we also compute the sentence probability a model assigns to a given statement by leveraging its output probability distribution over the model’s vocabulary (see Section 4.2).

While both direct elicitation and sentence probability have their respective advantages and disadvantages (as illustrated in Table 7), they bear key differences that should be considered. Sentence probability quantifies how probable a model considers a given string. This captures surface form rather than semantics and does not necessarily reflect how likely the model deems the statement to be true. Additionally, pure sentence probability is highly sensitive to paraphrasing and tokenization and exhibits a notable length bias. In contrast, direct elicitation allows the model to be asked directly for the probability that a given statement is true. For

Method	How It Works	Advantages	Disadvantages
Direct elicitation (ask the model “What is $P(B A)$?”)	Prompt the model to output a numeric probability for B under the condition A (optionally after reasoning)	<ul style="list-style-type: none"> • Uses global semantics & latent knowledge; can consider A and B holistically rather than token-by-token • Supports reasoning (e.g., via CoT) before committing to a number. • Black-box friendly: no logit access required • Empirically, verbalized probabilities can be well-calibrated—often better than token-likelihoods for instruct/chat models (Tian et al., 2023) 	<ul style="list-style-type: none"> • Less decomposable/explainable at the score level (gives a number, not a token-level breakdown) • Variance under sampling (for non-greedy decoding)
Sentence probability (likelihood of text B given text A)	Compute $P(B A)$ by multiplying (summing log) token probabilities of the exact string B under left-to-right factorization	<ul style="list-style-type: none"> • Well-defined and deterministic (given fixed model & settings) • Decomposable at token level; supports log-likelihood ratios and ablations. • Aligns with the model’s pre-training objective (good for base models) 	<ul style="list-style-type: none"> • Measures string form, not pure semantics; highly paraphrase- and tokenization-sensitive • Length bias (longer B penalized); normalization choices matter (per-token, per-byte, etc.) • Not directly comparable across models without normalization (different tokenizers, logit scales, priors) • Requires logit access • Can’t natively leverage the model’s reasoning (e.g., via CoT) during scoring

Table 7: Comparison of methods for eliciting (conditional) probability judgments with large language models.

instruct/chat models in particular, direct elicitation has been shown to be better calibrated than token likelihoods (Tian et al., 2023), as it leverages the model’s internal latent knowledge and reasoning.

A.5 Linear Mixed-Effects Model

Linear Mixed-Effects Models Put simply, a linear mixed-effects model can be viewed as an extension of linear regression: Whereas simple regression fits a single trend line through the data, linear mixed-effects models estimate an average trend while adjusting for group-level variation (e.g., due to different scenarios). This results in more accurate and generalizable findings, particularly in datasets with multiple co-existing sources of variation (Brown, 2021; Vasishth and Broe, 2010).

Predictors and Sources of Variation The main fixed effects are conditional probability $P(B | A)$, relation type (supporting, undermining, unrelated), and judgment metric (acceptability vs. if-probability). To account for potential context-specific variation, random effects are included for scenario and *respondent*, where the latter refers to either a human participant or an LLM sampling iteration.

Model Specification and Adaptation Each linear mixed-effects model uses *judgment*, i.e., the judgments of conditional statements, as the dependent variable. A single model is fit per respondent (LLM or human) with the same structure, enabling later comparisons. Following best practices (Bates et al., 2015; Brown, 2021), we balance model complexity and convergence. Due to singularity issues with full random-slopes structures, we simplify the models by including random intercepts for both *respondent* (sampling iteration) and scenario in the LLM linear mixed-effects models. The human

	IRR	NEG	POS
Llama 8B (vanilla)	-0.37	-0.27	0.01
Llama 70B (vanilla)	-0.26	-0.27	0.04
Qwen 7B (vanilla)	-0.22	-0.24	0.06
Qwen 72B (vanilla)	-0.27	-0.21	0.16
Llama 8B (few-shot)	-0.33	-0.22	0.07
Llama 70B (few-shot)	-0.30	-0.23	0.11
Qwen 7B (few-shot)	-0.20	-0.19	0.09
Qwen 72B (few-shot)	-0.28	-0.21	0.15
Llama 70B (CoT)	-0.28	-0.23	0.17
Qwen 72B (CoT)	-0.33	-0.20	0.14

Table 8: Means of conditional statement judgments for all LLMs (joint means of both acceptability and if-probability), divided by relation types.

mixed-effects model, however, is fit with respondent as a random slope instead, as for humans, this represents the individual participants and is expected to introduce much greater variation. This choice captures key variance across participants and model samples while retaining model stability.

A.6 ANOVA Tables

Tables 9 to 18 illustrate the results from the ANOVA analyses.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	79.15	1	21.73	< 0.0001	***
Conditional probability	43.63	1	1421.39	< 0.0001	***
Relation type	65.62	2	1414.24	< 0.0001	***
Judgment metric (P / A)	4.37	1	1413.00	0.037	*
Cond. prob. × Relation	7.98	2	1420.91	< 0.001	***
Cond. prob. × Metric	2.49	1	1413.00	0.115	
Relation × Metric	6.09	2	1413.00	0.002	**
Cond. prob. × Relation × Metric	3.71	2	1413.00	0.025	*

Table 9: Type III ANOVA table for fixed effects in the mixed-effects model of *Llama 8B (vanilla)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	35.46	1	30.41	< 0.0001	***
Conditional probability	244.54	1	1422.38	< 0.0001	***
Relation type	2.82	2	1415.34	0.060	.
Judgment metric (P / A)	3.61	1	1413.00	0.058	.
Cond. prob. × Relation	0.53	2	1418.53	0.591	
Cond. prob. × Metric	2.30	1	1413.00	0.130	
Relation × Metric	0.24	2	1413.00	0.783	
Cond. prob. × Relation × Metric	0.15	2	1413.00	0.861	

Table 10: Type III ANOVA table for fixed effects in the mixed-effects model of *Llama 70B (vanilla)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	72.03	1	16.77	< 0.0001	***
Conditional probability	94.36	1	1416.96	< 0.0001	***
Relation type	100.09	2	1413.35	< 0.0001	***
Judgment metric (P / A)	5.98	1	1413.00	0.015	*
Cond. prob. × Relation	10.72	2	1417.78	< 0.0001	***
Cond. prob. × Metric	24.02	1	1413.00	< 0.0001	***
Relation × Metric	1.51	2	1413.00	0.220	
Cond. prob. × Relation × Metric	0.69	2	1413.00	0.503	

Table 11: Type III ANOVA table for fixed effects in the mixed-effects model of *Qwen 7B (vanilla)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	224.65	1	18.65	< 0.0001	***
Conditional probability	34.03	1	1414.81	< 0.0001	***
Relation type	200.90	2	1414.06	< 0.0001	***
Judgment metric (P / A)	3.33	1	1413.00	0.068	.
Cond. prob. × Relation	59.37	2	1416.95	< 0.0001	***
Cond. prob. × Metric	72.79	1	1413.00	< 0.0001	***
Relation × Metric	0.49	2	1413.00	0.613	
Cond. prob. × Relation × Metric	7.53	2	1413.00	0.001	***

Table 12: Type III ANOVA table for fixed effects in the mixed-effects model of *Qwen 72B (vanilla)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	79.67	1	27.72	< 0.0001	***
Conditional probability	29.62	1	1423.62	< 0.0001	***
Relation type	68.20	2	1415.28	< 0.0001	***
Judgment metric (P / A)	15.88	1	1413.00	< 0.0001	***
Cond. prob. × Relation	9.86	2	1420.44	< 0.0001	***
Cond. prob. × Metric	8.19	1	1413.00	0.004	**
Relation × Metric	7.74	2	1413.00	< 0.001	***
Cond. prob. × Relation × Metric	1.98	2	1413.00	0.138	

Table 13: Type III ANOVA table for fixed effects in the mixed-effects model of *Llama 8B (few-shot)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	83.12	1	18.76	< 0.0001	***
Conditional probability	202.71	1	1419.33	< 0.0001	***
Relation type	63.57	2	1414.17	< 0.0001	***
Judgment metric (P / A)	14.26	1	1413.00	< 0.001	***
Cond. prob. × Relation	28.08	2	1416.22	< 0.0001	***
Cond. prob. × Metric	11.92	1	1413.00	0.001	***
Relation × Metric	2.03	2	1413.00	0.132	
Cond. prob. × Relation × Metric	1.42	2	1413.00	0.243	

Table 14: Type III ANOVA table for fixed effects in the mixed-effects model of *Llama 70B (few-shot)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	83.55	1	22.01	< 0.0001	***
Conditional probability	140.37	1	1422.45	< 0.0001	***
Relation type	51.58	2	1414.02	< 0.0001	***
Judgment metric (P / A)	6.02	1	1413.00	0.014	*
Cond. prob. × Relation	5.87	2	1419.68	0.003	**
Cond. prob. × Metric	5.06	1	1413.00	0.025	*
Relation × Metric	4.20	2	1413.00	0.015	*
Cond. prob. × Relation × Metric	3.98	2	1413.00	0.019	*

Table 15: Type III ANOVA table for fixed effects in the mixed-effects model of *Qwen 7B (few-shot)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	307.31	1	19.69	< 0.0001	***
Conditional probability	73.00	1	1417.90	< 0.0001	***
Relation type	175.10	2	1414.31	< 0.0001	***
Judgment metric (P / A)	42.24	1	1413.00	< 0.0001	***
Cond. prob. × Relation	48.50	2	1416.94	< 0.0001	***
Cond. prob. × Metric	59.29	1	1413.00	< 0.0001	***
Relation × Metric	5.51	2	1413.00	0.004	**
Cond. prob. × Relation × Metric	6.90	2	1413.00	0.001	**

Table 16: Type III ANOVA table for fixed effects in the mixed-effects model of *Qwen 72B (few-shot)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	175.71	1	19.94	< 0.0001	***
Conditional probability	165.37	1	1416.16	< 0.0001	***
Relation type	130.97	2	1414.13	< 0.0001	***
Judgment metric (P / A)	23.26	1	1413.00	< 0.0001	***
Cond. prob. × Relation	25.87	2	1416.33	< 0.0001	***
Cond. prob. × Metric	18.60	1	1413.00	< 0.0001	***
Relation × Metric	2.54	2	1413.00	0.079	.
Cond. prob. × Relation × Metric	6.18	2	1413.00	0.002	**

Table 17: Type III ANOVA table for fixed effects in the mixed-effects model of *Llama 70B (CoT)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	312.63	1	20.73	< 0.0001	***
Conditional probability	24.43	1	1416.32	< 0.0001	***
Relation type	188.00	2	1413.85	< 0.0001	***
Judgment metric (P / A)	14.43	1	1413.00	< 0.001	***
Cond. prob. × Relation	65.29	2	1416.77	< 0.0001	***
Cond. prob. × Metric	21.09	1	1413.00	< 0.0001	***
Relation × Metric	2.50	2	1413.00	0.083	.
Cond. prob. × Relation × Metric	1.60	2	1413.00	0.202	

Table 18: Type III ANOVA table for fixed effects in the mixed-effects model of *Qwen 72B (CoT)* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Effect	F value	Df	Residual Df	p-value	
(Intercept)	190.62	1	213.73	< 0.0001	***
Conditional probability	80.76	1	234.87	< 0.0001	***
Relation type	116.15	2	235.95	< 0.0001	***
Judgment metric (P / A)	1.94	1	1190.22	0.164	
Cond. prob. × Relation	22.73	2	238.67	< 0.0001	***
Cond. prob. × Metric	1.12	1	655.91	0.290	
Relation × Metric	2.07	2	786.15	0.127	
Cond. prob. × Relation × Metric	1.09	2	641.12	0.335	

Table 19: Type III ANOVA table for fixed effects in the mixed-effects model of *Human* judgments. Probability and acceptability ratings are predicted by conditional probability, relation type, and judgment type. Degrees of freedom are computed using the Kenward–Roger approximation. Significance codes: *** $p < .001$, ** $p < .01$, * $p < 0.05$.

Low Cond. Prob.	IRR × NEG		IRR × POS		NEG × POS	
	Estimate	<i>p</i>	Estimate	<i>p</i>	Estimate	<i>p</i>
Llama 8B (vanilla)	-0.09	< 0.001	-0.13	< 0.001	-0.04	0.148
Llama 70B (vanilla)	-0.07	0.009	0.01	0.827	0.08	0.065
Qwen 7B (vanilla)	-0.01	0.684	-0.13	< 0.0001	-0.12	< 0.0001
Qwen 72B (vanilla)	-0.02	0.232	-0.13	< 0.0001	-0.10	< 0.0001
Llama 8B (few-shot)	-0.06	0.030	-0.16	< 0.0001	-0.11	0.001
Llama 70B (few-shot)	-0.07	0.001	-0.07	0.025	0.00	0.906
Qwen 7B (few-shot)	-0.05	0.063	-0.18	< 0.0001	-0.13	< 0.0001
Qwen 72B (few-shot)	-0.03	0.263	-0.11	< 0.001	-0.08	0.007
Llama 70B (CoT)	-0.11	< 0.0001	-0.24	< 0.0001	-0.12	< 0.001
Qwen 72B (CoT)	-0.01	1.000	-0.10	0.003	-0.09	0.003
Human	0.01	0.421	-0.13	< 0.0001	-0.15	< 0.0001

Medium C. Prob.	IRR × NEG		IRR × POS		NEG × POS	
	Estimate	<i>p</i>	Estimate	<i>p</i>	Estimate	<i>p</i>
Llama 8B (vanilla)	-0.14	< 0.0001	-0.27	< 0.0001	-0.13	< 0.0001
Llama 70B (vanilla)	-0.09	0.001	-0.02	0.598	0.06	0.033
Qwen 7B (vanilla)	-0.07	< 0.0001	-0.24	< 0.0001	-0.17	< 0.0001
Qwen 72B (vanilla)	-0.21	< 0.0001	-0.30	< 0.0001	-0.09	< 0.0001
Llama 8B (few-shot)	-0.16	< 0.0001	-0.26	< 0.0001	-0.09	< 0.0001
Llama 70B (few-shot)	-0.25	< 0.0001	-0.16	< 0.0001	0.09	< 0.0001
Qwen 7B (few-shot)	-0.18	< 0.0001	-0.24	< 0.0001	-0.06	0.001
Qwen 72B (few-shot)	-0.20	< 0.0001	-0.25	< 0.0001	-0.04	0.004
Llama 70B (CoT)	-0.27	< 0.0001	-0.28	< 0.0001	-0.01	0.459
Qwen 72B (CoT)	-0.27	< 0.0001	-0.32	< 0.0001	-0.05	0.002
Human	-0.08	< 0.0001	-0.32	< 0.0001	-0.24	< 0.0001

High Cond. Prob.	IRR × NEG		IRR × POS		NEG × POS	
	Estimate	<i>p</i>	Estimate	<i>p</i>	Estimate	<i>p</i>
Llama 8B (vanilla)	-0.19	< 0.0001	-0.41	< 0.0001	-0.22	< 0.0001
Llama 70B (vanilla)	-0.10	0.065	-0.05	0.331	0.05	0.540
Qwen 7B (vanilla)	-0.14	< 0.0001	-0.35	< 0.0001	-0.21	< 0.0001
Qwen 72B (vanilla)	-0.40	< 0.0001	-0.47	< 0.0001	-0.07	0.001
Llama 8B (few-shot)	-0.27	< 0.0001	-0.35	< 0.0001	-0.08	0.036
Llama 70B (few-shot)	-0.44	< 0.0001	-0.25	< 0.0001	0.19	< 0.0001
Qwen 7B (few-shot)	-0.30	< 0.0001	-0.30	< 0.0001	0.00	0.972
Qwen 72B (few-shot)	-0.38	< 0.0001	-0.39	< 0.0001	-0.01	0.714
Llama 70B (CoT)	-0.42	< 0.0001	-0.33	< 0.0001	0.09	0.001
Qwen 72B (CoT)	-0.52	< 0.0001	-0.54	< 0.0001	-0.01	1.000
Human	-0.17	< 0.0001	-0.52	< 0.0001	-0.34	< 0.0001

Table 20: Pairwise comparisons of estimated marginal means at low, medium, and high conditional probability for all LLMs and humans. SE: 0.01 – 0.04 for all. Statistical significance at $p < 0.05$.

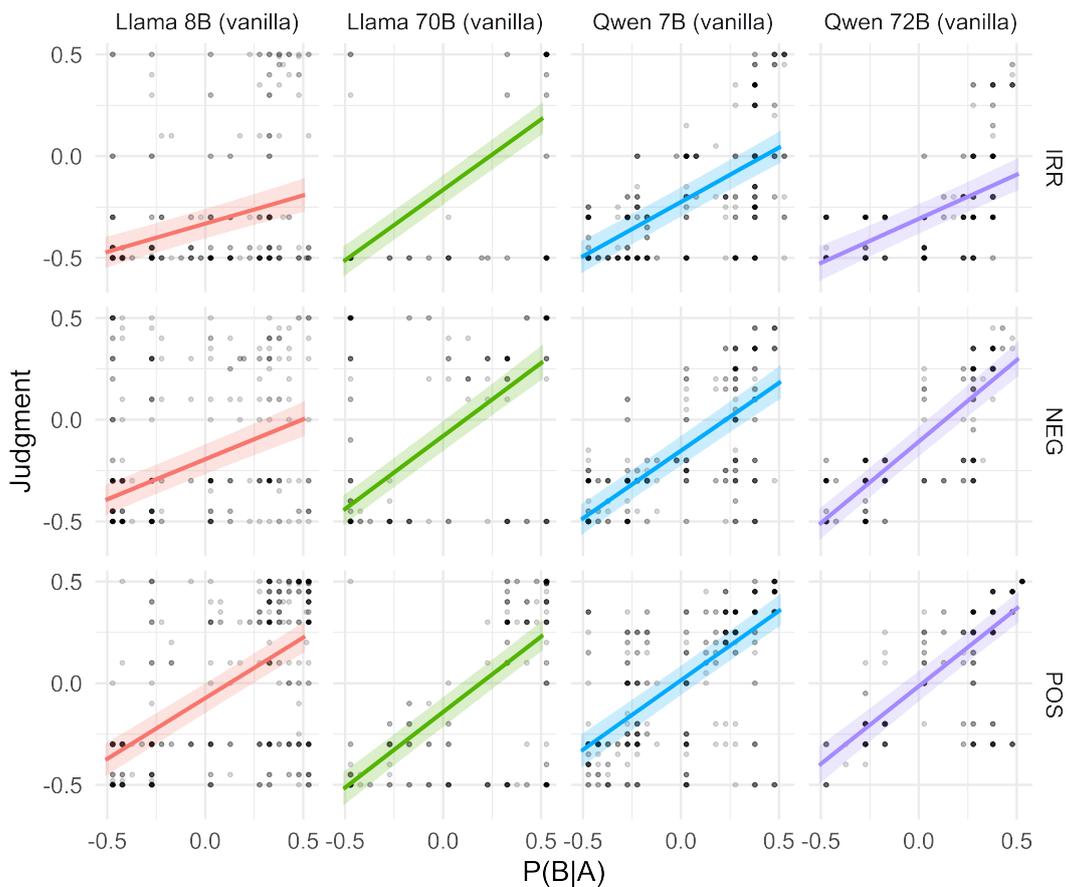


Figure 7: Scatterplots of data points and trend lines for all vanilla LLMs, divided by relation type.

Instruction:

You will be given an assumption and a sentence:

- Assumption: a supposition to treat as true
- Sentence: the claim to evaluate

Given the assumption, how probable is it that the sentence is true? Your answer must be a single number between 0 and 100, where:

- 0 means "completely improbable"
- 100 means "highly probable"

Respond only with this number and nothing else.

Output Format:

Answer: <number>

Input:

Context: Nicole's skin is pale and she wants a suntan with a mild brown color. It is summer, the weather is nice, and there is a beach nearby. Unfortunately, Nicole's skin is very sensitive to the sun, and she is somewhat bad at controlling the amount of her exposure to the sun.

Assumption: Suppose Nicole sunbathes on the beach.

Sentence: She will get sunburned

Figure 8: Prompt for eliciting conditional probability judgments (vanilla version).

Instruction:

Rate the probability of the following statement. Your answer must be a single number between 0 and 100, where:
- 0 means "completely improbable"
- 100 means "highly probable"

Respond only with this number and nothing else.

Output Format:

Answer: <number>

Input:

Context: Nicole's skin is pale and she wants a suntan with a mild brown color. It is summer, the weather is nice, and there is a beach nearby. Unfortunately, Nicole's skin is very sensitive to the sun, and she is somewhat bad at controlling the amount of her exposure to the sun.

Statement: If Nicole sunbathes on the beach, then she will get sunburned.

Figure 9: Prompt for eliciting If A , then B probability judgments (vanilla version).

Instruction:

If a statement is considered "acceptable", it means that the information expressed by it is reasonable given some context or prior knowledge about the world.

Rate the acceptability of the following statement. Your answer must be a single number between 0 and 100, where:
- 0 means "completely unacceptable"
- 100 means "highly acceptable"

Respond only with this number and nothing else.

Output Format:

Answer: <number>

Input:

Context: Nicole's skin is pale and she wants a suntan with a mild brown color. It is summer, the weather is nice, and there is a beach nearby. Unfortunately, Nicole's skin is very sensitive to the sun, and she is somewhat bad at controlling the amount of her exposure to the sun.

Statement: If Nicole sunbathes on the beach, then she will get sunburned.

Figure 10: Prompt for eliciting If A , then B acceptability judgments (vanilla version).

Context: Luisa wants to pick up some groceries on her way home from work because she needs the ingredients to cook her favorite meal tonight. However, the supermarket is very small and does not always carry all the things that other supermarkets do. Driving to a bigger supermarket would take a very long time, but she is already pretty hungry.

Example 1 (High probability):

Statement: If Luisa drives to a bigger supermarket, she will find all the ingredients she needs.

Answer: 92

Example 2 (Moderate probability):

Statement: If Luisa drives to the small supermarket, she will find all the ingredients she needs.

Answer: 57

Example 3 (Low probability):

Statement: If Luisa goes straight home, she will find all the ingredients she needs.

Answer: 12

Figure 11: Example context and statements used in the few-shot prompt variants.

Example 1 (High probability):

Statement: If Luisa drives to a bigger supermarket, she will find all the ingredients she needs.

Reasoning:

1. The context tells us that Luisa is looking for specific ingredients to cook her favorite meal.
2. Bigger supermarkets are generally well-stocked and carry a wider variety of products.
3. It is reasonable to expect that she will find the necessary items there.
4. Although there's a small chance that something might be out of stock, this doesn't undermine the general plausibility.
5. Therefore, the statement expresses a reasonable and plausible scenario.

Answer: 95

Example 2 (Moderate probability):

Statement: If Luisa drives to the small supermarket, she will find all the ingredients she needs.

Reasoning:

1. The context tells us that the small supermarket does not always carry everything other supermarkets do.
2. This implies that its inventory is limited or inconsistent.
3. This makes it uncertain whether it has the full range of ingredients she needs.
4. While the statement is not impossible, it's not strongly supported by what we know.
5. Therefore, the statement is somewhat acceptable, but questionable.

Answer: 51

Example 3 (Low probability):

Statement: If Luisa goes straight home, she will find all the ingredients she needs.

Reasoning:

1. The context tells us that Luisa wants to stop for groceries because she needs ingredients, implying that she does not already have the ingredients at home.
2. If she skips the store and goes straight home, she is unlikely to have everything required for the meal.
3. The statement contradicts her reason for going to the store in the first place.
4. While there's always a slim chance she forgot she already has everything, this is implausible.
5. Therefore, the statement is largely unacceptable.

Answer: 8

Figure 12: Example statements and reasoning used in the acceptability CoT prompt variants.