Evaluating Communicative Belief Updates in Large Language Models via Implicature Recognition and Cancellation

Cesare Spinoso-Di Piano[1] **Verna Dankers**[1]☕ **Marius Mosbach**[1]☕ **Jackie Chi Kit Cheung**[1,2]
[1]Mila - Quebec AI Institute & McGill University, [2]Canada CIFAR AI Chair
{cesare.spinoso, cheungja}@mila.quebec

Abstract

Human language is driven by unspoken beliefs and belief updates, making these critical to model for successful communication between large language models (LLMs) and their users. In this paper, we evaluate the ability of LLMs to recognize unspoken beliefs made through *implicatures* and to understand their updates through *implicature cancellation*: the pragmatic phenomenon whereby an utterance's implied meaning is *weakened* or *negated*. We create the first expert-annotated implicature cancellation dataset, IMPLICATUREX, crowdsourced for human judgements of implicatures and their corresponding cancellations. We find that LLM belief update understanding lags behind that of humans, especially in more naturally-occurring scenarios. Additional control experiments suggest that successes in LLM belief updates may stem in part from a reliance on prior beliefs, and that failures in belief updates may depend on their type and on their form. Overall, our study suggests that current LLMs have not yet reached human-level understanding of unspoken beliefs and belief updates.[1]

1 Introduction

Natural language communication consists of threads of beliefs negotiated between interlocutors in a conversational common ground (Lewis, 1979; Heim, 1982; Clark and Brennan, 1991; Stalnaker, 1998; Kamp, 1981). These beliefs are often introduced and updated implicitly through *implicatures* and *implicature cancellations*, whereby a previously implicated belief is *negated* or *weakened* (Grice, 1975; Sperber and Wilson, 1986). For instance, when Bo asks their friend Aya whether they can help with something, the response "I think Cai was looking for someone to buy drinks." triggers an implicature that Cai needs Bo to buy drinks (Figure 1). However, Aya can cancel this implicature by saying "Though they must have gotten around to it by now.", at which point Bo should update their beliefs and act accordingly, e.g., by asking Cai to confirm. Interlocutors' beliefs can thus change from one utterance to the next. Understanding these beliefs and how they are updated is thus of fundamental importance for successful interactions between large language models (LLMs) and human system users.

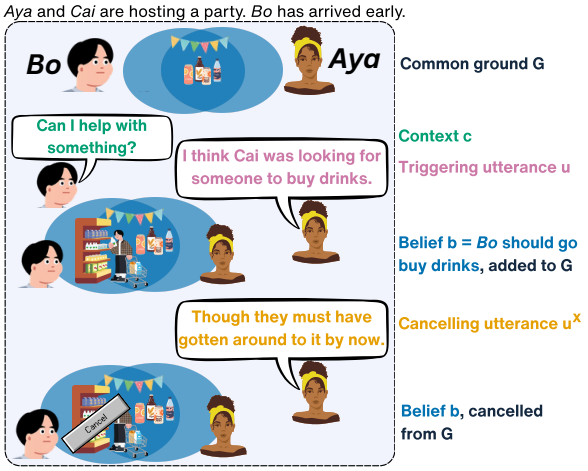

Figure 1: An example of belief negotiation: A belief b in the common ground G is updated as an implicature is triggered by u and then cancelled by $u^{\times}$.

In this work, we evaluate LLMs' ability to understand such dynamically changing beliefs in language. We focus on implicature recognition and cancellation, two of the most widely studied and fundamental phenomena involving belief updates in linguistics and cognitive science (Grice, 1975; Sperber and Wilson, 1986). While previous studies have evaluated the ability of LLMs to perform belief updates of world knowledge and logical reasoning (Rudinger et al., 2020; Hwang et al., 2021), similar evaluations in the communicative space re-

[1]Code and data available at https://github.com/cesare-spinoso/ImplicatureX.
☕ Equal contribution.

main underexplored. Our work fills this gap by evaluating LLMs' ability to identify beliefs triggered by implicatures and to update those implicated beliefs as a result of implicature cancellation.

We create the first implicature cancellation dataset, IMPLICATUREX, consisting of 271 expert-annotated implicatures and corresponding cancellations. Beyond synthetic two-turn conversational implicatures, it contains naturally occurring scalar implicatures, discourse implicatures and multi-turn dialogue conversational implicatures. In addition, we conduct and include a crowdsourcing annotation of our IMPLICATUREX items to measure implicature and cancellation recognition accuracies.

Our results show that LLMs' pragmatic reasoning falls short of a human understanding of unspoken beliefs conveyed by implicatures. While the strongest LLMs we tested (e.g., GPT-5.4 Thinking) match human performance on implicature recognition of scalar and discourse implicatures, LLMs struggle to perform at a level above random chance on naturally-occurring conversational implicatures. Moreover, a control experiment reveals that, in many cases, LLMs successfully recognize implicatures without observing the context or the utterance. We, therefore, question the extent to which their success is due to legitimate pragmatic reasoning.

In a similar vein to our implicature recognition findings, we find that belief updates following implicature cancellation—which are readily made by humans—are especially challenging for LLMs in naturally-occurring and realistic scenarios. Additional control experiments reveal that the type of belief update—cancelling, leaving unchanged, or strengthening—and the way in which the update is triggered—explicitly or implicitly—affect the extent to which LLMs are able to revise their existing beliefs. For instance, our results demonstrate that even when presented with explicit belief negations, strong LLMs (e.g., Qwen 3 32B Thinking) cannot match human accuracy in cancellation recognition.

To summarize, we study the extent to which LLMs are able to understand unspoken beliefs via implicature recognition and belief updates via implicature cancellation, two fundamental properties of human communication. We create the first implicature cancellation dataset, IMPLICATUREX, verified by linguistic experts and annotated with crowdsourced judgements. Our evaluation reveals that LLMs fall short of both a human understanding of unspoken beliefs and of belief updates. More broadly, our work suggests that LLMs may still not be able to understand the nuances of belief productions and negotiations, especially when they are unspoken and naturally-occurring.

2 Related Work

Philosophy of language Communication has long been thought to be an exercise of *belief negotiation* (Grice, 1957; Lewis, 1979). In this vein, several seminal studies have posited that we communicate by making updates to an ever-evolving tacitly agreed upon set of propositions, i.e., a *common ground* (Heim, 1982; Clark and Brennan, 1991; Stalnaker, 1998; Kamp, 1981). As such, implicatures—beliefs about a possible intended meaning suggested by a speaker—and implicature negotiations—the cancellation and strengthening of these beliefs—are a core mechanism by which we add and make updates to the common ground (Grice, 1975). In this work, we offer an experimental account of belief updates through implicature cancellations which, unlike implicature strengthenings (Benotti, 2010), remain an understudied area of belief negotiation.

Experimental pragmatics Belief negotiation in human communication has received widespread attention in experimental pragmatics. Initially studied in the context of referring expressions (Clark and Wilkes-Gibbs, 1986; Isaacs and Clark, 1987; Selten and Warglien, 2007; Deemter et al., 2012), negotiations of common ground beliefs have continued to be of central relevance to other pragmatic phenomena including non-verbal communication (Veinott et al., 1999; Clark and Krych, 2004), discourse relations (Fetzer, 2018) and scalar implicatures (Noveck, 2001). In particular, studies have shown that the inferences made from scalar implicatures are more easily and more readily accepted into the conversational common ground based on the context in which they appear and the prior beliefs held by interlocutors (Breheny et al., 2006; Grodner et al., 2010; Degen, 2015; Yang et al., 2018; Huang and Snedeker, 2018). Our study provides a continued experimental investigation of communicative belief updates through the phenomenon of implicature cancellation.

Communicative beliefs in NLP Notions surrounding communicative beliefs have long served as inspiration for language generation systems, including dialogue systems (Allen and Perrault, 1980; Grosz and Sidner, 1986; Dale and Reiter, 1995),

image captioning (Andreas and Klein, 2016), and conversational agents (Körner et al., 2025). Furthermore, implicatures and unspoken beliefs have been used extensively to evaluate the communicative competence of LLMs (Jeretic et al., 2020; Kabbara and Cheung, 2022; Ruis et al., 2023; Cho and mook Kim, 2024; Yue et al., 2024; Cong, 2024). Our contribution differs in that we leverage the cancellability of implicatures as a tractable space to evaluate the ability of LLMs to *update* unspoken beliefs. Finally, while processes similar to implicature cancellation, such as abductive and defeasible reasoning, have been studied extensively (Lascarides and Asher, 1991; Rudinger et al., 2020; Hwang et al., 2021), they differ from our focus of communicative belief updates.

3 Operationalizing Implicature and Implicature Cancellation

We view any communicative exchange (spoken, written, signed, etc.) as consisting of a sequence of *utterances* $u \in \mathcal{U}$, i.e., any unit of language produced by a participant of the exchange. Further, let $\mathcal{U}^*$ denote the set of all possible utterance sequences and $\mathbf{u} = \langle u_1, \ldots, u_n \rangle \in \mathcal{U}^*$ a communicative history. Each exchange takes place in a *context* $c \in \mathcal{C}$, which captures situational and communicative information relevant to interpreting the utterances.[2]

Participants share a set of *beliefs* $\mathcal{B}$, i.e., propositional statements such as "Cai needs help buying drinks.", which represent what they mutually take to be true. Formally, the *common ground* is a function $G : \mathcal{B} \times \mathcal{C} \times \mathcal{U}^* \rightarrow [0, 1]$ that, given c and $\mathbf{u}$, assigns to each $b \in \mathcal{B}$ a probability reflecting how strongly that belief is mutually held by the participants.[3]

A belief b becomes part of the common ground via an *implicature* when utterance u, produced in context c, makes b more probable than its negation according to the common ground, i.e., $G(b \mid c, \langle u \rangle) > G(\neg b \mid c, \langle u \rangle)$. A belief b is subsequently *cancelled* from the common ground via an *implicature cancellation* when a speaker extends the communicative history from $\langle \ldots, u \rangle$ to $\langle \ldots, u, u^{\times} \rangle$ with a cancelling utterance $u^{\times}$, which weakens or negates the implicature triggered by u, i.e., $G(b \mid c, \langle u, u^{\times} \rangle) < G(b \mid c, \langle u \rangle)$. Thus, we define a *belief update* in the context of implicature cancellation as $G(b \mid c, \langle u, u^{\times} \rangle) < G(b \mid c, \langle u \rangle)$ when provided with $u^{\times}$ for an utterance u for which $G(b \mid c, \langle u \rangle) > G(\neg b \mid c, \langle u \rangle)$ holds.

[2] c is assumed to be fixed for the duration of the exchange.

[3] Prior to any exchange, the common ground is given by $G(b \mid \emptyset, \emptyset)$, which may assign uniform probability to all $b \in \mathcal{B}$, or reflect presupposed propositions based on the participants' shared history and prior beliefs (Anderson, 2018).

4 The IMPLICATUREX Dataset

Here, we describe the composition of the IMPLICATUREX implicature cancellation dataset along with its expert and crowdsourcing annotation.

4.1 Dataset Composition

IMPLICATUREX is an expert-annotated dataset of 271 items which include ❶ SCALAR IMPLICATURES, ❷ DISCOURSE IMPLICATURES and ❸ CONVERSATIONAL IMPLICATURES. Figure 2 provides example stimuli per implicature type. Below, we describe the dataset's composition.

❶ SCALAR IMPLICATURES. Our items are based on the well-known pragmatic phenomenon where, given a pair of lexical items $\langle w_1, w_2 \rangle$ with w_2 semantically entailing w_1, the use of w_1 in an utterance indicates the negation of w_2. For instance, the triggering utterance u = "*Some* places nail them for tax." increases the probability of the belief b = "*Not all* places nail them for tax." belonging to the common ground. This belief can be revised by the speaker using a cancelling utterance $u^{\times}$ = "I think they *all* do that now.". We sample 50 naturally-occurring $\langle$*some*, *all*$\rangle$ scalar implicatures from the Switchboard Corpus (Godfrey et al., 1992) as originally collected by Degen (2015) and manually add an implicature cancellation to each sampled item.

❷ DISCOURSE IMPLICATURES. We posit that discourse relations, especially those which involve implicit causal relations, may be recast as *discourse* implicatures—i.e., implicatures found in traditional forms of discourse. In particular, we identify and extract discourse implicatures from Wall Street Journal (WSJ) articles, using the corresponding implicit discourse relations annotated in the Penn Discourse Tree Bank (PDTB) corpus (Prasad et al., 2019). Using the implicit PDTB causal discourse relations, we are able to identify excerpts such as c = "The Fuji apple has been extensively researched." and u = "Strains have been developed to age as gracefully as the Granny apples." which implicate the belief b = "Fuji apples last as long as Granny apples." This belief is negated—or at least weakened—with the cancelling utterance $u^{\times}$ =

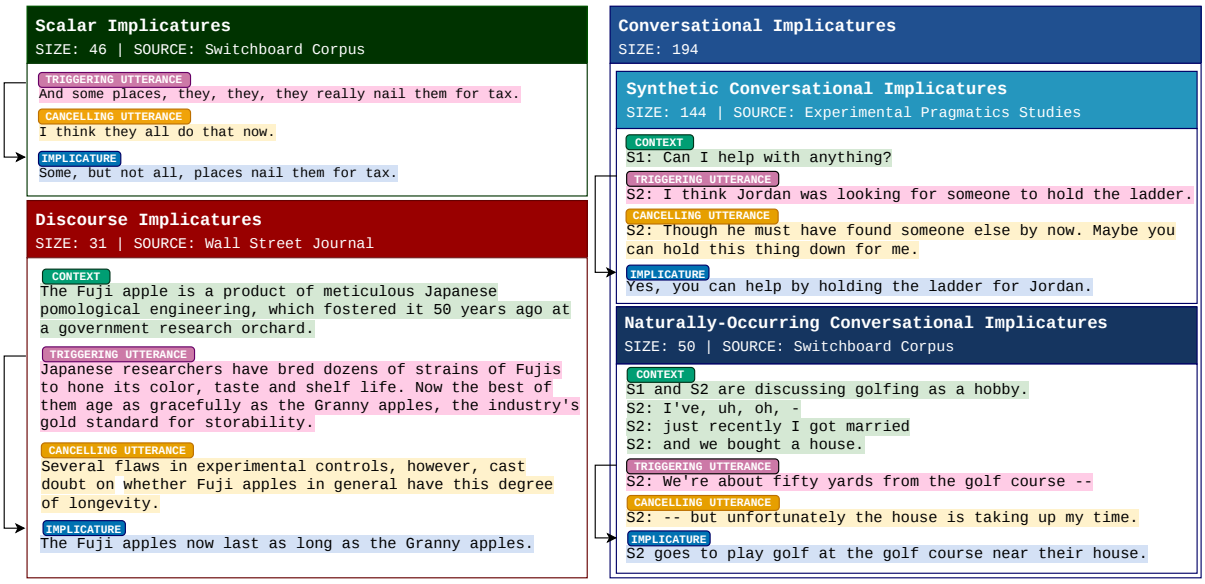

Figure 2: Composition of the IMPLICATUREX dataset with sizes, sources and illustrative examples. Note that the sizes reported in this figure reflect the dataset post cleaning (Section 4.2).

"Though flaws in experimental controls now cast doubt on the apple's longevity." We identify and annotate 31 WSJ article excerpts using implicit causal relations from the PDTB corpus.

3 CONVERSATIONAL IMPLICATURES. Our conversational implicatures consist of implicatures which are triggered by utterances produced in a transcribed exchange between two conversational participants. These 197 conversational implicatures are further subdivided into two categories which differ principally by their source: **A** Synthetic conversational implicatures and **B** Naturally-occurring conversational implicatures.

A The synthetic conversational implicatures consist of an exchange between two participants, S_1 and S_2. The exchange begins typically with a question (c = "Do you need help with anything?") which is followed by a response by the other participant (u = "I think C was looking for someone to do X."), introducing an implicated belief into the common ground (b = "You can help by doing X for C."). This belief is updated via a cancelling utterance ($u^{\times}$ = "Though they must be done by now."). We collect 146 such two-turn conversational implicatures from several existing implicature understanding datasets (George and Mamidi, 2020; Louis et al., 2020; Wilson and Bishop, 2021; Hu et al., 2023) and manually add implicature cancellations.

B Our scenario-based conversational implicatures also consist of an exchange between S_1 and S_2. In this case, the exchange is a naturally-occurring discussion between two Switchboard Corpus participants (Godfrey et al., 1992). We collect 51 naturally-occurring conversational implicatures. The cancelling utterances are produced by one of the participants within the exchange. For instance, u = "We're close to the golf course." (b = "They go golfing near their home") is cancelled by $u^{\times}$ = "Unfortunately the house is taking up time."

4.2 Expert Annotation

Annotation details To validate the plausibility of the implicatures and the correctness of their cancellations in IMPLICATUREX, we ran an expert annotation of our implicature cancellation items. We elicit expert judgements for the implicatures' plausibility as well as the cancellations' correctness. In addition, we elicit alternatives for implausible implicatures or incorrect cancellations (where possible).

The stimuli were subdivided into six batches of 53 items each of which included seven attention checks, i.e., items with either a clearly implausible implicature or incorrect cancellation. The seven attention checks were different for each batch and their frequency per batch reflected the makeup of the dataset: one scalar implicature, one discourse implicature, four synthetic conversational implicatures and one naturally-occurring conversational implicature.

We hired two professionally trained linguists for two rounds of expert annotation. In the first annotation round, both expert annotators were given the same batch in order to compute inter-annotator

agreement of implicature plausibility and cancellation correctness. In the second round of annotations, one of the expert annotators was tasked to annotate the five remaining batches.

Additional details regarding the expert annotation are presented in Appendix A.

Annotation results For the first round of annotation, we report the raw agreement, Cohen's kappa (κ) and the prevalence-adjusted bias-adjusted kappa (PABAK) for implicature plausibility and cancellation correctness[4] in Table 1. The raw agreement and PABAK are relatively high. The relatively low Cohen's κ is explained by the imbalance in class distributions for both the implicature plausibility and cancellation correctness annotation (Byrt et al., 1993) (See confusion matrices in Tables 3 and 4 in Appendix A).

	Raw Agreement	Cohen's κ	PABAK
Implicature Plausibility	0.85	0.34	0.70
Cancellation Correctness	0.86	0.32	0.72

Table 1: Inter-annotator agreement scores per annotation task.

The second round of annotation generated 18 implicature replacements and 27 cancellation replacements. In seven cases, the items were flagged but the annotator was not able to provide alternatives (e.g., the cancellations for the naturally-occurring conversational implicatures). After dropping and replacing the flagged items, this left us with a total of 271 expert-annotated implicature cancellation items which are distributed as follows: 1: 46 items, 2: 31 items, 3A: 144 items, 3B: 50 items.

4.3 Crowdsourcing Annotation

While our expert annotation served to validate the plausibility and correctness of IMPLICATUREX, we perform a crowdsourcing annotation to provide a realistic topline comparison to LLM understanding of implicature and implicature cancellation.

Annotation details To obtain aggregated human judgments of IMPLICATUREX belief updates, we run a crowdsourcing annotation of the 271 expert annotated items. To do so, we elicit likelihood judgements for each implicature item given its corresponding context, triggering utterance, and cancelling utterance. In particular, we create 16 stimuli batches of 30 items and two stimuli batches of 32 items. For each batch, half of the items contain the cancelling utterance and half do not. We shuffle batch items in random order and ensure that for every batch no overlap exists between the items with and without the cancelling utterance. In addition, we include and reuse 10 attention checks for every batch, five whose likelihood rating should be high and five whose likelihood rating should be low.

To implement our likelihood judgement elicitation, we generalize the approach from experimental pragmatics for studying the strength of scalar implicatures to our entire set of stimuli. In particular, we generalize Degen (2015)'s approach and ask participants to rate the likelihood of the implicature on a seven point Likert scale with endpoints labeled as "absolutely impossible" and "absolutely certain" and individual points labeled as 1, 2, ..., 7.

We recruit 90 participants from the Prolific platform based in Canada, the U.S. and the U.K. whose self-reported native language is English. In addition, we select participants with an undergraduate degree and who have taken part in at least 100 studies with a hit rate greater than or equal to 99/100. Each batch is annotated by 5 participants. We exclude a participant's responses if they fail 3 or more attention checks.

Additional details regarding the crowdsourcing annotation are presented in Appendix B.

Annotation results Since human annotators are known to interpret Likert scales differently (Cliff, 1993), we apply per-participant z-scoring to the

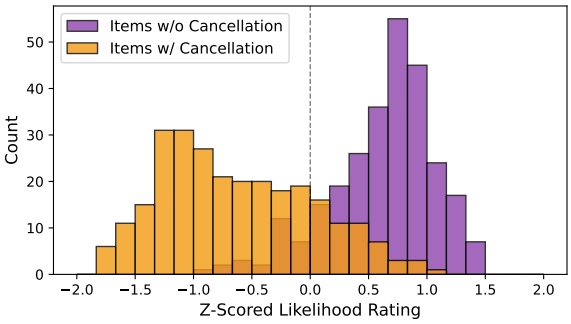

Figure 3: Histogram of average likelihood judgements z-scores from IMPLICATUREX with and without the cancelling utterance.

[4]To compute this second agreement, we exclude items where at least one of the annotators marked the implicature as implausible.

Likert scale likelihood judgements. We present z-scored likelihood judgements averaged per item in Figure 3. Figure 3 shows a clear belief update given cancelling utterances: items without a cancelling utterance tend to have a positive z-score while items with a cancelling utterance tend to have a negative z-score. Additional results related to our crowdsourcing annotation can be found in Appendix B.

5 Task Definitions and Setup

We leverage the IMPLICATUREX dataset to evaluate LLM understanding of implicature, implicature cancellation, and of belief updates induced by these two pragmatic phenomena. To do so, we use a multiple-choice-question prompt template to estimate specific values of the common ground function $G : \mathcal{B} \times \mathcal{C} \times \mathcal{U}^* \rightarrow [0, 1]$. Given a belief b and a conversational history $\mathbf{u} \in \mathcal{U}^*$, we operationalize $G(b \mid c, \mathbf{u})$ via an LLM $\mathcal{M}$'s next-token probability distribution over its token space $\mathcal{V}$, such that: $G(b \mid c, \mathbf{u}) \approx P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \mathbf{u}))$. Here, $\mathtt{t}_b(c, \mathbf{u})$ is a prompt containing the verbalization of b, the context c, the conversational history $\mathbf{u}$, and a question about the truthfulness of b in a multiple-choice format using "True" and "False" as options (Example prompt in Figures 17 and 18 in Appendix D). To reduce positional bias, we follow Shi et al. (2025) and shuffle the order of the option choices, creating two copies for each prompt.

Implicature recognition We evaluate an LLM $\mathcal{M}$'s *implicature recognition accuracy* by computing the proportion of items for which $\mathcal{M}$ favors the implicated belief over its negation, i.e.,

$$P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \mathbf{u})) > P_{\mathcal{M}}(\neg b \mid \mathtt{t}_b(c, \mathbf{u})).$$

Cancellation recognition We compute $\mathcal{M}$'s *cancellation recognition accuracy* as the proportion of items for which $\mathcal{M}$ decreases its probability of b upon observing $u^{\times}$, i.e.,

$$P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle)) > P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\times} \rangle)).$$

Belief update Lastly, we define a model's *belief update accuracy* as the proportion of items for which both

$$P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle)) > P_{\mathcal{M}}(\neg b \mid \mathtt{t}_b(c, \langle u \rangle)) \quad \text{and}$$
$$P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle)) > P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\times} \rangle))$$

hold, i.e., $\mathcal{M}$ both recognizes the implicature triggered by u and its subsequent cancellation by $u^{\times}$.

We take these accuracy definitions to be necessary (but not necessarily sufficient) conditions of an understanding of implicature, implicature cancellation, and of the belief updates these phenomena, taken together, entail.

Models We conduct our experiments on both open and closed instruction-tuned LLMs. The open-weight LLMs we evaluate include Gemma 3 (4B, 12B, 27B; Gemma-Team et al., 2025), Llama 3.1 (8B, 70B), Llama 3.2 (3B), Llama 3.3 (70B; Grattafiori et al., 2024), Qwen 2.5 (3B, 7B, 14B, 32B, 72B; Qwen-Team et al., 2025) and Qwen 3 (0.6B, 1.7B, 4B, 8B, 14B, 32B; Yang et al., 2025). For Qwen 3, we also experiment with the model's reasoning functionality. The closed-source LLMs we evaluate include GPT-5.2 and GPT-5.4 with and without their reasoning functionality.[5]

Computing $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \mathbf{u}))$ For open-weight LLMs, $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \mathbf{u}))$ is computed by extracting the logits corresponding to each option and renormalizing them, averaging over both order shuffles. For closed-source models, $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \mathbf{u}))$ is computed by sampling and using the resulting relative frequencies, sampling five times for each order shuffle. In our experimental setup, $\mathbf{u}$ is instantiated as a specific utterance sequence, e.g., $\langle u \rangle$ to estimate $G(b \mid c, \langle u \rangle)$, or $\langle u, u^{\times} \rangle$ to estimate $G(b \mid c, \langle u, u^{\times} \rangle)$, allowing us to evaluate how each additional utterance affects the common ground probability assigned to b.

6 Implicature Recognition Experiment

Here, we present the experiments and results for the implicature recognition task.

6.1 Human Topline and Prior Common Ground Control

We use our crowdsourcing annotation results from Section 4.3 to compute an implicature recognition *human accuracy topline*. To do so, we compute the proportion of items for which the z-scored Likert ratings averaged across participants are above 0.

We run an additional experiment estimating the *prior common ground* $G(b \mid \emptyset, \emptyset)$ for every item, i.e., $P_{\mathcal{M}}(b \mid \mathtt{t}_b(\emptyset, \emptyset))$, by removing c and $\mathbf{u}$ from the prompt and rewording the question to ask the model about its prior knowledge. We run this experiment to isolate the effect of the LLM's prior

[5] All open-weight models are downloaded from HuggingFace and all closed-source models are accessed via OpenAI's API. See Table 5 for full model names.

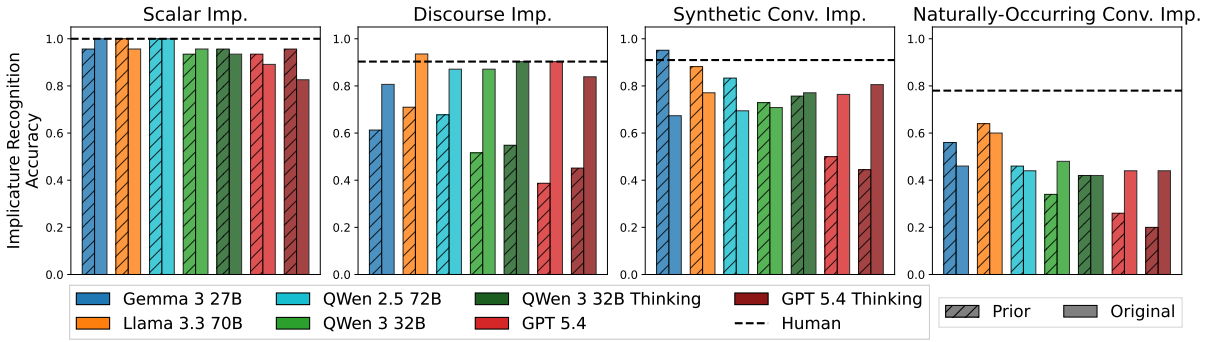

Figure 4: Implicature recognition accuracies of the largest and most modern models for each class of models tested. Original denotes the accuracy using the $\mathtt{t}_b(c, \mathbf{u})$) prompt and prior denotes the accuracy using the $\mathtt{t}_b(\emptyset, \emptyset)$) prompt.

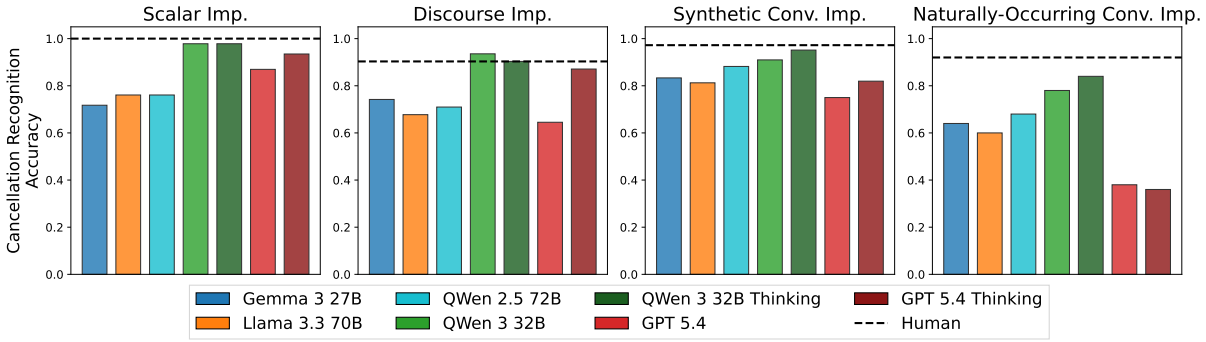

Figure 5: Cancellation recognition accuracies of the largest and most modern models for each class of models tested.

belief on implicature recognition. For instance, in the case of scalar implicatures, this prior control may reveal to what extent the "not all" pragmatic interpretation of "some" has been "memorized".

6.2 Results

The results for the implicature recognition task experiment as well as the prior belief control experiment are shown in Figure 4. Overall, we see that models are able to recognize scalar implicatures and discourse implicatures similarly to humans with Gemma 3 27B matching human scalar implicature recognition accuracy of 1.0 and Llama 3.3 70B outperforming the human discourse implicature recognition accuracy of 0.90 by 0.04. On the other hand, models tend to perform worse on conversational implicatures. Compared to the human accuracies of 0.91 and 0.78 on synthetic and naturally-occurring implicatures, the best-performing models achieve implicature recognition accuracies of 0.81 (GPT-5.4 Thinking) and 0.60 (Llama 3.3 70B), respectively. Furthermore, all of the models except for Llama 3.3 70B perform worse than random at identifying naturally-occurring implicatures, demonstrating that understanding unspoken beliefs remains a challenging task for LLMs. Full results are shown in Table 6 (Appendix F).

LLMs have strong priors for certain types of implicatures In addition, our prior belief experiment reveals that models have a moderately strong prior for synthetic conversational implicatures and an extremely strong prior for $\langle$*some*, *all*$\rangle$ scalar implicatures. For instance, both Llama 3.3 70B and Qwen 2.5 72B are able to achieve perfect implicature recognition accuracy on scalar implicature items *without* having access to their corresponding utterances. This result suggests that the performance of models on implicature recognition may not just stem from an understanding of implicated beliefs, but also from the prior knowledge these models may have about these implicated beliefs. We leave a thorough investigation of the interaction between LLMs' pragmatic understanding and their prior knowledge to future work.

7 Cancellation Recognition and Belief Update Experiment

Next, we present the experiments and results for the cancellation recognition and belief update tasks.

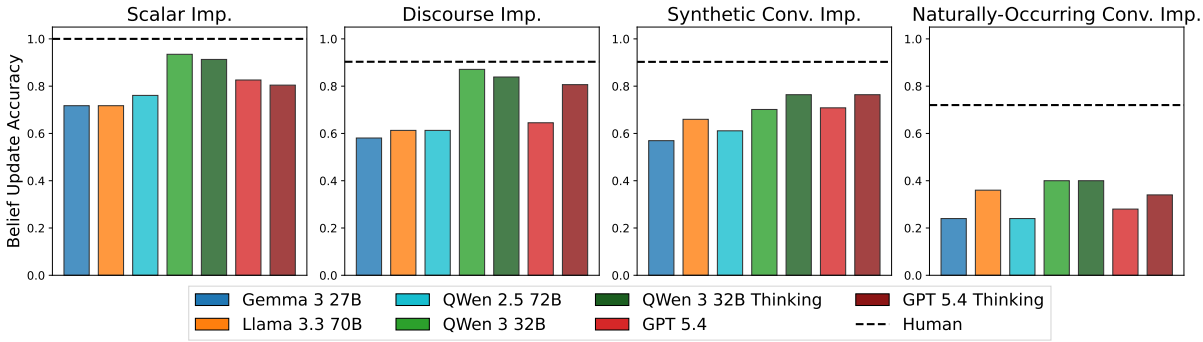

Figure 6: Belief update accuracies of the largest and most modern models for each class of models tested.

7.1 Human Topline and Controls

We use our crowdsourcing annotation results from Section 4.3 to compute cancellation recognition and belief update *human accuracy topline*. For the cancellation accuracy, we compute the proportion of items for which the average per-item z-scored Likert rating decreases when the cancelling utterance is introduced. For the *belief update accuracy*, we compute the proportion of items for which the average per-item z-scored Likert rating is above 0 given the triggering utterance and subsequently decreases given the cancelling utterance.

Form control We evaluate the extent to which the form of the cancellation affects an LLM's success in performing cancellation recognition. To do so, we create a new set of stimuli, IMPLICATURE$^{\perp}$, in which cancelling utterances have the form $u^{\perp}$ ="$d + \neg b$" where d is a discourse marker (e.g., *in fact*, *actually*, etc.) and $\neg b$ denotes an explicit negation of the implicated belief b triggered by u.

Update type control We also evaluate the extent to which LLM belief update accuracy is affected by the type of follow-up utterance, i.e., whether the utterance that follows u cancels, strengthens, or leaves b unchanged. To do so, we create two new sets of stimuli: (1) IMPLICATURE^{+}, in which cancelling utterances $u^{\times}$ are replaced with strengthening utterances u^{+}, i.e., utterances that strengthen b by providing information reinforcing its plausibility, and (2) IMPLICATURE$^{\approx}$, in which $u^{\times}$ is replaced with a randomly sampled utterance $u^{\approx}$ of the same stimuli type that leaves the belief b unchanged. We evaluate a model's ability to perform belief strengthening by computing the proportion of IMPLICATURE^{+} items for which $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{+}\rangle)) > P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u\rangle))$, and its ability to leave its belief unchanged by computing the proportion of IMPLICATURE$^{\approx}$ items for which

$$P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\approx}\rangle)) > P_{\mathcal{M}}(\neg b \mid \mathtt{t}_b(c, \langle u, u^{\approx}\rangle))$$
$$\iff$$
$$P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u\rangle)) > P_{\mathcal{M}}(\neg b \mid \mathtt{t}_b(c, \langle u\rangle)).$$

We provide examples for the different experimental controls in Table 2. Additional details regarding the composition of these control datasets can be found in Appendix E.

Control Type	Follow-Up Utterance u
Original	$u^{\times}$: "Though they must be done by now."
Strengthening	u^{+}: "Maybe go over and ask them?"
Unchanging	$u^{\approx}$: "Thankfully, I submitted before my laptop broke."
Negation	$u^{\perp}$: "Though, to be clear, I'm not saying that you could help by doing X for C."

Table 2: Control conditions for the belief update task, with fixed context c = "Do you need help with anything?", utterance u = "I think C was looking for someone to do X." and belief b = "You can help by doing X for C."

7.2 Results

We present model performance on cancellation recognition in Figure 5. Overall, LLMs perform cancellation recognition at levels close to human performance, with some models slightly surpassing human cancellation recognition accuracy—e.g., Qwen 3 32B achieves a cancellation accuracy of 0.94 on discourse implicatures whereas humans achieve an accuracy of 0.90. However, models perform substantially worse when presented with naturally-occurring cancellations. In this setting, the best model, Qwen 3 32B Thinking, achieves

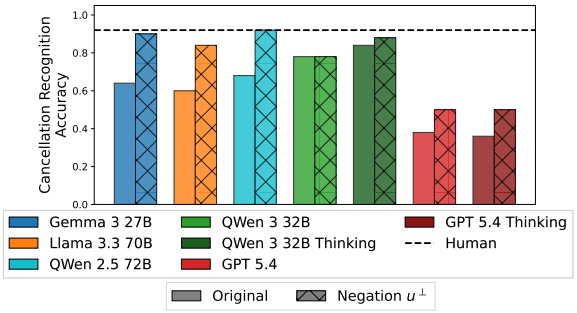

Figure 7: Cancellation recognition accuracy of models on the naturally-occurring conversational implicatures using the original cancelling utterances ($u^{\times}$) and the cancelling utterances with explicit negation ($u^{\perp}$).

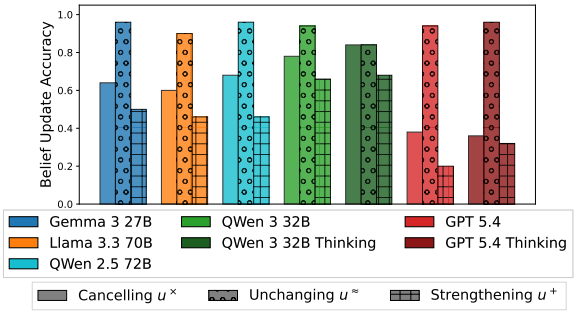

Figure 8: Update recognition accuracy of models on the naturally-occurring conversational implicatures using follow-up utterances of different types: cancelling ($u^{\times}$), unchanging ($u^{\approx}$), and strengthening (u^{+}).

a cancellation recognition accuracy of 0.84, compared to human performance of 0.92.

We also report results for belief updating in Figure 6. When evaluating the joint tasks of implicature and cancellation recognition, model performance falls sharply, especially for conversational implicatures. For instance, the model with highest belief update accuracy, Qwen 3 32B Thinking, achieves belief update accuracies of 0.76 and 0.40 on synthetic and naturally-occurring conversational implicatures, respectively, whereas humans achieve 0.90 and 0.72. Full results can be found in Tables 7 and 8 in Appendix F.

Explicit negation does not lead to perfect LLM cancellation recognition We compare model cancellation recognition accuracy when using the original IMPLICATUREX stimuli and the IMPLICATURE$^{\perp}$ variant, which contain explicitly negating cancelling utterances. We show the cancellation recognition accuracies on the naturally-occurring conversational implicatures in Figure 7 and on the full set of implicature types in Figure 19 in Appendix F. While the implicature cancellations with explicit negations lead to higher cancellation recognition accuracies, we find that in most cases — aside from scalar implicatures — the cancellation recognition accuracy falls short of its theoretical upper bound of one. This suggests that factors beyond pragmatic competence, such as difficulties in negation interpretation, may also contribute to observed cancellation recognition errors.

LLMs are better at maintaining existing beliefs than changing them We show the belief update accuracies using the cancelling, strengthening and unchanging utterances on naturally-occurring conversational implicatures in Figure 8 and on the entire set of implicature types in Figure 20 in Appendix F. Overall, we observe that LLMs struggle to update their beliefs both with strengthening and cancelling utterances. However, they have relatively little difficulty maintaining their beliefs when presented with an irrelevant follow-up utterance. For instance, while GPT-5.4 maintains its belief with an accuracy of 0.96 on naturally-occurring conversational implicatures, its update accuracies when presented with strengthening and cancelling utterances are both at 0.32. This result suggests that current LLMs are better at maintaining beliefs than at changing them.

8 Conclusion

In conclusion, in this paper we study LLM understanding of communicative beliefs and belief updates via implicature and cancellation recognition. To this end, we develop the first implicature cancellation dataset, IMPLICATUREX, and show that LLMs fall short of a human understanding of unspoken beliefs and belief updates. Control experiments reveal key weaknesses; for instance, a reliance on prior biases, an inability to reconcile explicit updates and a dependence on update type. Our results highlight a critical gap between current LLM capabilities and human-level pragmatic reasoning which calls for further research into how models manage evolving communicative beliefs.

Limitations

We identify three main limitations with our work. Firstly, although we thoroughly reviewed whether the implicatures and corresponding cancellations are, in fact, annotated as such by both experts and lay annotators, whether or not an utterance cancels a prior statement remains open to interpretation. With five annotations per stimulus, we consider our

results to be reliable, but we encourage future work to further explore this, particularly focusing on the examples for which our annotators demonstrate disagreement. Expanding the number of annotations could clarify whether or not such disagreement is due to legitimate stimulus ambiguity or instead to noise in our data.

Secondly, we identified that models' implicature recognition performance is strongly affected by their prior beliefs: without knowing utterance u for $\langle$*some*, *all*$\rangle$ statements, they will promptly score corresponding belief b highly. As a result, we cannot say with certainty that their implicature recognition accuracy is not affected by this; we leave disentangling scalar implicature reasoning from prior beliefs to future work.

Lastly, we elicited LLM judgements by extracting probabilities from their output distributions. Alternatively, one could provide the LLM with utterances and cancelling utterances, and ask them to further expand the provided discourse. Consistency vs. inconsistency with the beliefs implicitly conveyed in such generated text could reveal whether the LLMs *really* manage to remain consistent with those beliefs. We did not opt for this type of evaluation due to its lack of scalability, i.e., high-quality human annotations of generated text would be needed to assess the nuances behind the LLM responses.

Acknowledgements

The authors would like to thank the reviewers for their valuable comments. We would also like to thank Gaurav Kamath and Austin Kraft for their helpful comments on an earlier version of this paper. This work was supported by the Fonds de Recherche du Québec – Nature et Technologies (FRQNT), the Natural Sciences and Engineering Research Council of Canada (NSERC), the IVADO Postdoctoral Research Funding Program, and the Canada CIFAR AI Chair Program. We acknowledge material support from NVIDIA Corporation in the form of computational resources provided to Mila.

References

James F Allen and C Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial intelligence*, 15(3):143–178.

Catherine Anderson. 2018. *Essentials of linguistics*. McMaster University.

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1173–1182.

Luciana Benotti. 2010. *Implicature as an Interactive Process*. Ph.D. thesis, Université Henri Poincaré-Nancy I.

Richard Breheny, Napoleon Katsos, and John Williams. 2006. Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition*, 100(3):434–463.

Ted Byrt, Janet Bishop, and John B Carlin. 1993. Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on Socially Shared Cognition*.

Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1):62–81.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Norman Cliff. 1993. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological bulletin*, 114(3):494.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Kees van Deemter, Albert Gatt, Ielka van der Sluis, and Richard Power. 2012. Generation of referring expressions: Assessing the incremental algorithm. *Cognitive science*, 36(5):799–836.

Judith Degen. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1.

Anita Fetzer. 2018. The linguistic realization of contrastive discourse relations in context: Contextualization and discourse common ground. *Modélisation et utilisation du contexte*.

Gemma-Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Elizabeth Jasmi George and Radhika Mamidi. 2020. Conversational implicatures in english dialogue: Annotated dataset. *Procedia Computer Science*, 171:2316–2323.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing - Volume 1*, ICASSP'92, page 517–520, USA. IEEE Computer Society.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

H Paul Grice. 1957. Meaning. *The philosophical review*, 66(3):377–388.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Daniel J Grodner, Natalie M Klein, Kathleen M Carbary, and Michael K Tanenhaus. 2010. "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1):42–55.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Irene Roswitha Heim. 1982. *The semantics of definite and indefinite noun phrases*. University of Massachusetts Amherst.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213.

Yi Ting Huang and Jesse Snedeker. 2018. Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive psychology*, 102:105–126.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (Comet-)Atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6384–6392.

Ellen A Isaacs and Herbert H Clark. 1987. References in conversation between experts and novices. *Journal of experimental psychology: general*, 116(1):26.

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.

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. 1981. A theory of truth and semantic representation. *Proceedings of the Third Amsterdam Colloquium*, pages 277–322.

Anita Körner, Antonia Tolzin, Andreas Janson, Jan Marco Leimeister, and Ralf Rummer. 2025. Common ground improves learning with conversational agents. *Behaviour & Information Technology*, pages 1–17.

Alex Lascarides and Nicholas Asher. 1991. Discourse relations and defeasible knowledge. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 55–62.

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

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425.

Ira A Noveck. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2):165–188.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. Web Download. LDC2019T05.

Qwen-Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rachel Rudinger, Vered Shwartz, Jena D Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. *Advances in Neural Information Processing Systems*, 36:20827–20905.

Reinhard Selten and Massimo Warglien. 2007. The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, 104(18):7361–7366.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

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

Robert Stalnaker. 1998. On the representation of context. *Journal of logic, language and information*, 7(1):3–19.

Elizabeth S Veinott, Judith Olson, Gary M Olson, and Xiaolan Fu. 1999. Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 302–309.

Alexander C Wilson and Dorothy VM Bishop. 2021. "Second guessing yourself all the time about what they really mean...": Cognitive differences between autistic and non-autistic adults in understanding implied meaning. *Autism Research*, 14(1):93–101.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xiao Yang, Utako Minai, and Robert Fiorentino. 2018. Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in psychology*, 9:1720.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature - A case study with a chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1270–1285.

A Expert Annotation

Additional Annotation Details The expert annotation interface is illustrated in Figures 9, 10, 11, and 12. The expert annotators were first presented with a landing page containing general instructions along with a reminder of the notions of implicature and implicature cancellation (Figure 9). The annotation task was organized into several sub-batches based on stimulus type and the stimuli were presented in order of increasing cognitive load: scalar implicatures, synthetic conversational implicatures, naturally-occurring conversational implicatures, and discourse implicatures.

Before each stimulus type, annotators are given stimulus-type-specific instructions (Figure 10) and representative examples to familiarize them with the stimuli and task format (Figure 11). For each annotation item (Figure 12), annotators are asked to judge (i) whether the proposed interpretation constitutes a plausible implicature and (ii) whether the provided cancellation is correct. If the implicature is judged implausible, annotators are asked to provide an alternative implicature. Similarly, annotators are asked to provide an alternative cancellation when the given cancellation is incorrect or when the implicature itself is deemed implausible. We elicit such alternatives for all data types except for our naturally-occurring conversational implicatures.

Expert annotators were paid at their affiliated university's teaching assistant hourly rate. This study received approval from the Research Ethics Board at McGill University (REB File #: 21-06-019).

Additional Annotation Results We provide confusion matrices for our first round of expert annotation during which two expert annotators annotated the same batch of 53 items. We report the two-by-two confusion matrix for the implicature plausibility responses in Table 3 and for the implicature cancellation correctness in Table 4. For the cancellation correctness confusion matrix, we only consider the items for which there was agreement in terms of the implicature's plausibility.

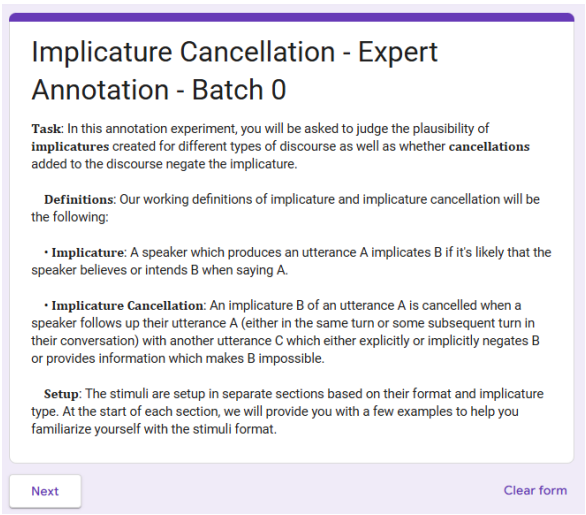

Figure 9: Landing page shown to expert annotators at the beginning of the annotation task. The landing page includes a reminder of the definitions of implicature and implicature cancellation.

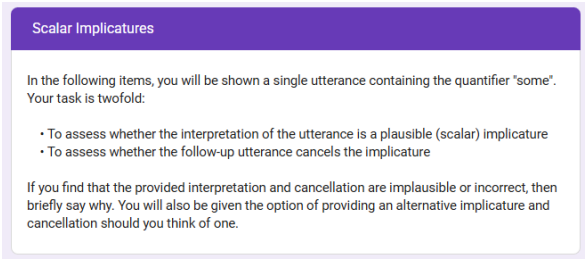

Figure 10: Instructions provided to expert annotators for the scalar implicature items.

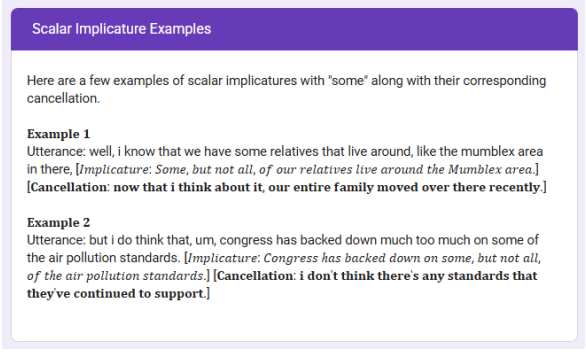

Figure 11: Example scalar implicatures shown to expert annotators prior to the annotation of the scalar implicature items. Examples like these are shown before each implicature type.

A_1 \ A_2	Plausible	Implausible
Plausible	43	4
Implausible	4	3

Table 3: Confusion matrix of annotators A_1 and A_2 on the implicature plausibility annotation task.

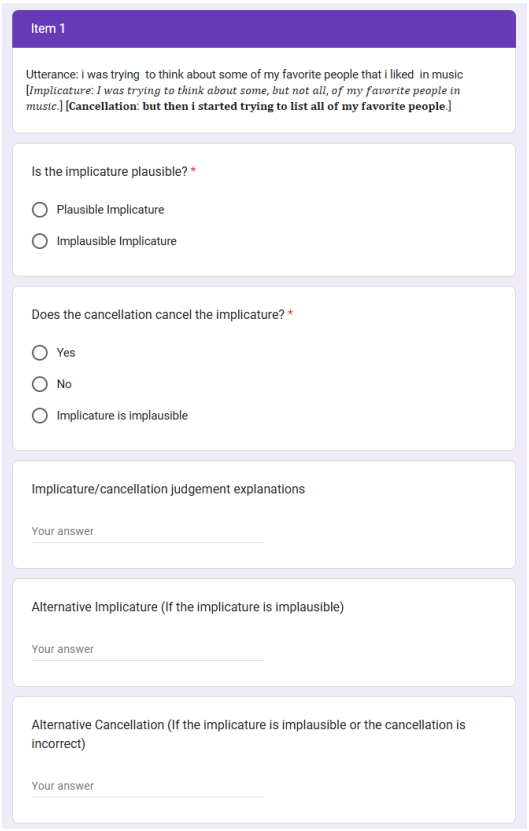

Figure 12: One of the scalar implicature items to annotate. The context, utterance, candidate implicature, and candidate implicature cancellation are shown. The expert annotator decides on the plausibility of the candidate implicature and on the correctness of the implicature cancellation.

B Crowdsourcing Annotation

Additional Crowdsourcing Details The crowdsourcing annotation interface is shown in Figures 13 and 14. Each annotator began the annotation from the same landing page (Figure 13) which contains general information about the study and its structure. The crowdworkers were then shown six example stimuli (Figure 14): one scalar implicature, two discourse implicatures, two synthetic conversational implicatures and one naturally-occurring conversational implicature. They were asked to

A_1 \ A_2	Correct	Incorrect
Correct	35	3
Incorrect	3	2

Table 4: Confusion matrix of annotators A_1 and A_2 on the implicature cancellation correctness annotation task.

try again if they provided an unlikely Likert score rating (e.g., 1, 2, or 3) for an item which did not contain a cancellation or if they provided a likely Likert score rating (e.g., 5, 6, or 7) for an item which did contain a cancellation.

After completing the training examples, the crowdworkers were presented with either 40 or 42 items depending on the batch assigned to them. For the 16 batches containing 40 items, 10 were attention checks (five items which should be rated likely and five items which should be rated unlikely), 15 were items without the cancelling utterance and 15 were items with the cancelling utterance. For the two batches of 42 items, 10 were attention checks, 16 were items without the cancelling utterance and 16 were items with the cancelling utterance. The 10 attention checks were the same for all batches. We ensured that no batch contained the same implicature more than once.

All crowdworkers were paid 15 USD/hour via the Prolific platform[6]. Data ingestion was done through Proliferate[7]. This study received approval from the Research Ethics Board at McGill University (REB File #: 21-06-019).

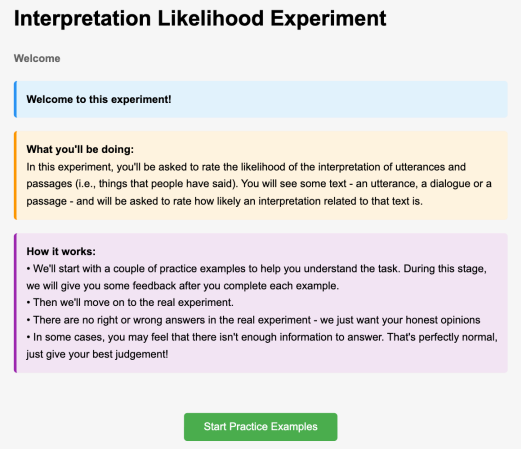

Figure 13: Landing page of the crowdsourcing experiment, outlining the instructions and interface for participants.

Additional Crowdsourcing Results We provide the per-implicature-type disaggregated histograms of the average human likelihood z-scores in Figure 15. Overall, the disaggregated histograms reveal that, across all implicature types, cancelling utterances weaken the z-scored human likelihoods of implicatures.

[6] https://www.prolific.com/
[7] https://docs.proliferate.alps.science/

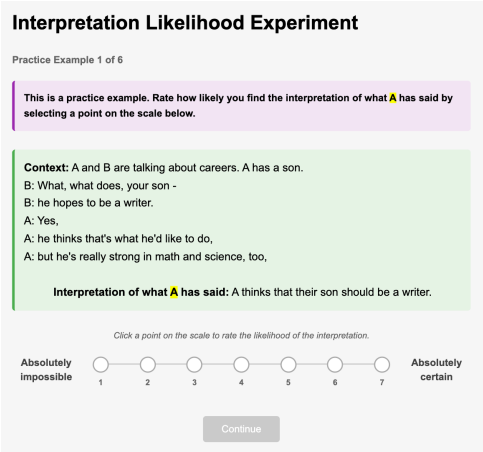

Figure 14: Example stimulus from the experiment, demonstrating the Likert-scale interface used to collect interpretation likelihood ratings.

We also provide disaggregated scatter plots with the human likelihood z-scores with and without the cancelling utterance on the y-axis and x-axis respectively in Figure 16. We also report the Pearson correlation r for each of the implicature types in their corresponding subplot. Overall, we find no statistically significant correlation between an implicature's likelihood with and without a cancelling utterance for scalar implicature, discourse implicatures and synthetic conversational implicatures. However, we do find a statistically significant correlation for naturally-occurring conversational implicatures ($r = 0.63$, $p < 0.001$) which suggests that "stronger" naturally-occurring implicatures are more "difficult" to cancel.

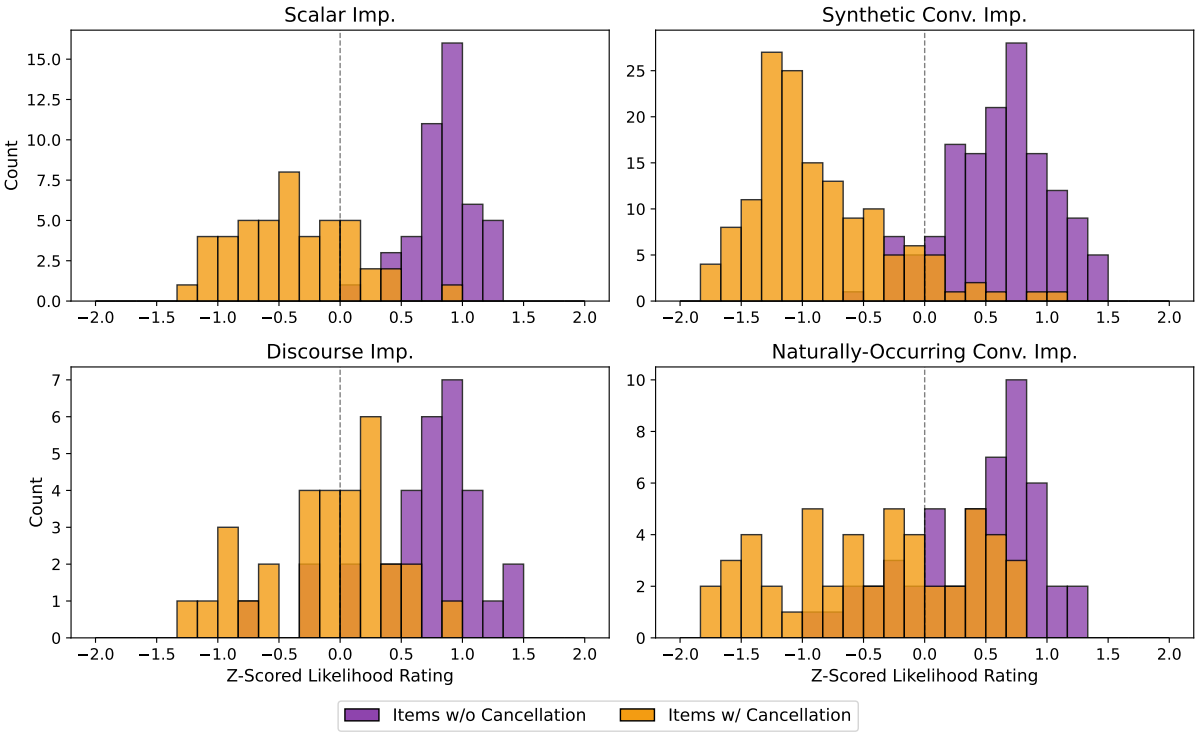

Figure 15: Disaggregated histograms of the average human likelihood z-scores with and without the cancelling utterance.

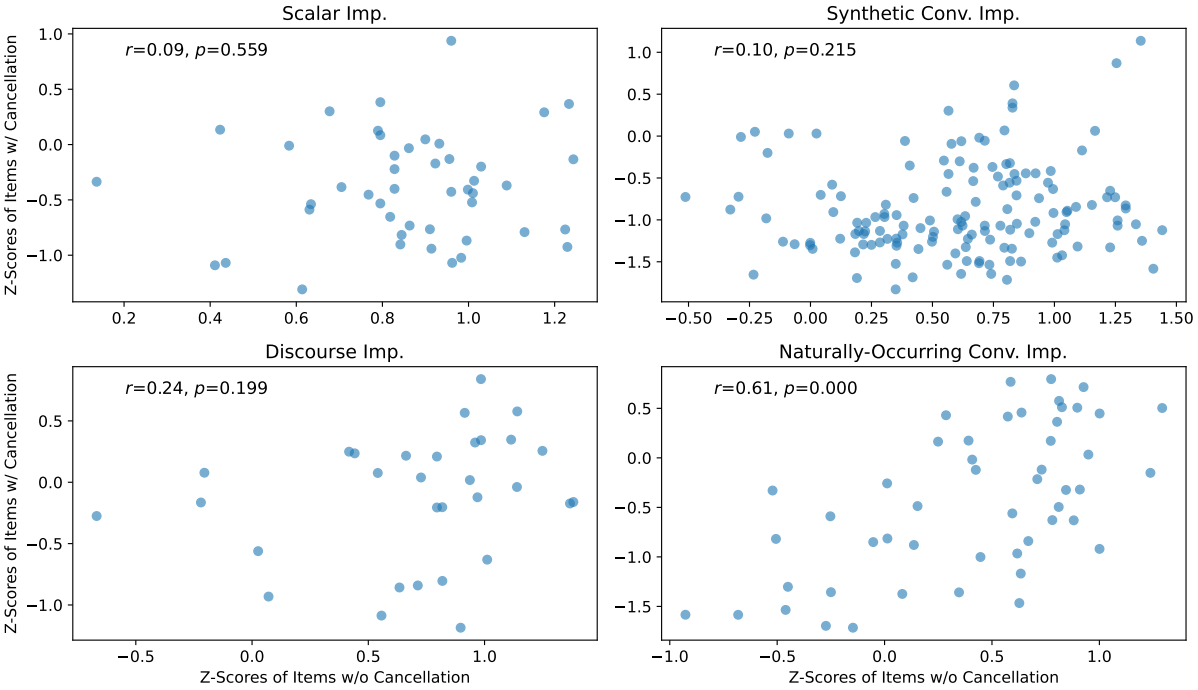

Figure 16: Scatter plots of the average human likelihood z-scores with and without the cancelling utterance. The Pearson correlation coefficient r and p-value p are reported at the top left.

C Model Details

Model Name	HuggingFace or OpenAI Identifier
Gemma 3 (4B)	`google/gemma-3-4b-it`
Gemma 3 (12B)	`google/gemma-3-12b-it`
Gemma 3 (27B)	`google/gemma-3-27b-it`
Llama 3.1 (8B)	`meta-llama/Llama-3.1-8B-Instruct`
Llama 3.1 (70B)	`meta-llama/Llama-3.1-70B-Instruct`
Llama 3.2 (3B)	`meta-llama/Llama-3.2-3B-Instruct`
Llama 3.3 (70B)	`meta-llama/Llama-3.3-70B-Instruct`
Qwen 2.5 (3B)	`Qwen/Qwen2.5-3B-Instruct`
Qwen 2.5 (7B)	`Qwen/Qwen2.5-7B-Instruct`
Qwen 2.5 (14B)	`Qwen/Qwen2.5-14B-Instruct`
Qwen 2.5 (32B)	`Qwen/Qwen2.5-32B-Instruct`
Qwen 2.5 (72B)	`Qwen/Qwen2.5-72B-Instruct`
Qwen 3 (0.6B)	`Qwen/Qwen3-0.6B`
Qwen 3 (1.7B)	`Qwen/Qwen3-1.7B`
Qwen 3 (4B)	`Qwen/Qwen3-4B`
Qwen 3 (8B)	`Qwen/Qwen3-8B`
Qwen 3 (14B)	`Qwen/Qwen3-14B`
Qwen 3 (32B)	`Qwen/Qwen3-32B`
Qwen 3 Thinking (0.6B)	`Qwen/Qwen3-0.6B` & `# reasoning tokens: 512`
Qwen 3 Thinking (1.7B)	`Qwen/Qwen3-1.7B` & `# reasoning tokens: 512`
Qwen 3 Thinking (4B)	`Qwen/Qwen3-4B` & `# reasoning tokens: 512`
Qwen 3 Thinking (8B)	`Qwen/Qwen3-8B` & `# reasoning tokens: 512`
Qwen 3 Thinking (14B)	`Qwen/Qwen3-14B` & `# reasoning tokens: 512`
Qwen 3 Thinking (32B)	`Qwen/Qwen3-32B` & `# reasoning tokens: 512`
GPT-5.2	`gpt-5.2-2025-12-11`
GPT-5.2 Thinking	`gpt-5.2-2025-12-11` & `reasoning effort: medium`
GPT-5.4	`gpt-5.4-2026-03-05`
GPT-5.4 Thinking	`gpt-5.4-2026-03-05` & `reasoning effort: medium`

Table 5: Model names as referenced in the paper and their corresponding HuggingFace or OpenAI identifiers. We also provide the thinking budget for the models that leverage their reasoning capabilities.

D Prompt Templates

Model Prompt

SYSTEM PROMPT

You are an expert in pragmatic inference and in identifying the intended meaning of utterances.

PROMPT TEMPLATE

You will be given an utterance as well as a potential interpretation for this utterance. Your task is to decide whether the interpretation provided for the utterance is true or false. You will be given a numbered list containing true and false. Read the utterance and the potential interpretation and pick the number corresponding to whether the interpretation is true or false. `{scenario}` Interpretation: `{implicature}` `{question}` `{options}` Answer:

Figure 17: Model prompt used for pragmatic inference evaluation. Variables `{scenario}`, `{implicature}`, `{question}`, and `{options}` are filled per instance.

Example Instance

u	and some places, they, they, they really nail them for tax.
b	Some, but not all, places nail them for tax.
$u^{\times}$	i think they all do that now

FILLED PROMPT (WITHOUT $u^{\times}$)

You will be given an utterance as well as a potential interpretation for this utterance. Your task is to decide whether the interpretation provided for the utterance is true or false. You will be given a numbered list containing true and false. Read the utterance and the potential interpretation and pick the number corresponding to whether the interpretation is true or false.

Utterance: and some places, they, they, they really nail them for tax.

Interpretation: Some, but not all, places nail them for tax.

Is the interpretation of the previous utterance true or false?
1: True
2: False

Answer:

FILLED PROMPT (WITH $u^{\times}$)

You will be given an utterance as well as a potential interpretation for this utterance. Your task is to decide whether the interpretation provided for the utterance is true or false. You will be given a numbered list containing true and false. Read the utterance and the potential interpretation and pick the number corresponding to whether the interpretation is true or false.

Utterance: and some places, they, they, they really nail them for tax. i think they all do that now

Interpretation: Some, but not all, places nail them for tax.

Is the interpretation of the previous utterance true or false?
1: True
2: False

Answer:

Figure 18: Example instance for the prompt in Figure 17, shown without (top) and with (bottom) the cancellation $u^{\times}$ appended to the utterance in `{scenario}`.

E Control Datasets

To conduct our control experiments in Section 7, we created three additional datasets: IMPLICATURE$^{\perp}$, IMPLICATURE^{+} and IMPLICATURE$^{\approx}$. All of these datasets have the same size as the original IMPLICATUREX dataset. They share the same context, triggering utterance and implicature as the original dataset but differ in their follow-up utterance (i.e., the utterance which follows the cancelling utterance). We provide details regarding how their follow-up utterances were created.

The IMPLICATURE$^{\perp}$ follow-up utterance, $u^{\perp}$, is a cancelling utterance with a discourse marker and an explicit negation of the implicature. We first determined which discourse markers best preserved coherence in each of the dataset splits and then applied formulaic negations to the implicature. For instance, to explicitly negate a scalar implicature like "...some of them showed up." we use discourse markers like "in fact", "actually" or "as a matter of fact" followed by the negation of the scalar implicature "all of them showed up." On the other hand, conversational implicatures suffer in coherence when negated using the same discourse markers. Thus, in this case, we use the discourse marker "Though, I don't mean to imply that" construction. For instance, "I'm more behind than you. Though, I don't mean to imply that I can't help you with this problem."

The IMPLICATURE^{+} follow-up utterance, u^{+}, is a strengthening utterance: an utterance designed to confirm the implicature triggered by the triggering utterance. We manually created these utterances while controlling for coherence. For instance, to strengthen a scalar implicature, we reinforce the fact that not the entire set of elements being discussed should be included e.g., "and some places, they, they, they really nail them for tax. though some get that exemption i was talking about". For the other types of implicatures, we create similar utterances which implicitly reinforce the implicature. For instance, in the case of synthetic conversational implicatures, the implicature "I cannot help you with this question." can be reinforced with the utterance "Is there no one else you can ask?"

The IMPLICATURE$^{\approx}$ follow-up utterance, $u^{\approx}$, is a neutral utterance: an utterance which is irrelevant to the triggering utterance and the implicature. The neutral utterance is selected by randomly sampling a cancelling utterance from the same corresponding dataset split.

F Additional Results

Full Results We provide full implicature recognition, cancellation recognition and belief update accuracies in Tables 6, 7 and 8 respectively. We also include the full results for our control experiments in Tables 9, 10, 11, 12. In addition, the form and update type controls on all the implicature types are plotted in Figures 19 and 20.

Does length explain implicature and cancellation recognition difficulty? To determine whether our results are confounded by length and the ability of LLMs to perform under longer prompts, we run a correlation analysis between different length variables and implicature and cancellation recognition probabilities. In particular, we compute Kendall tau's concordances between different space-separated lengths, $|\cdot| : \mathcal{V}^* \to \mathbb{N}$, and the implicature and cancellation probabilities of Llama 3.3 70B, the model which performed the best overall in this study. We provide Kendall tau's concordances τ between $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle))$ and the space-separated lengths of c, u and b in Table 13. We also provide Kendall tau's concordances τ between $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\times} \rangle))$ and the space-separated lengths of c, u, b and $u^{\times}$ in Table 14.

Our results in Tables 13 and 14 indicate weak concordances between the tested length variables and the LLM-induced probabilities $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle))$ and $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\times} \rangle))$. In addition, in most cases, the computed concordances are not statistically significant. The only statistically significant Kendall's tau values between $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle))$and the length variables are 0.27 (context length of the naturally-occurring implicatures) and 0.28 (utterance length of the scalar implicatures). The only statistically significant Kendall's tau between $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\times} \rangle))$and the length variables is 0.22 (triggering utterance length of the scalar implicatures). All other concordances are near zero and not statistically significant. While a lack of concordance is not evidence that there is no relation between length and implicature and cancellation difficulty, we believe it suggests that the difficulty of the items in IMPLICATUREX cannot be explained by length alone.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	1.000	0.871	0.708	0.700
Gemma 3 (12B)	1.000	0.903	0.722	0.560
Gemma 3 (27B)	1.000	0.806	0.674	0.460
Llama 3				
Llama 3.1 (8B)	0.913	0.677	0.229	0.160
Llama 3.1 (70B)	0.978	0.935	0.806	0.560
Llama 3.2 (3B)	0.478	0.452	0.271	0.100
Llama 3.3 (70B)	0.957	0.935	0.771	0.600
Qwen 2.5				
Qwen 2.5 (3B)	0.522	0.419	0.076	0.080
Qwen 2.5 (7B)	0.761	0.613	0.222	0.160
Qwen 2.5 (14B)	0.891	0.806	0.479	0.340
Qwen 2.5 (32B)	0.935	0.871	0.556	0.380
Qwen 2.5 (72B)	1.000	0.871	0.694	0.440
Qwen 3				
Qwen 3 (0.6B)	0.000	0.000	0.000	0.000
Qwen 3 (1.7B)	0.848	0.871	0.715	0.600
Qwen 3 (4B)	1.000	0.903	0.590	0.780
Qwen 3 (8B)	1.000	0.839	0.326	0.420
Qwen 3 (14B)	0.891	0.903	0.674	0.420
Qwen 3 (32B)	0.957	0.871	0.708	0.480
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.978	0.613	0.444	0.500
Qwen 3 (1.7B)	0.435	0.774	0.292	0.400
Qwen 3 (4B)	1.000	0.839	0.521	0.460
Qwen 3 (8B)	0.804	0.742	0.347	0.380
Qwen 3 (14B)	0.587	0.839	0.528	0.420
Qwen 3 (32B)	0.935	0.903	0.771	0.420
GPT				
GPT 5.2	0.804	0.806	0.674	0.500
GPT 5.4	0.891	0.903	0.764	0.440
GPT Thinking				
GPT 5.2	0.674	0.871	0.743	0.460
GPT 5.4	0.826	0.839	0.806	0.440
Human (avg.)	1.000	0.903	0.910	0.780

Table 6: Full implicature recognition accuracy results.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	0.565	0.710	0.826	0.720
Gemma 3 (12B)	0.326	0.548	0.743	0.560
Gemma 3 (27B)	0.717	0.742	0.833	0.640
Llama 3				
Llama 3.1 (8B)	0.870	0.839	0.944	0.660
Llama 3.1 (70B)	0.913	0.774	0.944	0.820
Llama 3.2 (3B)	0.848	0.806	0.729	0.540
Llama 3.3 (70B)	0.761	0.677	0.812	0.600
Qwen 2.5				
Qwen 2.5 (3B)	0.674	0.774	0.438	0.560
Qwen 2.5 (7B)	0.891	0.710	0.819	0.620
Qwen 2.5 (14B)	0.826	0.677	0.847	0.720
Qwen 2.5 (32B)	0.870	0.710	0.917	0.740
Qwen 2.5 (72B)	0.761	0.710	0.882	0.680
Qwen 3				
Qwen 3 (0.6B)	0.457	0.742	0.556	0.520
Qwen 3 (1.7B)	0.609	0.516	0.688	0.680
Qwen 3 (4B)	0.304	0.645	0.875	0.660
Qwen 3 (8B)	0.826	0.710	0.840	0.640
Qwen 3 (14B)	0.674	0.613	0.750	0.640
Qwen 3 (32B)	0.978	0.935	0.910	0.780
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.826	0.742	0.660	0.680
Qwen 3 (1.7B)	0.870	0.613	0.736	0.600
Qwen 3 (4B)	0.826	0.710	0.868	0.760
Qwen 3 (8B)	0.957	0.871	0.938	0.740
Qwen 3 (14B)	0.978	0.871	0.931	0.880
Qwen 3 (32B)	0.978	0.903	0.951	0.840
GPT				
GPT 5.2	0.891	0.613	0.715	0.320
GPT 5.4	0.870	0.645	0.750	0.380
GPT Thinking				
GPT 5.2	0.913	0.871	0.785	0.460
GPT 5.4	0.935	0.871	0.819	0.360
Human (avg.)	1.000	0.903	0.972	0.920

Table 7: Full cancellation recognition accuracy results.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	0.565	0.645	0.590	0.540
Gemma 3 (12B)	0.326	0.484	0.576	0.280
Gemma 3 (27B)	0.717	0.581	0.569	0.240
Llama 3				
Llama 3.1 (8B)	0.804	0.645	0.222	0.160
Llama 3.1 (70B)	0.891	0.742	0.792	0.460
Llama 3.2 (3B)	0.457	0.452	0.271	0.100
Llama 3.3 (70B)	0.717	0.613	0.660	0.360
Qwen 2.5				
Qwen 2.5 (3B)	0.370	0.387	0.069	0.040
Qwen 2.5 (7B)	0.674	0.484	0.215	0.140
Qwen 2.5 (14B)	0.717	0.484	0.438	0.200
Qwen 2.5 (32B)	0.804	0.581	0.521	0.280
Qwen 2.5 (72B)	0.761	0.613	0.611	0.240
Qwen 3				
Qwen 3 (0.6B)	0.000	0.000	0.000	0.000
Qwen 3 (1.7B)	0.565	0.484	0.535	0.440
Qwen 3 (4B)	0.304	0.581	0.542	0.480
Qwen 3 (8B)	0.826	0.613	0.292	0.240
Qwen 3 (14B)	0.609	0.548	0.604	0.340
Qwen 3 (32B)	0.935	0.871	0.701	0.400
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.804	0.581	0.403	0.380
Qwen 3 (1.7B)	0.413	0.516	0.257	0.320
Qwen 3 (4B)	0.826	0.613	0.479	0.360
Qwen 3 (8B)	0.761	0.677	0.340	0.280
Qwen 3 (14B)	0.587	0.774	0.507	0.380
Qwen 3 (32B)	0.913	0.839	0.764	0.400
GPT				
GPT 5.2	0.804	0.484	0.646	0.300
GPT 5.4	0.826	0.645	0.708	0.280
GPT Thinking				
GPT 5.2	0.652	0.806	0.715	0.400
GPT 5.4	0.804	0.806	0.764	0.340
Human (avg.)	1.000	0.903	0.903	0.720

Table 8: Full belief update accuracy results.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	1.000	0.774	1.000	0.920
Gemma 3 (12B)	1.000	0.806	0.965	0.660
Gemma 3 (27B)	0.957	0.613	0.951	0.560
Llama 3				
Llama 3.1 (8B)	1.000	0.677	0.854	0.620
Llama 3.1 (70B)	0.978	0.710	0.868	0.580
Llama 3.2 (3B)	1.000	0.806	0.979	0.820
Llama 3.3 (70B)	1.000	0.710	0.882	0.640
Qwen 2.5				
Qwen 2.5 (3B)	0.087	0.032	0.021	0.020
Qwen 2.5 (7B)	0.957	0.194	0.465	0.200
Qwen 2.5 (14B)	0.891	0.290	0.167	0.160
Qwen 2.5 (32B)	0.957	0.484	0.806	0.420
Qwen 2.5 (72B)	1.000	0.677	0.833	0.460
Qwen 3				
Qwen 3 (0.6B)	0.000	0.000	0.000	0.000
Qwen 3 (1.7B)	1.000	0.742	0.910	0.620
Qwen 3 (4B)	1.000	0.710	0.910	0.580
Qwen 3 (8B)	0.978	0.452	0.715	0.380
Qwen 3 (14B)	1.000	0.323	0.549	0.260
Qwen 3 (32B)	0.935	0.516	0.729	0.340
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.913	0.355	0.611	0.380
Qwen 3 (1.7B)	0.500	0.323	0.292	0.100
Qwen 3 (4B)	0.935	0.419	0.396	0.140
Qwen 3 (8B)	0.913	0.258	0.486	0.180
Qwen 3 (14B)	0.935	0.258	0.576	0.260
Qwen 3 (32B)	0.957	0.548	0.757	0.420
GPT				
GPT 5.2	0.978	0.290	0.389	0.260
GPT 5.4	0.935	0.387	0.500	0.260
GPT Thinking				
GPT 5.2	0.935	0.323	0.340	0.220
GPT 5.4	0.957	0.452	0.444	0.200
Human (avg.)	1.000	0.903	0.910	0.780

Table 9: Full implicature recognition accuracy results for the prior common ground control experiment.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	0.891	0.968	0.812	0.820
Gemma 3 (12B)	0.804	0.871	0.812	0.760
Gemma 3 (27B)	0.978	0.903	0.854	0.900
Llama 3				
Llama 3.1 (8B)	1.000	1.000	0.938	0.840
Llama 3.1 (70B)	0.978	0.935	0.951	0.940
Llama 3.2 (3B)	0.978	1.000	0.833	0.840
Llama 3.3 (70B)	0.978	0.871	0.875	0.840
Qwen 2.5				
Qwen 2.5 (3B)	0.957	0.806	0.479	0.540
Qwen 2.5 (7B)	1.000	0.903	0.868	0.820
Qwen 2.5 (14B)	0.957	0.871	0.889	0.900
Qwen 2.5 (32B)	1.000	0.871	0.965	0.940
Qwen 2.5 (72B)	0.978	0.903	0.889	0.920
Qwen 3				
Qwen 3 (0.6B)	0.413	0.806	0.542	0.500
Qwen 3 (1.7B)	0.913	0.871	0.701	0.820
Qwen 3 (4B)	0.870	0.968	0.833	0.880
Qwen 3 (8B)	1.000	0.968	0.847	0.900
Qwen 3 (14B)	0.978	0.871	0.812	0.680
Qwen 3 (32B)	1.000	0.968	0.924	0.780
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.957	0.839	0.792	0.740
Qwen 3 (1.7B)	0.978	0.871	0.785	0.880
Qwen 3 (4B)	0.978	0.968	0.882	0.960
Qwen 3 (8B)	1.000	0.968	0.910	0.940
Qwen 3 (14B)	0.978	0.935	0.924	0.920
Qwen 3 (32B)	0.978	0.903	0.958	0.880
GPT				
GPT 5.2	0.891	0.839	0.750	0.480
GPT 5.4	0.957	0.806	0.771	0.500
GPT Thinking				
GPT 5.2	0.935	0.839	0.799	0.500
GPT 5.4	0.957	0.871	0.833	0.500
Human (avg.)	1.000	0.903	0.972	0.920

Table 10: Full cancellation recognition accuracy results on the IMPLICATURE$^{\perp}$ items.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	0.174	0.226	0.424	0.620
Gemma 3 (12B)	0.022	0.129	0.382	0.540
Gemma 3 (27B)	0.065	0.226	0.389	0.500
Llama 3				
Llama 3.1 (8B)	0.522	0.484	0.535	0.680
Llama 3.1 (70B)	0.261	0.161	0.528	0.600
Llama 3.2 (3B)	0.522	0.613	0.458	0.580
Llama 3.3 (70B)	0.087	0.000	0.264	0.460
Qwen 2.5				
Qwen 2.5 (3B)	0.761	0.226	0.604	0.600
Qwen 2.5 (7B)	0.565	0.226	0.674	0.560
Qwen 2.5 (14B)	0.196	0.097	0.521	0.520
Qwen 2.5 (32B)	0.130	0.161	0.465	0.500
Qwen 2.5 (72B)	0.022	0.032	0.319	0.460
Qwen 3				
Qwen 3 (0.6B)	0.478	0.419	0.444	0.440
Qwen 3 (1.7B)	0.630	0.226	0.521	0.380
Qwen 3 (4B)	0.000	0.065	0.583	0.520
Qwen 3 (8B)	0.196	0.194	0.625	0.620
Qwen 3 (14B)	0.217	0.129	0.653	0.620
Qwen 3 (32B)	0.478	0.548	0.625	0.660
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.370	0.484	0.493	0.580
Qwen 3 (1.7B)	0.804	0.194	0.528	0.680
Qwen 3 (4B)	0.065	0.097	0.444	0.380
Qwen 3 (8B)	0.674	0.290	0.576	0.640
Qwen 3 (14B)	0.674	0.323	0.528	0.580
Qwen 3 (32B)	0.413	0.258	0.444	0.680
GPT				
GPT 5.2	0.261	0.065	0.319	0.140
GPT 5.4	0.152	0.000	0.229	0.200
GPT Thinking				
GPT 5.2	0.630	0.097	0.306	0.300
GPT 5.4	0.457	0.129	0.208	0.320

Table 11: Full belief strengthening accuracy results on the IMPLICATURE$^+$ items.

Model	Scalar Implicature	Discourse Implicature	Synthetic Conversational Implicature	Naturally-Occurring Conversational Implicature
Gemma 3				
Gemma 3 (4B)	1.000	1.000	0.757	0.860
Gemma 3 (12B)	0.913	0.968	0.792	0.940
Gemma 3 (27B)	0.826	0.935	0.812	0.960
Llama 3				
Llama 3.1 (8B)	0.674	0.806	0.868	0.900
Llama 3.1 (70B)	0.913	0.935	0.847	0.900
Llama 3.2 (3B)	0.609	0.742	0.819	0.940
Llama 3.3 (70B)	0.913	0.968	0.847	0.900
Qwen 2.5				
Qwen 2.5 (3B)	0.761	0.806	0.931	0.920
Qwen 2.5 (7B)	0.674	0.871	0.868	0.980
Qwen 2.5 (14B)	0.630	0.806	0.757	0.900
Qwen 2.5 (32B)	0.587	0.903	0.792	0.920
Qwen 2.5 (72B)	0.761	0.968	0.757	0.960
Qwen 3				
Qwen 3 (0.6B)	1.000	1.000	1.000	1.000
Qwen 3 (1.7B)	0.783	0.935	0.847	0.900
Qwen 3 (4B)	1.000	0.968	0.826	0.920
Qwen 3 (8B)	0.826	0.903	0.826	0.940
Qwen 3 (14B)	0.870	1.000	0.799	0.940
Qwen 3 (32B)	0.848	1.000	0.840	0.940
Qwen 3 Thinking				
Qwen 3 (0.6B)	0.870	0.774	0.674	0.660
Qwen 3 (1.7B)	0.457	0.903	0.771	0.840
Qwen 3 (4B)	1.000	0.903	0.806	0.880
Qwen 3 (8B)	0.739	0.871	0.771	0.940
Qwen 3 (14B)	0.674	0.903	0.743	0.920
Qwen 3 (32B)	0.870	0.935	0.771	0.840
GPT				
GPT 5.2	0.696	0.968	0.812	0.940
GPT 5.4	0.761	0.935	0.785	0.940
GPT Thinking				
GPT 5.2	0.717	0.871	0.847	0.960
GPT 5.4	0.826	0.968	0.847	0.960

Table 12: Full belief unchanging accuracy results on the IMPLICATURE$^{\approx}$ items.

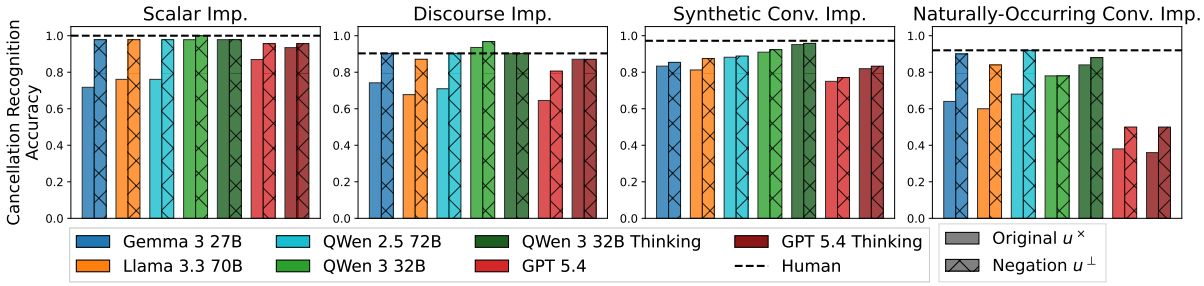

Figure 19: Cancellation recognition accuracy using original cancelling utterances from IMPLICATUREX and the explicit cancelling utterances from IMPLICATURE$^{\perp}$ which contain negation.

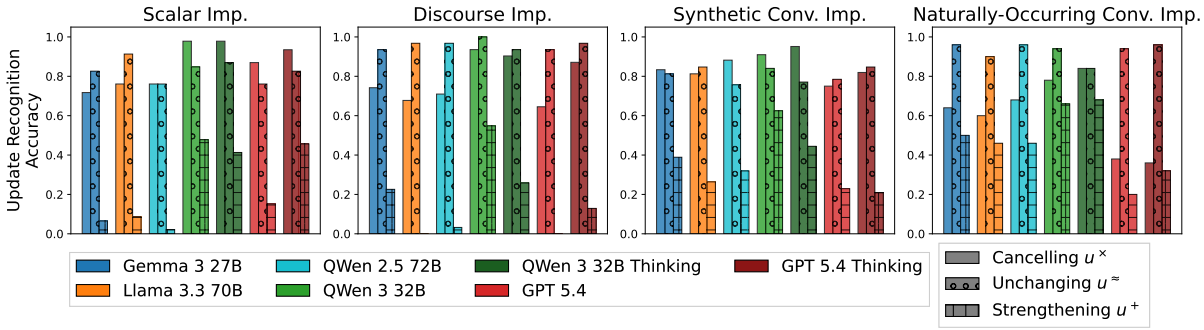

Figure 20: Update belief accuracies given different update types. Cancelling denotes the update belief as triggered by implicature cancellation while unchanging denotes the belief update accuracy using IMPLICATURE$^{\approx}$ items and strengthening denotes the belief update accuracy using IMPLICATURE^{+} items.

	Scalar Imp.	Discourse Imp.	Synthetic Conv. Imp.	Naturally-Occurring Conv. Imp.
$\lvert c \rvert$	—	-0.06	-0.01	0.27^{*}
$\lvert b \rvert$	-0.01	-0.26	0.07	-0.10
$\lvert u \rvert$	0.28^{*}	-0.13	0.06	-0.01

Table 13: Kendall's tau τ between $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u \rangle))$ and length variables for the dataset splits. Values with * indicate statistical significance ($p < 0.05$).

	Scalar Imp.	Discourse Imp.	Synthetic Conv. Imp.	Naturally-Occurring Conv. Imp.
$\lvert c \rvert$	—	-0.21	0.02	0.16
$\lvert b \rvert$	0.08	-0.00	0.07	0.02
$\lvert u \rvert$	0.22^{*}	-0.06	0.01	0.14
$\lvert u^{\times} \rvert$	0.09	-0.05	-0.07	-0.10

Table 14: Kendall's tau τ between $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^{\times} \rangle))$ and length variables for the different dataset splits. Values with * indicate statistical significance ($p < 0.05$).