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

Anonymous ACL submission

Abstract

Human language is driven by unspoken beliefs and belief updates which are thus of critical importance for successful interactions 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 recognize 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 in part stem 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 communicative belief updates.

1 Introduction

Natural language communication consists of threads of beliefs negotiated between interlocutors in a conversational common ground (Lewis, 1979; Heim, 1982; Clark, 1991; Stalnaker, 1998; Kamp, 2013). 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 (as is illustrated in Figure 1). However, Aya can cancel this implicature by saying "Though they must have gotten around to it by now.", at which point

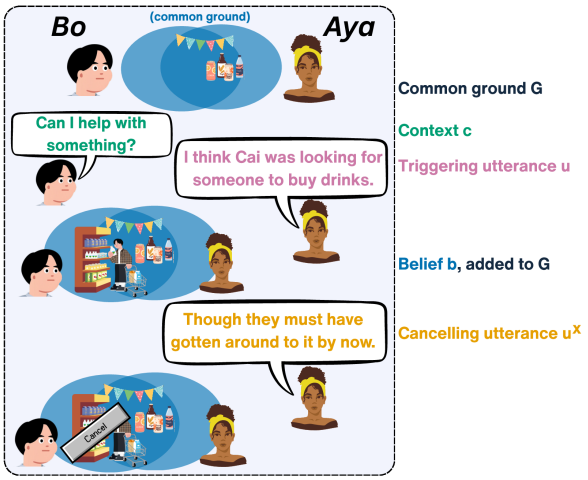

Figure 1: An example of belief negotiation: beliefs in the common ground are updated as an implicature is triggered by u then cancelled by $u^{\times}$.

Bo should update their beliefs and act accordingly, e.g., by asking Cai to confirm. Interlocutors' beliefs, and the corresponding appropriate actions, can thus change wildly following an implicature cancellation. 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.

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 remain 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 272 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 of implicature recognition on scalar and discourse implicatures, LLMs struggle to perform 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 to what extent their success can be attributed to true 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 32B Thinking) cannot match human accuracy in cancellation recognition.

To summarize, we study the extent to which LLMs are able to perform implicature recognition and belief updates via implicature cancellation, a fundamental property of human communication. We create the first implicature cancellation dataset, IMPLICATUREX, verified by linguistic experts and annotated with crowdsourced judgements of implicatures and their corresponding cancellations. Our evaluation reveals that LLMs fall short of both a human understanding of implicatures and of belief updates as triggered by implicature cancellations. More broadly, our work suggests that LLMs may still not be able to understand the nuances of belief productions and negations 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, 1991; Stalnaker, 1998; Kamp, 2013). As such, implicatures—beliefs communicated implicitly—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.

Evaluating beliefs in NLP Notions of beliefs and common grounds have long served as inspiration for language generation systems, including dialogue systems (Allen and Perrault, 1980; Grosz and Sidner, 1986; Dale and Reiter, 1995), neural-based image captioning (Andreas and Klein, 2016), and conversational agents (Körner et al., 2025). Furthermore, additions to the common ground, as triggered by conversational implicatures, 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* common ground beliefs. Finally, while processes similar to implicature cancellation, such as abductive and defeasible reasoning, have received extensive consideration (Lascarides and Asher, 1991; Rudinger et al., 2020; Hwang et al., 2021), they differ from our focus, which is primarily about communicative belief updates.

3 Theoretical Setup

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 $\mathcal{U}^*$ denotes 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, which captures situational and communicative information relevant to interpreting the utterances.[1]

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 $G : \mathcal{B} \times \mathcal{U}^* \rightarrow [0, 1]$ that, given $\mathbf{u}$, assigns to each $b \in \mathcal{B}$ a probability reflecting how strongly that belief is mutually held by the participants.[2]

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.

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

[2] Prior to any exchange, the common ground is given by $G(b \mid \emptyset, \emptyset)$, which may either assign zero 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 272 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." adds the belief b = "*Not all* places nail them for tax." 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 items.

❷ DISCOURSE IMPLICATURES. We posit that 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 (Webber 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 $u^{\times}$ = "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.

❸ CONVERSATIONAL IMPLICATURES. Our conversational implicatures consist of implicatures which are triggered by utterances produced in

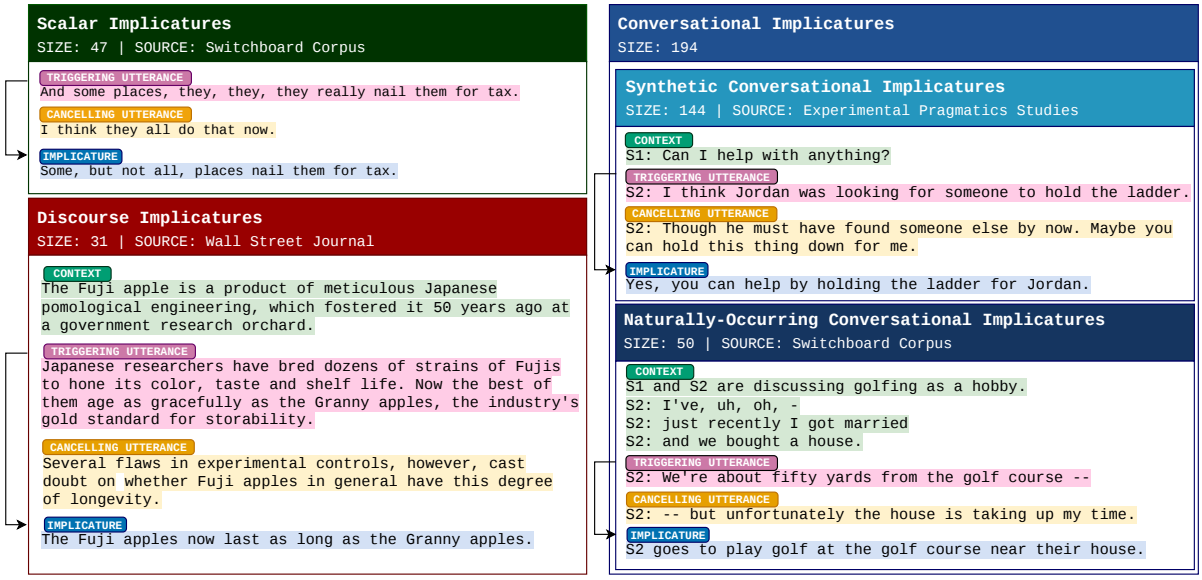

Figure 2: Composition of the IMPLICATUREX dataset with sizes, sources and illustrative examples.

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 ($b =$ "B needs help with X."). 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 study participants (Godfrey et al., 1992). We collect 51 naturally-occurring conversational implicatures which feature dialogues from 1 to 10 turns. The cancelling utterances are produced by one of the participants within the exchange. In addition, the cancelling utterances may either immediately follow the triggering utterance, like in previous implicature types, or may be introduced a few turns later $u =$ "We're close to the golf course." [...] $u^{\times} =$ "Unfortunately the house is taking up time." ($b =$ "They go golfing near their home").

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 (Appendix A). We elicit expert judgements for the implicatures' plausibility as well as the cancellations' correctness. 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 implicature and one naturally-occurring conversational implicature.

We hired two professionally trained linguists for our expert annotation. Our first expert annotator was tasked to annotate all six batches and to provide replacements for implausible implicatures and incorrect cancellations. Our second annotator was given one of the six batches so we could compute inter-annotator agreement of implicature plausibility and cancellation correctness.

Annotation results The agreement between the annotators was 0.85 for implicature plausibility

and 0.86 for cancellation correctness[3]. The expert annotation generated 18 implicature replacements and 27 cancellation replacements. In six 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 272 expert-annotated implicature cancellation items which are distributed as follows: 1: 47 items, 2: 31 items, 3 A: 144 items, 3 B: 50 items. Additional details can be found in Appendix A.

4.3 Crowdsourced 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 performance of belief updates via implicature cancellation.

Annotation details To obtain aggregated human judgments of IMPLICATUREX belief updates, we run a crowdsourcing annotation of the 272 expert annotated items (Appendix A). To do so, we elicit likelihood judgements for each implicature item given its corresponding context, triggering utterance, and cancelling utterance. In particular, we create 18 stimuli batches, each containing 15 items with cancelling utterances and 15 items without.[4] We ensure that batches do not contain overlapping implicature likelihood elicitations. 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 and with a hit rate greater than or equal to 99/100. Each batch is annotated by 5 participants.

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

[4]One batch includes two extra items.

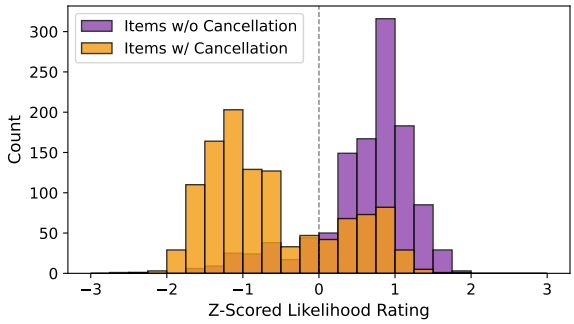

Figure 3: Histogram of per-participant z-scored likelihood judgements of the items from IMPLICATUREX with and without the cancelling utterance.

Annotation results Since human annotators are known to interpret Likert scales differently (Cliff, 1993), we apply per-participant z-scoring to the Likert scale likelihood judgements. 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 details and results about the crowdsourcing annotation can be found in Appendix B.

5 Task Definitions and Setup

We use a multiple-choice-question prompt template to estimate specific values of the common ground function $G : \mathcal{B} \times \mathcal{U}^* \rightarrow [0, 1]$. 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: $G(b \mid c, \mathbf{u}) \approx P_{\mathcal{M}}(b \mid \texttt{t}_b(c, \mathbf{u}))$. Here, $\texttt{t}_b(c, \mathbf{u})$ 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 (see Appendix D for the full prompt and examples). 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.,

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

Cancellation recognition We compute $\mathcal{M}$'s *cancellation recognition accuracy* as the proportion of items for which $\mathcal{M}$ decreases its probability of b

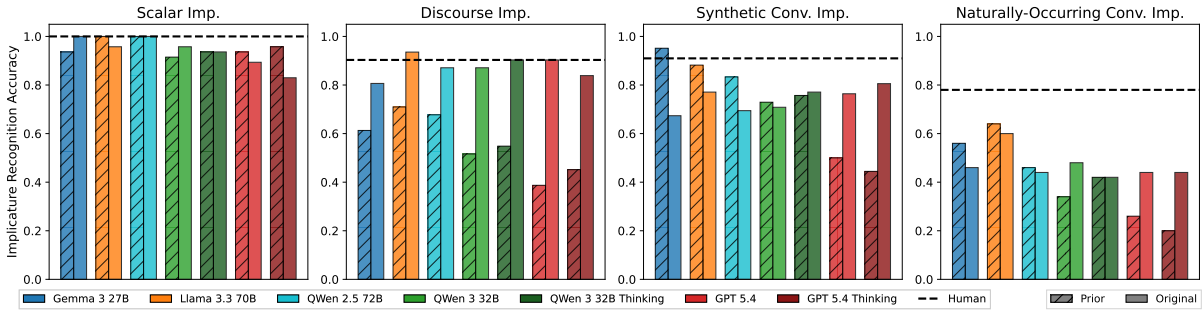

Figure 4: Implicature recognition accuracies of the largest and most modern models for each class of models tested. Original denotes the accuracy using the $\mathtt{t}_b(c, \mathbf{u})$) prompt and prior denotes the accuracy using the $\mathtt{t}_b(\emptyset, \emptyset)$) prompt.

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 accuracy 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 $u^{\times}$.

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 specific utterance sequences, 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.

[5]All open-weight models are downloaded from HuggingFace and all closed-source models are accessed via OpenAI's API. See Table 1 for full model names.

6 Implicature Recognition Experiment

Here, we present the experiments, evaluations 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 baseline*. To do so, we compute the proportion of items for which the average z-scored likelihood ratings 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 about prior knowledge. We run this experiment to isolate the effect of the LLM's prior belief on implicature recognition (See Appendix E for an example). 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" by models.

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 ❶ SCALAR IMPLICATURES and ❷ DISCOURSE IMPLICATURES similarly to humans with Gemma 3 27B matching human scalar implicature recognition accuracy of 1.0 and Llama 3.3 70B outperforming human discourse implicature accuracy of 0.90 by 0.04. On the other hand, models tend to perform worse on ❸ CONVERSATIONAL IMPLICATURES. Compared

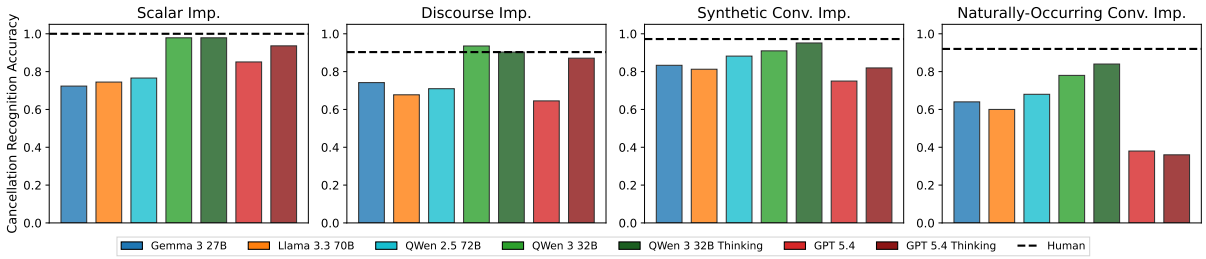

Figure 5: Cancellation recognition accuracies of the largest and most modern models for each class of models tested.

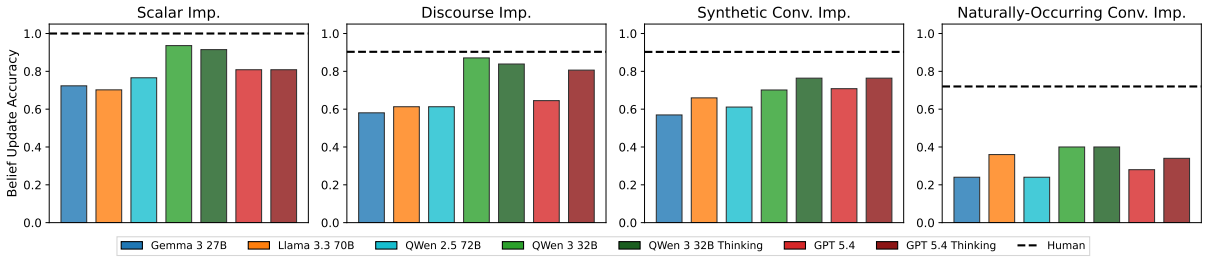

Figure 6: Belief update accuracies of the largest and most modern models for each class of models tested.

to the human accuracies of 0.91 and 0.78 on synthetic and naturally-occurring implicatures, the best performing models achieve implicature 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 tasks for LLMs. The full implicature recognition results are in Table 3.

Control results In addition, our prior belief experiment reveals that models have a moderately strong prior for synthetic conversational implicatures and 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 uniquely stem from a true understanding of implicated beliefs, but rather, and in some cases more importantly, due to interactions with LLMs' prior knowledge. We leave a thorough investigation of the interaction between pragmatic understanding and prior beliefs reasoning to future work.

7 Cancellation Recognition and Belief Update Experiment

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

7.1 Human Topline and Controls

We use our crowdsourcing annotation results from Section 4.3 to compute cancellation recognition and update belief *human accuracy topline*. For the cancellation accuracy, we compute the proportion of items for which the average z-scored likelihood ratings with the cancelling utterance is less than the average z-scored likelihood rating with only the triggering utterance. For the *belief update accuracy*, we compute the proportion of items for which humans provide an average z-scored likelihood rating above 0 and for which items average z-scored likelihood rating decrease when accompanied by the cancelling utterance.

Form control We evaluate to what extent the form of the cancellation affects an LLM's success in performing cancellation recognition. To do so, we create a new set of stimuli, IMPLICATUREINFACT, in which cancelling utterances have the form $u^{\times}$ ="$d + \neg b$" 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 (See Appendix E for an example).

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 strength-

ening utterances u^+, i.e., utterances that strengthen b by providing additional information confirming its plausibility, and (2) IMPLICATURE^O, in which $u^\times$ is replaced with a randomly sampled utterance u^O of the same stimuli type that neither strengthens nor cancels b (See Appendix E for an examples). 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 the belief unchanged by computing the proportion of IMPLICATURE^O items for which $P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u, u^O\rangle)) \approx P_{\mathcal{M}}(b \mid \mathtt{t}_b(c, \langle u\rangle))$.

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 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, human performance remains relatively stable, with accuracy only decreasing on naturally occurring conversational implicatures and only slightly from 0.78 to 0.72. In contrast, model performance on the belief update task falls considerably short of human performance, particularly for conversational implicatures. The best model, 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. The full cancellation recognition and belief update accuracies can be found in Tables 4 and 5.

Form control results We compare model cancellation-recognition accuracy when using the original IMPLICATUREX stimuli and the IMPLICATUREINFACT variant, which contain explicitly negating cancelling utterances. In all cases, cancellation accuracy increases Figures 17 and 7. Surprisingly, however, we find that in most cases — aside from scalar implicatures — the belief update

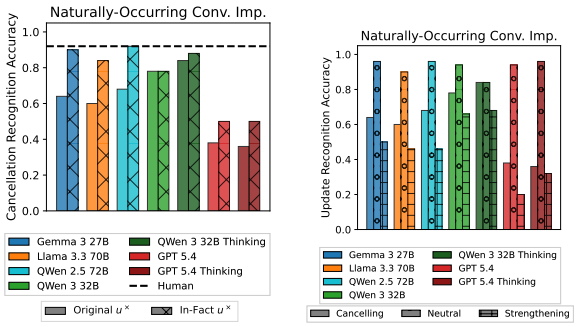

Figure 7: Lcancellation recognition accuracy using IMPLICATUREX and IMPLICATUREINFACT cancelling utterances (left) and update belief accuracies across different update types: cancellation ($u^\times$), neutrality (u^O), and strengthening (u^+) (right).

accuracy obtained with IMPLICATUREINFACT cancelling utterances 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.

Update type control results We show the effect of different update types on LLM belief-update accuracy in Figures 7 and 16. Overall, we observe that LLMs find implicature strengthening and cancellation similarly difficult, but have relatively little difficulty maintaining their beliefs when presented with an irrelevant follow-up utterance. For instance, while GPT-5.4 achieves an update-recognition accuracy of 0.96 on naturally occurring conversational implicatures, its strengthening and cancellation accuracies both remain at 0.32. This result suggests that LLMs perform worse at belief updates when the beliefs contradict some previously presented piece of information.

8 Conclusion

In conclusion, in this paper we study the tasks of implicature recognition and update beliefs via implicature cancellation. 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 over contextual evidence and an inability to reconcile explicit updates that contradict established common ground. 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 true ambiguity or 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 left 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.

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.

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. 1991. Brennan (1991) grounding in communication.

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.

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*, volume 35, 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. 2013. A theory of truth and semantic representation. In *Meaning and the Dynamics of Interpretation*, pages 329–369. Brill.

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.

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*, volume 142. Harvard University Press Cambridge, MA.

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.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

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

The expert annotation interface is illustrated in Figures 8, 9, 10, and 11. Annotators are first presented with a landing page (Figure 8) containing general instructions, along with a reminder of the notions of implicature and implicature cancellation (Figure 9). The annotation task is organized into several sub-batches based on stimulus type and is presented in order of increasing cognitive load: scalar implicatures, synthetic conversational implicatures, naturally occurring conversational implicatures, and finally discourse implicatures.

Before each stimulus type, annotators are shown representative examples to familiarize them with the task format (Figure 10). For each annotation item (Figure 11), 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 but, annotators are asked to provide an alternative if possible. 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 hourly rate. For anonymity purposes, we only report that this was over three times the minimum wage.

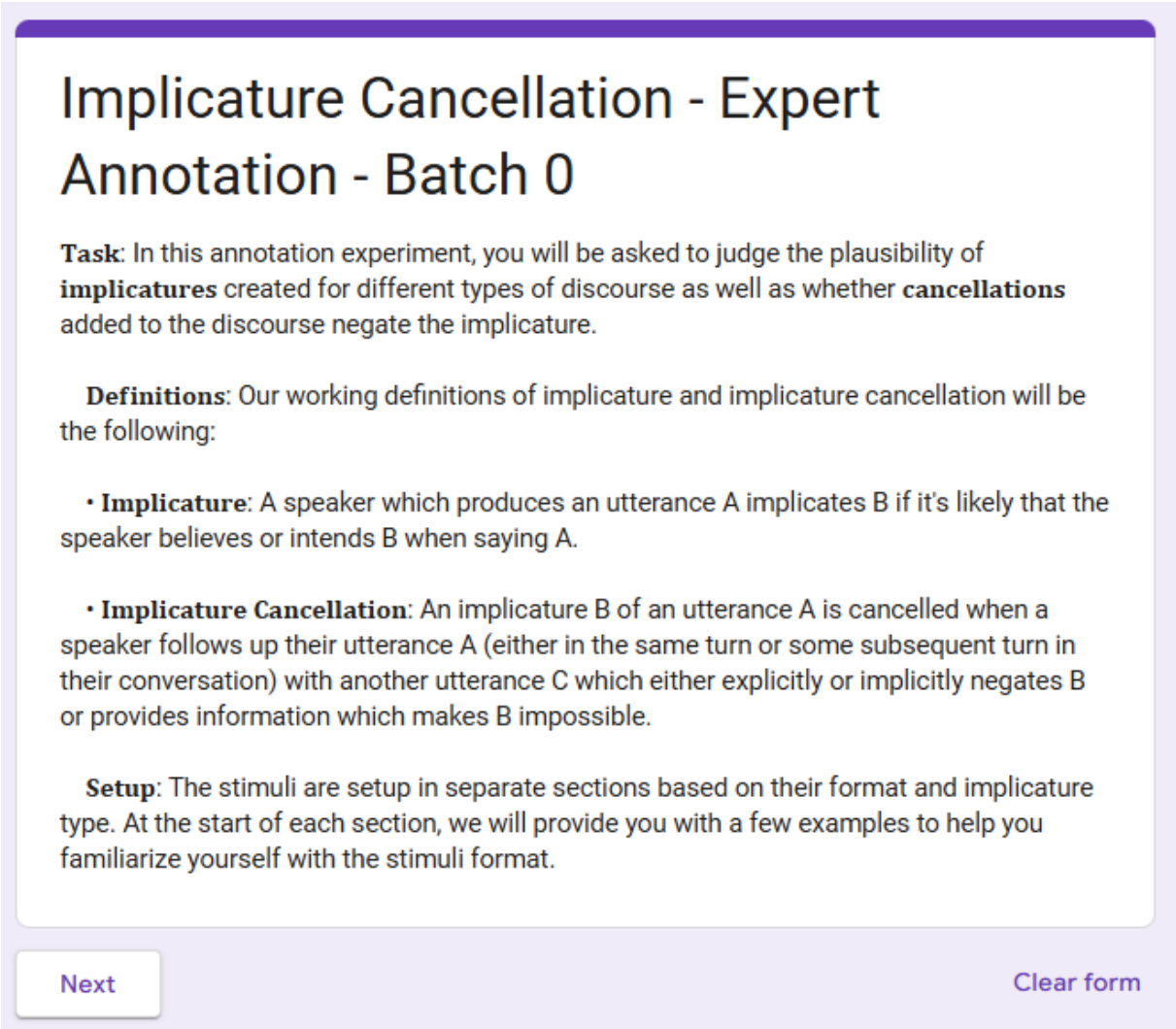

Figure 8: Landing page shown to participants at the beginning of the annotation task.

Scalar Implicatures

In the following items, you will be shown a single utterance containing the quantifier "some". Your task is twofold:

- To assess whether the interpretation of the utterance is a plausible (scalar) implicature
- To assess whether the follow-up utterance cancels the implicature

If you find that the provided interpretation and cancellation are implausible or incorrect, then briefly say why. You will also be given the option of providing an alternative implicature and cancellation should you think of one.

Figure 9: Instructions provided to participants, including definitions of implicature and cancellation.

Scalar Implicature Examples

Here are a few examples of scalar implicatures with "some" along with their corresponding cancellation.

Example 1
Utterance: well, i know that we have some relatives that live around, like the mumblex area in there, [*Implicature: Some, but not all, of our relatives live around the Mumblex area.*] [**Cancellation: now that i think about it, our entire family moved over there recently.**]

Example 2
Utterance: but i do think that, um, congress has backed down much too much on some of the air pollution standards. [*Implicature: Congress has backed down on some, but not all, of the air pollution standards.*] [**Cancellation: i don't think there's any standards that they've continued to support.**]

Figure 10: Example items shown to participants prior to the main task to illustrate the annotation format.

Item 1

Utterance: i was trying to think about some of my favorite people that i liked in music [*Implicature: I was trying to think about some, but not all, of my favorite people in music.*] [**Cancellation: but then i started trying to list all of my favorite people.**]

Is the implicature plausible? *

- ◯ Plausible Implicature
- ◯ Implausible Implicature

Does the cancellation cancel the implicature? *

- ◯ Yes
- ◯ No
- ◯ Implicature is implausible

Implicature/cancellation judgement explanations

Your answer

Alternative Implicature (If the implicature is implausible)

Your answer

Alternative Cancellation (If the implicature is implausible or the cancellation is incorrect)

Your answer

Figure 11: Example annotation item, including context, utterance, candidate interpretation, and cancellation.

B Crowdsourcing Annotation

We show images of the landing page (Figure 12) of our crowdsourcing experiment and an example stimuli (Figure 13). Crowdworkers were paid 15USD/hour.

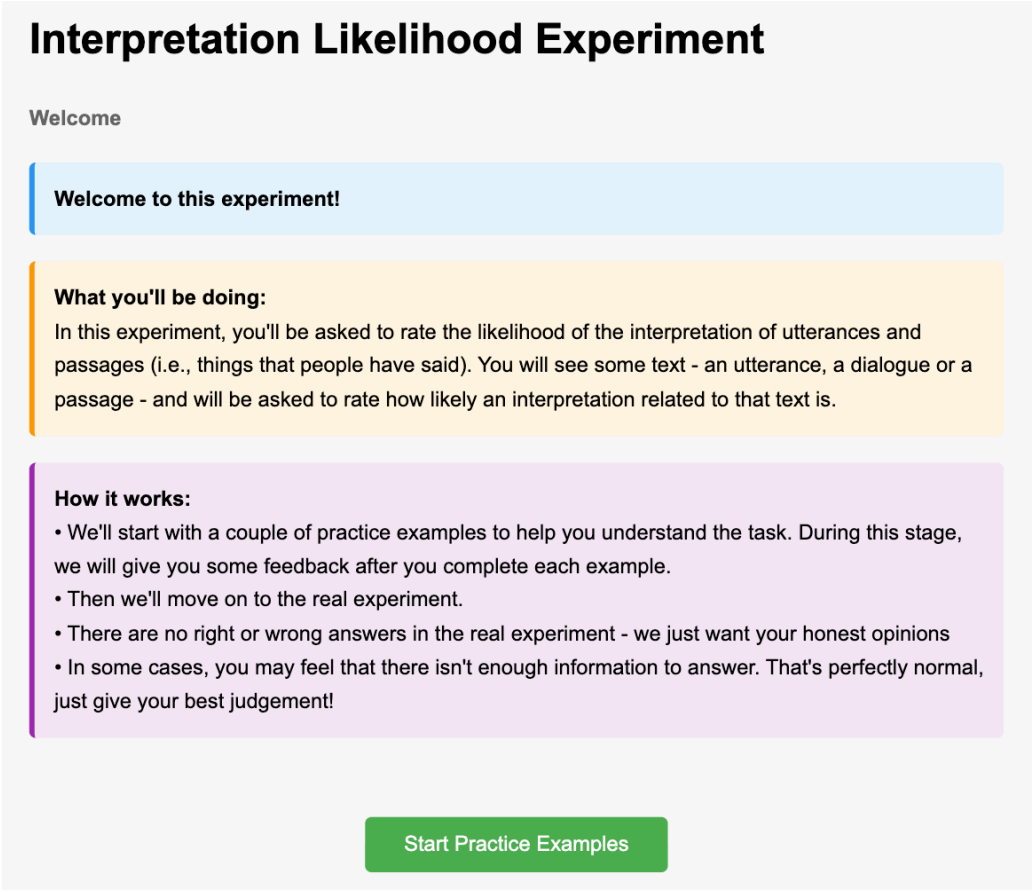

Figure 12: Landing page of the crowdsourcing experiment, outlining the instructions and interface for participants.

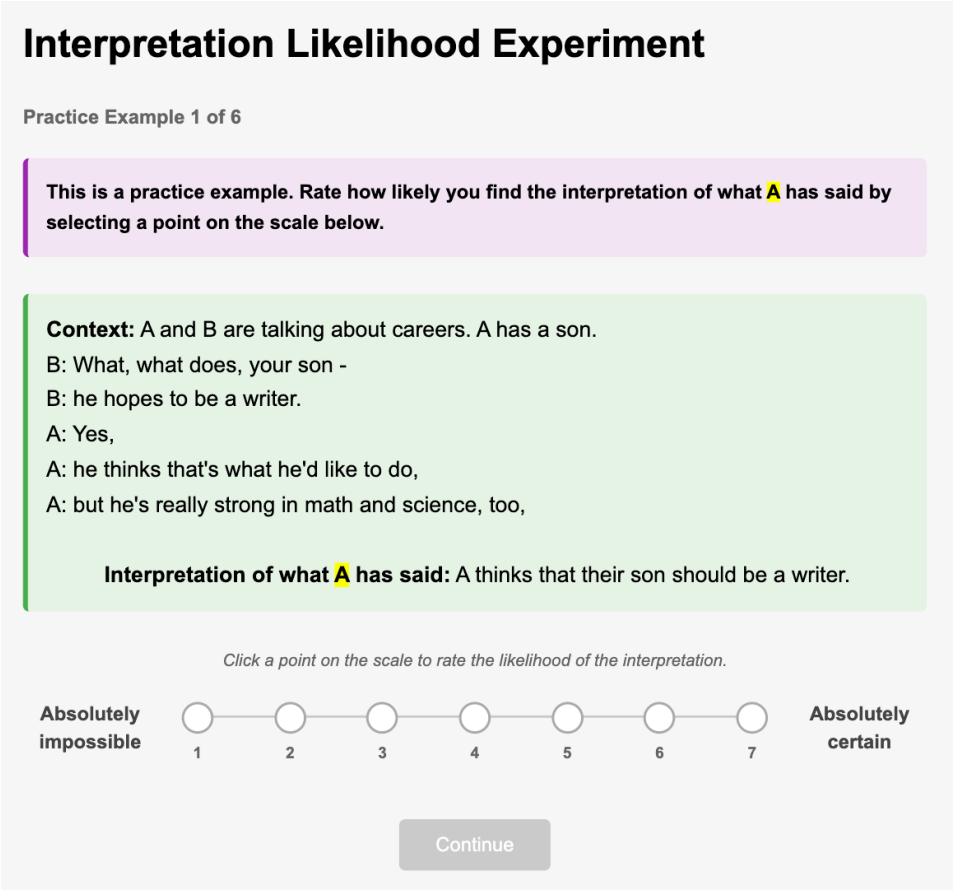

Figure 13: Example stimulus from the experiment, demonstrating the Likert-scale interface used to collect interpretation likelihood ratings.

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 3.2 (3B)	meta-llama/Llama-3.2-3B-Instruct
Llama 3.1 (8B)	meta-llama/Llama-3.1-8B-Instruct
Llama 3.1 (70B)	meta-llama/Llama-3.1-70B-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 (0.6B) + Reasoning	Qwen/Qwen3-0.6B-thinking
Qwen 3 (1.7B) + Reasoning	Qwen/Qwen3-1.7B-thinking
Qwen 3 (4B) + Reasoning	Qwen/Qwen3-4B-thinking
Qwen 3 (8B) + Reasoning	Qwen/Qwen3-8B-thinking
Qwen 3 (14B) + Reasoning	Qwen/Qwen3-14B-thinking
Qwen 3 (32B) + Reasoning	Qwen/Qwen3-32B-thinking
GPT-5.2	gpt-5.2-2025-12-11
GPT-5.2 + Reasoning	gpt-5.2-2025-12-11-thinking
GPT-5.4	gpt-5.4-2026-03-05
GPT-5.4 + Reasoning	gpt-5.4-2026-03-05-thinking

Table 1: Model names as referenced in the paper and their corresponding HuggingFace or OpenAI identifiers.

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 14: 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 15: Example instance for the prompt in Figure 14, shown without (top) and with (bottom) the cancellation $u^{\times}$ appended to the utterance in `{scenario}`.

E Control Experiments

Control Type	Context c	Utterance u	Follow-up Utterance
Original $(c, \langle u, u^{\times} \rangle)$	Can I help with anything?	I think Jordan was looking for someone to hold the ladder.	Though he must have found someone else by now.
Prior belief $(\emptyset, \emptyset)$	$\emptyset$	$\emptyset$	$\emptyset$
Neutral $(c, \langle u, u^{O} \rangle)$	Can I help with anything?	I think Jordan was looking for someone to hold the ladder.	Maybe go over and ask him?
Strengthening $(c, \langle u, u^{+} \rangle)$	Can I help with anything?	I think Jordan was looking for someone to hold the ladder.	Thankfully, I was able to submit the paper before my laptop broke.
Cancelling utterance from IMPLICATUREINFACT ($u^{\times}$ = "$d + \neg b''$)	Can I help with anything?	I think Jordan was looking for someone to hold the ladder.	To be clear, I'm not saying that you could help by holding the ladder for Jordan.

Table 2: Control conditions illustrating the values of the context, utterance, and follow-up utterance used to fill the prompt $\texttt{t}_b$, where, in this case, b = "I think Jordan was looking for someone to hold the ladder."

F Additional Results

Table 3: Full implicature recognition accuracies.

Model	Scalar Imp.	Discourse Imp.	Synthetic Conv. Imp.	Naturally-Occurring Conv. Imp.
GPT				
gpt-5.2-2025-12-11	0.809	0.839	0.681	0.500
gpt-5.4-2026-03-05	0.894	0.903	0.764	0.440
GPT (Reasoning)				
gpt-5.2-2025-12-11-thinking	0.660	0.871	0.743	0.460
gpt-5.4-2026-03-05-thinking	0.830	0.839	0.806	0.440
Gemma 3				
google/gemma-3-4b-it	1.000	0.871	0.708	0.700
google/gemma-3-12b-it	1.000	0.903	0.722	0.560
google/gemma-3-27b-it	1.000	0.806	0.674	0.460
Llama 3				
meta-llama/Llama-3.2-3B-Instruct	0.489	0.452	0.271	0.100
meta-llama/Llama-3.1-8B-Instruct	0.915	0.677	0.229	0.160
meta-llama/Llama-3.1-70B-Instruct	0.979	0.935	0.806	0.560
meta-llama/Llama-3.3-70B-Instruct	0.957	0.935	0.771	0.600
Qwen 2.5				
Qwen/Qwen2.5-3B-Instruct	0.511	0.419	0.076	0.080
Qwen/Qwen2.5-7B-Instruct	0.766	0.613	0.222	0.160
Qwen/Qwen2.5-14B-Instruct	0.894	0.806	0.479	0.340
Qwen/Qwen2.5-32B-Instruct	0.936	0.871	0.556	0.380
Qwen/Qwen2.5-72B-Instruct	1.000	0.871	0.694	0.440
Qwen 3				
Qwen/Qwen3-0.6B	0.000	0.000	0.000	0.000
Qwen/Qwen3-1.7B	0.830	0.871	0.715	0.600
Qwen/Qwen3-4B	1.000	0.903	0.590	0.780
Qwen/Qwen3-8B	1.000	0.839	0.326	0.420
Qwen/Qwen3-14B	0.894	0.903	0.674	0.420
Qwen/Qwen3-32B	0.957	0.871	0.708	0.480
Qwen 3 (Reasoning)				
Qwen/Qwen3-0.6B-thinking	0.979	0.613	0.444	0.500
Qwen/Qwen3-1.7B-thinking	0.447	0.774	0.292	0.400
Qwen/Qwen3-4B-thinking	1.000	0.839	0.521	0.460
Qwen/Qwen3-8B-thinking	0.809	0.742	0.347	0.380
Qwen/Qwen3-14B-thinking	0.574	0.839	0.528	0.420
Qwen/Qwen3-32B-thinking	0.936	0.903	0.771	0.420
Human (avg.)	**1.000**	**0.903**	**0.910**	**0.780**

Table 4: Full cancellation recognition accuracies.

Model	Scalar Imp.	Discourse Imp.	Synthetic Conv. Imp.	Naturally-Occurring Conv. Imp.
GPT				
gpt-5.2-2025-12-11	0.894	0.645	0.736	0.340
gpt-5.4-2026-03-05	0.851	0.645	0.750	0.380
GPT (Reasoning)				
gpt-5.2-2025-12-11-thinking	0.915	0.871	0.785	0.460
gpt-5.4-2026-03-05-thinking	0.936	0.871	0.819	0.360
Gemma 3				
google/gemma-3-4b-it	0.553	0.710	0.826	0.720
google/gemma-3-12b-it	0.319	0.548	0.743	0.560
google/gemma-3-27b-it	0.723	0.742	0.833	0.640
Llama 3				
meta-llama/Llama-3.2-3B-Instruct	0.851	0.806	0.729	0.540
meta-llama/Llama-3.1-8B-Instruct	0.872	0.839	0.944	0.660
meta-llama/Llama-3.1-70B-Instruct	0.915	0.774	0.944	0.820
meta-llama/Llama-3.3-70B-Instruct	0.745	0.677	0.812	0.600
Qwen 2.5				
Qwen/Qwen2.5-3B-Instruct	0.681	0.774	0.438	0.560
Qwen/Qwen2.5-7B-Instruct	0.894	0.710	0.819	0.620
Qwen/Qwen2.5-14B-Instruct	0.830	0.677	0.847	0.720
Qwen/Qwen2.5-32B-Instruct	0.872	0.710	0.917	0.740
Qwen/Qwen2.5-72B-Instruct	0.766	0.710	0.882	0.680
Qwen 3				
Qwen/Qwen3-0.6B	0.468	0.742	0.556	0.520
Qwen/Qwen3-1.7B	0.617	0.516	0.688	0.680
Qwen/Qwen3-4B	0.298	0.645	0.875	0.660
Qwen/Qwen3-8B	0.830	0.710	0.840	0.640
Qwen/Qwen3-14B	0.660	0.613	0.750	0.640
Qwen/Qwen3-32B	0.979	0.935	0.910	0.780
Qwen 3 (Reasoning)				
Qwen/Qwen3-0.6B-thinking	0.830	0.742	0.660	0.680
Qwen/Qwen3-1.7B-thinking	0.872	0.613	0.736	0.600
Qwen/Qwen3-4B-thinking	0.830	0.710	0.868	0.760
Qwen/Qwen3-8B-thinking	0.957	0.871	0.938	0.740
Qwen/Qwen3-14B-thinking	0.957	0.871	0.931	0.880
Qwen/Qwen3-32B-thinking	0.979	0.903	0.951	0.840
Human (avg.)	**1.000**	**0.903**	**0.972**	**0.920**

Table 5: Full belief update accuracies.

Model	Scalar Imp.	Discourse Imp.	Synthetic Conv. Imp.	Naturally-Occurring Conv. Imp.
GPT				
gpt-5.2-2025-12-11	0.809	0.548	0.653	0.320
gpt-5.4-2026-03-05	0.809	0.645	0.708	0.280
GPT (Reasoning)				
gpt-5.2-2025-12-11-thinking	0.638	0.806	0.715	0.400
gpt-5.4-2026-03-05-thinking	0.809	0.806	0.764	0.340
Gemma 3				
google/gemma-3-4b-it	0.553	0.645	0.590	0.540
google/gemma-3-12b-it	0.319	0.484	0.576	0.280
google/gemma-3-27b-it	0.723	0.581	0.569	0.240
Llama 3				
meta-llama/Llama-3.2-3B-Instruct	0.468	0.452	0.271	0.100
meta-llama/Llama-3.1-8B-Instruct	0.809	0.645	0.222	0.160
meta-llama/Llama-3.1-70B-Instruct	0.894	0.742	0.792	0.460
meta-llama/Llama-3.3-70B-Instruct	0.702	0.613	0.660	0.360
Qwen 2.5				
Qwen/Qwen2.5-3B-Instruct	0.362	0.387	0.069	0.040
Qwen/Qwen2.5-7B-Instruct	0.681	0.484	0.215	0.140
Qwen/Qwen2.5-14B-Instruct	0.723	0.484	0.438	0.200
Qwen/Qwen2.5-32B-Instruct	0.809	0.581	0.521	0.280
Qwen/Qwen2.5-72B-Instruct	0.766	0.613	0.611	0.240
Qwen 3				
Qwen/Qwen3-0.6B	0.000	0.000	0.000	0.000
Qwen/Qwen3-1.7B	0.553	0.484	0.535	0.440
Qwen/Qwen3-4B	0.298	0.581	0.542	0.480
Qwen/Qwen3-8B	0.830	0.613	0.292	0.240
Qwen/Qwen3-14B	0.596	0.548	0.604	0.340
Qwen/Qwen3-32B	0.936	0.871	0.701	0.400
Qwen 3 (Reasoning)				
Qwen/Qwen3-0.6B-thinking	0.809	0.581	0.403	0.380
Qwen/Qwen3-1.7B-thinking	0.426	0.516	0.257	0.320
Qwen/Qwen3-4B-thinking	0.830	0.613	0.479	0.360
Qwen/Qwen3-8B-thinking	0.766	0.677	0.340	0.280
Qwen/Qwen3-14B-thinking	0.574	0.774	0.507	0.380
Qwen/Qwen3-32B-thinking	0.915	0.839	0.764	0.400
Human (avg.)	**1.000**	**0.903**	**0.903**	**0.720**

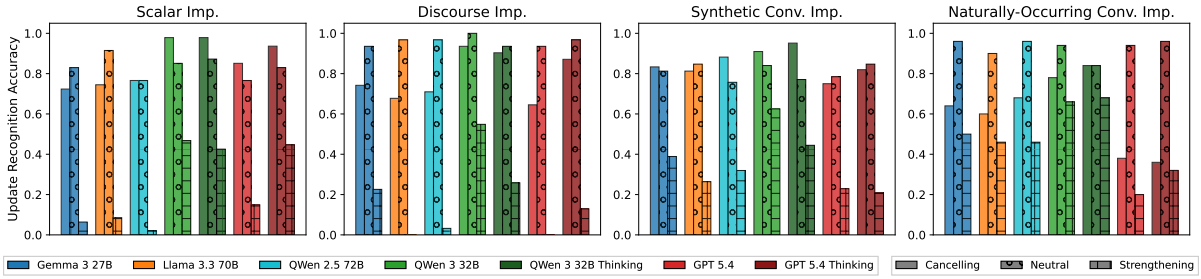

Figure 16: Update belief accuracies given different update types. Cancelling denotes the update belief as triggered by implicature cancellation while neutral denotes the belief update accuracy using IMPLICATURE^{O} items and strengthening denotes the belief update accuracy using IMPLICATURE^{+} items.

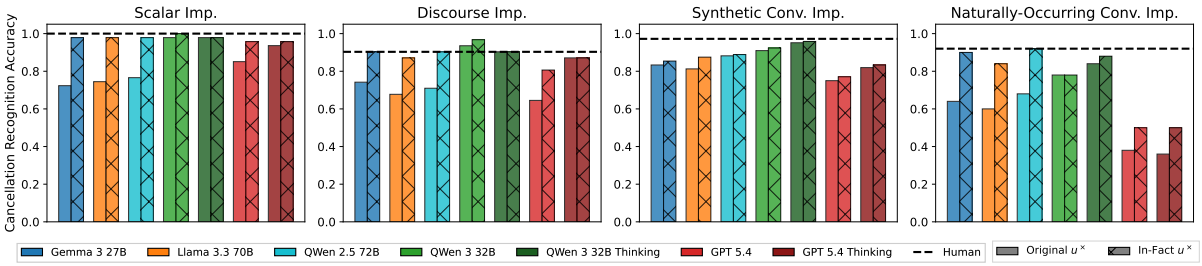

Figure 17: Cancellation recognition accuracy using original cancelling utterances from IMPLICATUREX and the cancelling utterances from IMPLICATUREINFACT.