DRInQ: Evaluating Conversational Implicature with Controlled Context Variation

Hirona Jacqueline Arai and **Xiang Ren**
University of Southern California
{hjarai, xiangren}@usc.edu

Abstract

Human conversation relies heavily on *conversational implicature*, in which speakers convey meanings that are suggested rather than explicitly stated. Although recent large language models (LLMs) exhibit strong conversational fluency, they remain unreliable when interpretation depends on reasoning that integrates social and contextual cues, a process rarely articulated in text. We introduce **DRinQ**, a benchmark for evaluating pragmatic reasoning about conversational implicature in question utterances, designed to isolate pragmatic variation while holding each question's surface form fixed.

To support scalable evaluation, we propose a semi-automated pipeline that produces question-context-interpretation instances with systematic variation. Across evaluations, we find a consistent generation-inference asymmetry: while state-of-the-art models can generate plausible pragmatic scenarios when guided, they often fail to recover the intended implication at inference time. For smaller models, structured prompting improves alignment with human judgments. A comparative writing study further reveals complementary strengths: human authors tend to produce safer, predictable contexts, whereas models generate varied scenarios with interpretations that sometimes exceed contextual support. These findings highlight persistent challenges in modeling conversational implicature and motivate more context-sensitive evaluation frameworks.

1 Introduction

A fundamental aspect of human communication is the ability to infer meaning that goes beyond what is explicitly stated. Speakers routinely rely on shared common sense, social knowledge, and cooperative assumptions to encode and decipher these "unspoken" messages (Grice, 1989; Searle, 1975). These inferences are typically predictable and reliable, allowing communication to remain efficient even in unfamiliar or unexpected situations (Levinson, 2000).

Figure 1: How conversational context reshapes implicature. The main takeaway from the utterance shifts from a genuine inquiry (scene 1) to a rhetorical rebuke (scene 2) or a desperate plea (scene 3).

As conversational agents and AI assistants are increasingly deployed in everyday settings, they are expected to handle communication styles that extends beyond mere information delivery (Srikanth et al., 2024). While pretrained large language models (LLMs) acquire substantial social commonsense understanding from training data alone (Talmor et al., 2019), they remain less competent at utilizing that information to make necessary inferences in veiled language use (Setlur and Tory, 2022; Miehling et al., 2024). One such phenomenon, *conversational implicature*, arises from how context reshapes meaning of an otherwise ordinary phrase, making it difficult to formalize and scale into an annotated resource (George and Mamidi, 2020).

We introduce **DRInQ** (Dataset for Recovering Implicature in Questions), a context-controlled

[0]Dataset available at https://github.com/hjarai/drinq

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34594–34611

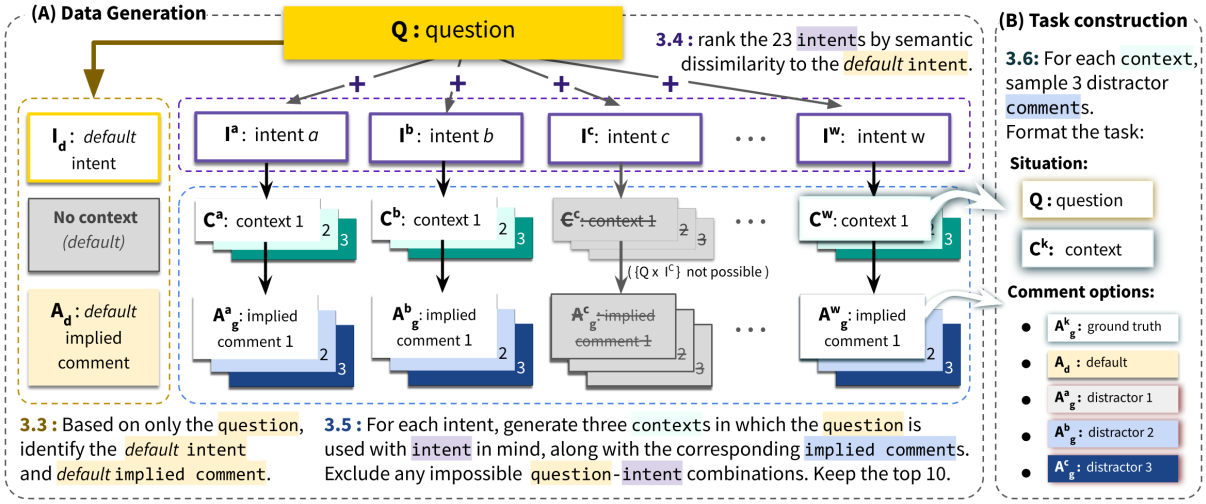


Figure 2: Pipeline for generating DRInQ task datapoints.

benchmark for probing discriminative pragmatic reasoning in question utterances. Canonical examples such as the indirect request "*Can you pass the salt?*" motivate our focus on questions, since interrogative forms routinely serve functions beyond literal inquiry (Searle, 1975). As illustrated in Fig 1, circumstance alone can shift the impression listeners would draw from the same utterance, making such context-driven variation difficult to capture at scale. To address this challenge, we compose multiple contextual framings for each base question using a semi-automated, structured generation pipeline that supports scalable, context-sensitive data construction. This formulation enables a controlled study of how contextual variation reshapes implicature. We evaluate with multiple-choice contrasts as a diagnostic probe: they offer minimal pairs, efficient, low-variance scoring, and clear attribution, complementing recent generation-based measures of pragmatic competence.

Evaluating 12 state-of-the-art language models on **DRInQ**, we find that while models perform well overall, their inferences diverge systematically from human judgments. In particular, models frequently select interpretations that humans judge as overly strong or insufficiently supported by the context. On a challenging human-verified subset, the best-performing model (GPT-4o) reaches 76% accuracy, compared to a human annotator average of 89%. We further observe that structured prompting improves alignment with human judgments for smaller models, narrowing the performance gap with larger models. A comparative writing study reveals complementary strengths in data generation: human authors tend to produce conservative, underspecified contexts, whereas LLMs generate more diverse scenarios but often over-specify the implied interpretation. Together, these results highlight both the value of controlled, context-sensitive benchmarks for studying conversational implicature and the persistent challenges LLMs face in calibrating pragmatic inferences.

2 Related Works

Computational pragmatics is theoretically rooted in Grice's Cooperative Principle, which characterizes communication as guided by shared assumptions of cooperation (Grice, 1989; Krause and Vossen, 2024). Text records of non-cooperative language (e.g., sarcasm, deception) rarely explain the inferential process for recovering intended meaning (Shwartz and Choi, 2020; Sravanthi et al., 2025; Banou et al., 2025), thus models often struggle when interpretation depends on unspoken conversational norms rather than surface semantics (Ruis et al., 2022; Chang and Bergen, 2023; Barattieri di San Pietro et al., 2023; Tao et al., 2024). Across pragmatic benchmarks, a consistent finding is that LLMs exhibit a strong *literalist bias*, defaulting to surface-level interpretations even when pragmatic alternatives are licensed (Tong et al., 2024; Saakyan et al., 2025). Empirically, partial linguistic cues used as scaffolding help SOTA LLMs on difficult indirect-meaning tasks (e.g., discourse relations, common ground)(Miao et al., 2024; Qiu et al., 2025; Wan et al., 2025), motivating continued development of resources that systematically provide and evaluate such cues.

Pragmatic Benchmarks and Tasks To address the scarcity of annotated pragmatic data, prior work has introduced datasets targeting non-literal language. Early efforts adapted natural language inference (NLI) formulations to probe pragmatic understanding, including social commonsense (Sap et al., 2019; Zellers et al., 2019), implicature and presupposition (Jeretic et al., 2020; Stowe et al., 2022), scalar inference (Schuster et al., 2020), and indirect answers to polar questions (Louis et al., 2020).

Subsequent datasets expanded both scope and size, though these resources still focus on phenomena that are relatively well-defined or tied to identifiable lexical cues. IMPRES (Jeretic et al., 2020) and GRICE (Zheng et al., 2021) provide controlled tests of implicature and presupposition via linguistically guided generation, while FLUTE (Chakrabarty et al., 2022; Kulkarni et al., 2024; Park et al., 2025) evaluates figurative language understanding including sarcasm, metaphor, and idiom. Aggregated evaluations show that instruction-tuned models approach human performance on some benchmarks, but results vary substantially across phenomena (Sravanthi et al., 2024).

Linguisically-guided data generation Most existing dataset classifications rely on coarse labels (i.e., literal vs. non-literal) which oversimplify meaningful distinctions, while more comprehensive studies require costly expert involvement to source, annotate, and validate (Hu et al., 2023). Given the cost and difficulty of manual annotation, many pragmatic benchmarks rely on (semi-)synthetic data generation. Prior work has explored linguistically-grounded rule-based generation (Jeretic et al., 2020; Zheng et al., 2021), pattern-based extraction (Parrish et al., 2021; Yue et al., 2024), and human-AI collaboration (Chakrabarty et al., 2022). While such approaches enable scale, surveys note persistent challenges in maintaining pragmatic diversity and realism particularly for implicature (George and Mamidi, 2020; Ma et al., 2025).

Our work builds on this literature by focusing specifically on *question utterances*, where the divergence between surface form and communicative intent is often widest (Yusupujiang and Ginzburg, 2023). Unlike prior benchmarks that emphasize categorical labels and isolated phenomena, DRInQ targets fine-grained variation in conversational implicature induced solely by contextual differences. By combining a linguistically grounded intent framework with human-verified, model-in-the-loop generation, we aim to support scalable evaluation while preserving the multi-faceted characteristic of real-world pragmatic inference.

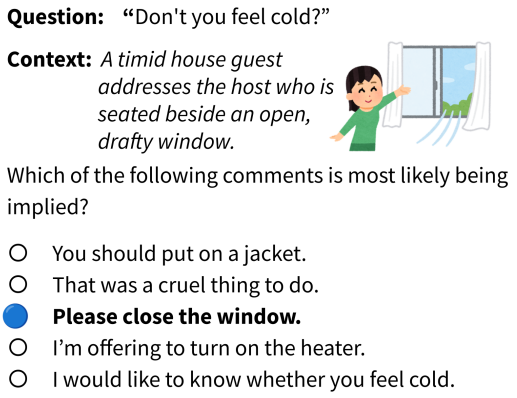


Figure 3: DRInQ task example. A single question utterance is paired with a contextual framing and multiple candidate interpretations.

3 The DRInQ Task

In natural conversation, implied meaning is often underspecified by utterance alone. A question like "*Don't you feel cold?*" rarely functions as purely a literal inquiry; it may signal a rhetorical critique or indirectly prompt a specific action (Searle, 1975) (see Fig. 3). While a clarifying follow-up could resolve the ambiguity, such interventions are often less desirable, particularly when indirectness is motivated by politeness, face-saving, or efficiency (Brown and Levinson, 1987).

3.1 Evaluation Desiderata

The reliance of conversational implicature on indirectness presents a distinct evaluation challenge for LLM inference. The difficulty is not in recognizing that an utterance *can* convey implied meaning, but in determining *which* contextual cues license a particular interpretation. While recent LLMs can readily track explicit state changes, they often struggle to separate pragmatically salient contextual information from incidental detail (Shi et al., 2023).

The failure to make this distinction amplifies the difficulty of designing and generating evaluation scenarios that probe this contrast. Although surface form wordings are easy to perturb, only a subset of contextual changes meaningfully affect interpretation (Ma et al., 2025). For example, "Can you pass the salt?" remains an indirect request whether it is asked indoors or outdoors, while for other questions these same factors carry significant pragmatic

Category	Definition	Examples
Directive	Commits the listener to some future task.	*prohibit; request; seek information*
Assertive	Conveys some information.	*predict; rhetorical; report; conclude*
Commissive	Commits the speaker to some future task.	*promise; warn; invite; offer*
Expressive	Expresses the speaker's emotion.	*thank; complain; apologize; insult*

Table 1: Intent labels (derived from speech act categories) used in the DRInQ generation pipeline

weight.

To systematically control such interpretive shifts during dataset generation, we draw on the notion of *speech acts* as a coarse-grained representation of a speaker's intended action. Following classic work in speech act theory (Searle, 1975, 1976; Austin et al., 1975), we treat question utterances as capable of realizing distinct functions (e.g., requesting, criticizing, reassuring) depending on context, even when their surface form remains *unchanged*. We use a structured inventory of speech act verbs (subsequently referred to as *intent* labels) to guide the construction of contexts that license meaningfully different interpretations. For example, for the question in Figure 1, scene 2 might have been produced with the intent to *scold*, and scene 3 to *confess*. Table 5 presents a subset of the 23 intent labels we use; Appendix 8.1 provides additional details on the taxonomy and its role in the generation pipeline.

3.2 Task Formulation

To translate the desiderata outlined above into a concrete evaluation setting, we formalize a multi-choice task that tests a model's ability to recover a speaker's intended meaning from a context. Each instance consists of a question utterance, a contextual framing, and a set of candidate interpretations, exactly one of which is licensed by the given context. Formally, each instance is represented as a tuple $(Q, C, \mathcal{A})$, where:

- Q is a *question utterance*, i.e., the surface form of the query.
- C is the *context*, describing relevant situational and social factors relevant to the interpretation.
- $\mathcal{A} = \{A_i\}_{i=1}^{5}$ is a set of *implied comments* or candidate interpretations, with exactly one correct answer.

The remaining options are not arbitrary distractors: each corresponds to a plausible reading of Q under

Question: "Have you done your homework?"

C : Context	I : Intent	A : Implied comment
Default: *The question is interpreted without a context*		
– –	to seek information	*I want to know if you've done your homework.*
Alternative: *Using intent, the question is interpreted with context*		
A dad sternly asks his daughter, who is gaming	to scold	*Stop playing and go do your homework now!*
A panicked student asks her concerned classmates	to assert	*I haven't finished my homework!*
A tutee is asked after calling to cancel a session.	to confirm	*I doubt you've finished your homework.*
During office hours, the TA realizes that they'd posted the wrong assignment.	to express regret	*I didn't want you to do that homework.*
Impossible: *Abstained, as the question is unlikely to serve these intents*		
– –	to prohibit	– –
– –	to promise	– –

Figure 4: Example of text generated in Steps 3.3-3.6 of the pipeline on the base question "Have you done your homework?" The cells outlined in orange indicate the (C, A) task components.

a different contextual framing. As a result, success on this task cannot rely on lexical cues alone, but instead requires reasoning about how contextual factors modulate communicative intent.
The task design is guided by three principles:

1. **Context sufficiency:** The context C must provide enough information to uniquely determine the intended interpretation of Q.
2. **Distractor plausibility:** Each incorrect A_i must represent a reasonable interpretation of Q in some alternative context.
3. **Controlled variation:** Contextual manipulations should isolate pragmatically meaningful factors rather than incidental variation.

To satisfy these constraints, we adopt a minimal-contrast strategy common in linguistic analysis: the question Q is held fixed while the context C is systematically varied. This allows us to probe how different contextual factors license different implied meanings and to capture the range of interpretations a single utterance may convey.

These principles motivate a semi-automated generation pipeline (Fig. 2). In the following, we detail each stage of this process, including base question curation, intent selection, context generation, task formatting, and human verification.

3.3 Base Question Curation

Most existing question datasets are QA- or task-oriented rather than casual conversation. As ensure

natural, colloquial surface forms, we hand-selected 30 seed questions from everyday settings. We then used GPT-4o (OpenAI, 2024) to expand these to 300 base questions (the fixed surface forms Q), constraining generation to an informal register, a mix of wh- or polar questions, and non–knowledge-heavy topics. A brief manual pass removed unnatural outputs and duplicated paraphrases. For each Q, we elicit a context-independent baseline interpretation with a default intent I_d and default implied comment A_d, which serve as references when varying context.

3.4 Intent Ranking and Selection

For each question Q, we rank the remaining 22 intent labels by cosine distance between their semantic embeddings (Reimers and Gurevych, 2019) to those of the default implied comment A_d. These ranked intents are used to seed the generation of novel contexts that license alternative interpretations (Figure 9).

3.5 Context Generation

For each selected pair (Q, I), we generate three context-interpretation pairs $\{(C_j, A_j)\}_{j=1}^{3}$. GPT-4o is provided with in-context examples sampled from a manually curated pool and is prompted to construct a scenario in which the speaker uses Q to indirectly realize the target intent I. To maintain quality, the model is instructed to abstain from answering if a specific $Q \times I$ combination is pragmatically implausible. We retain only the ten highest-ranked (Q, I) for which valid (C, A) pairs are produced. To ensure non-trivial implicature, we further filter out cases where the contextualized interpretation A_g is semantically too similar to the default interpretation A_d. The final dataset contains 300 questions, each associated with at least five distinct intents and three context-interpretation variants per intent. Example few-shot example prompts are provided in Appendix 8.

3.6 Translating into Task Format

We next map each generated context C into the multiple-choice task format defined in Section 3.2. Each instance is represented as a tuple $(Q, C, \mathcal{A})$, where $\mathcal{A} = \{A_g^k, A_d\} \cup \{A_g^j \mid j \neq k\}$. Here, A_g^k denotes the originally-generated implied comment conditioned on question Q and target intent I^k, A_d denotes the default implied comment for Q, and A_g^j denotes implied comments corresponding to alternative intents $I^j \neq I^k$. Superscripts index the conditioning intent, while subscripts distinguish default (d) from originally-generated, context-licensed (g) interpretations.

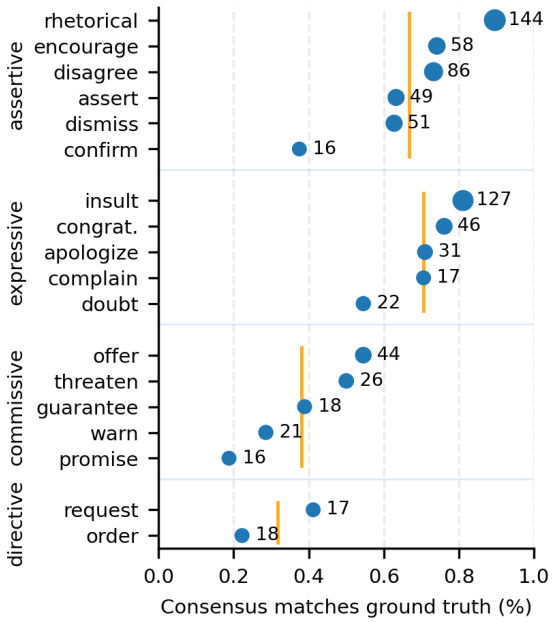


Figure 5: Proportion of question-context pairs (Q, C) for which the human consensus A_c matched the originally-generated comment A_g by intent category. Dot size and numeric labels indicate the number of datapoints per intent. Vertical lines denote the category-wise mean.

3.7 Human Verification and Consensus

To validate and measure human agreement on the dataset, we present the generated instances to 62 pre-screened annotators recruited via Prolific. We retain only datapoints for which at least 4 out of 5 annotators agree on the same implied comment, yielding 819 instances.

From the validated pool, we further identify a subset of 400 challenging instances for benchmarking. These tasks are characterized either by lower annotator-model agreement or by disagreement among annotators relative to the originally generated label, yielding a high-quality evaluation set for testing model performance on difficult pragmatic reasoning.

Under standard sampling, across this subset, agreement between the originally generated implied comment A_g and the human consensus A_c is 81%. However, the final validated dataset is intentionally skewed towards more challenging instances, resulting in an overall agreement of 67%. Within this set, agreement drops to 27% on the hard baseline subset, but reaches 94% on the remaining

Model	Vanilla	Explanation
GPT-5-Nano	0.45 ± 0.07	0.43 ± 0.02
GPT-5-Mini	0.56 ± 0.05	0.62 ± 0.06
GPT-4o	0.62 ± 0.03	0.63 ± 0.03
GPT-4.1	0.61 ± 0.02	0.61 ± 0.03
GPT-5	0.62 ± 0.01	0.63 ± 0.01
OpenAI-o3	0.67 ± 0.02	0.67 ± 0.03
Claude-Sonnet-4.5	0.65 ± 0.01	0.62 ± 0.02
Claude-Haiku-4.5	0.64 ± 0.01	0.63 ± 0.02
Llama-3.3-70B	0.58 ± 0.02	0.58 ± 0.03
Qwen2.5-72B	0.56 ± 0.03	0.57 ± 0.03
DeepSeek-V3	0.60 ± 0.03	0.61 ± 0.01
DeepSeek-R1	0.62 ± 0.04	0.60 ± 0.04
Human Avg	0.88 ± 0.10	–

Table 2: Hard subset (400 items) accuracy: mean ± SD over 3 runs (randomly sampling 3/6 in-context examples), vanilla vs. explanation prompting. 95% item-level bootstrap CIs and paired significance in Appendix 8.6

Error Type	Malintent Attribution	Over-fixation on Detail
Question	Do you need me to carry some of your shopping?	Could you please turn down the volume?
Context	*An older sibling confidently asks their younger sibling, who is struggling with a heavy load of groceries.*	*A parent who ignored the child's earlier request for louder music asks with a sheepish expression.*
Human Consensus	I will help you, don't worry.	Please lower the volume.
LLM Choice	You clearly can't handle that by yourself.	I apologize for not considering your love for music.

Table 3: Representative error cases for LLMs (gpt-o3 and claude-haiku-5.4), contrasted with human consensus judgments.

instances.

Figure 5 summarizes agreement stratified by intent labels. Agreement is highest for questions functioning as rhetorical commentary (assertive) or as insults (expressive), where implied meaning is typically overt, approaching 90%. In contrast, agreement is lower for commissive and directive uses, which require more nuanced reasoning about indirect offers or requests.

4 Evaluation on Model Inference

We evaluate LLM performance on **DRInQ** to assess their ability to recover conversational implicature from context. We report baseline results across state-of-the-art models, analyze systematic error patterns relative to human judgments, and examine whether targeted prompting scaffolds improve pragmatic inference.

4.1 Baseline comparison of SOTA models

We benchmark twelve SOTA models from five model families on the DRInQ task. under two prompting conditions: `vanilla`: a standard few-shot prompt with three in-context examples sampled from six manually written instances, and `explanation`: an enhanced prompt that explicitly requires models to produce a written justification before answering. We adopt explanation-based prompting rather than chain-of-thought, following evidence that explicit step-by-step reasoning can hinder performance on tasks requiring intuitive inference (Yao et al., 2025; Liu et al., 2025). We omit zero-shot or one-shot settings, as preliminary experiments indicate that minimal prompting fails to convey task conditions.

We report the results[1] in Table 2. Compared to the human annotator average baseline of 89%, most SOTA models cluster within a narrow performance range around 60%. Within the closed models, GPT-4o, which was used to generate the data, does well, but is out-performed by GPT-o3, as well as the models from Anthropic. Of the open-source models, DeepSeek-V3 and Llama-3.3-70B-Instruct-Turbo performed best with the `vanilla` and `explanation` variations, respectively. Overall, the reasoning-oriented variants from both the GPT and DeepSeek families outperform their chat-oriented counterparts, suggesting that enforcing explicit deliberation might aid in reasoning about conversational implicature.

Across the benchmark, we observe systematic differences in how human annotators and LLMs resolve implied meaning.

Malintent attribution: Human readers tend to adopt a more charitable interpretation of speaker intent and tone: in the absence of explicit cues, they are reluctant to attribute hostile or morally negative intentions. A recurring model error arises when LLMs overemphasize isolated negative details, selecting a harsher interpretation than one supported by the human consensus (Table 3, left). In this ex-

[1]Accuracies are computed on a fixed test set of 400 instances; differences of 1–2% fall within statistical uncertainty. Model names are abbreviated for readability; full model identifiers are listed in Appendix 7

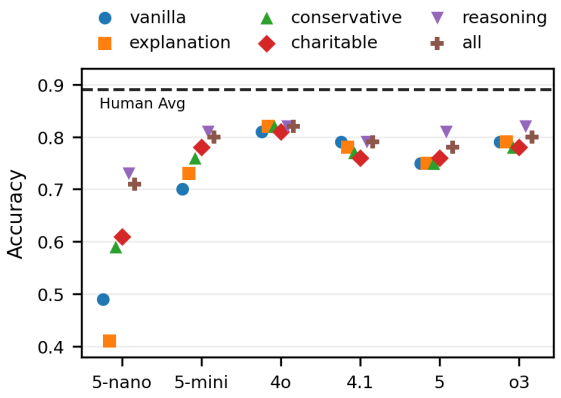


Figure 6: Benchmark performance of the four proposed structured prompting strategies on six GPT models.

ample, models appear to infer condescension from the speaker's "confident demeanor," whereas annotators infer cooperative intent given the sibling dynamic.

Over-fixation on detail: A second error pattern concerns *contextual sufficiency*. In some cases, the originally-generated implied comment is logically compatible with the context but requires additional assumptions that human listeners typically do not make. While LLMs correctly identify relevant semantic cues, they often miscalibrate the strength of the inference, endorsing interpretations that are possible but pragmatically over-committed. In the situation shown in Table 3 (right), the model places disproportionate weight on the parent's "sheepish" expression, attributing it to regret over the earlier decision, while annotators instead interpret it as social awkwardness in making a repeated request. The divergence reflects not a a failure of semantic reasoning, but a mismatch in estimating which implicatures are warranted by the available evidence.

5 Improved Prompts for the task

Setup To address the failure modes identified in the previous section, we evaluate prompt-based interventions aimed at eliciting more human-like pragmatic inference. Prior work suggests that errors on judgments that are intuitive for humans can be mitigated by constraining how models reason through the presuppositions and inference (Sravanthi et al., 2025; Weston and Sukhbaatar, 2023). Building on these insights, we test four targeted prompts that either directly counter the observed error patterns or explicitly scaffold pragmatic reasoning.

The *Conservativeness Constraint* prompt (`conservative`) discourages over-commitment to merely compatible interpretations, while the *Charitable Interpretation* prompt (`charitable`) discourages unsupported attribution of hostile intent. We additionally test a *Pragmatic Reasoning Scaffold* prompt (`reasoning`), inspired by prior work on constrained, task-aware prompting for intuitive linguistic judgments (Yue et al., 2024; Yao et al., 2025; Lee et al., 2025), which models a cooperative listener's reasoning process, requiring the model to sequentially consider the surface question, contextual evidence, and inferred speaker intent before selecting an interpretation. The *Combined Scaffold* (`all`) condition combines all three. The detailed differences in the final prompts are available in Appendix 8.4.

Results As shown in Figure 6, prompt-based interventions yield consistent improvements for smaller-capacity models, while having little to no effect on larger models; On the full annotated dataset, `gpt-5-nano` improves from 41% to 73% accuracy, and `gpt-5-mini` improves from 71% to 81%, narrowing the gap to the performance ceiling achieved by `gpt-4o` (82%). Among the targeted strategies, `conservative` and `charitable` produce comparable gains, while the `reasoning` prompt yields the largest performance increase. This pattern suggests that explicitly scaffolding pragmatic reasoning is especially beneficial under limited model capacity. Nevertheless, across all prompting conditions, performance remains below the human annotator average, indicating that prompt-based interventions mitigate but do not fully overcome capacity-related limitations in pragmatic inference.

6 Comparing LLM and human writing

6.1 Human Baseline and Generation Study

To assess the comparative quality of our machine-generated data, we conducted a human writing study on Prolific. The study mirrors the second stage of our generation pipeline: given question-intent pair (Q, I), annotators were asked to write a contextual framing and a corresponding implied comment (C, A_g). Annotators were screened using a qualification task assessing their ability to meet criteria of novelty, faithfulness, and sensibility. From our qualified pool, 16 annotators contributed writing for 64 (Q, I) pairs drawn from the human-consensus subset of the dataset, with each

pair written by three annotators, yielding a human-written baseline of 192 context-implied comment pairs[2]. Full instructions are provided in the Appendix (Fig 11). On average, annotators required 2 minutes and 4 seconds per datapoint (SD: 55s), underscoring the cognitive demands of devising pragmatically appropriate conversational contexts.

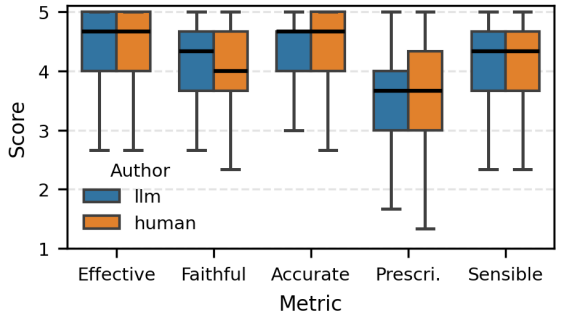


Figure 7: Distribution of ratings for LLM and human-authored context-interpretation pairs across five pragmatic quality metrics.

6.2 Context Quality Evaluation Methods

Metrics To assess how well individual datapoints satisfy the desiderata outlined in Section 3.2, we conducted a separate human perception study assessing both human- and LLM-authored context-interpretation pairs (C, A) on a 5-point Likert scale. Each metric operationalizes a distinct aspect of pragmatic quality targeted by our generation procedure: *Effectiveness* (whether the context shifts interpretation away from the default baseline A_d), *Faithfulness* (support for the specified intent I), *Accuracy* (likelihood of inferring the target implied comment A_g), *Non-prescriptiveness* (whether intent is implied rather than explicitly stated), and *Sensibility* (internal coherence and commonsense plausibility). Because novelty is inherently comparative, we assess *perceived novelty* using pairwise human judgments rather than absolute ratings. We define *Novelty* as the degree to which a response is unexpected relative to plausible alternatives. Appendix 8.7 details the full metric definitions and annotation prompts, and Figure 7 shows the distribution of scores across these dimensions for both human- and LLM-authored datapoints.

For each of the 192 (Q, I) pairs, we compare three human-written and three LLM-generated contexts. Annotators compare randomly sampled cross-group pairs and judge which context is more novel. The resulting nine pairwise judgments per (Q, I) are aggregated to produce a group-level outcome. Overall, LLM-authored contexts are judged more novel in 37% of cases, compared to 22% for human-authored contexts, with no clear preference in 40% of cases.

[2] 32 questions × 2 intents × 3 annotations

"Are we having dinner at home tonight?" *(Insult)*	
Human	A husband asks his wife after a failed home-cooked meal → *The cooking is of a low standard.*
LLM	An adult child mockingly asks their parent after seeing a messy dining area → *This place is a mess; you can't manage the house.*

Table 4: Human and LLM (GPT-4o) writings of context and interpretation for the same question and intent.

6.3 LLM vs. human-written situations

Across evaluation dimensions, LLM-generated contexts largely overlap with human-written ones on *Effectiveness*, *Faithfulness*, and *Sensibility*. Human writers score higher on *Accuracy*, indicating closer alignment between their authored context and intended implied comment. Both groups exhibit similar central tendencies on *Non-prescriptiveness*, though human-written contexts show higher variance, suggesting less consistent control over keeping internal motivations implicit. Overall, these results indicate a trade-off between contextual leverage and interpretive precision. LLM-generated contexts more reliably push interpretations towards the specified intent, whereas human-written contexts more consistently support the intended comment. This pattern is illustrated in the human-authored example in Table 9: although the intended interpretation is plausible, it depends on unstated assumptions about tone, relationship dynamics, and prior events. In the absence of these details, alternative readings (e.g., consoling a disappointed partner or expressing gratitude for the attempt) remain viable, reducing the perceived necessity of the intended implicature.

Novelty judgments further reveal complementary strengths. While many comparisons result in ties, LLM-generated contexts are more often judged novel, reflecting a greater willingness to explore non-canonical social configurations. In the LLM-authored example (Table 9), a grown child mocks their parent – an atypical arrangement that departs from familiar power dynamics. Human authors, by contrast, tend to default to more conventional role relations, producing scenarios that are less novel but more pragmatically stable.

Overall, these findings suggest that human authors prioritize interpretive plausibility grounded in shared social knowledge, whereas LLMs favor explicit contextual cues and exploratory configurations, yielding more overt but sometimes less intuitive implicatures.

7 Conclusion

We introduced **DRInQ**, a benchmark and semi-automated pipeline for constructing implicature-driven question-context-interpretation triples that enables controlled evaluation of pragmatic reasoning under contextual variation. By holding surface form constant while systematically manipulating context, **DRInQ** reveals interpretive distinctions that are difficult to observe with existing pragmatic benchmarks. Across evaluations, we find that many LLM errors reflect systematic over-commitment to inferred meaning, particularly in calibrating how much the context reasonably warrants beyond the literal question. For smaller models, structured prompting narrows the performance gap relative to the strongest benchmarks. Our human-LLM writing study further highlights complementary strengths and weaknesses: human writers tend to produce more prototypical, low-risk contexts, while LLMs generate more diverse scenarios that over-specify implied meaning. Together, these results suggest that progress on pragmatic reasoning may require not only stronger models, but also evaluation frameworks that explicitly encode the underspecification inherent to everyday communication.

Limitations

A growing body of recent work has established the inadequacies of multiple-choice task setups for measuring pragmatic reasoning in LLMs (Yerukola et al., 2024; Yu et al., 2026). While **DRInQ** aims to capture a broader range of context-dependent implicatures, it remains a discriminative evaluation with a fixed interpretation space, which may obscure alternative but reasonable pragmatic inferences. A further limitation concerns the relationship between our evaluation setup and the broader goal of conversational competence. Ultimately, pragmatic understanding should be assessed through a model's ability to generate appropriate responses in dialogue. Our multiple-choice formulation can of only approximate this ability, serving as a controlled proxy rather than a direct measure of generative implicature recovery. MCQ offers: (i) controlled minimal contrasts (ii) scalable evaluation and low-variance scoring and (iii) attribution (to identify *which* alternative interpretation a model preferred). Thus, we view this task as a minimal-threshold diagnostic probe which complements generation-based evaluations of pragmatic competence. In addition, our benchmark assumes that intent-conditioned interpretations generated during data construction constitute valid ground truth. While quality analysis and human agreement support this assumption in the majority of cases, we find that only 80.8% of instances achieve our human consensus threshold of 4/5 annotators. This indicates that such interpretations cannot yet be treated as indisputable targets and underscores the need for future work that integrates uncertainty-aware or generative evaluation of pragmatic inference.

Ethics Statement

Human Subjects and Compensation We recruited 78 participants via Prolific for dataset verification and writing tasks. All annotators were compensated at an hourly rate of at least $12, in accordance with Prolific's fair payment guidelines. We collected no personally identifiable information during this process, and all data were anonymized prior to analysis.

Cultural Limitations We acknowledge that our dataset is English-centric and reflects Anglophone social norms (Li et al., 2024). The annotation was conducted by our vetted Prolific participants, whose subjective judgments have been informed by their culture and language; we report the annotators' demographic distribution in Table 6. Because conversational implicature is deeply culturally situated, this resource is unlikely to generalize to non-Western or low-context communicative settings.

Downstream Harm Improving LLM inference of user intent increases dual-use risks, including the amplification of *dog whistles* or coded hate speech that evades moderation (Mendelsohn et al., 2023), as well as *implicit profiling*, where models might infer sensitive user attributes from obstensively innocuous dialogue (Weidinger et al., 2021). Despite the risk of misuse enabled by improvements in downstream models, we argue that the long-term benefits for pragmatic capabilities and defensive research outweigh the associated risks.

Environmental Impact While our model-in-the-loop framework relies on API-inference rather than pretraining, we acknowledge the aggregate carbon footprint of large-scale querying (Strubell et al., 2019). By releasing this benchmark as a static source, we aim to reduce the need for redundant high-volume data generation by future researchers.

Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200006, the Defense Advanced Research Projects Agency with award HR00112220046, and NSF IIS 2048211. We thank our annotators for their contribution, and would like to thank the collaborators at the INK research lab at USC for their constructive feedback on the work.

References

J.L. Austin, J.O. Urmson, and M. Sbisà. 1975. *How to Do Things with Words: Second Edition*. Harvard paperback. Harvard University Press.

Zouheir Banou, Sanaa El Filali, El Habib Benlahmar, Fatima-Zahra Alaoui, Laila El Jiani, and Hasnae Sakhi. 2025. A systematic review of figurative language detection: Methods, challenges, and multilingual perspectives. *Natural Language Processing Journal*, 13:100192.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi Intelligenti*, XXXV:379–400.

P. Brown and S.C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Politeness: Some Universals in Language Usage. Cambridge University Press.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50:293–350.

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

Herbert Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press, Cambridge.

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, Toronto, Canada. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Lea Krause and Piek T.J.M. Vossen. 2024. The Gricean maxims in NLP - a survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 470–485, Tokyo, Japan. Association for Computational Linguistics.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the FigLang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. Pragmatic metacognitive prompting improves llm performance on sarcasm detection. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 63–70.

Stephen C Levinson. 2000. *Presumptive meanings*. Language, Speech, and Communication. Bradford Books, Cambridge, MA.

Huihan Li, Liwei Jiang, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *CoRR*, abs/2404.10199.

Ryan Liu, Jiayi Geng, {Addison J.} Wu, Ilia Sucholutsky, Tania Lombrozo, and {Thomas L.} Griffiths. 2025. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *Proceedings of Machine Learning Research*, 267:38489–38517. Publisher Copyright: © 2025, by the authors.; 42nd International Conference on Machine Learning, ICML 2025 ; Conference date: 13-07-2025 Through 19-07-2025.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "i'd rather just go to bed": Understanding indirect answers. In *Conference on Empirical Methods in Natural Language Processing*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M. Daly, David Piorkowski, and John T. Richards. 2024. Language models in dialogue: Conversational maxims for human-AI interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14420–14437, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card.

Seoyoon Park, Hyeji Choi, Minseon Kim, Subin An, Xiaonan Wang, Gyuri Choi, and Hansaem Kim. 2025. FLUID QA: A multilingual benchmark for figurative language usage in dialogue across English, Chinese, and Korean. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30268–30282, Suzhou, China. Association for Computational Linguistics.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*.

Linlu Qiu, Cedegao E. Zhang, Joshua B. Tenenbaum, Yoon Kim, and Roger P. Levy. 2025. On the same wavelength? evaluating pragmatic reasoning in language models across broad concepts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19913–19935, Suzhou, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktaschel, and Edward Grefenstette. 2022. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

John R. Searle. 1975. *Indirect Speech Acts*, pages 59 – 82. Brill, Leiden, The Netherlands.

John R Searle. 1976. A classification of illocutionary acts. *Language in society*, 5(1):1–23.

Vidya Setlur and Melanie Tory. 2022. How do you converse with an analytical chatbot? revisiting gricean maxims for designing analytical conversational behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

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

Settaluri Lakshmi Sravanthi, Kishan Maharaj, Sravani Gunnu, Abhijit Mishra, and Pushpak Bhattacharyya. 2025. Understand the implication: Learning to think for pragmatic understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23778–23790, Vienna, Austria. Association for Computational Linguistics.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024. Chatgpt role-play dataset: Analysis of user motives and model naturalness. In *International Conference on Language Resources and Evaluation*.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Stephen Wan, Wei Liu, and Michael Strube. 2025. On the role of context for discourse relation classification in scientific writing. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 96–106, Suzhou, China. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Jason Weston and Sainbayar Sukhbaatar. 2023. System 2 attention (is something you might need too). In *Advances in Neural Information Processing Systems*.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2025. Is sarcasm detection a step-by-step reasoning process in large language models? *Proceedings of the AAAI Conference on Artificial Intelligence*.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. 2026. The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models.

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*.

Zulipiye Yusupujiang and Jonathan Ginzburg. 2023. Unravelling indirect answers to wh-questions: Corpus construction, analysis, and generation. pages 336–348.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

8 Appendix

8.1 Using Speech Acts as Intent Labels

Drawing on Searle's speech acts (Searle, 1976; Austin et al., 1975), we adopt a coarse-grained taxonomy of illocutionary acts as a way to characterize the functional roles a question utterance may serve in context. Although speech act theory is a rich and contested linguistic framework, we use it here as a practical abstraction to structure pragmatic variation during data generation.

When a speaker produces an utterance, its *illocutionary force* corresponds to the action performed by speaking – what they are *doing* – rather than the propositional content alone. For example, an utterance may function to insult, warn, or invite, even when its surface form is interrogative. Individual act verbs can be grouped into broader **illocutionary points** which reflect their its essential communicative purpose.

Act Category	Definition	Examples
Directive	Commits the listener to some future action.	*prohibit; request; seek information; order; advise*
Assertive	Represents a state of affairs as true.	*predict; report; conclude; inform; make rhetorical comment*
Commissive	Commits the speaker to some future action.	*promise; warn; invite; threaten; offer; guarantee*
Expressive	Expresses the speaker's psychological state or emotion.	*thank; complain; apologize; insult; greet; support*

Table 5: Speech act categories used

We consider four illocutionary points, **Directives, Assertives, Commissives, and Expressives**, from which we aggregate 23 representative act verbs to support controlled variation in pragmatic interpretation. (Table 5). We omit the fifth category **Declarations** (brings about an institutional change in the world, e.g. *"I resign"* or *"You are fired"*), as such acts rarely present as a question.

8.2 Prompt Templates for Generating context-interpretation (C, I) pairs:

Instructions:

```
You are an insightful, creative expert in
    pragmatics. Often, people will say one
    thing, but mean something else.
```

```
You are given a "question" and "intent".

Imagine a scenario in which someone asks this
     "question," but with a deeper meaning
    because they had the goal to "intent".
    Your job is to generate a "context" that
    would make the "intent" true and obvious,
     then identify the "implied comment".

"context" is a description of the situation
    in which the question is asked (e.g. the
    speaker's tone or expression, the speaker
     and listeners' relationship or power
    dynamic, social conventions, history of
    the conversation, past events, location
    etc.), which makes the "intended comment"
     apparent. "context" should not include
    interpretations/implications of the
    situation, nor the speaker's internal
    hopes, nor information about the future.

"Implied comment" is the message being
    implied by the "question," as a brief
    statement.

Abstain from generating (respond with an
    empty list) if the "question" could never
     be used to imply "implied comment".
    Otherwise, generate 3 sets of these
    answers, formatted as a list of
    dictionaries."
```

In-context Examples

```
Question 1: Are you going out?
Context: The speaker points at a full garbage
     can.
Implied meaning: Please take out the trash on
     your way out.
Context: The speaker looks at the clock,
    which shows midnight.
Implied meaning: Why are you leaving the
    house so late?
Context: The listener is wearing pajamas.
Implied meaning: Why are you leaving the
    house in pajamas?

Question 2: Do you have a pencil?
Context: The speaker is in class without a
    pencil.
Implied meaning: Please let me borrow your
    pencil.
Context: The listener needs to write
    something down.
Implied meaning: Would you like to borrow my
    pencil?
Context: The listener is writing on a plastic
     bottle.
Implied meaning: Why are you trying to write
    on plastic with a pencil?
```

8.3 Intent Selection for Context Generation

To induce diverse pragmatic interpretations while holding the surface form of each question fixed, we seed context generation with intents that are

semantically distant from the question's default reading. Figure 9 provides supporting statistics on the resulting intent distributions.

8.4 Prompt Templates for Evaluation

Core Prompt

```
Instructions:
Based on the conversational context, choose
    the option from the given list that most
    accurately reflects the speaker's
    intended meaning when they ask the
    question.

## Output Format
Return your response as a JSON object in the
    following format:
‘‘‘
{
  "selected_option": "<letter and full text
    of the selected option>",
  "explanation": "<concise explanation for
    why this option was selected>"
}
‘‘‘
Replace ‘<letter and full text of the
    selected option>‘ with your choice, using
     the format provided in the list. The
    explanation should briefly justify your
    selection. Ensure the output strictly
    matches the specified JSON structure. If
    the output structure deviates, self-
    correct before returning your final
    answer.
```

***Pragmatic Reasoning* prompt**

```
Add instructions:
1. Identify the intent of the question as
    stated, without considering context.
2. Now incorporate context, reasoning whether
     and how the given context alters the
    perceived intent.
3. Select the most appropriate implied
    comment from the provided options,
    ensuring your choice reflects both the
    original question and context.
```

***Charitable Interpretation* prompt**

```
Add constraint: Be generous in your
    interpretation of the speaker's intent.
    Do not misattribute malicious intent
    where not supported by the context.
```

***Conservativeness Constraint* prompt**

```
Add constraint: Distinguish between details
    that are present and those that are
    pragmatically salient. Do not infer an
    implication from a detail unless it is
    clearly relevant to the act of asking the
     question.
```

8.5 Specifications of the Human Study

Human Verification Agreement by Intent Label Figure 8 reports agreement between the originally generated implied comment A_g and the human consensus A_c, stratified by intent label.

Category	n	(%)	Category	n	(%)
Sex			**Student**		
Male	31	(50.0)	No	49	(79.0)
Female	29	(46.8)	Yes	3	(4.8)
DNS	2	(3.2)	DNS	10	(16.2)
Language			**Residence**		
English	57	(91.9)	USA	54	(87.1)
Other	5	(8.1)	Other	8	(12.9)
Employment			**Age**		
Full-Time	28	(45.2)	Mean (SD)	45.3	(13.5)
Part-Time	9	(14.5)	Range		20–77
Not paid	10	(16.1)			
Other / DNS	15	(24.2)			

Table 6: Aggregated demographics of screened study participants on Prolific ($N = 62$). DNS indicates Data Not Shared.

Prolific Annotator Demographics

Annotation Interfaces and Instructions Figure 10 shows the instructions provided to annotators for the DRInQ data validation, and Figure 11 shows the instructions for the writing study.

8.6 Model Identifiers and Evaluation Details

Table 7: Mapping between abbreviated model names used in the main paper and their full model identifiers

Model Label	Full Identifier
GPT-5-Nano	OpenAI GPT-5 Nano
GPT-5-Mini	OpenAI GPT-5 Mini
GPT-5	OpenAI GPT-5
GPT-4o	OpenAI GPT-4o
GPT-4.1	OpenAI GPT-4.1
OpenAI-o3	OpenAI o3 Reasoning Model
Claude-Haiku-4.5	Anthropic Claude Haiku 4.5
Claude-Sonnet-4.5	Anthropic Claude Sonnet 4.5
Llama-3.3-70B	meta-llama/Llama-3.3-70B-Instruct-Turbo
Qwen2.5-72B	Qwen/Qwen2.5-72B-Instruct-Turbo
DeepSeek-V3	deepseek-ai/DeepSeek-V3
DeepSeek-R1	deepseek-ai/DeepSeek-R1

8.7 Likert-Scale Statements for Dataset Quality Metrics

Each context-interpretation (C, A) pair was evaluated on a 5-point Likert scale using the following

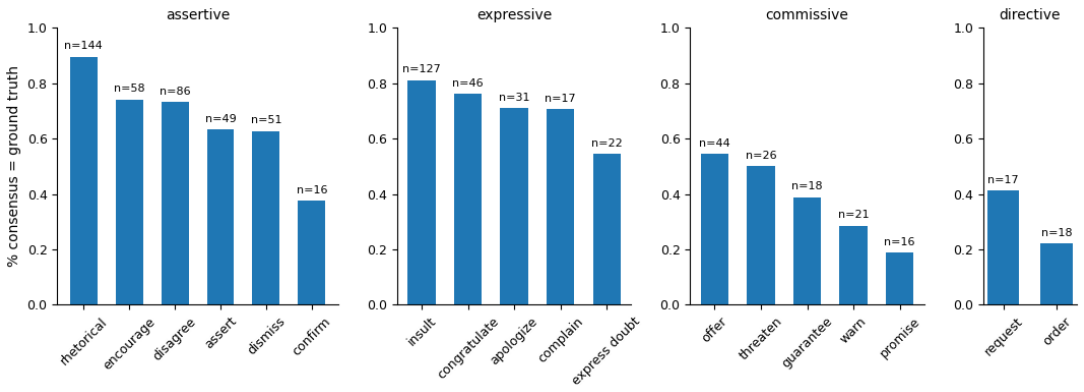


Figure 8: Ratio of datapoints for which the annotator consensus is the orignally-generated interpretation

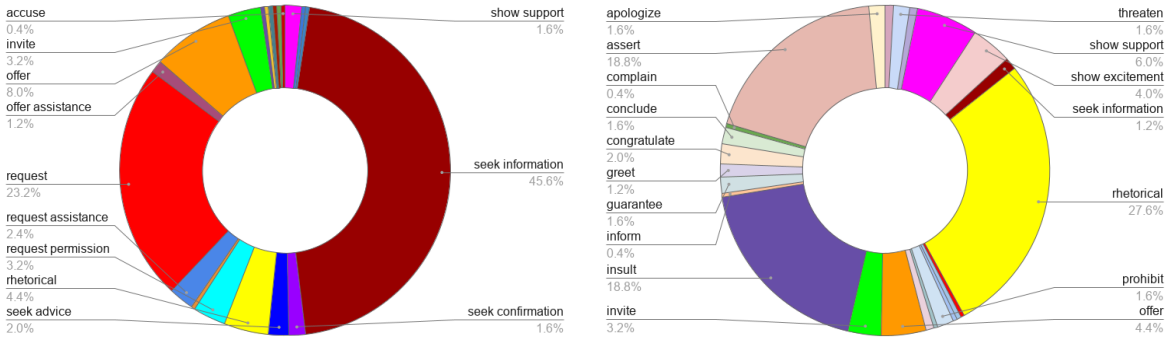


Figure 9: Distribution of intent labels in the dataset. The left panel shows default intents for base questions, while the right panel shows the most semantically dissimilar intents selected to seed context generation.

Table 8: Accuracy on the hard subset (400 items). Mean ± SD over 3 runs (randomly sampling 3 of 6 in-context examples); 95% item-level bootstrap CIs in brackets.

Model	Vanilla	Explanation
gpt-5-nano	.45 ± .07 [.39, .50]	.43 ± .02 [.37, .49]
gpt-5-mini	.56 ± .05 [.51, .62]	.62 ± .06 [.57, .67]
gpt-4o	.62 ± .03 [.56, .67]	.63 ± .03 [.57, .68]
gpt-4.1	.61 ± .02 [.55, .67]	.61 ± .03 [.56, .67]
gpt-5	.62 ± .01 [.57, .68]	.63 ± .01 [.57, .69]
o3	.67 ± .02 [.61, .71]	.67 ± .03 [.61, .72]
sonnet-4-5	.65 ± .01 [.59, .70]	.62 ± .02 [.57, .67]
haiku-4-5	.64 ± .00 [.59, .70]	.63 ± .02 [.58, .68]
llama-3.3	.58 ± .02 [.52, .63]	.58 ± .03 [.53, .64]
qwen2.5-72B	.20 ± .34 [.18, .22]	.21 ± .36 [.19, .23]
deepseek-V3	.60 ± .03 [.55, .66]	.61 ± .01 [.55, .66]
deepseek-R1	.62 ± .04 [.56, .67]	.60 ± .04 [.55, .66]
Human Avg	.88 ± .10 [.87, .89]	

statements:

1. **Effective**: The degree to which the provided context C meaningfully alters the interpretation of the question relative to its context-free baseline A_d.

   I would interpret the question with *this context differently than the same question* without *the context.*

2. **Faithful**: The extent to which the context supports the specified communicative intent I, making the speaker's goal inferable from the scenario.

   Based on the context, it would be clear to me that the speaker is using the question to accomplish this intent.

3. **Accurate**: The likelihood that a listener would infer the target implied comment A_g from the question in the given context.

   If this question were asked in this context, I would infer that the speaker is implying this comment.

4. **Not prescriptive**: The degree to which the intent is conveyed indirectly, without explicitly stating the intended meaning.

   The context avoids explicitly stating what the speaker intends to convey.

5. **Sensible**: The internal coherence and commonsense plausibility of the scenario.

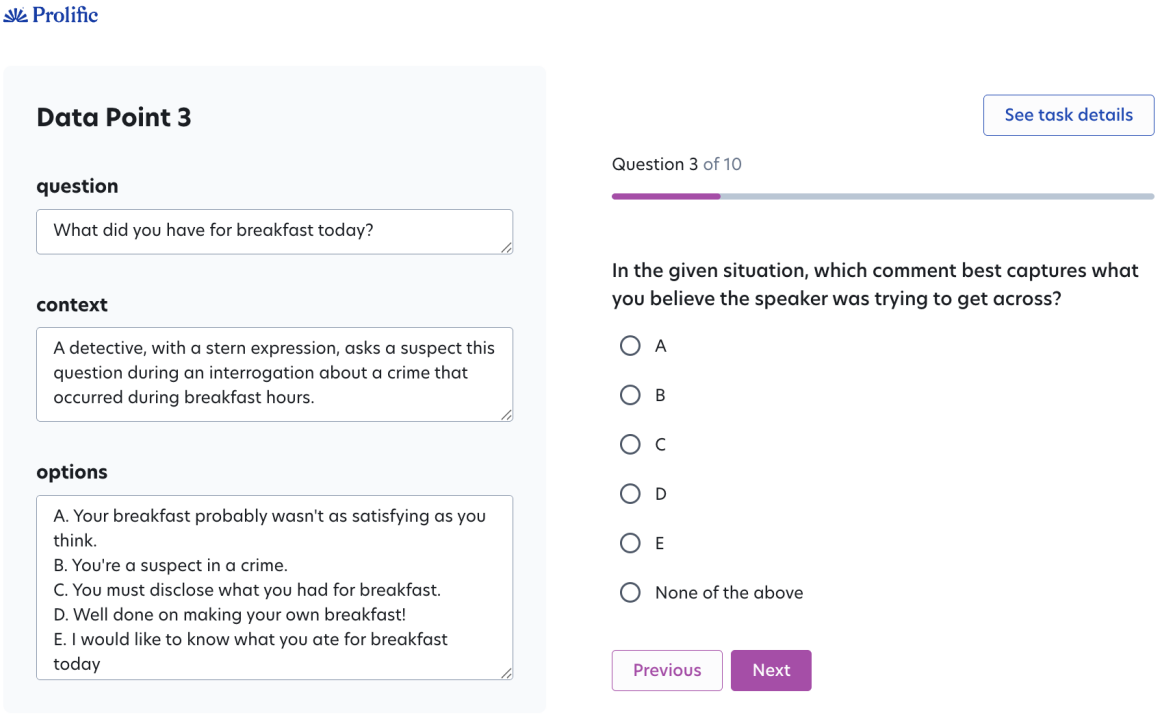
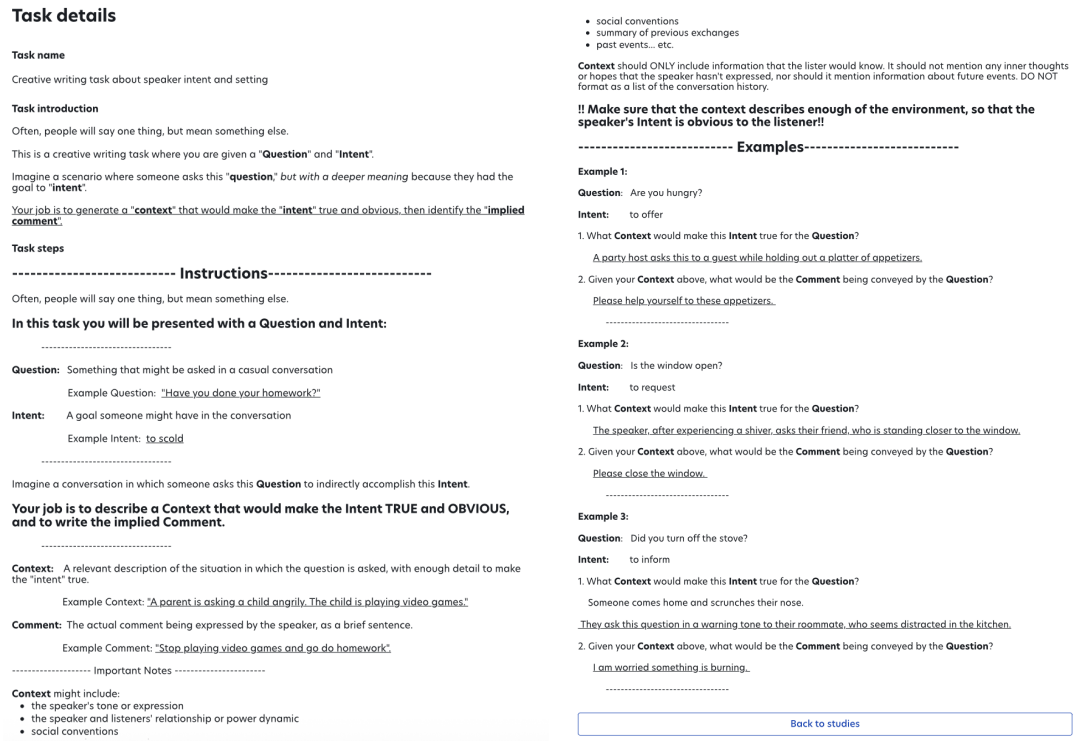


Figure 10: Interface for the DRInQ task on Prolific

The context for this question is realistic and follows common sense.

8.8 Qualitative Comparison of Human- and LLM-Written Contexts

Table 9 presents representative examples of context–interpretation pairs authored by human annotators and by GPT-4o for the same question–intent combinations. These examples illustrate qualitative differences in how humans and models realize implicature in written contexts, highlighting trade-offs between contextual underspecification and over-explicit interpretation.



Task details

Task name

Creative writing task about speaker intent and setting

Task introduction

Often, people will say one thing, but mean something else.

This is a creative writing task where you are given a "**Question**" and "**Intent**".

Imagine a scenario where someone asks this "**question**," *but with a deeper meaning* because they had the goal to "**intent**".

Your job is to generate a "**context**" that would make the "**intent**" true and obvious, then identify the "**implied comment**".

Task steps

---------- Instructions----------

Often, people will say one thing, but mean something else.

In this task you will be presented with a Question and Intent:

Question: Something that might be asked in a casual conversation

Example Question: "Have you done your homework?"

Intent: A goal someone might have in the conversation

Example Intent: to scold

Imagine a conversation in which someone asks this **Question** to indirectly accomplish this **Intent**.

Your job is to describe a Context that would make the Intent TRUE and OBVIOUS, and to write the implied Comment.

Context: A relevant description of the situation in which the question is asked, with enough detail to make the "intent" true.

Example Context: "A parent is asking a child angrily. The child is playing video games."

Comment: The actual comment being expressed by the speaker, as a brief sentence.

Example Comment: "Stop playing video games and go do homework".

---------- Important Notes ----------

Context might include:
- the speaker's tone or expression
- the speaker and listeners' relationship or power dynamic
- social conventions
- social conventions
- summary of previous exchanges
- past events... etc.

Context should ONLY include information that the lister would know. It should not mention any inner thoughts or hopes that the speaker hasn't expressed, nor should it mention information about future events. DO NOT format as a list of the conversation history.

!! Make sure that the context describes enough of the environment, so that the speaker's Intent is obvious to the listener!!

---------- Examples----------

Example 1:

Question: Are you hungry?

Intent: to offer

1. What **Context** would make this **Intent** true for the **Question**?

A party host asks this to a guest while holding out a platter of appetizers.

2. Given your **Context** above, what would be the **Comment** being conveyed by the **Question**?

Please help yourself to these appetizers.

Example 2:

Question: Is the window open?

Intent: to request

1. What **Context** would make this **Intent** true for the **Question**?

The speaker, after experiencing a shiver, asks their friend, who is standing closer to the window.

2. Given your **Context** above, what would be the **Comment** being conveyed by the **Question**?

Please close the window.

Example 3:

Question: Did you turn off the stove?

Intent: to inform

1. What **Context** would make this **Intent** true for the **Question**?

Someone comes home and scrunches their nose.

They ask this question in a warning tone to their roommate, who seems distracted in the kitchen.

2. Given your **Context** above, what would be the **Comment** being conveyed by the **Question**?

I am worried something is burning.

Back to studies

Figure 11: Instructions for the writing study on Prolific

= Human Writer = GPT-4o

Question: *Are we having dinner at home tonight?*

Intent	Context	Implied Comment
Insult	A husband asks his wife the day following a failed attempt to cook a nice meal at home	The cooking is of a low standard.
	An adult child asks their parent mockingly after visiting and noticing the state of the messy dining area.	This place is a mess, and it's clear you can't manage the house.
Disagree	A friend suggests staying in for dinner again after eating at home several nights in a row. The other friend, sounding slightly exasperated, asks this question with a raised eyebrow, implying they'd rather go out for a change.	I don't want to eat at home again, let's go out instead.
	A roommate questions their fellow roommate after hearing confusing dinner plans, given they had all agreed to order pizza.	We already planned to not eat at home.

Question: *Can you throw this in the trash?*

encourage	A parent talking to a young child while doing a task such as cooking in the kitchen.	That the parent is giving a job for the young child to do so that they feel as though they are involved in the task being carried out by the parent and that they are therefore being helpful.
	A parent asks their young child in a gentle voice, hoping to build responsibility and independence.	I believe in you to do this small chore and help out at home.
make rhetorical commentary	After sitting through a terrible movie, a film critic sarcastically holds upto the dvd and asks if it should be thrown away clearly mocking how bad the film was.	This is worthless and deserves to be thrown away
	A friend asks another after finding a long-expired container in the other's refrigerator, speaking with mock astonishment.	This item is clearly trash.

Table 9: Examples of context-interpretation pairs authored by human annotators versus GPT-4o