# Understanding Conversational Implicatures in Humans and LLMs

**Daeun Kang**
Department of Linguistics, Purdue University
kang584@purdue.edu

## Abstract

In conversational implicatures, speakers convey hidden intended meanings beyond the literal content of their utterances, and hearers are expected to infer *what is implied*. This study examines how Large Language Models (LLMs) interpret conversational implicatures, using human interpretation as a baseline and gold standard for comparison. The same experiments were conducted with two types of participants: humans and LLMs. Two metrics were adopted: a surprisal-based metric and a response-based metric. The results suggest that the response-based metric demonstrates higher accuracy, comparable to human responses, than the surprisal-based metric. In particular, humans and LLMs using the response-based metric performed better in the literal condition than in the implied condition. Additionally, they were more sensitive to capturing implied meanings for *some-all* trigger than for other triggers, whereas they showed lower performance on Manner implicatures. Overall, LLMs employing the response-based metric tend to exhibit human-like behavior, but still diverge from humans in their understanding of conversational implicatures.

## 1 Introduction

Large Language Models (LLMs), such as ChatGPT, have demonstrated strong performance in answering questions, providing explanations, and creating original writings (Douglas, 2023; Makridakis et al., 2023). LLMs can handle literal statements effectively, but everyday human communication relies not only on what is said but also on *the way what is said* (Grice, 1975). Speakers often convey intended meanings implicitly through conversational implicatures, and hearers are expected to infer these intended meanings beyond literal semantics. This raises a critical question: Do LLMs exhibit a human-like sensitivity to pragmatic inference, or are they limited to literal semantics?

Conversational implicatures were introduced by Grice (1975), grounded in the Cooperative Principle (CP) and its four conversational maxims. The CP proposed that interlocutors are expected to "make their conversational contribution as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which they are engaged" (p.45). Based on the CP, Grice suggested four conversational maxims: maxim of Quantity, Quality, Relation, and Manner (Grice, 1989, p.26), but he noted that interlocutors do not invariably adhere to these maxims. Conversational implicatures arise, or are intended, by observing, exploiting or flouting these maxims (Green, 2012). Therefore, to understand conversational implicature properly, hearers are required to infer the intended meaning beyond the literal semantics, assuming that the speaker is in general compliance with the CP.

A substantial body of empirical research has shown that interpretation of conversational implicature varies across populations, including children (Noveck and Reboul, 2008; Olkoniemi et al., 2023; Pouscoulous, 2023), adults (Bott and Chemla, 2016; Cowan and Katsos, 2023; Katsos et al., 2023), and L2 learners (Alharbi, 2022; Lou and Noels, 2016; Rabab'ah et al., 2024; Zhang and Wu, 2023; Zhao et al., 2021). Even among adult native English speakers, different implicature types show variation in their interpretation. For example, Manner implicatures are interpreted with less sensitivity and conventionally, compared to Quantity implicatures (Wilson and Katsos, 2016). These findings suggest that implicature interpretation is a complex process influenced by context, shared knowledge, and individual perspective.

Recent advances of LLMs have spurred research assessing the models' linguistic capabilities to capture the subtleties of human language. While some studies show that LLMs can infer basic scalar implicatures, such as interpreting *some* as *some but*



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 741–753

*not all* in straightforward contexts, they struggle with more complex or context-dependent situations (Jeretic et al., 2020; Cho and mook Kim, 2024), or Manner implicatures (Cong, 2024). However, prior research has been limited in that it primarily evaluates LLMs' technical performance rather than LLMs' human-likeness. As concerns grow that training on synthetic data may cause LLMs to drift from real-world language patterns and degrade in performance (del Rio-Chanona et al., 2024; Shumailov et al., 2024), evaluating their human-likeness has become increasingly critical (Duan et al., 2024). The present study addresses this gap by investigating how LLMs interpret conversational implicatures, using human interpretation as a baseline and gold standard for comparison.

## 2 Related Work

### 2.1 Grice's Four Conversational Maxims and Conversational Implicatures

Conversational implicatures arise when speakers intend to convey meaning beyond the literal semantic content of an utterance, and the hearer infers this meaning by relying on pragmatic reasoning, the CP, and contextual information. Among Grice's four conversational implicature types, this study focuses on three: Quality implicature triggered by *sarcasm*, Quantity implicature triggered by *numerals* and *some-all*, and Manner implicature triggered by *causality*, *modality*, and *double negation*. These implicatures were selected because they have been commonly discussed in the pragmatic literature (Olkoniemi et al., 2023; Pouscoulous, 2023; Wilson and Katsos, 2016) and are amenable to experimental manipulation. The following subsections describe each implicature type in turn, illustrating how they are instantiated in the experimental materials used in this study.

**Quality Implicatures** Quality implicatures arise when a speaker observes or flouts the maxim of Quality: "do not say what you believe to be false", and "do not say that for which you lack adequate evidence" (Grice, 1975, p.46). Based on this maxim, *sarcasm* is a typical example of a Quality implicature, where speakers deliver an attitude or evaluation that contrasts sharply with the literal content of the utterance. Specifically, *sarcasm* is a subtype of irony, a form of nonliteral language that often conveys the opposite of what is literally said, and is widely used in everyday conversation to humorously criticize a situation or object (Filik et al., 2016; Kreuz and Glucksberg, 1989). In addition, Bamman and Smith (2015), drawing on Kreuz and Caucci (2007), argue that "sarcastic utterances often contain lexical indicators (such as interjections and intensifiers) and other linguistic markers (such as nonveridicality and hyperbole) that signal irony." These text-based cues allow sarcasm to be detected with relatively high accuracy across different domains (Davidov et al., 2010; González-Ibánez et al., 2011; Riloff et al., 2013; Lukin and Walker, 2013; Reyes et al., 2013). This suggests that sarcasm can be recognized through textual information alone, without relying on nonverbal cues such as gestures, facial expressions, and tone of voice. In the present study, Quality implicatures are triggered by sarcastic utterances in which the literal interpretation conflicts with contextual expectations, requiring participants to infer the speaker's intended attitude.

**Quantity Implicatures** Quantity implicatures arise from maxim of Quantity, which requires speakers to provide as much information as is necessary for the purposes of the conversation, but no more than is required (Grice, 1975, p.45). Quantity implicatures are among the most commonly studied types of conversational implicatures. In particular, implicatures triggered by *numerals* and scalar implicatures, which is a subtype of Quantity implicature, triggered by the quantifier *some* have been investigated as canonical cases. *Numerals* and *some* often give rise to upper-bounded interpretations. These two types of Quantity implicatures are examined in the present study.

For *numerals*, such as *five* in the stimulus *"I'm taking five courses,"* (see Appendix A) the conversational implicature arises from observing the first submaxim of Quantity, leading to the interpretation that the speaker is taking exactly five courses, rather than at least five or possibly more. For *some-all*, the word *some* in the stimulus *"I have lots of homework, but... I've done some of it."* semantically entails that at least a subset and possibly all of the homework have been completed. Pragmatically, however, it is typically interpreted as *not all*, with the upper-bounded meaning enriched by the implicature (Cummins and Katsos, 2010). Therefore, the experimental materials included two types of stimuli: one that conveyed an implied meaning using the quantifier *some*, and another that explicitly expressed a literal meaning by stating *all*. Accordingly, the present study includes both *numerals* and *some-all* triggers to capture variation in Quantity implicature interpretation.

**Manner Implicatures** Manner implicatures arise when a speaker intends to convey additional meaning through marked or unusual expressions, thereby flouting the maxim of Manner: "avoid obscurity of expression", "avoid ambiguity", "be brief", and "be orderly" (Grice, 1975, p.46). Manner implicatures require hearers to reason why a speaker selected a less straightforward or more complex form, which may increase processing demands and lead to variability in interpretation. In this study, Manner implicatures are triggered by *causality*, *modality*, and *double negation*, all of which require additional inferential steps beyond literal meaning. The complex and prolix form of each stimulus (see Appendix A), such as saying *"The intruders were the reason behind the passing of the victims"* rather than *"The intruders killed the victims," "Carolyn had the ability to paint the board"* rather than *"Carolyn was able to paint the board," "In my view, this book isn't uninteresting"* rather than *"In my view, this book is interesting,"* pragmatically enriches meaning and conveys the hidden meaning intended by the speaker.

### 2.2 Experimental Studies in Humans

Conversational implicature is one of the most widely investigated phenomena requiring pragmatic reseasoning in human language comprehension. Previous studies using truth-value judgment tasks, offline comprehension tasks, and reaction time measures have demonstrated that adult native speakers showed variability in interpreting conversational implicatures across different implicature types (Cummins and Katsos, 2010; Katsos et al., 2023; Zhang and Wu, 2023).

In particular, previous research has shown that adult native English speakers tend to interpret Quantity implicatures more robustly and consistently than Manner implicatures (Wilson and Katsos, 2016). Quantity implicatures often rely on clearly defined alternatives, such as *some* versus *all* or exact *numerals* interpretations, which makes the relevant pragmatic inference relatively straightforward (Franke, 2011; Rett, 2020). In contrast, Manner implicatures are typically more context-dependent and rely on subtler cues such as markedness, verbosity, or unusual expression (Horn, 1991; Rett, 2020). Moreover, Bott and Chemla's (2016) findings revealed that linguistic enrichments such as quantifiers and numerals rely on shared cognitive mechanisms for pragmatic reasoning. On the other hand, Cowan and Katsos, 2023 explored whether Quantity and Manner implicatures share an underlying mechanism. Their results indicate that exposure to Manner implicatures influences the subsequent understanding of Manner implicatures, but Quantity implicatures, particularly the *some* trigger, did not significantly facilitate the interpretation of Manner implicatures. These findings suggest that while all implicature types are grounded in the Gricean framework, they may differ in their processing demands and interpretative mechanisms.

Previous studies have primarily focused on implicature types rather than trigger types, leaving gaps in understanding how specific triggers contribute to variability in human performance and error patterns. Investigating such variability is crucial not only for understanding human pragmatic competence, but also for evaluating whether LLMs exhibit human-like pragmatic behavior. Human performance thus provides a critical empirical baseline and gold standard for assessing LLMs' pragmatic abilities, as human-likeness involves not only overall accuracy but also comparable patterns of success and failure on interpretation of conversational implicatures.

### 2.3 LLMs' Understanding

With the rapid development of large language models (LLMs), recent studies have begun to investigate whether these models exhibit pragmatic competence. Beyond high performance on syntactic and semantic tasks such as question answering and cloze task (Brown et al., 2020; Bommasani, 2021), LLMs have been shown to capture contextual sensitivity and inference ability in interaction with humans (Bubeck et al., 2023). Conversational implicature presents a particularly challenging test for LLMs, as successful interpretation requires integrating linguistic processing with contextual reasoning, speaker intentions, and conversational maxims. A growing body of work has evaluated LLMs' ability to interpret conversational implicatures (Bojic et al., 2023; Yue et al., 2024; Zhang et al., 2025). For example, Zhang et al. (2025) reported that GPT-4 and Gemini 2.0 outperform other LLMs across various prompting methods. Research on scalar implicatures shows that BERT is sensitive to scalar alternatives such as *some* and *all* (Jeretic et al., 2020). Cho and mook Kim (2024) reported that BERT demonstrates consistent outcomes when contextual cues are offered, but GPT-2 does not. However, LLMs tend to struggle with more complex and context-dependent situations, particularly Manner

implicatures. Cong (2024) showed that most models perform at or below chance in distinguishing implicature from entailment or equivalence.

Despite these advances, several limitations remain in the previous literature on LLMs' understanding of conversational implicature. First, many studies evaluate LLM's performance technically, without comparing model behavior to human data collected using the same experimental materials. Consequently, it remains unclear whether high model accuracy reflects human-like pragmatic reasoning or merely task-specific success under trained conditions. Second, previous research has rarely examined how LLM performance varies systematically across specific implicature triggers, such as *sarcasm*, *numerals*, *some-all*, *causality*, *modality*, and *double negation*. With trigger-specific analyses, it is possible to obtain a fine-grained understanding of whether LLMs exhibit human-like sensitivity and error pattern in understanding conversational implicatures. Third, prior research rarely investigated how well LLMs can perform depending on different metrics, such as surprisal and prompting. This gap motivates our interest in exploring LLMs' pragmatic processing abilities across the different levels at which they operate.

Surprisal, represented as the negative log-probability of a word given its preceding context, P(*token*|*context*), is a fundamental unit of computation in LLMs (Hu and Levy, 2023). This measure has been widely used as a computational index of linguistic expectation and language model processing difficulty in a range of prior studies (Huber et al., 2024; Krieger et al., 2025; Linzen et al., 2016). Recently, research has increasingly adopted prompting to assess LLMs' abilities. Unlike surprisal, prompting methods require models to interpret and respond to task instructions, including metalinguistic judgment and task reasoning beyond the raw probability estimates from the model's internal distribution (Brown et al., 2020; Hu and Levy, 2023). Although surprisal has known limitations as a comprehensive explanatory measure (Krieger et al., 2025), prior work has demonstrated correlations between surprisal and human language processing measures, including event-related potential (ERP) and reading time (Huber et al., 2024). Nevertheless, most probability-based studies to date have concentrated on syntactic and semantic processing (Futrell et al., 2019; Linzen et al., 2016; Wilcox et al., 2018), with little attention paid to pragmatic inference in LLMs. Cong (2024) remains a rare exception in applying surprisal to the evaluation of pragmatic reasoning, but even there, LLMs exhibited consistently weak performance in interpreting Manner implicatures. By contrast, Hu et al. (2023) demonstrated that expectations computed by neural language models over strong alternatives can effectively predict human scalar implicature inference.

The present study addresses these gaps by comparing humans and LLMs. By examining both surprisal-based and response-based metrics, this study provides a more comprehensive assessment of LLMs' pragmatic behavior on conversational implicature. Furthermore, by systematically analyzing performance across different implicature triggers, the present work aims to determine not only whether LLMs can interpret conversational implicatures, but also whether they do so in a manner that aligns with human sensitivity patterns and error distributions.

## 3 Experiments

This study conducts experiments with two types of participants: human and LLMs. To enable direct comparison with human interpretation, the same experimental materials were used in both experiments. In the LLM experiment, two metrics were adopted: a probability-based surprisal metric and a response-based metric using prompt engineering. This study received approval from the Institutional Review Board (IRB) to ensure compliance with ethical research standards, and human participants provided written consent as a part of the survey.

### 3.1 Human Experiment

**Participants** The experiment included 32 participants, all of whom were over 19 years of age and native speakers of English with no reported reading difficulties.

**Materials** Participants were presented with 39 statements in total, including three attention check items designed to ensure that participants were paying attention and responding carefully throughout the experiment. The remaining 36 test items consisted of an equal number of items conveying implied and literal meanings, with the exception of items involving the *numerals* trigger. The items involving *numerals* were all implied meaning items, and there were no corresponding literal meaning items. The test items were modeled after Wilson and Katsos's (2016) dataset, augmented using

prompts generated with ChatGPT, and reviewed by three human raters: the author and two linguists. The experiment was administered online using Qualtrics. Each test item consisted of a short context, a stimulus sentence, and two interpretation options. The items were randomly arranged, and their order was manually adjusted by a human reviewer to minimize potential order effects on participants' responses. Responses were coded binarily, with correct interpretations labeled as 1 and incorrect interpretations as 0.

**Procedure** Participants completed the experiment independently in a single session, which took approximately 15 minutes. After providing consent, they read each statement and selected one of two interpretation options. Participants could not move on to the next item without selecting a response.

### 3.2 Experiment with LLMs

To evaluate LLMs' interpretation of conversational implicatures, the same materials used in the human experiment were used with two different metrics: a surprisal-based metric and a response-based metric. These measures allow examination of both token-level probabilistic expectations and explicit interpretation behavior under instruction-following conditions.

**Surprisal-based Metric** Surprisal scores were obtained using two transformer-based models, *GPT-Neo-1.3B* (Black et al., 2021) and *RoBERTa-large* (Liu et al., 2019). GPT-Neo is a decoder-based model designed for autoregressive text generation, predicting each upcoming token based on the preceding context. In contrast, RoBERTa is an encoder-based model that captures bidirectional contextual information and is trained using masked language modeling. These two models were selected because they represent foundational transformer models with distinct processing architectures. Following Levy (2008), the surprisal score was employed as a quantitative measure of how unexpected a sentence is in context, providing an index of the model's performance. Thus, higher surprisal scores indicate lower predictability. The surprisal scores were computed using *minicons* (Misra, 2022), and conditional probabilities were calculated for the question including the context, stimulus, instruction, and two possible interpretations (see Appendix A). If an LLM can interpret conversational implicature accurately, it should assign a lower surprisal score to the intended interpretation. Thus, when an LLM assigns a lower surprisal score to the intended interpretation than to the alternative interpretation, the response is considered correct and labeled as 1; otherwise, it is considered incorrect and labeled as 0.

**Response-based Metric** The LLMs examined in this study were *GPT-4o-mini* (Achiam et al., 2023) and *Gemini-flash-1.5* (Team et al., 2023). These models were selected because they are lightweight and recent language models, optimized for rapid response-based interaction and instruction-following, making them suitable for evaluating pragmatic interpretation in prompting-based experimental settings. Responses were collected via the models' official Python APIs using default parameter settings. Each input consisted of a structured prompt including the context, stimulus sentence, task instructions, and two interpretation options, mirroring the human experiment. A fixed prompt instruction was used across all items, formatted as follows: *There are two types of meaning reflected in the interpretations: (1) implied (2) literal. Here is the difference between (1) implied and (2) literal: In conversational implicature, speakers convey hidden intended meanings beyond the literal content of their utterances, and hearers are expected to infer what is implied. Implicatures can be canceled. Your task is to choose the interpretation that you believe the speaker intended. Which interpretation is more likely intended by the speaker? Please answer only 'Interpretation 1' or 'Interpretation 2'.* As in the human and surprisal-based experiments, each item had a single intended interpretation, and model responses were coded binarily as correct (1) or incorrect (0).

## 4 Results

### 4.1 Human Experiment

Data were initially collected from 32 participants. Eight participants did not complete the experiment and four failed at least one attention check item; their data were excluded. The final analysis therefore included 20 participants. Participants showed generally high performance (M = 35.90, SD = 2.47), corresponding to an overall accuracy rate of 92.1%. Accuracy differed across triggers (see Figure 1). Quality implicatures triggered by *sarcasm* showed highest accuracy (99.2%) and the Manner implicatures triggered by *causality* were the lowest (79.2%). A one-way ANOVA showed a significant difference among trigger types in ac-

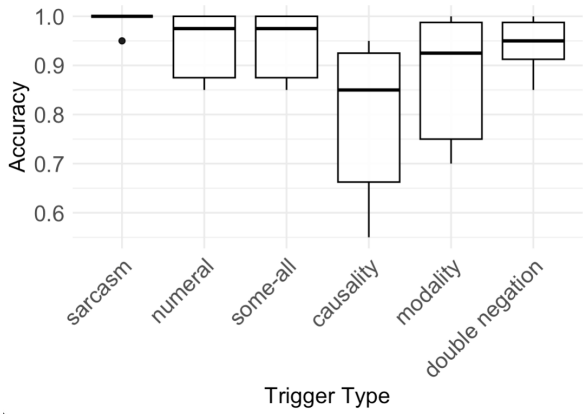


Figure 1: Boxplot for Accuracy by Trigger

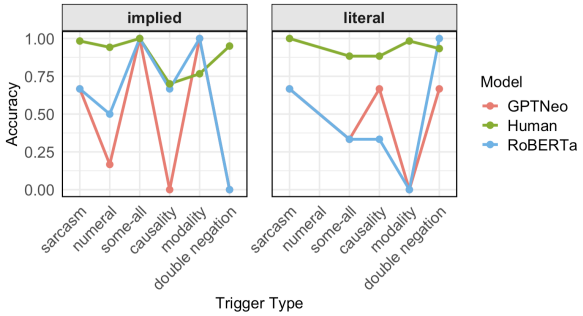


Figure 2: Distribution of Accuracy by Trigger and Meaning across GPT-Neo, RoBERTa, and Humans *(Note: The conversational implicature triggered by the numerals is interpreted as only implied meaning.)*

curacy rate [$F(5, 30) = 2.76$, $p = .036$], motivating further mixed-effects analyses.

In terms of meanings (see Figure 2), participants were slightly more accurate in the literal condition (93.7%) than in the implied condition (89.8%). Quality implicatures triggered by *sarcasm* reached perfect accuracy in the implied condition (100%) and ceiling-level performance in the literal condition (98.3%). Except for *numerals* and *double negation*, all triggers showed lower performance in the implied condition. Notably, the *some-all* trigger exhibited perfect accuracy in the implied condition (100%) but slightly lower in the literal condition (88.3%). Manner implicatures triggered by *causality* and *modality* showed greatest variability, particularly in the implied condition.

To examine the effects of trigger type and meaning (literal vs. implied), a generalized linear mixed-effects model (GLMM) with a binomial link was fitted, including participants and items as random effects. The model revealed significantly higher accuracy relative to the implied baseline for *sarcasm* ($p = .002$), *numerals* ($p = .014$), *double negation* ($p = .032$), and *some-all* ($p = .033$). However, meaning did not have a significant main effect ($p = .278$). Tukey's post-hoc comparisons using the *emmeans()* package in R (Lenth, 2023) showed that only the contrast between *sarcasm* and *causality* was significant, with higher accuracy for *sarcasm* ($p = .022$). However, all other trigger types contrasts were not significant, and differences between literal and implied meanings within each trigger were not significant (all $p > .27$).

### 4.2 Experiment with LLMs

**Surprisal-based Metric** To measure LLMs' interpretation using a surprisal-based metric, GPT-Neo and RoBERTa were evaluated. Assuming that surprisal scores reflect LLMs' understanding of conversational implicatures, both LLMs performed substantially worse than humans. GPT-Neo obtained an overall accuracy of 44.4%, while RoBERTa achieved 55.6%, compared to accuracy for humans (92.1%). For GPT-Neo, across trigger types, *sarcasm* and *some-all* showed the highest accuracy (66.7%), whereas *numerals* showed the lowest accuracy (16.7%). RoBERTa showed a similar pattern, with its highest accuracy for *sarcasm* and *some-all* (66.7%), and lower performance for *numerals*, *causality*, *modality*, and *double negation* (each 50%). GPT-Neo showed a slightly higher accuracy for the literal condition (46.7%) than the implied condition (42.9%). In contrast, RoBERTa showed a higher accuracy (61.9%) in the implied condition than the literal condition (46.7%). A GLMM with binary accuracy as a dependent variable and model type and items as random effects revealed no significant effects of either trigger type (all $p > .30$) or meaning ($p = .38$).

In more detail, accuracy for each trigger and meaning (see Figure 2) further illustrates that both LLMs performed poorly even in the literal condition, where interpretation is relatively straightforward. Notably, both models performed comparably to humans under the implied condition (100%) in the *some-all* trigger. Although RoBERTa showed slightly higher accuracy than GPT-Neo in the implied condition, this difference was not statistically significant ($p > .50$). Both LLMs scored higher than humans in the *modality* implied condition, but they scored 0 in the literal condition. Overall, their interpretation is not consistent across triggers and meanings.

To compare LLMs and humans, a linear mixed-

effects model using the *lme4* package in R (Kuznetsova et al., 2016) was fitted with participant type (LLMs and humans) as a fixed effect. The model revealed a significant difference in accuracy across participants [$F(2, 20) = 9.51$, $p = .001$]. Post-hoc comparisons showed that humans significantly outperformed both GPT-Neo and RoBERTa, while the difference between the two models was not significant (see Table 2 in Appendix B). To sum up, under the surprisal-based metric, GPT-Neo and RoBERTa performed significantly worse than humans and showed inconsistent performance across trigger types and meaning conditions.

**Response-based Metric** In order to examine LLMs' interpretation of test items using the response-based metric, Gemini-flash-1.5 and GPT-4o-mini were evaluated using the same binary-choice task as humans. Both LLMs showed lower accuracy in interpreting conversational implicatures than human participants, but higher accuracy than that observed with the surprisal-based metric using GPT-Neo and RoBERTa. Gemini-flash-1.5 achieved 29 correct responses out of 36 (80.6%). On the other hand, GPT-4o-mini correctly answered 31 items out of 36 (86.1%). Across trigger types, both models achieved ceiling accuracy on Quantity implicatures triggered by *numerals* and *some-all*. In contrast, *modality* showed the lowest accuracy for both models, unlike in the human results. Both models performed better on literal meanings than on implied meanings. Notably, Gemini-flash-1.5 showed a substantial drop in performance for implied meanings (66.7%) compared to literal meanings (100%), whereas GPT-4o-mini showed relatively stable performance even for implied meanings (81%) compared to the literal meanings (93.3%). A GLMM with a binomial link revealed a significant main effect of meaning with higher accuracy for literal meaning than implied meaning ($p = .046$), indicating that implied interpretations were relatively more difficult than literal ones across LLMs. However, no significant effects of trigger type were observed (all $p > .30$).

More detailed results by trigger and meaning (see Figure 3) demonstrated that both LLMs and humans showed ceiling-level performance for Quality implicatures triggered by *sarcasm* in literal condition. In contrast, Gemini-flash-1.5 underperformed in the implied condition (66.7%) compared to the GPT-4o-mini and humans (100%). For Quantity implicatures, LLMs achieved perfect accuracy in both *numerals* and *some-all* triggers across implied and literal conditions. In short, while they performed near ceiling on literal interpretations across triggers, their performance diverges from humans in the implied condition, particularly for Manner implicatures triggered by *modality*. By contrast, Quantity implicatures demonstrated high and consistent accuracy across both humans and LLMs. Additionally, the linear mixed-effects model revealed no statistically significant differences in accuracy across LLMs and human participants [$F(2, 20) = 1.67$, $p = .213$].

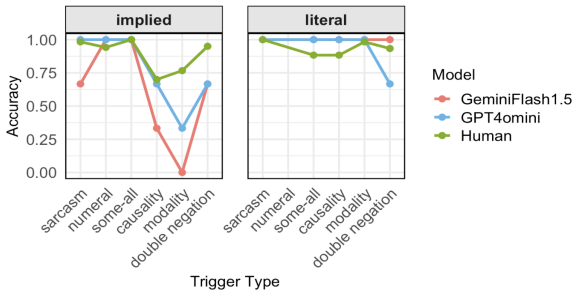


Figure 3: Distribution of Accuracy by Trigger and Meaning across Gemini-flash-1.5, GPT-4o-mini and, Humans *(Note: The conversational implicature triggered by the numerals is interpreted as only implied meaning.)*

Lastly, LLMs' human-likeness was evaluated by modeling the absolute difference between LLMs and human accuracy (see Table 1 and Figure 4). A linear mixed-effects model was fitted with the absolute difference from human accuracy as the dependent variable including trigger type as a random effect, and LLM and meaning were included as fixed effects. Compared to the intercept (Gemini-flash-1.5, implied meaning), GPT-Neo showed the largest deviation from human performance ($\beta = 0.292$, $p = .010$), followed by RoBERTa ($\beta = 0.214$, $p = .054$). In contrast, GPT-4o-mini tended to be more similar to human performance ($\beta = -0.070$, $p = 0.521$), although this difference did not reach statistical significance. Deviations increased for implied meanings across all models, indicating that implicature inference remains challenging.

## 5 Discussion

### 5.1 LLMs' Human-like Accuracy

Although none of the LLMs outperformed humans, Gemini-flash-1.5 and GPT-4o-mini using the response-based metric achieved accuracy comparable to humans, whereas GPT-Neo and RoBERTa using the surprisal-based metric did not. Humans achieved consistently high accuracy across impli-

| Fixed Effect | Estimate($\beta$) | SE | *p*-value |
|---|---|---|---|
| Intercept | 0.206 | 0.087 | 0.022* |
| GPT-4o-mini | -0.070 | 0.107 | 0.521 |
| GPT-Neo | 0.292 | 0.107 | 0.010* |
| RoBERTa | 0.214 | 0.107 | 0.054 |
| Meaningliteral | -0.031 | 0.077 | 0.686 |

Table 1: Linear Mixed-Effects Model Summary: LLMs and Human Performance *(Note: SE indicates standard error. Significance levels of p values *p < .05, **p < .01, ***p < .001.)*

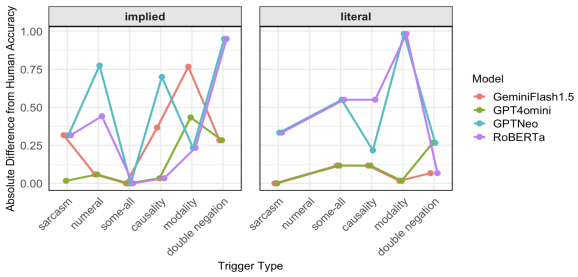


Figure 4: Human-LLMs Similarity by Trigger and Meaning *(Note: The conversational implicature triggered by the numerals is interpreted as only implied meaning.)*

cature types (> 90%), with Quality implicatures triggered by *sarcasm* showing the most robust performance, while Manner implicatures, particularly *causality*, were the most challenging and partially aligned with Wilson and Katsos (2016). Against this human baseline, LLMs' performance varied by model architectures and evaluation metrics. Under the surprisal-based metric, LLMs performed significantly worse than humans (< 60%) with no clear alignment to human interpretation pattern, suggesting token-level predictability fails to capture pragmatic inference (Cong, 2024). While *some-all* trigger showed partial alignment with humans in the implied condition (Hu et al., 2023), the results were unstable. In contrast, response-based metric achieved accuracy levels that were not statistically different from those of humans, supporting Zhang et al.'s (2025) claim that response-based evaluation paradigms can disclose pragmatic competencies in LLMs. In particular, the results from GPT-4o-mini showed the closest alignment with human accuracy across trigger types.

### 5.2 Error Patterns between Human and LLMs

Error pattern analysis offers further insight into the extent of human-model similarity. Do LLMs and humans make similar mistakes in their interpretations? First, humans performed better in the literal condition than the implied condition. This pattern was observed in the response-based metric but not in the surprisal-based metric. Both Gemini-flash-1.5 and GPT-4o-mini using the response-based metric achieved near ceiling performance for literal meanings, whereas they failed to do so for implied meanings. This asymmetry is consistent with prior research that LLMs, despite their effectiveness at reproducing surface-level linguistic patterns, still struggle with the tasks requiring pragmatic inference beyond literal contents. In contrast, the surprisal-based metric indicates the different pattern that GPT-Neo showed relatively higher performance for literal meanings, whereas RoBERTa better captured implied meanings.

Interestingly, the surprisal-based metric partially mirrors human error patterns in that they perform well on *sarcasm* but poorly on *causality*, but the result does not reveal a consistent pattern, making it difficult to conclude that the surprisal-based metric reliably replicates human error distributions. In contrast, under the response-based metric, both Gemini-flash-1.5 and GPT-4o-mini achieved ceiling performance on Quantity implicatures triggered by *numerals* and *some-all* in both implied and literal conditions, but they struggled with Manner implicatures, especially *modality*. Although, unlike human results, the *causality* trigger does not achieve the lowest accuracy in the response-based metric, both LLMs demonstrate performance comparable to human participants. Specifically, GPT-4o-mini largely mirrored human performance except for *modality* and *double negation*. In sum, LLMs using the surprisal-based metric should not be characterized as agents capable of fully performing pragmatic inference. In contrast, models using the response-based metric appear to demonstrate greater pragmatic competence. While these models approximate human accuracy, their behavior remains incomplete in terms of human-like behavior.

## 6 Conclusion

The present study examined how conversational implicatures are interpreted by humans and LLMs, focusing on trigger types. By comparing human comprehension with LLMs' performance using both surprisal-based and response-based approaches, this study aimed to assess the extent to which cur-

rent LLMs exhibit human-like pragmatic competence. All four LLMs using both metrics and human participants showed strong performance in interpreting Quantity implicatures triggered by *some-all* in the implied condition. LLMs' performance depended strongly on the evaluation metric. Under the response-based metric, Gemini-flash-1.5 and GPT-4o-mini approximated human-like accuracy, especially for literal meanings, but implied meanings and certain trigger types remained challenging. However, under the surprisal-based metric, GPT-Neo and RoBERTa exhibited low accuracy and highly inconsistent results. The lack of competence in conversational implicature interpretation suggests that the surprisal-based metric fails to capture human-like pragmatic behavior.

These findings indicate that LLMs using the response-based metric can approximate human-like pragmatic interpretation to some extent, while still showing huge variability. This research extends the previous literature on conversational implicatures by providing experimental evidence on how native English speakers interpret conversational implicature across the triggers. Moreover, the findings are expected to deepen our understanding of LLMs' pragmatic inference abilities and the extent to which their language comprehension aligns with that of humans through a fine-grained human-model comparison. Lastly, it establishes a benchmark for systematically evaluating LLMs' pragmatic reasoning.

## Limitations

This study has several limitations. First, the number of human participants was relatively small. Data from 20 human participants may not be sufficient to represent a general pattern of human interpretation. Future research should include a larger sample size. Second, this study focuses on a relatively limited set of models (e.g, GPT-Neo, RoBERTa, GPT-4o-mini, and Gemini-flash-1.5), which makes it difficult to generalize the findings on LLMs. Expanding the range of evaluated models based on both model family and size would help improve the robustness of the results. Third, only a limited set of implicature triggers was examined. This study focused on Quality, Quantity, and Manner implicatures triggered by six triggers, including *sarcasm*, *some-all*, *numerals*, *causality*, *modality*, and *double negation*. However, it did not investigate a broader range of linguistic phenomena that require pragmatic reasoning, such as metaphor or Relevance implicatures. Future research should extend the analysis to additional types of conversational implicatures and triggers, providing a more comprehensive picture of the model's human-like pragmatic inference abilities.

Lastly, the use of binary-choice tasks may overlook subtle aspects of interpretation. Unlike literal meanings, conversational implicatures often allow multiple plausible interpretations. This is especially true for Manner implicatures, which are more context-dependent and rely on subtle cues such as markedness, verbosity, or unusual expressions (Horn, 1991; Rett, 2020). Consequently, in the experimental setting, some items may lead participants to infer interpretations other than the intended meaning or may appear somewhat artificial, as they are designed to be amenable to controlled manipulation. For example, consider the following context and stimulus: *In a warehouse, Alex says, "Carolyn had the ability to paint the board."* While the intended implicature is that *Carolyn could have painted the board but she did not*, alternative interpretations are also possible, such as that *Carolyn had the ability, but it is not certain that she actually painted the board.* In this experiment, the corresponding literal meaning presented in the binary-choice task was that *Carolyn could paint the board and she did*, thereby encouraging participants to interpret the marked expression as conveying an implied meaning. However, if the task were designed in a multiple-choice or open-ended question, it might better capture human-like interpretive patterns in LLMs' pragmatic reasoning, underscoring the need for future research to adopt alternative methodologies.

## Acknowledgements

We would like to thank Atsushi Fukada, Yan Cong, Spencer Stewart, and Elsayed Issa for their valuable guidance and support in the development of this project. We are also deeply grateful to the anonymous ACL 2026 SRW mentors and reviewers for their detailed and insightful feedback and suggestions.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman,

Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohammed Abdullah Alharbi. 2022. Pragmatic awareness of conversational implicatures by l2 undergraduate students in saudi arabia. *East Asian Pragmatics*, 7(2):237–266.

David Bamman and Noah Smith. 2015. Contextualized sarcasm detection on twitter. In *proceedings of the international AAAI conference on web and social media*, volume 9, pages 574–577.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *Zenodo*.

Ljubisa Bojic, Predrag Kovacevic, and Milan Cabarkapa. 2023. Gpt-4 surpassing human performance in linguistic pragmatics. *arXiv preprint arXiv:2312.09545*.

Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Lewis Bott and Emmanuel Chemla. 2016. Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, 91:117–140.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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.

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

Joe Cowan and Napoleon Katsos. 2023. Investigating a shared mechanism in the priming of manner and quantity implicature. *Experiments in Linguistic Meaning*, 2:36–48.

Chris Cummins and Napoleon Katsos. 2010. Comparative and superlative quantifiers: Pragmatic effects of comparison type. *Journal of Semantics*, 27(3):271–305.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.

R Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2024. Large language models reduce public knowledge sharing on online q&a platforms. *PNAS nexus*, 3(9):pgae400.

Michael R Douglas. 2023. Large language models. *arXiv preprint arXiv:2307.05782*.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G Cai. 2024. Hlb: Benchmarking llms' humanlikeness in language use. *arXiv preprint arXiv:2409.15890*.

Ruth Filik, Alexandra Ţurcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. 2016. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69(11):2130–2146.

Michael Franke. 2011. Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, 4:1–1.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 581–586.

Georgia M Green. 2012. *Pragmatics and natural language understanding*. Routledge.

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

Paul Grice. 1989. In the way of words.

Laurence R Horn. 1991. Duplex negatio affirmat...: the economy of double negation. In *Regional Meeting of the Chicago Linguistic Society. Part Two*, volume 27, pages 80–106.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264*.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, 11:885–901.

Eva Huber, Sebastian Sauppe, Arrate Isasi-Isasmendi, Ina Bornkessel-Schlesewsky, Paola Merlo, and Balthasar Bickel. 2024. Surprisal from language models can predict erps in processing predicate-argument structures only if enriched by an agent preference principle. *Neurobiology of language*, 5(1):167–200.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. *arXiv preprint arXiv:2004.03066*.

Napoleon Katsos, Blanche Gonzales de Linares, Ekaterina Ostashchenko, and Elspeth Wilson. 2023. Perspective-taking in deriving implicatures: The listener's perspective is important too. *Cognition*, 241:105582.

Roger Kreuz and Gina Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, pages 1–4.

Roger J Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of verbal irony. *Journal of experimental psychology: General*, 118(4):374.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W Crocker. 2025. On the limits of llm surprisal as a functional explanation of the n400 and p600. *Brain Research*, page 149841.

Alexandra Kuznetsova, Per B Brockhoff, Rune HB Christensen, and Sofie Pødenphant Jensen. 2016. Tests in linear mixed effects models. *R package version*, 2:33.

Russell Lenth. 2023. emmeans: Estimated marginal means, aka least-squares means_. *R package version 1.8. 5*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nigel Mantou Lou and Kimberly A Noels. 2016. Changing language mindsets: Implications for goal orientations and responses to failure in and outside the second language classroom. *Contemporary Educational Psychology*, 46:22–33.

Stephanie Lukin and Marilyn Walker. 2013. Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the workshop on language analysis in social media*, pages 30–40.

Spyros Makridakis, Fotios Petropoulos, and Yanfei Kang. 2023. Large language models: Their success and impact. *Forecasting*, 5(3):536–549.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Ira A Noveck and Anne Reboul. 2008. Experimental pragmatics: A gricean turn in the study of language. *Trends in cognitive sciences*, 12(11):425–431.

Henri Olkoniemi, Sohvi Halonen, Penny M Pexman, and Tuomo Häikiö. 2023. Children's processing of written irony: An eye-tracking study. *Cognition*, 238:105508.

Nausicaa Pouscoulous. 2023. More than one path to pragmatics? insights from children's grasp of implicit, figurative and ironical meaning. *Cognition*, 240:105531.

Ghaleb Rabab'ah, Mariam Cheikh, and Mutasim Al-Deaibes. 2024. Unraveling conversational implicatures: A study on arabic efl learners. *Open Cultural Studies*, 8(1):20240006.

Jessica Rett. 2020. Manner implicatures and how to spot them. *International Review of Pragmatics*, 12(1):44–79.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Elspeth Wilson and Napoleon Katsos. 2016. In a manner of speaking: an empirical investigation of manner implicatures. *Pre-proceedings of Trends in Experimental Pragmatics*, pages 170–176.

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 *China National Conference on Chinese Computational Linguistics*, pages 402–418. Springer.

Jun Zhang and Yan Wu. 2023. Epistemic reasoning in pragmatic inferencing by non-native speakers: The case of scalar implicatures. *Second Language Research*, 39(3):697–729.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing*.

Shuyan Zhao, Jie Ren, Michael C Frank, and Peng Zhou. 2021. The development of quantity implicatures in mandarin-speaking children. *Language Learning and Development*, 17(4):343–365.

## A Example of Experimental Material

## B Tables

| Trigger | Context | Stimulus | Binary Choices |
|---|---|---|---|
| *Sarcasm* | After the driver missed a turn, with a raised eyebrow, Sarah said, | "Wow, you're the best driver I've ever seen!" | (a) Sarah thought the person was not a good driver. (b) Sarah praised the person's driving skills. |
| *Sarcasm* | After the driver missed a turn, with a raised eyebrow, Sarah said, | "Wow, you're the worst driver I've ever seen!" | (a) Sarah thought the person was not a good driver. (b) Sarah praised the person's driving skills. |
| *Numerals* | Mike, a college student, asks Alex, another student, "How many courses are you taking this semester?" to which Alex responds, | "I'm taking five courses." | (a) Alex is taking exactly five courses. (b) Alex is taking at least five courses, maybe more. |
| *Some-all* | Mike says, | "I have lots of homework, but... I've done some of it." | (a) Mike has done a portion of the homework. (b) Mike may have done all of the homework. |
| *Some-all* | Mike says, | "I have lots of homework, but... I've done all of it." | (a) Mike has done a portion of the homework. (b) Mike may have done all of the homework. |
| *Causality* | The history teacher explains, | "The intruders were the reason behind the passing of the victims." | (a) The intruders killed the victims directly. (b) The intruders contributed to the death of victims indirectly. |
| *Causality* | The history teacher explains, | "The intruders killed the victims." | (a) The intruders killed the victims directly. (b) The intruders contributed to the death of victims indirectly. |
| *Modality* | In a warehouse, Alex says, | "Carolyn had the ability to paint the board." | (a) Carolyn could paint the board (and she did). (b) Carolyn could have painted the board but she didn't. |
| *Modality* | In a warehouse, Alex says, | "Carolyn was able to show the machine." | (a) Carolyn could paint the board (and she did). (b) Carolyn could have painted the board but she didn't. |
| *Double negation* | Reviewing a book manuscript, Alex says, | "In my view, this book isn't uninteresting." | (a) This book is very interesting. (b) This book is somewhat interesting. |
| *Double negation* | Reviewing a book manuscript, Alex says, | "In my view, this book is interesting." | (a) This book is interesting. (b) This book is somewhat interesting. |

| Participants | Estimate ($\beta$) | SE | df | t | *p*-value |
|---|---|---|---|---|---|
| GPT-Neo vs. RoBERTa | -0.091 | 0.107 | 20 | -0.850 | 0.677 |
| GPT-Neo vs. Human | -0.442 | 0.107 | 20 | -4.129 | 0.001** |
| Human vs. RoBERTa | 0.351 | 0.107 | 20 | 3.279 | 0.010* |

Table 2: Tukey's Post-Hoc Comparison Between GPT-Neo, RoBERTa and Human *(Note: SE indicates standard error. df indicates degrees of freedom. Significance levels of p values *p < .05, **p < .01, ***p < .001.)*