Conversational Implicatures Through the Lens of LLMs

Agnese Lombardi, Alessandro Lenci
University of Pisa, CoLing Lab
agnese.lombardi@phd.unipi.it, alessandro.lenci@unipi.it

Abstract

Recent research has explored the capacity of Large Language Models (LLMs) to perform pragmatic reasoning and interpret complex pragmatic phenomena. However, such phenomena are inherently ambiguous, and even human evaluations are highly variable. Many existing studies directly compare human and model responses while assuming a single "correct" interpretation, thereby overlooking the natural variability that characterizes human pragmatic understanding. This raises two key issues: (1) the need for novel evaluation methods that account for interpretive variability and allow for meaningful comparison between humans and models, and (2) the potential limitations of current linguistic theories in capturing the richness of human pragmatic behavior. We propose that LLMs can serve not only as benchmarks for human-model alignment, but also as tools for investigating the nature of pragmatic phenomena and their relationship to linguistic theory. To this end, we developed a handcrafted dataset encompassing eight types of conversational implicatures. Our study addresses three main research questions: (1) Do LLMs process conversational implicatures differently from humans? (2) If so, how do these differences manifest? (3) What do these findings reveal about the cognitive capacities of LLMs and the explanatory adequacy of pragmatic theory?

Keywords: Conversational Implicatures, LLMs, Pragmatics

1. Introduction

Recent advances in computational modeling, along with their new emerging abilities, have brought to the forefront the challenge of evaluating these systems, particularly in relation to their cognitive and reasoning capabilities. Notably, Large Language Models (LLMs) appear predisposed to errors within the domain of functional competencies, tasks that demand extralinguistic knowledge, rather than formal competencies (Mahowald et al., 2024). In response to these developments, numerous benchmarks have been designed to rigorously assess these higher-level abilities (Sravanthi et al., 2024; Srivastava et al., 2023).

Benchmarks play a fundamental role in the evaluation process, as they enable the automation of assessments and facilitate comparative analysis across different models. However, most existing evaluations exhibit several limitations as they often lack a fine-grained analysis of model outputs and fail to thoroughly assess error patterns (Hu et al., 2023). Furthermore, benchmarks tend to prioritize aspects such as reasoning, computation, and knowledge, while neglecting the evaluation of intended non-literal meanings in linguistic expressions (Park et al., 2024). They also fall short in determining whether models are effectively aligned with human expectations and natural language capabilities. Additionally, many benchmarks rely on multiple-choice question-answering (MCQA) formats to streamline evaluation, which inadequately captures the nuanced aspects of language, particularly in the domain of pragmatics (Wu et al., 2024). Beyond their use as evaluation tools, we view LLMs as **theoretical instruments** that can contribute to fundamental questions about linguistic structure, language processing, and learning. In doing so, they compel us to reconsider long-standing assumptions and arguments within linguistic theory (Futrell and Mahowald, 2025).

To address some of these limitations, we focused on a fine-grained analysis on **Conversational Implicatures** (Grice, 1975), systematically deriving implicature classifications and stimulus patterns from the linguistic literature. Specifically, we conducted a comparison of eight distinct types of conversational implicatures across twelve different LLMs. We aim to analyse the following questions:

1. **RQ1**: Do LLMs process conversational implicatures in ways that differ from humans?
2. **RQ2**: If so, how do these differences manifest?
3. **RQ3**: What do these findings suggest about the capacities of LLMs and the explanatory scope of pragmatic theory?[1]

2. Related Work

Prior research has investigated LLMs' ability to handle various pragmatic phenomena. For instance, Hu et al. (2023) conducted a comprehensive comparison between LLMs and humans across seven pragmatic categories: deceit, indirect speech acts, irony, Gricean maxims, metaphor, humor, and coherence inferences. Similar analyses have been carried out for other languages, such as Italian (Barattieri di San Pietro et al., 2023) and Korean

[1]Code and dataset: https://github.com/agneselombardi/Conversational_Implicatures.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4955–4966
11–16 May 2026.

(Park et al., 2024). Other studies have focused on specific pragmatic aspects, such as scalar implicatures (Li et al., 2021; Cho and Kim, 2024) or figurative language (Stowe et al., 2022; Balestrucci et al., 2024). Closely related to our work, several papers have examined Gricean implicatures, exploring how LLMs interpret or generate them across different languages and data sources (Yue et al., 2024; Zheng et al., 2021; Jeretic et al., 2020). In addition, several benchmark datasets assess social commonsense reasoning, a key component underlying implicature understanding (Sap et al., 2019; Xu et al., 2025). Building on previous work, we create a linguistic-based dataset to test LLMs and explore whether their pragmatic behavior coheres with linguistic theory.

3. Conversational Implicatures

A **conversational implicature** is a type of nonconventional inference in which the speaker's intended meaning is implied rather than directly expressed. Its interpretation relies on the assumption that the speaker is guided by conversational norms, encapsulated in Grice (1975) *Cooperative Principle* and later elaborated by Levinson (1983) and Crystal (2008). According to Grice's Maxims, speakers are expected to provide the appropriate amount of information without redundancy (*Maxim of Quantity*), to be truthful and evidence-based (*Maxim of Quality*), to remain relevant to the ongoing discourse (*Maxim of Relation*), and to be clear, orderly, and unambiguous (*Maxim of Manner*). Conversational implicatures typically emerge when a speaker appears to violate one of these maxims, yet still upholds the *Cooperative Principle* at a deeper level, a process known as **flouting**. Through recognizing such apparent violations, interlocutors derive the implied meaning, illustrating the subtle and inferential character of pragmatic communication.

We classified implicatures in the following way:

1. **Particularized Conversational Implicatures (PCI)**. This is a context-dependent implicature that requires specific contextual cues for interpretation.

   A: What happened to the roast beef?
   B: The dog is looking very happy.

   The implicature is that the dog ate the roast beef, inferred through the Maxim of Relevance (Levinson, 1983).

2. **Generalized Conversational Implicatures (GCI)**. This is a type of implicature derivable without special contextual support, e.g., scalar implicatures.

   The water is warm.

   The implicature is that the water is not hot (Levinson, 1983).

3. **Indefinite Article (a form of GCI)**. Following Grice, the use of *a/an* implies that the referent is not closely associated with the speaker, i.e., it is not definite. Contextual manipulations can shift interpretation toward definiteness.

   John walked into a house yesterday and saw a tortoise.

   The implicature is that the house is not John's own (Grice, 1975).

4. **Together Implications**. Coordinated subjects in generic expressions imply collective action unless context specifies otherwise.

   John and Jerry bought a piano.

   The implicature is that they purchased it together (Harnish, 1976).

5. **Coreference**. According to Levinson, the choice between pronouns and definite descriptions influences local referential assumptions. A pronoun (e.g., he) strongly signals local coreference with an already mentioned referent. A definite description (e.g., the man) generally signals a new referent, so local coreference is not expected.

   i) Jerry Rich came in. He walked to the window.
   ii) Jerry Rich came in. The man walked to the window.

   While *i) he* refers to Jerry Rich, in *ii) the man* suggests a different referent (Levinson, 1983).

6. **Bridging**. A referent in one clause is inferred to belong to the preceding discourse, even without pronouns or explicit markers.

   John unpacked the picnic. The beer was warm.

   The implicature is that the beer was part of the picnic (Clark and Haviland, 1977).

7. **Indirect Speech Acts (ISAs)**. Utterances where the intended meaning differs from the literal meaning, relying on pragmatic inference.

   Do you know the time?

   The implicature is a request for the time rather than a literal inquiry into the listener's knowledge (Austin, 1975; Searle, 1975).

8. **Informativeness Principle (I-Principle)**. Levinson's I-Principle states that listeners enrich utterances beyond their literal meaning by drawing on world knowledge, preferring the most specific and informative interpretation that fits the context.

   He turned on the switch and the motor started.

   The implicature is that turning on the switch caused the motor to start (Levinson, 1983).

4. Task Structure

In total, we manually constructed 708 stimuli, organized by task. Each task targets one of the conversational implicatures introduced in Section 3. Examples of the tasks are illustrated in Figure 1.

For the **PCI** and **GCI** tasks, we employed a Multiple-Choice Question Answering (MCQA) format with four response options. In the PCI task, the options include both the context-dependent interpretation (PCI) and the context-independent interpretation (GCI), along with two distractors. In the PCI condition, the stimulus consists of a *Context utterance* and a *Target utterance* whose full meaning becomes apparent only when interpreted relative to the contextual background. In contrast, for the GCI condition, the *Context utterance* provides a neutral backdrop that merely frames the situation without constraining the implicature, it supplies minimal information to render the target reasonable.

The **Bridging** task is a completion task in which participants (human or model) must identify the most plausible completion among three alternatives. Each completion option elicits a bridging inference by implicitly explaining a contrast presented in the *Context utterance*. The context typically contains two contrasting propositions, and the possible completions provide implicit explanations for that contrast rather than direct answers.

In the **Coreference** task, we again use an MCQA format, focusing on the interpretation of a *Target sentence* following a *Context* that introduces a potential referent. The context primes the reader or listener to anticipate possible coreference in the target utterance, which contains a definite description (Levinson, 1983). This definite noun phrase (NP) may either refer back to the antecedent in the context or introduce a new referent, rendering the sentence potentially ambiguous in the absence of context. The answer options distinguish between coreferential and non-coreferential interpretations.

The **Together Implications** task is a binary (yes/no) task. Each *Target utterance* contains a coordinated subject joined by "and". The key ambiguity lies in whether the coordinated entities performed the action jointly or independently.

The **Indirect Speech Acts** task also employs the MCQA format, comprising a *Context utterance* and a *Target* that conveys an indirect request. Among the four answer options, both literal and non-literal interpretations are provided.

The Indefinite Article and Informativeness Principle tasks involve contextual manipulations (highlighted in light green in Figure 1).

The **Indefinite Article task** is an MCQA with three contextual variants of the same target utterance. In the first version, the target appears in an underspecified context, favoring an indefinite interpretation. In the second, the context biases interpretation toward a definite referent while retaining some uncertainty (see Figure 1). In the third, the context strongly enforces definiteness. Each set of answer options includes both definite and indefinite interpretations.

Finally, the **I-Principle task**, adapted from Foraker and Murphy (2012), is an MCQA comprising three versions of the same stimulus. Each version manipulates the *Target utterance* to bias the interpretation of an ambiguous term in the context. For instance, the context may include a polysemous noun such as *production*, which can have several interpretations. Each target version biases toward a specific sense: *the assembly line* implies an industrial context, *the chorus line* a theatrical one, and *the line* remains underspecified but plausibly suggests engineers repairing a production line. The four response options cover all these possible interpretations (workers, engineers, dancers), plus a fourth, underspecified option (it cannot be said).

5. Evaluation

We employed three distinct methods to query the models:

1. **Number generation**: The model was prompted to generate the numeral corresponding to the correct answer option.

2. **Probability measures**: We extracted both the probability assigned to the selected answer choice (i.e., its corresponding number) and the probability assigned to the full sentence representing that option.

3. **Free generation**: The task description and final question were provided without predefined response options. The model was allowed to freely generate an answer of up to 150 tokens. These responses were then evaluated using GPT-4o-mini, following the methodology and prompt design of Wu et al. (2024). Our approach diverged slightly from Wu et al. (2024). by treating the humans' preferred answer as the gold standard against which all generated outputs were assessed.

Figure 1: Examples of tasks corresponding to each type of implicature. Tasks highlighted in light green involve context manipulation.

5.1. Models

We evaluated a total of twelve models, including both base and instruction-tuned versions: **Mistral-7B** (Jiang et al., 2023), **Falcon-3-7B**, **Falcon-3-10B** (Almazrouei et al., 2023), **Llama-3.2-1B**, **Llama-3.2-3B**, and **Llama-3-8B** (AI@Meta, 2024).

5.2. Human Data

We collected human responses via the Prolific platform. The 708 stimuli were divided into 12 groups of approximately 60 items each, with every group assigned to 20 participants. In total, we obtained judgments from 240 speakers.

To assess inter-rater reliability, we computed **Krippendorff's alpha**, which measures the degree of agreement among annotators beyond chance.

Table 1 reports alpha values by phenomenon. Values of Krippendorff's alpha close to zero suggest that human annotators agreed only marginally beyond chance. Such low agreement is not unexpected for pragmatically ambiguous stimuli, where judgments are inherently noisy. We consider this case one of that in which disagreement between subjects may be the result of genuine differences in interpretation rather than of unclear guidelines

Phenomenon	α	**N**
Bridging	0.0197	100
Coreference	0.0017	100
GCI	0.0156	93
Indefinite Article	0.0087	93
I-Principle	0.0007	48
PCI	0.0087	144
ISA	0.0078	40
Together Implications	-0.0003	90

Table 1: Krippendorff's alpha and stimulus count for each phenomenon.

(Rizzi et al., 2024). This variability implies that any comparison between human and model probabilities must account for uncertainty.

To address this, we incorporated entropy as an evaluation measure, treating it as an index of human uncertainty. While the most frequently chosen human response was used as the gold standard for calculating accuracy, accuracy was employed only for comparing evaluation methods (Section 6). The primary analyses were conducted on distributions using weighted regression (Section 7), with entropy providing a principled way to assess whether model probability estimates align with the variability in human judgments.

6. Evaluation Comparison

Figure 2 shows the comparison of all evaluation methods in terms of accuracy, using the most frequently selected human response as the gold standard. The results show that **free generation** performs poorly across most tasks, with the exception of together implications, which take the form of a binary yes/no question. This suggests the unreliability of this method, as the GPT-based rating procedure only works when generated strings exactly match the expected output (yes/no task), contradicting the purpose of free response evaluation.

By contrast, **number generation** yields near-null results for the smallest models, reflecting a broader inability to follow instructions rather than a failure to encode pragmatic information. The method achieving the highest accuracy is the **probability** encoding the number of the option (prob choice in Figure 2). Consequently, we will use this as our primary evaluation measure in the next analysis.

The differences observed in the heatmap underscore the importance of applying and comparing multiple evaluation methods, as they capture distinct aspects of model behavior. For example, **number generation** highlights the limitations of the model's instruction-following, while probability-based measures reveal whether pragmatic competence is encoded (Hu and Levy, 2023). Even discrepancies between different probability measures provide insights into model functioning. Accuracy also varies substantially across phenomena. However, given the very low Krippendorff's α observed in the human data (see Section 5.2), accuracy against the human gold standard alone is not highly informative in this context. To more appropriately address RQ1, we therefore focus on **distributional probability comparisons** rather than discrete accuracy measures, adopting a soft-label evaluation approach (Uma et al., 2021).

6.1. Instruction Tuning

Figure 3 compares base models with their instruction-tuned counterparts. Although the number of models tested does not allow for strong generalizations, the results indicate that instruction tuning and model size do not lead to substantial performance gains. Notably, for larger models, the performance gap between base and instruction-tuned versions tends to diminish, suggesting that instruction tuning leads to a stronger different in smaller models.

7. Weighted Regression

Using accuracy would require designating the most frequent human response as the gold standard, thereby ignoring the distribution of alternative responses that reflect meaningful human variation (Basile et al., 2021). To better account for human uncertainty, we instead employed entropy derived from human response distributions as weights in an **Ordinary Least Squares (OLS) regression**, testing whether model probability estimates align with human preferences.

Our rationale was as follows:

- **Weighting by human certainty**: Items with low entropy reflect strong human consensus ("easier" items), while high-entropy items are more ambiguous. Models are expected to perform better on low-entropy items.
- **Comparison**: Model probabilities were used as the dependent variable, and human probabilities as the independent variable.
- **Weighted regression**: Weights were computed as $1 -$ entropy, assigning greater influence to items with clear human preference.

Ideally, if model probabilities align closely with human distributions, we would expect low-entropy (high-consensus) items to fall along the regression line, while high-entropy (uncertain) items should exhibit greater dispersion.

Variable	Coef.	Std.Err.	t	P> |t|
const	**0.2156**	0.004	52.18	0.000
human_prob	**0.3080**	0.005	56.23	0.000
Model Summary				
R^2 = 0.208, Adj. R^2 = 0.208, F = 3162, **Prob(F) = 0.000**				
LogLik = -4160.2, AIC = 8324, BIC = 8339, N = 12048				
DW = 2.297, JB = 2026.9 (p = 0.000), Skew = 0.624, Kurt = 4.575				

Table 2: WLS Regression Results

Model diagnostics indicated that the linear regression model was statistically significant ($F(1, 12046) = 3162$, $p < 0.001$), explaining approximately 20.8% of the variance in model probabilities ($R^2 = 0.208$). The predictor human probabilties had a significant positive association with model preferences ($\beta = 0.308$, $p < 0.001$), but the slope

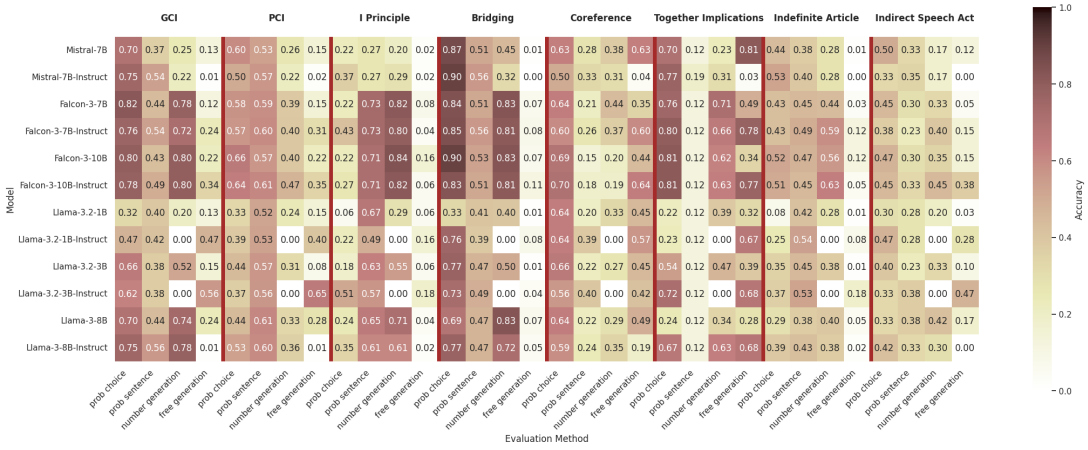

Figure 2: Accuracy comparison across all evaluation methods.

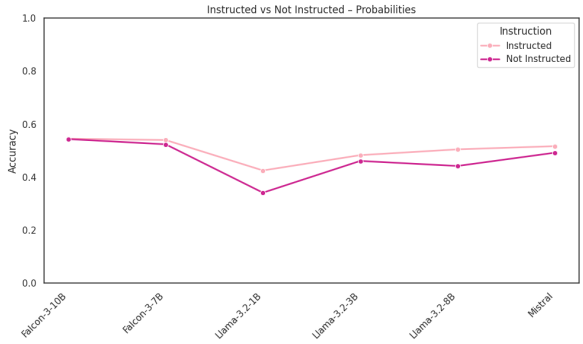

Figure 3: Probabilities comparison of instruction tuned vs non instructed models.

< 1, so the **model tends to underpredict human preferences**. Residual analysis showed no evidence of autocorrelation (Durbin–Watson $= 2.30$), although residuals exhibited mild deviations from normality (Jarque–Bera $p < 0.001$). These results support the adequacy of the linear model, though unmodeled nonlinear effects may account for some unexplained variance.

8. Distribution

Figure 4 shows violin plots used to visualize the spread of model probabilities across all answer options for each phenomenon. Wide sections indicate probability mass distributed across multiple answers, whereas narrow peaks reflect concentration on a few options. Human probability distributions are overlaid as grey dots.

When comparing across models, clear differences emerge in both calibration and human-likeness. Smaller models (e.g., *LLaMA 1B* and *LLaMA 2B*, both base and instruct versions) show larger deviations from human response distributions. In contrast, mid-sized models (*Mistral 7B*, *Falcon 7B*, and *LLaMA 8B*) produce distributions that better capture human variability, although their absolute probability estimates are often less accurate. Larger models (*Falcon 10B Base/Instruct* and *LLaMA 8B Instruct*) exhibit narrower violin plots, reflecting higher confidence. In some cases, such as *Bridging* and *Coreference*, this confidence aligns with human judgments; however, in others, particularly *I-Principle* and *Indirect Speech Acts*, it signals overconfidence and reduced flexibility. Across architectures, instruct-tuned variants consistently display slightly closer alignment with human responses, suggesting that instruction tuning enhances pragmatic calibration and helps models better approximate human uncertainty patterns.

We additionally use soft evaluation metric, comparing human distribution with model distribution. We computed:

- **Spearman and Pearson correlations** between model-assigned probabilities and human distributions.
- **KL Divergence** for each question, treating the human distribution as the *gold standard*.
- **Cross-Entropy** to measure how "surprised" humans would be if their distribution were replaced by the model's (Rizzi et al., 2024).
- **Entropy comparison** between humans and models to assess calibration of uncertainty.

Table 3 reports values of this comparison. Phenomena that show high correlations and low KL divergence, such as **Bridging**, **Coreference**, **Generalized Conversational Implicatures**, and **Together Implications**, are well captured by the models, suggesting that model predictions align closely with human judgments in these contexts.

In contrast, phenomena with low or negative correlations, including **I-Principle**, **Indefinite Article**, **Indirect Speech Acts**, and **Particularized Conversational Implicatures**, diverge more substantially

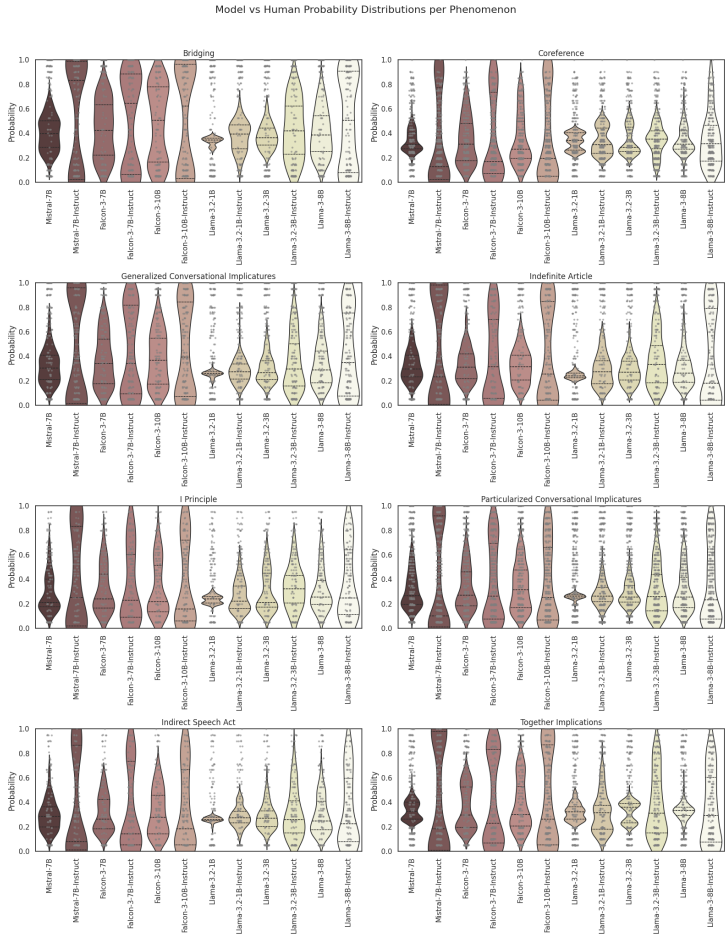

Figure 4: Violin plot displaying distribution of models probabilties and human probabilities.

Scenario	Pearson	Spearman	KL Div.	Cross-Ent.	Human Ent.	Model Ent.
Bridging	0.6855	0.6527	0.2109	5.0611	4.8502	4.9466
Coreference	0.5336	0.4652	0.2753	5.6670	5.3918	5.4011
GCI	0.5971	0.6025	0.2533	5.1689	4.9156	5.0055
I Principle	-0.0199	-0.0011	0.5261	5.0267	4.5006	4.5054
Indefinite Article	-0.0360	-0.0307	0.6469	5.4963	4.8494	4.9045
Indirect Speech Act	0.1943	0.2222	0.5333	4.9025	4.3692	4.4025
PCI	0.3530	0.3591	0.4456	5.9683	5.5227	5.6088
Together Implications	0.5323	0.4989	0.2857	5.3300	5.0443	5.0987

Table 3: Summary across scenarios: correlations, divergence, and entropy measures across linguistic phenomena.

from human distributions. Interestingly, most of these cases **might involve Theory of Mind (ToM) reasoning**, as they require inferring the speaker's intentions or mental states. The involvement of ToM in pragmatic reasoning remains a matter of debate, largely because it is difficult to isolate pragmatic inferences that do not rely on some form of mentalizing (Hu et al., 2025; Bosco et al., 2018).

The entropy analysis indicates that, while the models are generally well-calibrated in their overall uncertainty, they do not necessarily assign high probability to the same options that humans consider plausible. In other words, models tend to approximate human uncertainty patterns, but **they are often overconfident in ambiguous situations**, reflecting a mismatch in how uncertainty is distributed across possible interpretations.

9. Error Analysis

To address RQ3, we performed a qualitative error analysis to investigate the nature of model mistakes and their correspondence with human responses. This analysis complements the quantitative results by revealing where and how model behavior diverges from human reasoning, offering insight into the cognitive and computational underpinnings of these difference

The dataset was originally annotated according to linguistic theory, with the most probable answer reflecting theoretical predictions. These labels were not used in previous analyses, as our primary

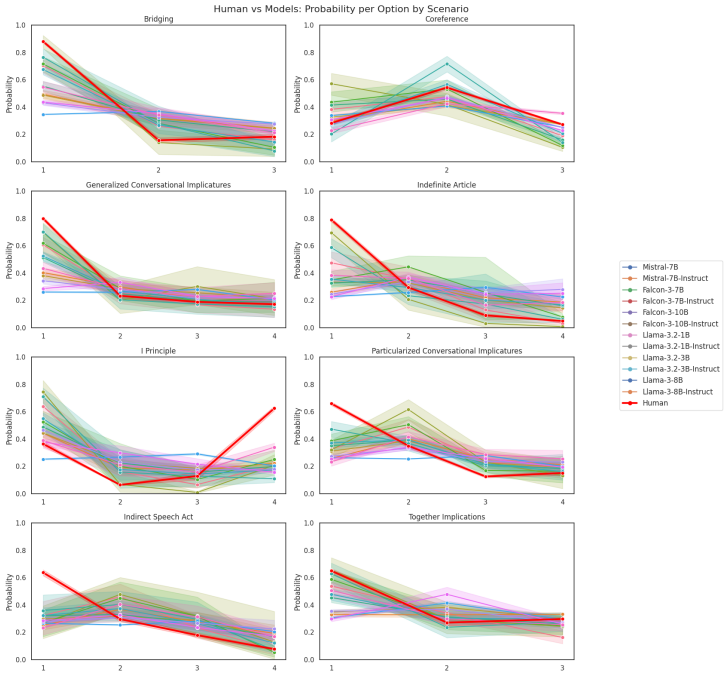

Figure 5: Error analysis reporting distributions of models and humans for every answer. The theory annotated answer is always the first option. Red line represents human's answer.

focus was on human-model alignment. In this section, we compare both humans and models against the **theory-annotated answers** to identify areas of agreement and divergence. In all stimuli, the theory-consistent answer is ordered first. Figure 5 illustrates result of this analysis.

For the **I-Principle**, humans frequently selected option 4 (*It cannot be said*), indicating that the task may have been challenging or underspecified. When a clear preference emerged, human choices clustered around option 1, the response predicted by linguistic theory, suggesting that participants nonetheless displayed sensitivity to the intended pragmatic inference. In **Particularized Conversational Implicatures**, models disproportionately favored option 2, corresponding to the **Generalized Conversational Implicature** that can be derived without engaging ToM reasoning. This bias reflects a systematic preference for general, surface-level inferences over deeper, intention-based interpretations. For **Indirect Speech Acts**, humans predominantly chose the theory-consistent option 1, while models tended toward the literal interpretation, again failing to access the ToM-based pragmatic layer required for indirect meaning resolution. In the case of **Together Implications**, some models diverged from human interpretations, though the deviations were less consistent and not clearly patterned across architectures. Finally, **Coreference** presents an interesting contrast: both humans and models occasionally depart from theoretical predictions. In our stimuli, contextual cues introduce a potential referent, biasing interpretation toward a coreferential reading even when, under Levinson (1983) definition, the sentence should not be considered coreferential. This pattern suggests that both humans and models rely primarily on semantic and discourse-level links rather than on purely grammatical pronoun cues, highlighting the influence of contextual coherence in reference resolution.

10. Conclusions and Future Directions

Taking into account the results and discussions presented across all experiments, we can now directly address the initial research questions.

RQ1. As discussed in Section 7, model probability estimates exhibit partial alignment with human preferences. However, the low R^2 and sublinear slope indicate that **model processing diverges from human pragmatic judgments**, and that model-derived probability distributions cannot be directly equated with human response patterns.

RQ2. Substantial variation emerges across models in how they process conversational implicatures and align with human distributions. As highlighted in Section 8, these differences are more pronounced for certain phenomena than for others. Overall, models capture aspects of human uncertainty patterns but fail to reproduce the full extent of human disagreement, particularly in ambiguous or ToM tasks.

RQ3. Not all pragmatic inferences are equally challenging. Patterns of divergence across phenomena suggest that certain types of implicatures demand more sophisticated inferential reasoning (e.g., ToM), offering insight into how models encode and resolve pragmatic meaning.

Future work should investigate the strategies models use to select responses and how these strategies differ from human reasoning. The observed distributional discrepancies, especially model overconfidence relative to human uncertainty, may reflect reliance on different cues or heuristics during inference. While processing differences clearly contribute to these effects, a deeper analysis could reveal shared and distinct mechanisms underlying human and model pragmatic reasoning.

It would also be informative to examine whether ToM underlies the differences observed across phenomena. Integrating neuroimaging methods, such as fMRI, could help identify which phenomena engage ToM processes in humans.

11. Bibliographical References

AI@Meta. 2024. Llama 3 model card.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

John L. Austin. 1975. *How to Do Things With Words*. Harvard University Press.

Pier Felice Balestrucci, Silvia Casola, Soda Marem Lo, Valerio Basile, and Alessandro Mazzei. 2024. I'm sure you're a real scholar yourself: Exploring ironic content generation by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14480–14494, Miami, Florida, USA. Association for Computational Linguistics.

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.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Francesca M. Bosco, Maurizio Tirassa, and Ilaria Gabbatore. 2018. Why pragmatics and theory of mind do not (completely) overlap. In *Front Psychol*.

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, Bangkok, Thailand. Association for Computational Linguistics.

Herbert H. Clark and Susan E. Haviland. 1977. Comprehension and the given-new contrast. *RoyO. Freedle (ed.), Discourse Production and Comprehension*.

David Crystal. 2008. *A Dictionary of Linguistics and Phonetics*, 6 edition. Blackwell Publishing, Oxford.

Stephani Foraker and Gregory L. Murphy. 2012. Polysemy in sentence comprehension: Effects of meaning dominance. *Journal of Memory and Language*, 67(4):407–425.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Behavioral and Brain Sciences*, page 1–98.

Paul H. Grice. 1975. Logic and conversation. *Speech Acts*, page 41–58.

Robert M. Harnish. 1976. Logical form and implicature. In Thomas G. Bever, Jerrold J. Katz, and D. Terence Langendoen, editors, *An Integrated Theory of Linguistic Ability*, pages 313–392. Thomas Y. Crowell, New York. Reprinted in Davis, S. (ed.), 1991, pp. 316–364.

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.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 380(1932):20230499.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University.

Elissa Li, Sebastian Schuster, and Judith Degen. 2021. Predicting scalar inferences from "or" to "not both" using neural sentence encoders. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 446–450, Online. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024. Pragmatic competence evaluation of large language models for the Korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266, Tokyo, Japan. Tokyo University of Foreign Studies.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

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)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

John R. Searle. 1975. Indirect speech acts. *P. Cole, and J. Morgan (Eds.), Syntax and Semantics 3: Speech Acts*, pages 59–82.

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*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi

Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao

Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

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.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. In *Journal of Artificial Intelligence Research*.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. 2025. Socialmaze: A benchmark for evaluating social reasoning 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 (Volume 1: Main Conference)*, pages 1270–1285, Taiyuan, China. Chinese Information Processing Society of China.

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.