Evaluating Pragmatic Reasoning in Large Language Models: Evidence from Scalar Diversity

Ye-eun Cho
Sungkyunkwan University
Seoul, South Korea
joyenn@skku.edu

Abstract

Evaluating pragmatic reasoning in large language models (LLMs) remains challenging because model behavior can vary depending on evaluation methods. Previous studies suggest that prompt-based judgments may diverge from models' internal probability distributions, raising questions about whether observed performance reflects underlying competence or task-induced behavior. This study examines this issue using scalar diversity as a graded diagnostic for pragmatic inference. Following Hu and Levy (2023), this study compares direct probability measurement and metalinguistic prompting across multiple models and experimental settings. The results show that neither evaluation method consistently outperforms the other and that pragmatic behavior varies substantially across model families, prompting strategies, and task structures. Moreover, scalar diversity gradients emerge only in specific model–condition combinations, suggesting that pragmatic reasoning in LLMs reflects an interaction between internal probabilistic representations and task-induced prompting behavior rather than a stable competence captured by a single evaluation paradigm. These findings highlight the central role of evaluation design in interpreting pragmatic abilities in LLMs.

1 Introduction

Recent advances in large language models (LLMs) have demonstrated remarkable performance across a wide range of linguistic tasks (Wang et al., 2019; Hu and Frank, 2024; Marjieh et al., 2024). Despite these successes, however, LLMs continue to show notable limitations in the domain of pragmatics, particularly in their ability to reason about implicit meaning (Mielke et al., 2022; Webson and Pavlick, 2022; Turpin et al., 2023; Cong, 2024). Interestingly, while many studies report limited pragmatic abilities in LLMs (Webson and Pavlick, 2022; Mielke et al., 2022; Turpin et al., 2023; Cong, 2024), other experiments have shown that certain prompt designs or task framings can substantially improve model performance on pragmatic reasoning tasks (Cong, 2024; Cho and Maeng, 2025; Cho, 2025; Shulginov et al., 2025). This raises a fundamental question: do LLMs genuinely possess pragmatic competence, or do they merely appear competent under specific evaluation conditions?

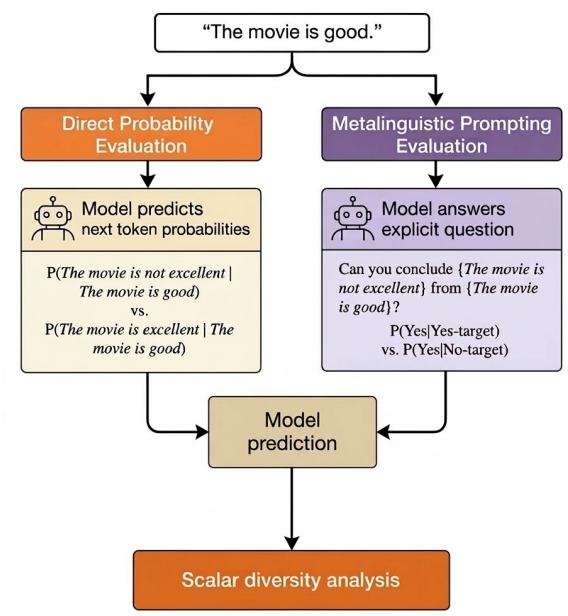

Figure 1: Overview of the two evaluation paradigms used in this study

To address this issue, the present study adopts the competence–performance framework proposed by Chomsky (1965). Competence refers to internalized linguistic knowledge, whereas performance reflects language use under particular cognitive and contextual conditions. Hu and Levy (2023) applied this distinction to LLM evaluation, emphasizing the importance of separating underlying model knowledge from prompt-induced behavior. How-

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 120–129
July 3, 2026 ©2026 Association for Computational Linguistics

ever, their work primarily focused on local linguistic phenomena, leaving pragmatic abilities largely unexplored.

We investigate this question through the phenomenon of scalar implicature, a widely studied form of pragmatic inference. Scalar implicature arises when, on the lexical scale <*some, all*>, a weaker term (e.g., *some*) leads to the inference that a stronger alternative (e.g., *all*) does not hold (Horn, 1972; Grice, 1975; Levinson, 2000). Importantly, the strength of such implicatures varies across lexical scales, a phenomenon known as scalar diversity (Van Tiel et al., 2016). For example, <*some, all*> reliably triggers implicatures in human interpretation, whereas pairs such as <*warm, hot*> often yield weaker inferences. This variability provides a useful diagnostic for evaluating models' sensitivity to pragmatic reasoning.

Following the evaluation framework proposed by Hu and Levy (2023), the present study examines LLMs' interpretation of scalar implicatures using two evaluation methods: direct probability measurement and metalinguistic prompting, as illustrated in Figure 1. In addition, we investigate how prompting strategies and model size influence the observed patterns.

2 Background

2.1 Direct vs. Metalinguistic Evaluation

Hu and Levy (2023) systematically compared two evaluation methods for assessing the linguistic knowledge of LLMs: direct probability measurement and metalinguistic prompting. Direct probability measurements evaluate LLMs by directly accessing their internal probabilistic representations, such as next-token probabilities or sentence-level pseudo-likelihoods. These probabilities are often interpreted as the most transparent reflection of what a model "knows," derived from its learned representations. In contrast, metalinguistic prompting presents models with natural-language queries that require them to make explicit judgments about linguistic stimuli. This approach asks the model to externalize its knowledge through prompted responses.

Using these two evaluation methods, Hu and Levy (2023) conducted a series of experiments across multiple linguistic domains, including word prediction, word comparison, and grammaticality judgment. They implemented one direct probability-based method and three metalinguistic prompting methods that varied in their structural similarity to the direct format. For instance, MetaQuestionSimple closely mirrors the direct method by placing the prediction target immediately adjacent to the query, whereas MetaInstruct introduces an instructional prompt format and MetaQuestionComplex embeds the target in a more elaborate context.

Their experiments revealed several key findings. First, metalinguistic judgments often diverge from direct probability measurements, indicating that models' explicit responses may not faithfully reflect their internal representations. Second, direct probability measurements generally outperform metalinguistic prompting in linguistic evaluation tasks. Third, presenting sentences as minimal pairs improves metalinguistic performance compared to evaluating sentences in isolation. Finally, the consistency between direct and metalinguistic responses decreases as the prompting format becomes more distant from the direct probability structure.

Based on these findings, Hu and Levy (2023) argued that LLM evaluation should distinguish between competence, reflected in internal probability distributions, and performance, reflected in prompt-based responses. From this perspective, models may possess linguistic knowledge that is not reliably expressed through metalinguistic prompts.

Subsequent study has extended this discussion to pragmatic reasoning. For example, Cong (2024) evaluates manner implicature using both probability-based and prompting-based methods, but does not directly compare these approaches within a unified model framework, limiting the comparability of the results. To address this limitation, the present study applies the competence–performance framework to scalar implicatures by systematically comparing direct probability measurements and metalinguistic prompting within the same models.

2.2 Scalar Implicature

Scalar implicature is a well-studied form of pragmatic inference in which the use of a weaker expression on a lexical scale leads the listener to infer that a stronger alternative does not hold (Horn, 1972; Grice, 1975; Levinson, 2000). Consider the example sentence (1).

(1) Some cookies were eaten.

Weak	Strong	Anchor Sentence	Yes-target	No-target
Good	Excellent	*The movie is good*	*The movie is not excellent*	*The movie is excellent*
Warm	Hot	*The soup is warm*	*The soup is not hot*	*The soup is hot*
...	...	...	...	...

Table 1: Sample items from the scalar diversity evaluation dataset

From a logical perspective, the quantifier *some* means *at least one and possibly all*. Therefore, the literal meaning of (1) is compatible with both the interpretation that *only one cookie was eaten* and the interpretation that *all cookies were eaten*. Nevertheless, in ordinary communication, listeners often infer (1) as *not all cookies were eaten*. This inference arises because the speaker chose the weaker expression (*some*) despite the availability of the more informative alternative (*all*), leading the listener to infer that the stronger alternative does not hold; otherwise, the speaker would have used the stronger expression instead.

Recent experimental research has shown that scalar implicatures do not occur uniformly across lexical scales. Instead, different scalar pairs vary in how strongly implicatures are inferred, a phenomenon known as scalar diversity (Van Tiel et al., 2016). Consider the examples in (2) and (3).

(2) a. John ate some cookies.
b. John did not eat all cookies.

(3) a. The coffee is warm.
b. The coffee is not hot.

Listeners readily infer the meaning in (2b) from the sentence in (2a), interpreting *some* as implying *not all*. In contrast, the inference from *warm* in (3a) to *not hot* in (3b) is much weaker and less consistently derived, illustrating that scalar implicatures vary in strength across different lexical scales (Van Tiel et al., 2016; Ronai and Xiang, 2021, 2024; Pankratz and Van Tiel, 2021; see also Degen and Tanenhaus, 2015).

This variability across scalar pairs provides a useful diagnostic for evaluating pragmatic reasoning. In the context of LLM evaluation, examining multiple scalar pairs rather than a single item allows for a more fine-grained assessment of models' pragmatic competence.

Previous work on LLMs has often evaluated scalar implicature using a single scalar pair such as <*some, all*> (Cho and Kim, 2024). However, focusing on a single item may overestimate models' pragmatic competence. To address this limitation, the present study adopts a scalar diversity framework for assessing models' pragmatic reasoning abilities.

3 Methods

3.1 Models

Following Hu and Levy (2023), the Flan-T5 family (Chung et al., 2024) was included, which is an instruction-tuned encoder–decoder language model. To examine whether the observed patterns generalize beyond this architecture, the Qwen2 family (Yang et al., 2024) was additionally evaluated. Qwen models are decoder-only instruction-tuned language models designed for autoregressive next-token prediction. This contrast enables a comparison based on different model architectures.

In addition, multiple model sizes were evaluated within each model family in order to examine potential scaling effects. For Flan-T5, three model sizes were used as: Small, Base, and Large. For Qwen, three parameter scales were evaluated as: 0.5B, 1.5B, and 7B.

3.2 Materials

The materials were adapted from the scalar diversity dataset developed by Ronai and Xiang (2024), which was originally designed to examine variability in the derivation of scalar implicatures (SIs) across a wide range of lexical scales. In the original dataset, each item consisted of a <weak, strong> scalar pair (e.g., <*good, excellent*>) embedded in a neutral carrier sentence (e.g., *The movie is good*), which could be interpreted either literally or with an SI-enriched meaning.

For the present study, the dataset was expanded using GPT-4o (OpenAI, 2024). While preserving the original scalar pairs, new carrier sentences were generated in the form "[something] *is good*", varying the surrounding content while keeping the SI trigger constant. Implausible or identical sentences were filtered and regenerated during the generation process. This procedure produced 100 distinct

Evaluation	Example
Direct	{The movie is good, The movie is not excellent}
MetaSimple	Can you conclude from {The movie is good} that {The movie is not excellent}? Respond with either Yes or No as your answer.
MetaInstruct	You are a helpful writing assistant. Tell me if you can conclude from {The movie is good} that {The movie is not excellent}. Respond with either Yes or No as your answer.
MetaComplex	Here is a sentence: {The movie is not excellent}. Can you conclude this from {The movie is good}? Respond with either Yes or No as your answer. Answer:

Table 2: Examples of direct probability and metalinguistic prompting in Experiment A

Evaluation	Example
Direct	{The movie is good, The movie is not excellent}
MetaSimple	Which sentence can you conclude from {The movie is good}?: 1) {The movie is not excellent} 2) {The movie is excellent}. Respond with either 1 or 2 as your answer.
MetaInstruct	You are a helpful writing assistant. Tell me which sentence you can conclude from {The movie is good}: 1) {The movie is not excellent} 2) {The movie is excellent}. Respond with either 1 or 2 as your answer.
MetaComplex	Here are two sentences: 1) {The movie is not excellent} 2) {The movie is excellent}. Which sentence can you conclude from {The movie is good}? Respond with 1 or 2. Answer:

Table 3: Examples of direct probability and metalinguistic prompting in Experiment B

carrier sentences for each of the 60 scalar pairs, yielding 6,000 sets of items in total.

As exemplified by Table 1, each evaluation item consists of an anchor sentence together with two candidate interpretations. The Yes-target represents the pragmatic interpretation associated with scalar implicature, while the No-target corresponds to the logical interpretation that does not require pragmatic enrichment. The dataset and experimental code used in this study are publicly available.[1]

3.3 Procedure

In the experiment, two experimental conditions were used. In Experiment A (sentence judgment), the model evaluates one candidate interpretation at a time and produces an absolute judgment. In Experiment B (sentence comparison), the model is presented with two candidate interpretations simultaneously and must select the preferred interpretation among the two.

In both experimental conditions, model preferences for pragmatic interpretations were evaluated using two paradigms: direct probability and metalinguistic prompting evaluation. For direct probability evaluation, the preference between the two candidate interpretations was estimated by comparing the conditional probabilities of the candidates given the anchor sentence. Specifically, the probability of each candidate continuation was computed as $P(\text{candidate} \mid \text{anchor})$, under the assumption that a continuation that forms a more acceptable interpretation of the anchor sentence will receive a higher conditional probability. For each item, the anchor sentence and each candidate sentence were concatenated and scored by the models. The candidate with the higher conditional probability was taken as the model's preferred interpretation and coded as TRUE; otherwise, it was coded as FALSE.

For metalinguistic evaluation, three prompting conditions were used: MetaSimple, MetaInstruct, and MetaComplex. Following Hu and Levy (2023), these prompt types vary in the distance between the prediction target and the question as well as in the degree of contextual framing. MetaSimple places the target sentence directly within the question. MetaInstruct adds an instructional frame (e.g., "You are a helpful writing assistant"), while MetaComplex embeds the target in a longer contextual frame. The interrogative form "Can you conclude {} from {}?" was adapted from Ronai and Xiang (2024) to elicit explicit inferential judgments.

In Experiment A, as in Table 2, the models were presented with an anchor sentence and a single candidate sentence (either Yes-target or No-target). The model responds with Yes or No. The log-probability of generating the token 'Yes' is computed for both candidates, and the difference between these probabilities is calculated as:

$$P(\text{Yes} \mid \text{Yes-target}) - P(\text{Yes} \mid \text{No-target})$$

In Experiment B, as in Table 3, the models receive the anchor sentence together with both candidate

[1] https://github.com/joyennn/pragmatic_reasoning

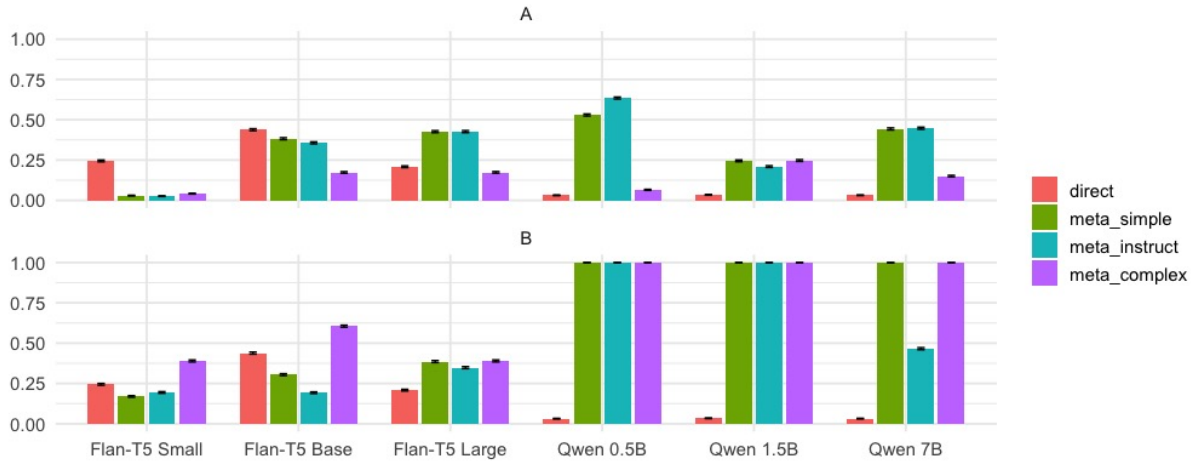

Figure 2: Overall accuracy of scalar inference predictions across models and prompting conditions for Experiments A (top) and B (bottom)

sentences simultaneously and must select between 1 (Yes-target) or 2 (No-target). The difference between the log-probabilities of the two responses is computed as:

$$P(1) - P(2)$$

In both experiments, a positive value indicates a preference for the Yes-target and is coded as TRUE. These TRUE responses were used for the subsequent analysis.

Finally, to examine the relationship between direct and metalinguistic evaluations, Pearson correlations were computed between the direct preference scores and the metalinguistic preference scores across items.

4 Results

4.1 Overall accuracy

Figure 2 presents the overall accuracy of scalar inference predictions across models, prompting conditions, and experimental paradigms. Accuracy was calculated as the proportion of TRUE responses across the entire item set. The results reveal systematic differences across prompting strategies, model families, model sizes, and experimental conditions as follows.

4.1.1 Direct probability vs. Metalinguistic prompting

First, a clear difference emerges between the direct probability and the metalinguistic prompting conditions. Across nearly all models and experimental settings, the direct probability condition consistently yielded lower accuracy than the metalinguistic prompting conditions. A small number of exceptions are observed in Experiment A for the smaller Flan-T5 models, where direct probability estimates slightly outperform some metalinguistic prompts, but the overall scores remain low.

To examine the relationship between the two evaluation paradigms, Pearson correlations were computed over overall accuracy scores obtained from direct probability estimation and metalinguistic prompting (see Table 4&5 in Appendix A). Overall, the correlations were generally weak (typically $r \approx .10 - .25$), indicating that the two evaluation strategies produce only partially overlapping patterns of overall accuracy. In several cases, correlations were near zero or even negative, particularly for Flan-T5 Large and some Qwen conditions. These results suggest that direct probability estimation and metalinguistic prompting capture substantially different aspects of model behavior in scalar inference tasks.

4.1.2 Metalinguistic prompting strategies

Differences are also observed among the three metalinguistic prompting strategies. While the metalinguistic prompts generally outperform the direct probability condition, the relative performance among the three varies across models and experimental settings. In many cases, the complex prompt (MetaComplex) shows noticeably different accuracy compared to the other two conditions, sometimes performing substantially lower and in other cases slightly higher. However, there are also instances in which all three metalinguistic prompts yield very similar accuracy, as well as occasional idiosyncratic patterns for particular models. Over-

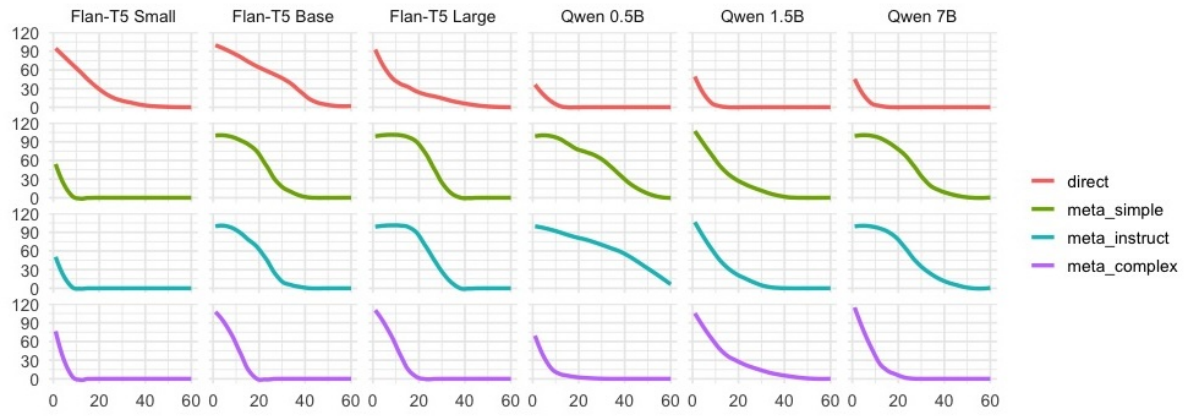

Figure 3: Item-level accuracy across scalar items for each model and evaluation condition in Experiment A

all, these observations make it difficult to identify a consistent ranking among the three prompting strategies, suggesting that the effectiveness of specific prompt formulations interacts with model characteristics and task conditions.

4.1.3 Model differences

In addition, clear differences emerge between the two model families, Flan-T5 and Qwen2. While both models show improvements under metalinguistic prompting relative to the direct probability condition, their overall behavioral patterns differ substantially. The Qwen models exhibit a highly polarized pattern: direct probability estimates are consistently very low, whereas metalinguistic prompting often yields extremely high accuracy, in several cases approaching ceiling levels in Experiment B. In contrast, the Flan-T5 models display more moderate patterns in which direct probability estimates remain relatively higher and occasionally even exceed the performance of some metalinguistic prompts.

One possible explanation for these contrasting behaviors lies in differences in model architecture and training objectives. Flan-T5 models are instruction-tuned encoder–decoder models, which may produce probability estimates that remain more sensitive to surface lexical cues in the input sentence. By contrast, Qwen models are decoder-only autoregressive models trained on large-scale instruction-following data, which may respond more strongly to explicit metalinguistic task framing.

4.1.4 Model size

Model size effects also differ across the two model families. Within the Flan-T5 family, a generally positive scaling trend can be observed under metalinguistic prompting conditions, with larger models typically achieving higher accuracy than smaller ones. In contrast, the Qwen models do not exhibit a consistent scaling trend. Across several conditions, larger Qwen models do not outperform smaller ones and in some cases even show slightly lower accuracy. This pattern suggests that increasing model size within this family does not systematically improve scalar inference performance.

4.1.5 Experimental settings

Moreover, there are clear differences between Experiment A and B. Across nearly all models and prompting conditions, accuracy is substantially higher in Experiment B than in Experiment A. This pattern suggests that the comparison-based evaluation used in Experiment B provides stronger cues for selecting the pragmatically enriched interpretation than the single-sentence judgment task used in Experiment A.

4.2 Item-level accuracy

While the overall accuracy analysis provides a useful summary of model performance, it does not directly capture the central phenomenon of scalar diversity. Scalar diversity refers to the graded nature of scalar implicature, where different items exhibit different likelihoods of pragmatic enrichment. That is, some scalar expressions strongly favor the pragmatic interpretation, whereas others do not.

To examine whether models capture this graded tendency, it is necessary to analyze item-level accuracy across scalar items. If models are sensitive to scalar diversity, their predictions should exhibit a gradual pattern across the item spectrum, rather

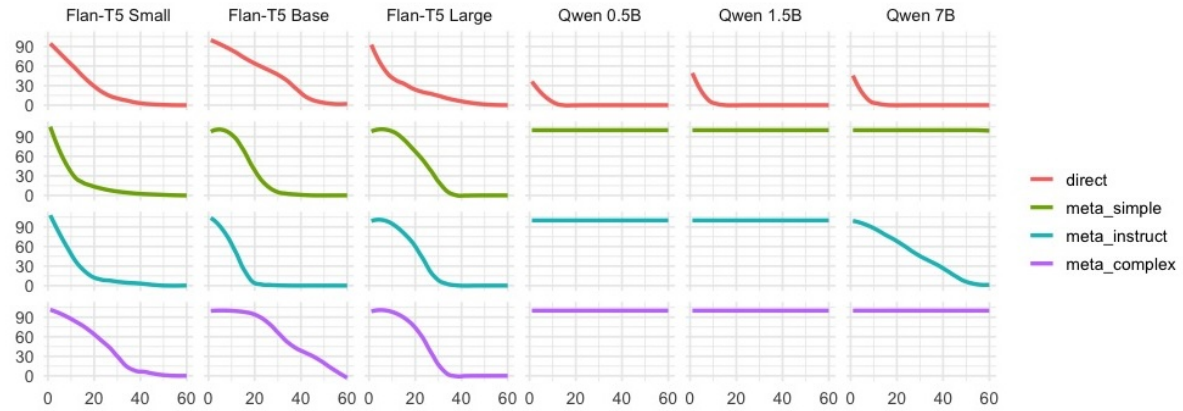

Figure 4: Item-level accuracy across scalar items for each model and evaluation condition in Experiment B

than a flat pattern or an abrupt threshold-like drop. Figures 3 and 4 therefore visualize the gradient patterns across items for each model and prompting condition in Experiments A and B. In addition, the steepness of each gradient was quantified using slope estimates (see Table 6&7 in Appendix B), which serve as a useful indicator for capturing the overall tendency of how model predictions vary across scalar items, although they do not provide a definitive measure of gradient structure.

4.2.1 Overall gradient

In Figures 3 and 4, scalar items are ordered according to their tendency to trigger pragmatic interpretations, and model accuracy is plotted across the item spectrum. If models capture this tendency, their predictions should show gradual variation across items.

Across the models, three broad patterns can be observed. First, several conditions show gradual gradients, with slope values typically between approximately -1 and -2, indicating continuous variation in model predictions across scalar items. Second, some conditions exhibit weaker gradients, with slopes close to zero or within the range of approximately 0 to -1, suggesting limited differentiation among scalar expressions. Finally, a small number of conditions produce nearly flat curves, with slopes effectively at zero, indicating that models assign the same interpretation to most items.

4.2.2 Direct probability vs. Metalinguistic prompting

In the direct probability condition, several models exhibit relatively weak gradients across scalar items. In particular, the Qwen models show curves that remain relatively flat across the item spectrum (-0.28 to -0.31) in both experiments. The Flan-T5 models display a somewhat different pattern. Although the curves also show declines across the item spectrum, the slopes are substantially larger in magnitude (-1.17 to -1.90), indicating stronger changes in predictions across scalar items. In other words, the direct probability condition produces clearer gradients in the Flan-T5 models than in the Qwen models.

Metalinguistic prompting often produces stronger gradients. In many cases, particularly within the Flan-T5 models, the metalinguistic prompts yield slope values between approximately -1 and -2 or lower, indicating a more pronounced change in predictions across items. However, the effects of metalinguistic prompting are not uniform across models. While some conditions show clearer gradients, others produce nearly constant predictions in the Qwen models, particularly in Experiment B. This variability indicates that the influence of prompting strategies interacts strongly with model architectures and experimental settings.

4.2.3 Model differences

Also, Flan-T5 and Qwen model families show substantial differences. Within the Flan-T5 models, the larger models (Base and Large) frequently exhibit stronger gradients across scalar items under metalinguistic prompting conditions. In these cases, slope values often fall between -1 and -2 or slightly lower, corresponding to smooth declines across the item spectrum. These patterns indicate that the models reflect item-level variation more clearly.

The smallest Flan-T5 model (Flan-T5 Small) displays weaker gradients in Experiment A under metalinguistic prompting, with slope values around

-0.25 to -0.40. These values correspond to relatively shallow curves with limited variation across items. In Experiment B, however, the same model shows stronger gradients under some prompting conditions, suggesting that the emergence of item-level variation depends partly on the evaluation paradigm.

The Qwen models display a different pattern. In the direct probability condition, the models consistently show relatively weak gradients across both experiments. Under metalinguistic prompting conditions, however, the Qwen models behave differently depending on the experimental paradigm. In Experiment A, several conditions produce relatively strong gradients across scalar items, with slope values around -1.5 to -2.2. In Experiment B, by contrast, several Qwen conditions produce almost perfectly flat curves across the item spectrum, corresponding to slope values of zero. These results indicate that the two model families differ not only in overall accuracy but also in item-level accuracy.

4.2.4 Experimental settings

Finally, two experimental paradigms show the different patterns. In Experiment A, several models, particularly under metalinguistic prompting, exhibit noticeable gradients across scalar items. In Experiment B, however, some models show substantially weaker variation across items. Most notably, the Qwen models under several metalinguistic prompting conditions produce nearly flat curves, indicating that pragmatic interpretations are assigned uniformly across scalar items. This behavior results in slope values that are effectively zero. These findings suggest that the extent to which scalar diversity is reflected in model predictions depends not only on model architectures and prompting strategies but also on the evaluation paradigm used to elicit scalar interpretations.

5 Discussion

5.1 Competence-Performance

The results suggest that pragmatic inference in LLMs cannot be attributed to a single evaluation paradigm. Across models, prompting strategies, and experimental settings, neither direct probability measurement nor metalinguistic prompting consistently produced superior performance. Moreover, the graded patterns predicted under scalar diversity did not appear uniformly across conditions: while some model–condition combinations exhibited gradual gradients across scalar items, others showed relatively flat or weakly varying patterns.

From a competence–performance perspective, these findings suggest that pragmatic reasoning in LLMs emerges through an interaction between internal probabilistic representations and prompting-based responses rather than mapping cleanly onto a single evaluation method. Direct probability measurements may reflect aspects of the model's underlying probabilistic competence, but this internal representation does not always manifest as scalar implicature behavior in isolation. Conversely, metalinguistic prompting can elicit explicit pragmatic responses, but these may reflect task-driven strategies rather than underlying knowledge. As a result, pragmatic competence cannot be reliably inferred from either evaluation paradigm alone. Instead, the observed patterns indicate that pragmatic reasoning depends jointly on various factors, such as model architectures, evaluation paradigms, and task-framing.

5.2 Comparison with Hu & Levy (2023)

Compared with the findings of Hu and Levy (2023), several similarities and differences emerge. First, consistent with their results, direct probability measurements and metalinguistic prompting often produced divergent predictions, indicating that the two evaluation methods capture different aspects of model behavior. Second, unlike Hu and Levy's finding that direct probability measurements generally outperform metalinguistic prompting, the present study shows no consistent performance advantage for either method. Third, consistent with prior findings, the comparison-based paradigm in Experiment B yielded higher accuracy than the single-sentence task in Experiment A. However, higher accuracy in the comparison setting does not necessarily indicate stronger pragmatic reasoning, as seen in the results of Experiment B. Finally, whereas Hu and Levy reported that prompt formats increasingly distant from the direct probability structure reduce consistency, the present results show no stable ordering among metalinguistic prompt types, suggesting that pragmatic inference is jointly shaped by model, condition, and prompt type.

6 Conclusion

This study examined how pragmatic reasoning in LLMs varies across evaluation methods. Using scalar implicature as a test case and scalar diversity as a graded diagnostic, this study compared direct probability estimation and metalinguistic prompting across models and conditions. The results show that pragmatic inference cannot be captured by a single evaluation paradigm: neither method consistently outperformed the other, and patterns varied across models and tasks. Moreover, graded patterns predicted by scalar diversity emerged only in certain model–condition combinations. From a competence–performance perspective, these findings suggest that pragmatic reasoning in LLMs reflects an interaction between internal probabilistic representations and prompt-based responses, underscoring the role of evaluation design.

Limitations

This study has several limitations. First, the analysis focuses on scalar implicature as a single pragmatic phenomenon. While scalar implicature provides a well-established test case for pragmatic inference, the findings may not generalize to other forms of pragmatic reasoning. Second, the evaluation design is limited to a specific set of prompting strategies and task paradigms. Different prompt formulations or evaluation settings may lead to different patterns of model behavior. Finally, the experiments were conducted on a limited set of model families. Although both encoder–decoder and decoder-only architectures were included, the results may not fully generalize to other LLMs with different training objectives or alignment strategies.

References

Ye-eun Cho. 2025. Prompting strategies of generative ai for korean pragmatic inference. *Korean Journal of Linguistics*, 50(2):423–455.

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.

Ye-eun Cho and Yunho Maeng. 2025. Can vision-language models infer speaker's ignorance? the role of visual and linguistic cues. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 298–308, Suzhou, China. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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

Judith Degen and Michael K Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science*, 39(4):667–710.

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

Laurence Robert Horn. 1972. *On the semantic properties of logical operators in English*. University of California, Los Angeles.

Jennifer Hu and Michael C Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. *arXiv preprint arXiv:2404.02418*.

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.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. 2024. Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

OpenAI. 2024. *Hello GPT-4o*. OpenAI.

Elizabeth Pankratz and Bob Van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4):562–594.

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.

Eszter Ronai and Ming Xiang. 2021. Pragmatic inferences are qud-sensitive: An experimental study. *Journal of Linguistics*, 57(4):841–870.

Eszter Ronai and Ming Xiang. 2024. What could have been said? alternatives and variability in pragmatic inferences. *Journal of Memory and Language*, 136:104507.

Valery Shulginov, Hasan Berkcan Şimşek, Sergei Kudriashov, Renata Randautsova, and Sofya A Shevela. 2025. Evaluating the pragmatic competence of large language models in detecting mitigated and unmitigated types of disagreement. In *Proceedings of the International Conference "Dialogue*, volume 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of semantics*, 33(1):137–175.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 2300–2344.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

A Pearson Correlations between Direct and Metalinguistic Accuracy

Model	MetaSimple	MetaInstruct	MetaComplex
Flan-T5 Small	.20	.20	.24
Flan-T5 Base	.18	.16	.16
Flan-T5 Large	−.10	−.11	.04
Qwen 0.5B	.20	.07	.08
Qwen 1.5B	.07	.12	.18
Qwen 7B	.24	.22	.27

Table 4: Pearson correlations between direct and metalinguistic accuracy (Experiment A)

Model	MetaSimple	MetaInstruct	MetaComplex
Flan-T5 Small	.11	.07	.12
Flan-T5 Base	.16	.14	.19
Flan-T5 Large	−.03	−.01	−.04
Qwen 0.5B	−.15	−.05	.24
Qwen 1.5B	.19	.11	−.15
Qwen 7B	.08	.14	−.14

Table 5: Pearson correlations between direct and metalinguistic accuracy (Experiment B)

B Slopes of Item-Level Accuracy across Scalar Items

Model	Direct	MetaSimple	MetaInstruct	MetaComplex
Flan-T5 Small	−1.52	−0.278	−0.248	−0.392
Flan-T5 Base	−1.90	−2.21	−2.15	−1.40
Flan-T5 Large	−1.17	−2.37	−2.37	−1.40
Qwen 0.5B	−0.278	−2.06	−1.52	−0.539
Qwen 1.5B	−0.308	−1.58	−1.46	−1.49
Qwen 7B	−0.285	−2.25	−2.23	−1.21

Table 6: Slopes of item-level accuracy across scalar items (Experiment A)

Model	Direct	MetaSimple	MetaInstruct	MetaComplex
Flan-T5 Small	−1.52	−1.15	−1.34	−2.05
Flan-T5 Base	−1.90	−2.00	−1.49	−2.01
Flan-T5 Large	−1.17	−2.24	−2.16	−2.28
Qwen 0.5B	−0.278	0	0	0
Qwen 1.5B	−0.308	0	0	0
Qwen 7B	−0.285	−0.003	−1.91	0

Table 7: Slopes of item-level accuracy across scalar items (Experiment B)