Pragmatic Theories Enhance Understanding of Implied Meanings in LLMs

Takuma Sato[1,2], **Seiya Kawano**[2,3], and **Koichiro Yoshino**[1,2,4]

[1]Nara Institute of Science and Technology, Nara, Japan
[2]Guardian Robot Project, RIKEN
[3]Kyoto Institute of Technology
[4]Institute of Science Tokyo
[a]sato.takuma.sq6@naist.ac.jp
[b]seiya.kawano@riken.jp
[c]yoshino.k.ai@m.titech.ac.jp

Abstract

The ability to accurately interpret implied meanings plays a crucial role in human communication and language use, and language models are also expected to possess this capability. This study demonstrates that providing language models with pragmatic theories as prompts is an effective in-context learning approach for tasks to understand implied meanings. Specifically, we propose an approach in which an overview of pragmatic theories, such as Gricean pragmatics and Relevance Theory, is presented as a prompt to the language model, guiding it through a step-by-step reasoning process to derive a final interpretation. Experimental results showed that, compared to the baseline, which prompts intermediate reasoning without presenting pragmatic theories (0-shot Chain-of-Thought), our methods enabled language models to achieve up to 9.6% higher scores on pragmatic reasoning tasks. Furthermore, we show that even without explaining the details of pragmatic theories, merely mentioning their names in the prompt leads to a certain performance improvement (around 1-3%) in larger models compared to the baseline.

1 Introduction

Language often contains implicit meanings, unstated intentions, and context-dependent interpretations, collectively referred to as **implied meanings**. The ability to correctly understand and interpret implied meanings play a crucial role in human communication (Grice, 1989; Levinson, 1983; Carston, 2002). In order to correctly interpret the implied meanings, people use background knowledge of communication, such as dialogue context, common sense, and cultural background, and achieve appropriate communication in their daily activities. The ability to understand implied meanings is important not only for humans but also for AI systems or robots based on them. For example, when a human's intention is ambiguous, inferring their intentions and taking actions proactively while considering the situation and common sense is necessary (Tanaka et al., 2024). The ability to interpret implied meaning, as well as understanding the nuances and intentions of language, is essential for AI and robots to be needed in human society.

In recent years, following the success of Large Language Models (LLMs), research has been conducted on employing them to develop systems that have abilities to interpret language and situation (Ahn et al., 2022; Ma et al., 2025). Numerous benchmark tasks and datasets have been proposed to assess these capabilities (Jeretic et al., 2020; Zheng et al., 2021; Takayama et al., 2021; Hu et al., 2023; Li et al., 2024a; Sravanthi et al., 2024; Yerukola et al., 2024; Yue et al., 2024), facilitating comparisons between current language models and human performance, as well as analyses of the challenges these models face.Numerous benchmark tasks and datasets have been proposed to assess these capabilities (Jeretic et al., 2020; Zheng et al., 2021; Takayama et al., 2021; Hu et al., 2023; Li et al., 2024a; Sravanthi et al., 2024; Yerukola et al., 2024; Yue et al., 2024), facilitating comparisons between current language models and human performance, as well as analyses of the challenges these models face.

Two major approaches have been explored to enhance the understanding of implied meanings in LLMs: post-training involving parameter updates (Wu et al., 2024; Sravanthi et al., 2024), and in-context learning (ICL), which draws out specific capabilities of LLMs through carefully designed prompts (Yerukola et al., 2024; Ruis et al., 2024; Kim et al., 2023). ICL is particularly important, as it enables the full utilization of the model' s abilities without additional training costs, and has been addressed with extensive approaches regarding its generalization capabilities. In the context of

This is the preprint version of the paper accepted to the main conference of IJCNLP-AACL 2025.

understanding implied meanings, there is a growing need to explore not only instance-specific methods but also more comprehensive, instance-agnostic methods.

In this study, we propose an in-context learning method that enhances model performance on tasks to understand implied meanings in a zero-shot setting without relying on prompts specific to particular problem formats or providing top-down hints from correct answers. Implied meanings are handled in the field of linguistics known as **pragmatics**, where various theories have been proposed regarding their properties and the mechanisms of their interpretation. The proposed method inserts summaries of theoretical frameworks from linguistic pragmatics into the model's prompt and instructs the model to generate intermediate reasoning processes by following those theories. It is known that existing LLMs have a variety of knowledge in their parameters; however, making them recall and operate knowledge appropriately and for the task is still challenging. By giving the rough outline of pragmatic theories as a bootstrap, we expected that these models could recall the related knowledge to be used and manipulate them for solving such tasks to understand the implied meanings.

Experimental results demonstrated that the proposed method consistently improved model performance on tasks to understand implied meanings without providing instance-dependent information, achieving an accuracy improvement of up to 0.096 in the experimental tasks. We tested the proposed method in both commercial-closed and open models and found that the proposed method contribute to a wide range of models. Notably, GPT-4o achieved higher scores than human scores by applying the proposed method. Additionally, slight score improvements were observed in many models even when the prompt did not include an overview of pragmatic theories but only mentioned their names while encouraging reasoning by following them. These investigations not only have engineering utility and can be expected to apply to upstream tasks such as dialogue, but also lead to clarifying the nature of the task of pragmatic understanding itself and what kind of thought processes are effective for executing such tasks. Our contributions can be summarized as follows:

- We proposed **a simple in-context learning method that incorporates summaries of pragmatic theories (namely Gricean and Relevance Theories) into prompts** and showed that this improves LLM performance on implied meaning understanding tasks without preparing task-dependent prompts.
- We showed that even without explaining the theories in detail, **simply referencing theory names** in zero-shot Chain-of-Thought prompting leads to performance improvements over baseline methods.
- Through detailed quantitative and qualitative analyses, we demonstrated that our method is particularly effective for tasks involving utterances that flout Grice's maxims and for interpreting irony.

2 Related Works

2.1 Pragmatic Reasoning by Language Models

Previous research has demonstrated that methods such as post-training, Chain-of-Thought (CoT) reasoning, and few-shot learning effectively enhance the pragmatic inference capabilities of language models. As post-training approaches using pragmatic reasoning datasets, policy optimization methods such as Direct Preference Optimization (DPO) are more effective than supervised fine-tuning (Wu et al., 2024). It has also been confirmed that instruction-tuned models achieve higher scores in pragmatic reasoning tasks compared to their corresponding base models (Ruis et al., 2024; Sravanthi et al., 2024). These methods involve updating the model's parameters.

For in-context learning methods that do not involve updating the model's parameters, previous research has shown that CoT prompting, where guidances for correct interpretation are included in the prompt (Yerukola et al., 2024), and few-shot prompting, where the prompts provide examples of problems and correct answers similar to the target task (Ruis et al., 2024), are effective approaches. As a method that combines these approaches, experiments have reported that providing reasoning steps based on Gricean theory (Grice, 1989) within few-shot examples can serve as effective guidance for correct interpretation (Kim et al., 2023).

However, there has been insufficient research on methods for enhancing the pragmatic reasoning capabilities of language models without ad hoc interventions, such as modifying model parameters or providing top-down hints specific to the task. For large-scale pretrained LLMs, conducting post-training with parameter updates (e.g., supervised fine-tuning or preference optimization) is not only costly and labor-intensive but also carries the risk of "catastrophic forgetting," where the performance on previously learned tasks deteriorates after fine-tuning on a specific task (Kirkpatrick et al., 2017; Li et al., 2024b). Additionally, existing in-context learning methods should be developed in light of the fact that pragmatic inference is often not an end goal in itself but rather a necessary component ability for higher-level tasks or objectives. Given this, the approaches in prior studies (Yerukola et al., 2024; Ruis et al., 2024; Kim et al., 2023), which rely on top-down and ad hoc applications of few-shot learning or CoT prompting tailored to specific pragmatic reasoning problems, may not be sufficient. We tackle their remaining challenge to maintain generalized prompts for interpreting implied meanings using pragmatic theories.

2.2 Pragmatic Theories in Linguistics

We proposes an in-context learning method based on two well-established pragmatic theories from the fields of linguistics and philosophy of language: Gricean pragmatics (Grice, 1989) and Relevance Theory (Sperbel and Wilson, 1995).

Gricean Theory Grice proposed a pragmatic theory in which he argued that the correct interpretation of what a speaker means in an utterance is achieved by assuming that participants in a conversation adhere to, or at least appear to adhere to, the *Cooperative Principle*. This principle serves as the foundation for the reasoning made by listeners.

> ***Cooperative Principle* (Grice, 1989)**
> Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged (p.26).

He also argued that adherence to the cooperative principle requires the observance of the four subordinate maxims, namely *Maxim of Quantity, Quality, Relation, and Manner*[1].

Grice explains the understanding and interpretation of implied meanings based on the idea that our conversations generally adhere to these maxims. When they do not, the listener infers meaning by recognizing that they are violating one or more maxims while assuming that the speaker still follows the *cooperative principle*. Grice's theory has had a significant impact not only in linguistics and philosophy of language but also in Natural Language Processing (NLP), where it has been widely applied (Krause and Vossen, 2024).

While Grice's contributions to pragmatics are undeniably foundational, subsequent research has pointed out various limitations and shortcomings in his theory, leading to the emergence of approaches that aimed to refine and extend his ideas (Levinson, 2000; Horn, 1984). Relevance Theory, which will be explained next, is one such development that seeks to further refine Grice's claims about utterance interpretation into a more cognitively plausible framework (Wearing, 2015).

Relevance Theory A key feature of Relevance Theory, proposed by Sperber and Wilson, is its cognitive approach to pragmatic meaning (Sperbel and Wilson, 1995). Relevance Theory defines *relevance* in terms of two factors: the *cognitive effect*, which refers to the degree to which an utterance influences the listener's thoughts, and the *processing effort* required to understand or interpret the utterance. All else being equal, an utterance is considered more relevant if it produces greater *cognitive effects* and requires less *processing effort*.

Relevance Theory asserts that, based on this notion of *relevance*, we expect the *presumption of relevance*[2] in communication and that pragmatic meaning is interpreted accordingly.

Our study proposes an in-context learning method that incorporates summaries of Gricean pragmatics and Relevance Theory into the model's prompt, guiding the model to generate reasoning processes aligned with these theories.

2.3 Zero-shot prompt templates

It is known that specifying certain thinking methods as prompts can generically improve model performance without giving language models problem-answer pairs (Brown et al., 2020) or providing explicit hints (Wei et al., 2022). Kojima et al. (2022) showed that simply inputting the text "Let's think step by step." to a model can significantly improve performance on benchmarks for various tasks, and this method is called *zero-shot Chain-of-Thought*. Such methods are not only convenient for practical use of language models, but also noteworthy for exploring the foundations and mechanisms that realize their capabilities, so various studies have

[1] For details of each Maxim, see the Appendix §A.

[2] For details, see §B in the Appendix.

been conducted subsequently (Wang et al., 2023a,b; Schulhoff et al., 2025).

In such contexts, several prompt templates have been discovered that aim to maximize LLM capabilities in specific tasks or domains. For example, (He et al., 2024) shows that prompt templates that make models extract keywords and topics from source sentences before performing final translation improve LLM performance on translation tasks. Additionally, (Sivarajkumar et al., 2024) demonstrated that zero-shot prompt templates based on medical domain knowledge are effective in clinical information extraction tasks. Methods of *jailbreak attacks* (Yi et al., 2024), which are intended to elicit harmful or inappropriate outputs from models by devising prompts, might also be positioned within this context.

3 Experiments

3.1 Dataset and Task

We use PRAGMEGA (Floyd, 2022; Hu et al., 2023) as a pragmatic reasoning dataset and task. This dataset covers seven broad pragmatic phenomena observed in English conversations. Each instance in the dataset requires correctly answering questions designed primarily to test understanding an utterance's implied meanings, considering the utterance itself and its conversational context.

Each problem in the dataset is classified into one of the seven pragmatic phenomena the dataset addresses. These seven pragmatic phenomena are *Deceits*, *Indirect speech*, *Irony*, *Maxims*, *Metaphor*, *Humor*, and *Coherence*. Since the present study specifically focuses on the task of correctly interpreting non-literal meanings in utterances, we conducted experiments targeting five of these pragmatic phenomena, excluding *Humor* and *Coherence*, which do not directly address such meanings. The following descriptions are provided in (Hu et al., 2023) for each phenomena[3]:

- ***Deceits***: *Polite deceits* used for social or personal relationships. In the questions, respondents must employ *Theory of Mind* and other reasoning skills to determine why the speaker used a particular utterance correctly.

- ***Indirect Speech***: Utterances with performative meanings, such as prompting others to take action. The questions test whether respondents understand what the speaker tries to convey in stories involving indirect requests.

- ***Irony***: Utterances that communicate the opposite of their literal meaning. The questions require respondents to correctly interpret what the characters intend to convey through the presented irony.

- ***Maxims***: Utterances that violate one of Grice's four maxims. Respondents determine why the characters made such utterances.

- ***Metaphor***: Utterances that depict comparisons between entities in a non-literal manner. The questions present metaphors within a story and ask respondents to interpret what the speaker is trying to convey.

The total number of problems is 520, comprising 100 instances each for *Deceits*, *Indirect Speech*, and *Metaphor*, 125 instances for *Irony*, and 95 instances for *Maxims*. Our experiments do not explicitly indicate the phenomenon to which each problem belongs within the prompt. Instead, a common prompt format is used across all phenomena to present the problem and response structure.

3.2 Methods of In-context Learning

We conducted comparative experiments on the following prompting methods (two baselines and two proposed methods) for in-context learning. Baseline-1/2 denote baseline methods, and Proposed-1/2 denote proposed methods. The baseline and proposed methods are instance-agnostic, meaning they do not provide models with instance-dependent information or top-down hints. Within these instance-agnostic methods, we hypothesize that our proposed methods, including summaries of pragmatic theory in prompts, would achieve higher performance on pragmatic reasoning tasks than the baseline methods. This is because, by indicating these theories, we expect the model to know which area of knowledge or reasoning contained in the model should be called. Note that all experiments adopt a zero-shot learning setup, meaning no task-solving examples are provided in the prompts.

Baseline-1: Simple A setting in which the prompts explicitly instruct the models to output **only** the final selected answer.

[3] Examples of problems corresponding to each phenomenon are shown in Table 2 in the appendix.

Baseline-2: Chain-of-Thought (`cot`) A setting in which the model is prompted with "`Firstly, think step-by-step and write down your process of thinking`," explicitly instructing it to output its reasoning process leading to the final answer, followed by the selected answer. This method is often referred to as *zero-shot Chain-of-Thought* ((Kojima et al., 2022)[a]).

Proposed-1: Gricean Prompting (`grice`) A setting in which the prompt provides the models with a brief overview of Gricean theory (Grice, 1989), explicitly instructing them to output a reasoning process aligned with this overview, followed by the final selected answer.

Proposed-2: Relevance Theory Prompting (`relevance`) A setting in which the prompt provides the models with a brief overview of Relevance Theory (Sperbel and Wilson, 1995; Carston, 2002), explicitly instructing them to describe a reasoning process aligned with this overview before outputting the final selected answer.

[a] Note that this method differs from the "Chain-of-Thought" approach presented in (Yerukola et al., 2024) and (Kim et al., 2023), as it does not provide instance-specific prompts or hints for correct interpretation (instance-agnostic).

3.3 Models

We conducted experiments using various LLMs implementing decoder-based Transformer architecture (Vaswani et al., 2017; Liu et al., 2018). We experimented with publicly available models (open models), whose source code and pre-trained parameters are released, and proprietary models (closed models), whose parameters are not publicly available. As Open models, we selected the LLaMa3 series (Grattafiori et al., 2024) and the Qwen2.5 series (Yang et al., 2025). As Closed models, we selected GPT-4o and GPT-4o mini (OpenAI, 2025)[4]. Considering the length of our prompts, we chose models with sufficiently large context lengths (16k tokens or more) for the experiments. All models were quantized to 8-bit before performing inference using vLLM (Kwon et al., 2023). Appendix §E shows the hyperparameters used in the experiments. About computation details, see §F in the Appendix.

4 Results

4.1 Experimental Results

The experimental results shown in Figure 1 confirm that the proposed methods perform better than the baseline methods on the pragmatic reasoning task[5]. Notably, with GPT-4o, the proposed methods enabled the model to achieve performance surpassing that of humans. Even in the case of phi-4, where there was a performance gap between the baseline methods and human scores, applying our methods allowed the model to reach human-level accuracy.

Using **Gricean prompting** (`grice`) resulted in higher scores than both baseline methods across all models tested. This method had the most significant effect on phi-4, where accuracy improved by 0.096 compared to the higher-scoring baseline method (`simple`). Even for Llama-3.1-8B-Instruct, where the most minor improvement was observed, accuracy increased by 0.032.

For **Relevance prompting** (`relevance`), no performance improvement was observed for Llama-3.1-8B-Instruct compared to the baseline, but all other models showed performance gains. When comparing `grice` and `relevance`, `grice` achieved higher scores in most models. While some models recorded higher scores with `relevance`, the difference from `grice` in those cases was minimal.

For **Short prompting** (`grice short`, `relevance short`), performance improvements over the baseline were observed in all models except Qwen2.5-7B-Instruct. Even in models where scores improved over the baseline, the magnitude of improvement was generally smaller than that achieved by the proposed methods.

4.2 Analysis

4.2.1 General Overview

When comparing `grice` and `relevance`, `grice` achieved higher scores in most models. While some models recorded higher scores with `relevance`, the difference from `grice` in those cases was minimal. However, this difference does not necessarily indicate that Gricean Theory is more valid than Relevance Theory. Instead, we speculate that the difference is due to Gricean Theory appearing more frequently in the training corpus of the models.

There was no clear trend indicating that longer input or output lengths consistently led to higher

[4] The experiments were conducted on March 2, 2025.

[5] The exact scores are presented in §G, I in the Appendix.

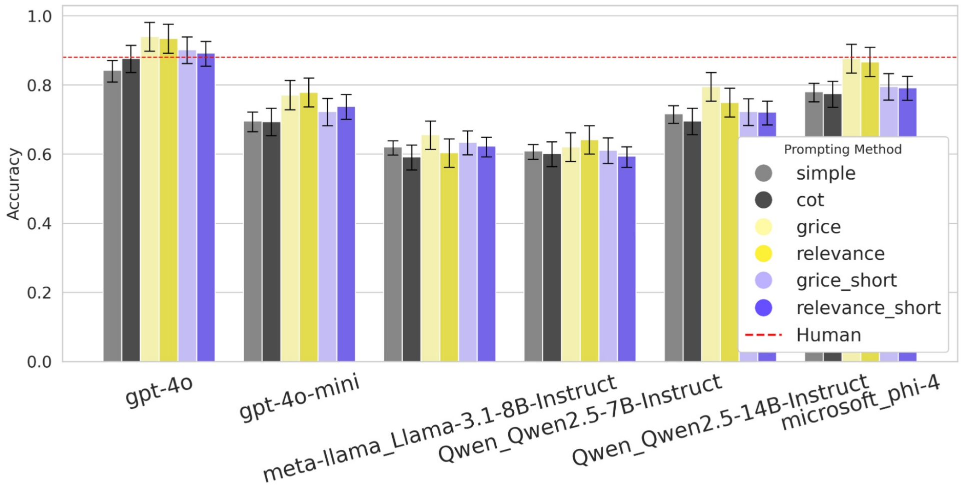

Figure 1: Accuracies on pragmatic inference task of PRAGMEGA. In most models, the proposed methods outperformed the baseline methods. The human scores indicate scores presented in the original paper by (Hu et al., 2023). Error bars represent 95% confidence intervals calculated using Wilson's method (Wilson, 1927). Even with a short prompt for the pragmatic theory, larger models showed improvements from the proposed methods; however, the extent of improvement was smaller compared to when the theory was explained in detail.

accuracy. An analysis of the relationships between input/output length and accuracy revealed little correlation between these factors. The Pearson correlation coefficients between accuracy and input length and accuracy and output length were 0.181 and 0.211, respectively. Additionally, the coefficient of determination (R^2) from simple regression analysis was 0.032 and 0.044, respectively (see Appendix §H for detailed analysis results).

4.2.2 Analyses by Pragmatic Phenomena

A summary of the analysis results for each of the five pragmatic phenomena included in PRAGMEGA, comparing different prompting methods using GPT-4o and Qwen2.5-7B-Instruct, is shown in Figure 2. [6] The results showed the most notable score improvement from the proposed methods in *Irony*. In the case of GPT-4o, baseline methods resulted in performance lower than that of humans; however, by applying the proposed methods, the model achieved human-level accuracy. For this phenomenon, regardless of whether the prompts provide the models with an overview of the theory, using Gricean Theory generally resulted in higher scores than Relevance Theory. In *Indirect Speech*, although the margin was smaller than in *Irony*, consistent improvements over the baseline methods were also observed. However, we found no clear superiority between Gricean Theory and Relevance Theory in this phenomenon. For *Maxims*, using Gricean Prompting, which includes an overview of Grice's maxims, did not always lead to improved scores over the baseline. However, for models with 14B or more parameters, the proposed methods generally resulted in higher accuracy. In most cases, models performed better when using Gricean Theory compared to Relevance Theory, and this trend was consistent even in the short prompts experiments. For *Metaphor*, models with 14B or more parameters tended to improve scores when using the proposed methods. In GPT-4o, GPT-4o-mini, and Qwen2.5-14B-Instruct, performance slightly improved with at least one of the proposed methods, even when using short prompts, though no performance improvement was observed with short prompts in phi-4. For *Deceits*, the effectiveness of the proposed methods varied depending on the model, and we observed no clear or consistent trend in relation to models' parameter size.

4.2.3 Error Analysis

We conducted error analysis using the results from GPT-4o, categorized errors into the following five patterns, and counted the number of instances corresponding to each phenomenon (Figure 3). For ① and ③, actual examples of errors are presented in Table 1. Examples of the other error patterns are provided in Appendix Table 11.

① Cases where the proposed methods were clearly effective This category includes cases where `simple` and `cot` produced incorrect answers,

[6] For exact scores, see §I in the appendix.

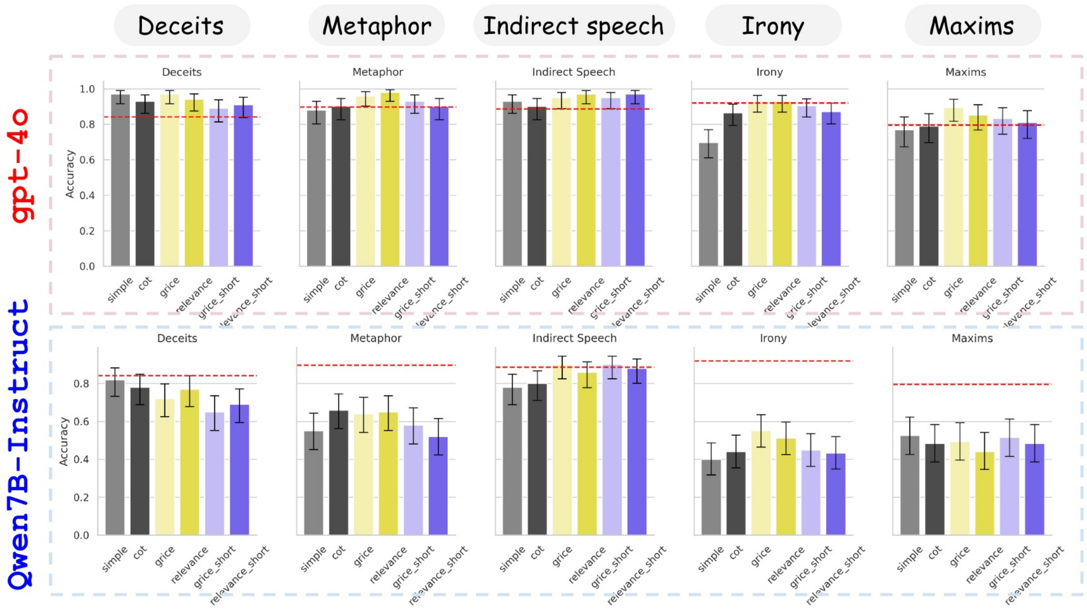

Figure 2: Accuracy of the model for each pragmatic phenomenon included in PRAGMEGA (Hu et al., 2023) when using different methods. Due to space constraints, we present the results for GPT-4o and Qwen2.5-7B-Instruct (for detailed results, including other models, see Appendix §I). The human score is based on (Hu et al., 2023).

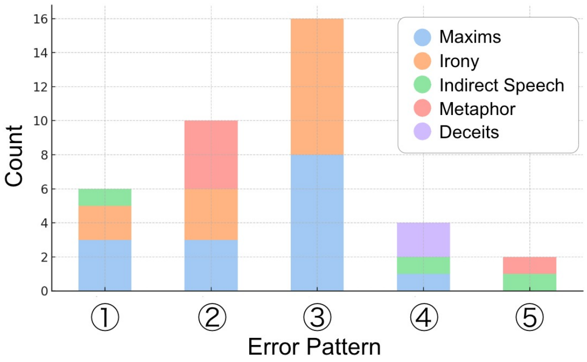

Figure 3: The number of instances for each error pattern by GPT-4o, as described in the main text. A cumulative bar chart represents these counts, including the distribution of each phenomenon within each pattern.

but all other methods resulted in correct answers. Many problems classified under *Maxims* fell into this category. The first example in Table 1 corresponds to this pattern. We hypothesize that including pragmatic theories in the prompt made the model more sensitive to the distinction between *what is said* and *what is implied*.

② Cases where *Short Prompting* was insufficient This category includes cases where `grice` and `relevance` resulted in correct answers, but all other methods produced incorrect answers. A relatively large proportion of problems in this pattern fell under *Metaphor*. This suggests that understanding metaphors may not benefit from a superficial grasp of pragmatic theories alone, but a more detailed comprehension of these theories could enable their application to interpretation in such cases.

③ Cases where all methods failed In this category, problems classified under *Maxims* and *Irony* appeared in equal numbers. Since Figure 2 also indicates that these two phenomena were the most challenging for the models, it is natural that they appear prominently in this error pattern. Indeed, the second example in Table 1 seems difficult even for humans. Furthermore, determining that option 4 is the correct answer might require additional contextual information. If such problems can be considered to lack a "definitive correct answer," then the performance of cutting-edge models like GPT-4o may already be saturated within our experimental setting.

④ Cases where only *Gricean Prompting* was effective This category includes cases where only `grice` and `grice_short` resulted in correct answers, while all other methods failed. The most frequent phenomenon observed in this pattern was *Deceits*. This phenomenon primarily deals with cases where individuals avoid directly conveying their thoughts due to social considerations (Hu et al., 2023). Gricean Theory explicitly explains this phenomenon through the concept of flouting the *Maxim of Quality*. This possibly explains why

Table 1: Actual question examples for some error patterns. **Bold** indicates the correct options. Due to space constraints, examples other than ① and ③ are provided in Appendix §J.

Pattern	Questions	Options
①	Samantha is talking with her dad about her fiance. Samantha notes: "John is an innocent person." Her dad replies: "Undoubtedly, as innocent as a saint." Why has Samantha's dad responded like this?	1. Samantha's dad is impressed with John's innocence. **2. Samantha's dad thinks that Samantha has an incorrect view of her fiance.** 3. Samantha's dad thinks that Samantha's fiance is a saint. 4. Samantha's dad thinks that John is too religious.
③	John is a teacher at an elementary school. When talking with the principal about a new student, who did poorly on her entrance examination, John said, "This one is really sharp." What did John want to convey?	1. The entrance exam is unfair. 2. The pencils need to be sharpened. 3. The student is smart. **4. The student is not very clever.**

Gricean Theory handles these cases more effectively than Relevance Theory. It could be valuable to explore how different results might emerge if theories that focus more on social aspects, such as Politeness Theory (Brown and Levinson, 1987), were incorporated into the prompts.

⑤ Cases where only *Relevance Prompting* was effective This category includes cases where only `relevance` and `relevance_short` resulted in correct answers, while all other methods failed. This pattern was relatively rare, with only two instances equally distributed between the *Metaphor* and *Indirect Speech* phenomena. Since the PRAGMEGA dataset used in this study primarily consists of single-turn utterances, we hypothesize that creating more challenging problems requiring longer contextual reasoning and background knowledge might lead to more cases where Relevance Theory proves effective.

5 Conclusion

In this paper, we proposed an instance-agnostic in-context learning method for pragmatic reasoning tasks by incorporating an overview of linguistic pragmatic theories into the model's prompt. Experimental results demonstrated that the proposed methods improve performance on pragmatic reasoning tasks without providing instance-dependent information to the model in a top-down manner. GPT-4o, in particular, achieved a score that surpassed human performance on the experimental task when using the proposed methods.

As a direction for future work, It is desirable to develop pragmatic reasoning tasks that require consideration of richer and more extended contexts, followed by experiments using such tasks. A limitation in PRAGMEGA and other previous studies is that the instances they handle often lack sufficient complexity compared to real-world pragmatic phenomena, or the contextual information relevant to interpretation is insufficiently rich. To address these issues, creating datasets that incorporate more complex and contextually rich pragmatic phenomena, along with tasks utilizing these datasets, is necessary for a more precise analysis of model capabilities and limitations. The dataset used in (Shisen et al., 2024), which is based on a Chinese sitcom, represents a promising approach to addressing these challenges. Focusing on this direction and developing more comprehensive datasets and benchmarks will contribute significantly to the further advancement of this research field.

It is also important to verify whether we can obtain similar results with datasets in languages other than English. Since the incorporation and degree of pragmatic meaning in language use vary across languages (Baumgarten, 2022), assessing the consistency of our method's effectiveness across diverse languages would be valuable.

Finally, exploring the approach of applying discussions from specific domains as prompt templates, as we have done in other domains and tasks, would contribute to optimizing model performance strategies and exploring the nature of the tasks themselves in those use cases. Even beyond other domains, we believe it would be an interesting direction to investigate how our prompt template using pragmatic theory affects model performance in more general tasks such as dialogue and intention estimation.

Limitations

One limitation of our research is that we have not been able to verify whether the proposed method is effective when applied to more upstream tasks or broader language domains using language models. For example, if we could demonstrate the generalized effectiveness of the proposed method in the performance of other tasks where pragmatic abilities such as question answering, dialogue, and instruction following could be useful, the value of this research would be further enhanced.

Additionally, a limitation of this research is that we have not sufficiently verified “why” the proposed method was effective in the experiments. More advanced verification is needed to elucidate why this method improves performance and why it does not improve all phenomena. For example, we stated in Sec. 4.2.1 that the hypothesis is that the differences in score trends across theories might be due to the frequency of data included in the training data; however, more detailed analysis is required to make definitive statements about this.

Acknowledgments

A part of this work was supported by JST PRESTO grant number JPMJPR24TC. Also, our work was partly supported by the Toyota Foundation, grant number D24-ST-0023.

Ethics Statements

AI Assistants In this study, AI assistants, including ChatGPT, Copilot, and DeepL, were used in accordance with the ACL Policy on AI Writing Assistance. We primarily used them to assist with coding and writing, but all code and text outputs were manually reviewed. The authors take full responsibility for all of them.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, and 26 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *Preprint*, arXiv:2204.01691.

Nicole Baumgarten. 2022. *Contrastive Pragmatics*, pages 172–189. Routledge.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness : some universals in language usage*. Number 4 in Studies in interactional sociolinguistics. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Robyn Carston. 2002. *Thoughts and Utterances: The Pragmatics of Explicit Communication*. Blackwell Publishing Ltd.

Sammy Floyd. 2022. Pragmega materials.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Laurence R. Horn. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In *Meaning, form, and use in context: linguistic application*. Georgetown University Press.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

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

Zae Myung Kim, David E. Taylor, and Dongyeop Kang. 2023. "is the pope catholic?" applying chain-of-thought reasoning to understanding conversational implicatures. *Preprint*, arXiv:2305.13826.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

S. C. Levinson. 1983. *Pragmatics*. Cambridge University Press.

S. C. Levinson. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. The MIT Press.

Hengli Li, Song-Chun Zhu, and Zilong Zheng. 2024a. DiPlomat: a dialogue dataset for situated pragmatic reasoning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024b. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *Preprint*, arXiv:2502.12378.

OpenAI. 2025. GPT-4o mini: advancing cost-efficient intelligence. Accessed: 2025-01-26.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: fine-tuning strategy matters for implicature resolution by LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23. Curran Associates Inc.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Yue Shisen, Song Siyuan, Cheng Xinyuan, and Hu Hai. 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.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Med Inform*, 12:e55318.

Dan Sperbel and Deirdre Wilson. 1995. *Relevance: Communication and Cognition (2nd Edition)*. Blackwell.

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. Association for Computational Linguistics.

Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. DIRECT: Direct and indirect responses in conversational text corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989. Association for Computational Linguistics.

Shohei Tanaka, Konosuke Yamasaki, Akishige Yuguchi, Seiya Kawano, Satoshi Nakamura, and Koichiro Yoshino. 2024. Do as i demand, not as i say: A dataset for developing a reflective life-support robot. *IEEE Access*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Catherine J. Wearing. 2015. Relevance theory: pragmatics and cognition. *WIREs Cognitive Science*, 6(2):87–95.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

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.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

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

Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *Preprint*, arXiv:2407.04295.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature – a case study with a chinese sitcom. *Preprint*, arXiv:2404.19509.

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. Association for Computational Linguistics.

A The Gricean Maxim

Grice argued that to adhere to the Cooperative Principle, the following four subordinate maxims must be observed.

***The Gricean Maxims* (Grice, 1989)**

1. *Maxim of Quantity*: Provide an amount of information that is neither too little nor too much.
2. *Maxim of Quality*: Do not say what you believe to be false or for which you lack sufficient evidence.
3. *Maxim of Relation*: Be relevant in your utterance.
4. *Maxim of Manner*: Avoid ambiguity and be clear, concise, and orderly in your expression.

B Presumption of Relevance

***Presumption of Relevance* (Sperbel and Wilson, 1995)**

1. An ostensive stimulus is relevant enough to make it worth the addressee's effort to process it.
2. An ostensive stimulus is the most relevant one, given the communicator's abilities and priorities.

C Examples of PRAGMEGA dataset

Table 2 shows examples of problems and choices for each phenomenon in the PRAGMEGA dataset.

D Each Prompting method

Table 3 shows each prompt used as an in-context learning method in the experiments of this study.

E Hyperparameters

We set `temperature=0.8`, `max_new_tokens=1500`, `repetition_penalty=1.2`, `do_sample=True`.

F Computation Details

The parameter counts for each open model we used are as follows:

- Llama-3.1-8B-Instruct: 8B
- Qwen2.5-7B-instruct: 7B
- Qwen2.5-14B-instruct: 14B
- phi-4: 14B

We used NVIDIA RTX™ A6000, and our experiments took around 200 GPU hours.

G Exact Experimental Results

In our experiments, the exact accuracy achieved by each model is shown in 4.

Table 2: Examples of problems from each pragmatic phenomenon that PRAGMEGA includes. **Bold** indicates the correct answers. In the actual dataset, the order of choices is randomly shuffled. We adapt the table from examples shown in (Hu et al., 2023).

Phenomenon	Questions	Options
Deceits	Henry is sitting at his desk and watching TV, and reluctantly switches off the TV with the remote control and picks up a textbook. Shortly after, his mother comes in the room and asks, "What have you been doing up here?" Henry responds: "Reading." Why has Henry responded in such a way?	**1. He does not want to get into trouble for not studying.** 2. He has been reading for some time. 3. He does not want to offend his mom by not reading the books that she gave him. 4. He wants his mom to believe that he has been watching TV.
Metaphor	Andrew and Bob were discussing the investment company where Andrew works. Bob said: "The investors are squirrels collecting nuts." What does Bob mean?	**1. They buy stocks hoping for future profit.** 2. Squirrels were hired to work in the company. 3. The investors dress and eat well. 4. Bob is allergic to nuts. 5. The investors enjoy picking nuts as squirrels do.
Indirect Speech	Nate is about to leave the house. His wife points at a full bag of garbage and asks: "Are you going out?" What might she be trying to convey?	**1. She wants Nate to take the garbage out.** 2. She wants to know Nate's plans. 3. She wants Nate to bring his friends over. 4. She wants Nate to spend more time with the family.
Irony	It is a holiday. Stefan and Kim are sitting in the backseat of the car. They are fighting all the time. Their father says: "Oh, it is so pleasant here." What did the father want to convey?	**1. He does not want to listen to his kids' arguments.** 2. He enjoys listening to his kids fighting. 3. AC gives them some needed cool. 4. He remembers about his wife's birthday.
Maxims	Leslie and Jane are chatting at a coffee shop. Leslie asks, "Who was that man that I saw you with last night?" Jane responds, "The latte is unbelievable here." Why has Jane responded like this?	**1. She does not want to discuss the topic that Leslie has raised.** 2. She thinks that it is the best latte in the town. 3. The man who Leslie saw makes unbelievable lattes. 4. A coffee break is not a good time to discuss men.

Table 3: Prompts used in the in-context learning methods compared in this study.

Method	Prompt
Simple (Baseline-1)	Write ONLY the option number of your final answer and its contents in the format like: [Answer] 2) hogehoge is hogehoge. Any additional output beyond this penalized.
Chain-of-Thought (Baseline-2)	Firstly, think step-by-step and write down your process of thinking. After that, select your final answer. Your final answer should be in the format like: [Answer] 2) hogehoge is hogehoge.
Gricean Prompting (Proposed-1)	Let's think in line with the Gricean theory. In Grice's framework, hearers arrive at the implied meanings (or "implicatures") through an inferential process guided by the Cooperative Principle and its associated conversational maxims (Quantity, Quality, Relation, and Manner). Specifically: Cooperative Principle: The assumption that speakers and hearers are cooperating with one another to communicate effectively. Conversational Maxims: Quantity: Be as informative as required (but not overly so). Quality: Do not say what you believe to be false or lack evidence for. Relation (Relevance): Be relevant. Manner: Be clear, avoid ambiguity and obscurity. When a hearer detects a potential mismatch between what is said (literally) and one of the maxims, they hypothesize a conversational implicature—that the speaker must mean something more or different than the literal meaning. The hearer then uses context, background knowledge, and reasoning about the speaker's intent and adherence to the maxims to infer the intended meaning. Write down your thinking process in the line with Gricean theory and ultimately decide on the final answer. Your final answer should be in the format like: [Answer] 2) hogehoge is hogehoge.
Relevance Theory Prompting (Proposed-2)	Let's think in line with the Relevance theory. According to Relevance Theory, the interpretation of utterance implicatures proceeds through the following processes, where the balance between **cognitive effects** and **processing effort** plays a crucial role.** — 1. **Starting Point of Utterance Interpretation**: - The linguistic meaning (logical form) of an utterance is merely a "clue" to the interpretation intended by the speaker. - The listener must infer the speaker's intended meaning behind the utterance using this linguistic clue as a basis. ... 6. **Interaction Between Explicit Meaning and Implicature**: - Explicit meaning (the overt content of the utterance) and implicature (implied content) influence each other during processing. - This interaction forms the overall interpretation of the utterance. In Relevance Theory, **"optimal relevance"** is achieved when an utterance delivers **"cognitive effects worth the processing effort"** to the listener. Thus, the balance between cognitive effects and processing effort is consistently emphasized in utterance interpretation. By seeking interpretations that maximize effects with minimal effort, listeners achieve efficient understanding of utterances. Write down your thinking process in the line with Relevance theory and ultimately decide on the final answer. Your final answer should be in the format like: [Answer] 2) hogehoge is hogehoge.
Short Gricean Prompting (Short-1)	Let's think in line with the Gricean theory. Write down your thinking process in the line with Gricean theory and ultimately decide on the final answer. Your final answer should be in the format like: [Answer] 2) hogehoge is hogehoge.
Short Relevance Prompting (Short-2)	Let's think in line with the Relevance theory. Write down your thinking process in the line with Relevance theory and ultimately decide on the final answer. Your final answer should be in the format like: [Answer] 2) hogehoge is hogehoge.

Table 4: Experimental results. The highest Accuracy among the four in-context learning methods is indicated in **bold**.

Model	Baseline		Proposed		Short	
	simple	cot	grice	relevance	grice	relevance
gpt-4o	0.842	0.877	**0.940**	0.935	0.902	0.892
gpt-4o-mini	0.696	0.694	0.771	**0.779**	0.723	0.738
Llama-3.1-8B-Instruct	0.621	0.592	**0.656**	0.604	0.635	0.623
Qwen2.5-7B-Instruct	0.610	0.612	0.660	**0.683**	0.592	0.604
Qwen2.5-14B-Instruct	0.717	0.696	**0.796**	0.750	0.723	0.721
phi-4	0.781	0.775	**0.877**	0.867	0.796	0.792

H Correlation Analyses between Accuracy and Input/output Length

The results of the analysis on the relationship between Accuracy scores and the length of Input to the model and the length of Output from the model are shown in Table 4 and 5, respectively.

I Exact Results for Each Phenomena

The exact experimental results aggregated for each phenomenon in PRAGMEGA are shown in Table 5, 6, 7, 8, 9, and 10.

J Examples of Each Error Pattern

Table 11 shows specific examples of each error pattern by GPT-4o presented in Sec. 4.2.3.

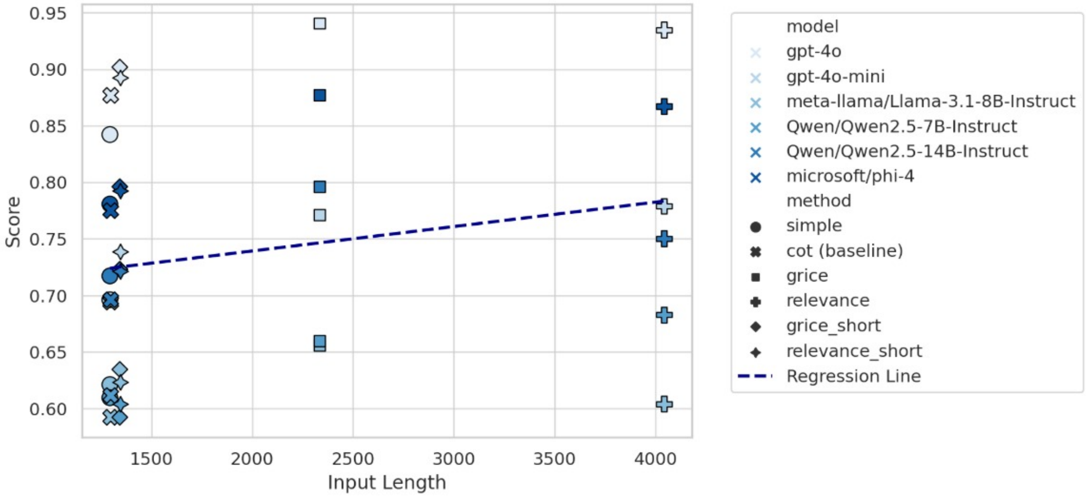

Figure 4: Correlation analysis between input length and accuracy.

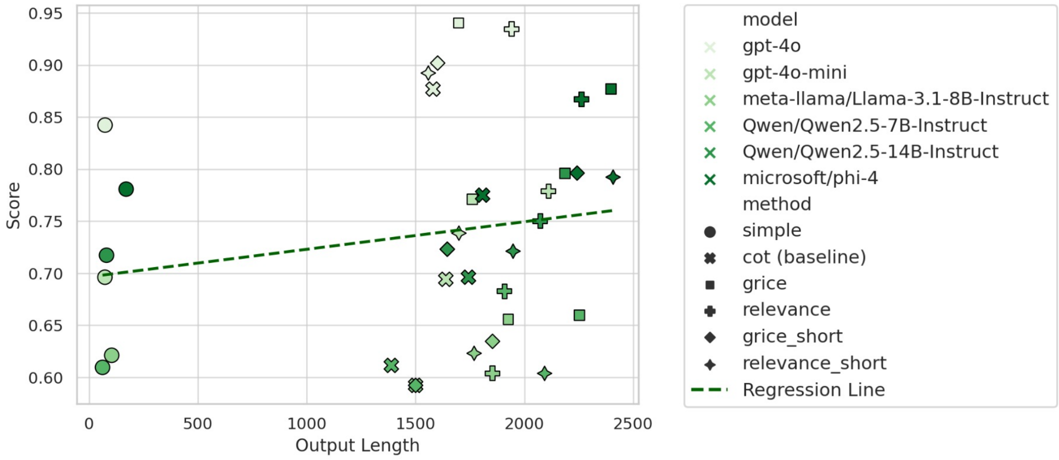

Figure 5: Correlation analysis between output length and accuracy.

Table 5: gpt-4o

		Deceits	**Metaphor**	**Indirect Speech**	**Irony**	**Maxims**
Baseline	`simple`	0.97	0.88	0.93	0.69	0.76
	`cot`	0.93	0.90	0.90	0.86	0.78
Proposed	`grice`	0.97	0.96	0.95	0.92	0.89
	`relevance`	0.94	0.98	0.97	0.92	0.85
Short	`grice_short`	0.89	0.93	0.95	0.90	0.83
	`relevance_short`	0.91	0.90	0.97	0.87	0.81

Table 6: gpt-4o-mini

		Deceits	**Metaphor**	**Indirect Speech**	**Irony**	**Maxims**
Baseline	simple	0.78	0.81	0.80	0.55	0.56
	cot	0.74	0.79	0.74	0.66	0.53
Proposed	grice	0.76	0.83	0.85	0.73	0.68
	relevance	0.82	0.84	0.90	0.71	0.63
Short	grice_short	0.69	0.83	0.83	0.64	0.64
	relevance_short	0.79	0.82	0.87	0.70	0.50

Table 7: Llama-3.1-8B-Instruct

		Deceits	**Metaphor**	**Indirect Speech**	**Irony**	**Maxims**
Baseline	simple	0.77	0.59	0.78	0.47	0.52
	cot	0.70	0.62	0.74	0.51	0.40
Proposed	grice	0.66	0.62	0.76	0.70	0.51
	relevance	0.67	0.58	0.77	0.60	0.37
Short	grice_short	0.68	0.57	0.76	0.65	0.49
	relevance_short	0.71	0.60	0.77	0.58	0.45

Table 8: Qwen2.5-7B-Instruct

		Deceits	**Metaphor**	**Indirect Speech**	**Irony**	**Maxims**
Baseline	simple	0.82	0.55	0.78	0.40	0.52
	cot	0.78	0.66	0.80	0.44	0.48
Proposed	grice	0.72	0.64	0.90	0.55	0.49
	relevance	0.77	0.65	0.86	0.51	0.44
Short	grice_short	0.65	0.58	0.90	0.44	0.51
	relevance_short	0.69	0.52	0.88	0.43	0.48

Table 9: Qwen2.5-14B-Instruct

		Deceits	**Metaphor**	**Indirect Speech**	**Irony**	**Maxims**
Baseline	simple	0.90	0.69	0.76	0.64	0.61
	cot	0.85	0.70	0.78	0.56	0.53
Proposed	grice	0.94	0.78	0.90	0.72	0.63
	relevance	0.92	0.70	0.83	0.68	0.62
Short	grice_short	0.84	0.72	0.87	0.63	0.56
	relevance_short	0.81	0.73	0.88	0.57	0.58

Table 10: phi-4

		Deceits	**Metaphor**	**Indirect Speech**	**Irony**	**Maxims**
Baseline	`simple`	0.88	0.79	0.84	0.74	0.65
	`cot`	0.87	0.85	0.86	0.67	0.64
Proposed	`grice`	0.93	0.86	0.94	0.90	0.73
	`relevance`	0.90	0.87	0.90	0.84	0.67
Short	`grice_short`	0.82	0.82	0.89	0.78	0.66
	`relevance_short`	0.81	0.84	0.92	0.77	0.61

Table 11: Actual question examples for some error patterns. **Bold** indicates the correct options. Due to space constraints, examples other than ① and ③ are provided in Appendix §J.

Pattern	**Questions**	**Options**
①	Samantha is talking with her dad about her fiance. Samantha notes: “John is an innocent person.” Her dad replies: “Undoubtedly, as innocent as a saint.” Why has Samantha’s dad responded like this?	1. Samantha’s dad is impressed with John’s innocence. **2. Samantha’s dad thinks that Samantha has an incorrect view of her fiance.** 3. Samantha’s dad thinks that Samantha’s fiance is a saint. 4. Samantha’s dad thinks that John is too religious.
②	Lenny comes to the kitchen and asks his wife, Marcie: “What will we have for breakfast?” Marcie responds: “A hard-boiled egg cooked in hot water and toast that is toasted evenly on both sides.” Why has Marcie responded in such a way?	1. Marcie is really good at cooking eggs and making toast. 2. Marcie thinks that breakfast is the main meal of the day. 3. Marcie wants Lenny to know how his breakfast was made. **4. Marcie thinks that her husband’s expectations about breakfast are too high.**
③	John is a teacher at an elementary school. When talking with the principal about a new student, who did poorly on her entrance examination, John said, “This one is really sharp.” What did John want to convey?	1. The entrance exam is unfair. 2. The pencils need to be sharpened. 3. The student is smart. **4. The student is not very clever.**
④	One day Jane comes home and is delighted to find her partner Anthony straightening up her apartment. Jane notices that Anthony threw out lots of things which were creating clutter, including an old photo that she had always kept on the coffee table. Anthony is worried that something is troubling Jane and asks if anything is wrong. Jane answers, “Everything is fine, dear. You did a great job of cleaning the apartment.” Why has Jane responded like this?	1. She is happy that Anthony has cleaned the apartment and does not care about the picture that got thrown away. 2. She wants to show that she is angry that Anthony has cleaned the apartment. **3. She wants to show that she appreciates that Anthony has cleaned the apartment.** 4. She shows him how angry she is with him for throwing out things without her consent.
⑤	Cindy wanted to paint a picture. She got her paints, paper and brushes ready. She has a meeting to go to in 10 minutes. Her dad said to her, “I am not sure that now is the best time for painting.” What might he be trying to convey?	**1. He does not want Cindy to start painting.** 2. He wants Cindy to create a sculpture. 3. He wants Cindy to paint a picture for the meeting. 4. He has some doubts whether Cindy should be painting.