

MultiPragEval: Multilingual Pragmatic Evaluation of Large Language Models

Dojun Park^{1*} Jiwoo Lee^{1*} Hyeyun Jeong¹ Seohyun Park¹
Youngeun Koo² Soonha Hwang³ Seonwoo Park¹ Sungeun Lee¹

¹Seoul National University, ²Sungkyunkwan University, ³Yonsei University

{dojun.parkk, seohyun.parkk88}@gmail.com

{lee9055, tosirihy, su3503, cristlo5}@snu.ac.kr

sarah8835@skku.edu, soonha.hwang@yonsei.ac.kr

Abstract

As the capabilities of Large Language Models (LLMs) expand, it becomes increasingly important to evaluate them beyond basic knowledge assessment, focusing on higher-level language understanding. This study introduces MultiPragEval, the first multilingual pragmatic evaluation of LLMs, designed for English, German, Korean, and Chinese. Comprising 1200 question units categorized according to Grice’s Cooperative Principle and its four conversational maxims, MultiPragEval enables an in-depth assessment of LLMs’ contextual awareness and their ability to infer implied meanings. Our findings demonstrate that Claude3-Opus significantly outperforms other models in all tested languages, establishing a state-of-the-art in the field. Among open-source models, Solar-10.7B and Qwen1.5-14B emerge as strong competitors. By analyzing pragmatic inference, we provide valuable insights into the capabilities essential for advanced language comprehension in AI systems. The test suite is publicly available on our GitHub repository at <https://github.com/DojunPark/MultiPragEval>.

1 Introduction

Understanding a language involves not only the ability to process explicit information but also an awareness of the context that influences the meaning of each utterance (Sperber and Wilson, 1986). In human communication, context acts as a critical element as it provides a foundation upon which dialogue participants can understand and interact with each other more efficiently. With a shared context, communication becomes more facilitated, allowing subtle nuances to be successfully conveyed, which is essential for engaging in meaningful conversations (Krauss and Fussell, 1996).

With recent advancements in generative AI, current LLMs have demonstrated capabilities that ex-

*These authors contributed equally to this work.

Aspect	Details
Utterance	"There's the door."
Literal Meaning	A door is located over there.
Contextual Implication	Context: An interviewer says it to the interviewee after finishing an interview. Implied Meaning: The interview has concluded and the interviewee is free to leave the room.

Table 1: Literal and contextual implications of the utterance "There's the door" in an interview scenario.

tend far beyond traditional natural language processing (NLP) tasks (Brown et al., 2020; Achiam et al., 2023). These models are increasingly becoming integral to our daily lives as AI assistants, closely engaging with human users in diverse conversational setups that demand a rapid understanding of the users’ needs and intentions, far surpassing mere literal interpretation of text (Roller et al., 2021). Given the growing importance of LLMs, accurately evaluating their ability to comprehend context-dependent meanings and demonstrate human-like language comprehension has become crucial (McCoy et al., 2019; Xu et al., 2020).

Pragmatics is a branch of linguistics that studies how language is used to achieve specific goals, where the interpretation of utterances depends not only on their literal meaning but also, crucially, on the surrounding context (Grice, 1975). Consider the example in Table 1, which demonstrates both the literal and implied meanings of the utterance, "There's the door." Literally, this phrase simply indicates the presence of a door in the specified direction. However, from a pragmatic standpoint, it conveys an additional implied meaning in the context of its usage by an interviewer to an interviewee after an interview has concluded. In this

scenario, the speaker is subtly suggesting that the interviewee is free to leave the room. This example underscores the critical role that context plays in shaping the interpretation of human language.

Despite the clear need for studies analyzing the pragmatic competence of current LLMs, there is not only a lack of systematic evaluation across various models (Chang et al., 2024) but also a strong bias towards English (Guo et al., 2023; Bommasani et al., 2023), leaving the pragmatic abilities of LLMs in other languages largely unexplored and difficult to compare. Such oversight demonstrates a significant gap in current evaluation practices, particularly given the multilingual nature of today’s state-of-the-art LLMs (Kwon et al., 2023).

To address these challenges, our study introduces **MultiPragEval**, the first multilingual test suite designed for the pragmatic evaluation of LLMs in English, German, Korean, and Chinese. Our suite comprises 300 question units per language, totaling 1200 units. These questions are divided into five categories based on Grice’s Cooperative Principles and the corresponding four conversational maxims: quantity, quality, relation, manner, and an additional category dedicated to assessing mere literal meaning understanding, independent of context.

Our main contributions are as follows:

- **Development of MultiPragEval:** We introduce MultiPragEval, a comprehensive test suite specifically designed to evaluate the pragmatic abilities of LLMs across English, German, Korean, and Chinese.
- **Systematic Evaluation of LLMs:** We conduct a thorough evaluation of 15 state-of-the-art LLMs, including both proprietary and open-source models, assessing their contextual awareness and pragmatic understanding capabilities.
- **In-depth Performance Analysis:** We offer a detailed analysis of LLM performance, systematically categorized according to Grice’s Cooperative Principle and its maxims, highlighting critical patterns and implications for further enhancements in LLM capabilities.

2 Related Work

Current Practices in LLM Evaluation. Benchmarks serve as critical tools for standardized evaluation in the field of LLM studies, enabling fair

and systematic comparisons across models trained with diverse architectures and strategies (Guo et al., 2023). These benchmarks span a wide range of domains, from general reasoning (Zellers et al., 2019) to specialized fields such as mathematics (Cobbe et al., 2021), coding (Chen et al., 2021), and biomedical sciences (Jin et al., 2019). While comprehensive, they primarily focus on assessing knowledge and logical reasoning, emphasizing explicit semantic meanings over the contextual and implied meanings that can vary in different scenarios (Sileo et al., 2022).

Leaderboards further enhance the field of LLM evaluation by providing a transparent platform where the performance of various models can directly compete with each other. The Open LLM Leaderboard (Beeching et al., 2023), featuring a range of rigorous benchmarks, establishes a venue for open-source models to showcase their capabilities, thereby fostering engagement in LLM development among both individual developers and tech companies. Meanwhile, Chatbot Arena (Chiang et al., 2024) is gaining recognition as a crowd-sourced evaluation platform. It leverages real-time feedback from users who vote on outputs from two randomly selected models. Models are then ranked on the leaderboard based on their Elo rating (Elo and Sloan, 1978), thus filling the gaps left by automatic benchmarks.

Recently, efforts have been made to create benchmarks specifically targeted at measuring the capabilities of LLMs in languages such as Chinese (Li et al., 2023) and Korean (Son et al., 2024). This development contributes to advancing a more inclusive multilingual evaluation landscape.

Pragmatic Evaluation of LLMs. As LLMs continue to evolve, it has become crucial to evaluate how effectively they consider context, which crucially shapes meanings beyond their literal interpretations. Bojic et al. (2023) examined multiple LLMs under the framework of Grice’s Cooperative Principle and its conversational maxims to assess their capabilities in understanding implicature. The results demonstrated that GPT-4 (Achiam et al., 2023) outperformed other models, including human performance. However, the human participants were not native English speakers but educated individuals from Serbia, which potentially limits the impact of the findings.

di San Pietro et al. (2023) conducted a comparable study focusing on GPT-3.5, leveraging the

Language	Context	Utterance	MCQ
English	While visiting Charlie’s house, Emily saw a large pile of oranges in the kitchen and asked why there were so many. Charlie responded:	"My uncle lives in Florida."	Choose the most appropriate meaning of the above utterance from the following options. (A) Charlie’s uncle sent the oranges. (B) Charlie’s uncle resides in Florida. (C) People in Florida do not like oranges. (D) Charlie’s uncle lives in a rural house. (E) None of the above.
German	Anna, die Felix besuchte, sah, dass es bei Felix viel Wein gab, und als sie fragte, warum es so viel Wein gab, wie er zu so viel Wein komme, sagte Felix:	"Mein Onkel betreibt ein Weingut in Freiburg."	Wählen Sie die passendste Bedeutung der obigen Äußerung aus den folgenden Aussagen aus. (A) Felix hat den Wein von seinem Onkel. (B) Der Onkel von Felix lebt in Freiburg. (C) Freiburger lieben keinen Wein. (D) Der Onkel von Felix wohnt in einem Landhaus. (E) Keine der obigen Aussagen ist richtig.
Korean	철수 집에 놀러 간 영희는 주방에 많은 껌이 쌓여 있는 것을 보고 껌이 왜 이렇게 많은지 물었고 철수는 다음과 같이 말했다.	"우리 작은 아버지께서 제주도에 사셔."	다음 보기에서 위 발화가 갖는 가장 적절한 의미를 고르세요. (A) 작은 아버지께서 껌을 보내주었다. (B) 작은 아버지의 거주지는 제주도이다. (C) 제주도 사람들은 껌을 좋아하지 않는다. (D) 작은 아버지께서 전원 주택에 사신다. (E) 정답 없음.
Chinese	王芳去张伟家看到厨房里堆放着几大袋葡萄干，便问为什么有这么多，张伟回答说：	"我叔叔住在新疆。"	请在以下选项中选择最恰当地表达上述话语含义的选项。 (A) 叔叔给张伟邮了葡萄干。 (B) 张伟的叔叔住在新疆。 (C) 新疆人不喜欢葡萄干。 (D) 张伟的叔叔住在乡间别墅里。 (E) 没有正确答案。

Table 2: Multilingual test units from the test suite on the maxim of relation, comprising a context, an utterance, and a multiple-choice question (MCQ) to assess the understanding of implied meanings. Charlie’s response indirectly addresses Emily’s question, thereby violating the maxim of relation. Assuming adherence to the cooperative principle, the most appropriate interpretation is option (A), indicating that Charlie’s uncle sent the oranges.

APACS test set (Arcara and Bambini, 2016), which consists of various subtasks such as interviews, descriptions, and narratives. The tests were conducted in both English and Italian, with results reported for Italian due to no notable differences between the two. The findings indicate that GPT-3.5 comes close to human ability but reveals weaknesses in understanding physical metaphors and jokes.

Focusing on Korean, Park et al. (2024) employed 120 test questions aligned with the four Gricean maxims to further probe the capabilities of various LLMs. The findings demonstrate that GPT-4 excelled in both multiple-choice and open-ended question setups, with HyperCLOVA X (Yoo et al., 2024), a Korean-specific LLM, closely following. The study also explored in-context learning, demonstrating that the few-shot learning technique consistently leads to positive outcomes across all tested models.

Sravanthi et al. (2024) introduce a comprehensive pragmatic benchmark that evaluates LLMs across 14 distinct tasks, including implicature, presupposition and deictic detection. Comprising 28k

data points, this benchmark aims to provide a nuanced assessment of LLMs’ pragmatic abilities, marking a substantial contribution to the field. Yet, there remains a significant need to extend these evaluations to multiple languages to thoroughly assess the multilingual capabilities of LLMs.

3 Methodology

3.1 Theoretical Foundations of Pragmatics

To accurately assess the contextual awareness of LLMs, we primarily focus on implicature, based on Grice’s theory (Grice, 1975). Implicature refers to a specific way language is used, in which the literal meaning of an utterance differs from the intended meaning of the speaker, requiring the listener to infer the intended meaning from the surrounding context. This concept is critical for evaluating how well LLMs understand human language, particularly in their ability to capture nuanced meanings beyond the explicit words.

Grice introduced the Cooperative Principle that explains how speakers and listeners cooperate to achieve mutual understanding, and its four conver-

Maxim	Description	Specific Cases Covered
Quantity	Make your contribution as informative as is required.	Tautology, insufficient information, excessive information, and cases where the maxim is abided by.
Quality	Try to make your contribution one that is true.	Irony, hyperbole, and misinformation.
Relation	Ensure that all the information you provide is relevant to the current conversation.	Unrelated information and cases where the maxim is abided by.
Manner	Be perspicuous; Be brief and orderly, and avoid obscurity and ambiguity.	Ambiguity, vagueness, double negation, verbosity, improper order, complicated expressions, and cases where the maxim is abided by.

Table 3: Grice’s maxims and their principles with related linguistic phenomena

sational maxims, which suggest how an utterance should desirably be conducted. Detailed in Table 3, the maxim of quantity requires information to be as informative as necessary—neither more nor less. The maxim of quality emphasizes the importance of offering truthful contributions. The maxim of relation ensures all information is pertinent to the current conversation. The maxim of manner demands clarity and brevity, avoiding obscurity and ambiguity.

Considering the critical role of understanding implicated meanings in communication, this study investigates LLMs’ comprehension of conversational implicatures. Specifically, we evaluate LLMs’ capabilities in inferring implied meanings that arise from either abiding by or violating these maxims.

3.2 Development of the Test Suite

To develop our test suite, we followed a structured process divided into three key phases: describing the initial dataset, expanding its scope, and translating it into the target languages and verifying the translations. Table 2 showcases an example of a test unit focused on the maxim of relation from our complete test suite, presented in English, German, Korean, and Chinese.

Initial Dataset. The development of the Multi-PragEval test suite began with the foundational work by (Park et al., 2024), who crafted a set of 120 question units designed to assess LLMs in terms of four conversational maxims. Each maxim was represented by 30 units, which included a structured scenario setting the conversational context, an utterance by a participant, and a set of questions comprising both a multiple-choice question and an open-ended question. We adopted the context, utterance, and multiple-choice question components from this test set as our starting point.

Expansion. Next, we expanded the number of question units from 120 to 300 to encompass a

broader range of pragmatic contexts. Each conversational maxim, originally represented by 30 units, was doubled to 60 to deepen the evaluative scope, including more diverse linguistic phenomena as shown in Table 3. Additionally, we introduced a new category specifically designed to assess the understanding of literal meanings, which allows us to explore potential trade-offs between performances in understanding literal versus implied meanings. To further enhance the complexity of our test suite, we included units that do not have a correct answer by adding a ‘None of the above’ option to the multiple-choice setups.

Translation and Verification. In the subsequent phase, we translated the Korean test set into English, German, and Chinese using DeepL¹ for the initial conversion. Then, Korean-native linguistic experts with CEFR C1² level proficiency in the target languages refined the translations to ensure that these translations preserved the intended meanings and nuances. They also adapted cultural elements by substituting the names of characters and setting details to reflect the local context of each language. Finally, native speakers of each target language, who hold degrees in linguistics and related fields, conducted a thorough verification of the translations. This process confirmed that the quality and accuracy of the translations were on par with the original Korean versions.

3.3 Experimental Setup

Models. Our study includes 15 LLMs, categorized into two types: proprietary LLMs accessed via API, and open-source LLMs where we have direct access to the model weights. As detailed in Table 4, the proprietary models comprise two GPT models (Achiam et al., 2023) by OpenAI,

¹<https://www.deepl.com>

²<https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions>

Type	Model	Version
Proprietary	GPT-3.5	turbo-0125
	GPT-4	turbo-2024-04-09
	Claude3-Haiku	haiku-20240307
	Claude3-Sonnet	sonnet-20240229
	Claude3-Opus	opus-20240229
	Mistral-small	small-2402
	Mistral-medium	medium-2312
	Mistral-large	large-2402
Open-Src.	Llama-2-13B	chat-hf
	Llama-2-7B	chat-hf
	Llama-3-8B	Instruct
	Gemma-7B	1.1-7b-it
	Solar-10.7B	Instruct-v1.0
	Qwen-14B	1.5-14B-Chat
	Qwen-7B	1.5-7B-Chat

Table 4: Overview of proprietary and open-source LLMs evaluated in the study

along with three different sizes of both Claude3 (Anthropic, 2024) by Anthropic and Mistral by Mistral AI³. We exclude Gemini by Google from our analysis due to its limited accessibility via API.

Additionally, we evaluate publicly available open-source models, each with approximately 10 billion parameters. These models were selected based on two criteria: their architecture (Transformer decoder-based models) and their performance on publicly accessible benchmarks. The selected models include three Llama models (Touvron et al., 2023) by Meta, Gemma (Team et al., 2024) by Google, Solar (Kim et al., 2023) by Korean company Upstage, and two Qwen models (Bai et al., 2023) by Chinese firm Alibaba, with consideration also given to the diversity of languages represented in our study.

LLM Response Generation. To generate answers from each LLM, we set the temperature hyperparameter at 0.5 across models to balance coherence and creativity in their responses. For inference on the open-source LLMs, we utilized a single H100-80GB unit. Each model was queried three times to account for the inherent randomness in responses. We then computed the average score for each model across these trials to ensure a robust assessment of performance for each LLM iteration. Scores were calculated based on the ratio of cor-

rect answers to the total number of test units across all three trials. The actual prompt for the experiment and inter-rater agreement across three trials are detailed in the Appendix B.

4 Result

4.1 Analysis of LLM Performance

Overall Performance. Table 5 presents the results from the evaluation of the selected LLMs on the MultiPragEval test suite. It demonstrates that Claude3-Opus significantly outperforms all other models across four languages, with GPT-4 trailing by approximately 6-10 points. This performance gap underscores Claude3-Opus’s exceptional ability to capture the subtle nuances of language that are highly context-dependent. These findings highlight its position as the most proficient among the current state-of-the-art LLMs across English, German, Korean, and Chinese.

Mistral-Large and Claude3-Sonnet are closely matched for the next tier of performance; Mistral-Large outperforms Claude3-Sonnet in German, Korean, and Chinese. However, Claude3-Sonnet achieves a higher score in English, registering 66.39 compared to Mistral-Large’s 61.39. Interestingly, while Mistral-Large generally shows improved scores across languages compared to Mistral-Medium, it scores lower in English, dropping to 61.39 from the medium-sized model’s 66.25.

Solar-10.7B demonstrates stable performance, consistently outperforming GPT-3.5 across all four languages. It is the only open-source model that surpasses GPT-3.5 in both English and German. In English, it closely follows Mistral-Large with a score of 59.31 and is just behind Claude3-Sonnet in German, with a score of 55.69.

Qwen-14B also stands out among other open-source LLMs, outperforming its counterparts with scores of 50.00 in Chinese and 49.72 in Korean. In contrast, both Llama2-13B and Llama2-7B demonstrate a strong bias towards literal interpretations yielding poor scores, while Llama3-8B shows enhanced performance compared to its earlier versions. Notably, Llama2-13B achieves a significant leap in Korean, scoring 47.50 compared to Llama2-7B’s 3.06, while exhibiting a more gradual increase in other languages.

Performance Gap Across Languages. We observed that the models generally achieve higher per-

³<https://mistral.ai/>

	English					German	Korean	Chinese
	Quan.	Qual.	Rel.	Man.	Avg.	Avg.	Avg.	Avg.
GPT-4	65.00	83.89	82.22	70.00	75.28	72.50	81.25	68.75
GPT-3.5	51.11	66.67	52.78	42.89	53.61	52.92	38.89	43.61
Claude3-Opus	81.11	88.89	88.89	81.11	85.00	82.78	87.08	76.67
Claude3-Sonnet	62.22	81.67	67.22	54.44	<u>66.39</u>	<u>60.14</u>	<u>63.33</u>	<u>48.61</u>
Claude3-Haiku	56.67	67.78	58.89	43.33	56.67	45.14	38.47	40.83
Mistral-Large	61.11	71.11	61.11	52.22	<u>61.39</u>	<u>63.75</u>	<u>65.56</u>	<u>54.72</u>
Mistral-Medium	61.11	69.44	72.22	62.22	<u>66.25</u>	53.61	52.92	38.89
Mistral-Small	57.22	57.78	54.44	35.00	51.11	51.11	40.42	33.61
Llama3-8B	54.44	68.89	44.44	<u>45.56</u>	53.33	40.00	32.50	46.81
Llama2-13B	26.67	32.22	16.67	32.22	26.94	16.39	47.50	<u>8.75</u>
Llama2-7B	31.11	26.67	11.11	<u>18.33</u>	21.81	<u>4.44</u>	<u>3.06</u>	<u>4.17</u>
Gemma-7B	37.78	36.67	35.00	30.56	35.00	27.22	20.83	25.28
Solar-10.7B	58.33	65.56	62.22	51.11	59.31	55.69	49.03	46.39
Qwen-14B	52.22	61.67	56.11	43.33	53.33	43.06	49.72	50.00
Qwen-7B	53.89	62.22	47.22	37.78	50.28	39.44	35.14	41.11

Table 5: Performance of LLMs on the MultiPragEval test suite: scores across four languages and by maxims with overall averages; Leading scores among proprietary and open-source models are highlighted in bold. The scores for each maxim are color-coded in shades of blue to represent the relative ranking within each model.

formance scores in English than in other languages, likely due to larger English training datasets enhancing reasoning capabilities. Interestingly, flagship proprietary models like GPT-4, Claude-Opus, and Mistral-large show slightly better performance in Korean. We believe there could be two possible reasons for this performance gap. First, it is possible that the initial Korean dataset, from which we extended our test suite (Park et al., 2024), was used in model training, allowing the models to better understand newly created Korean questions that follow the same template. Secondly, the gap could stem from the test suite being initially developed in Korean and then translated into other languages. Cultural nuances and conventions embedded in each language may lead to subtle differences in how the same expressions are interpreted, with the implications being understood differently depending on the language region.

Significant performance discrepancies were also observed across models. Claude-Haiku scored 56.7 in English but only 38.4 in Korean, while Mistral-small dropped from 51.1 in English to 33.6 in Chinese. Llama2-13B showed the largest gap, with scores of 47.5 in Korean versus 8.7 in Chinese. These differences highlight language-specific biases in the models, indicating a need for improvements to boost multilingual capabilities.

Closer Look at Individual Maxims. Table 5 also shows the performance scores of LLMs on individual maxims in the English test suite. We observe a consistent pattern across LLMs where scores for the maxim of quality generally rank highest, while scores for the maxim of manner rank lowest. This pattern is not unique to English but is also observable in other languages, suggesting a universal trend (see Appendix A). This outcome is expected because expressions governed by the maxim of quality, which become untrue statements when interpreted literally, make it easier for LLMs to infer the appropriate implied meanings. Conversely, the maxim of manner, involving verbose or ambiguous expressions, poses more subtle challenges that likewise pose difficulties for humans (Hoffmann, 2010).

Another noteworthy observation is that as the overall performance increases, the scores for the maxim of relation also generally improve. This pattern is more evident among proprietary models, where the maxim of relation mostly ranks second. Similarly, Solar-10.7B and Qwen-14B, which perform comparably to GPT-3.5, achieve higher scores in the maxim of relation compared to those of quantity and manner. Conversely, other open-source models with lower average scores tend to have lower rankings in the maxim of relation, falling

	English			German			Korean			Chinese		
	Avg.	Opt. None	Literal	Avg.	Opt. None	Literal	Avg.	Opt. None	Literal	Avg.	Opt. None	Literal
GPT-4	75.28	90.00	100.00	72.50	90.56	97.22	81.25	75.00	96.67	68.75	79.44	98.33
GPT-3.5	53.61	55.00	85.56	52.92	69.44	85.56	38.89	31.11	83.33	43.61	62.78	88.33
Claude3-Opus	85.00	92.78	98.89	82.78	85.00	93.33	87.08	70.56	99.44	76.67	83.33	95.56
Claude3-Sonnet	66.39	81.11	91.67	60.14	67.22	91.67	63.33	28.33	84.44	48.61	34.44	87.78
Claude3-Haiku	56.67	63.89	91.11	45.14	37.22	90.00	38.47	9.44	80.00	40.83	8.33	80.56
Mistral-Large	61.39	66.11	95.56	63.75	77.22	87.78	65.56	58.33	91.11	54.72	54.44	88.33
Mistral-Medium	66.25	80.56	98.33	53.61	61.11	91.11	52.92	45.00	86.11	38.89	16.11	81.11
Mistral-Small	51.11	47.78	92.22	51.11	43.33	87.22	40.42	31.11	85.00	33.61	18.33	82.78
Llama3-8B	53.33	43.89	85.00	40.00	56.11	87.22	32.50	21.67	80.00	46.81	28.33	89.44
Llama2-13B	26.94	65.00	70.00	16.39	9.44	69.44	47.50	2.22	67.78	8.75	7.78	64.44
Llama2-7B	21.81	13.33	70.56	4.44	1.11	<u>45.56</u>	3.06	0.00	<u>42.22</u>	4.17	0.00	<u>49.44</u>
Gemma-7B	35.00	23.33	77.22	27.22	7.28	80.00	20.83	0.56	79.44	25.28	0.00	80.00
Solar-10.7B	59.31	81.11	97.78	55.69	38.33	86.11	49.03	22.22	78.89	46.39	26.67	88.89
Qwen-14B	53.33	78.33	93.33	43.06	52.78	85.00	49.72	41.67	87.78	50.00	79.44	94.44
Qwen-7B	50.28	31.67	80.00	39.44	10.00	76.67	35.14	0.00	73.33	41.11	43.33	86.67

Figure 1: Breakdown of LLM scores for ‘No Correct Answers’ and literal meaning tests across four languages; the heatmap uses two colors—blue indicating higher scores and yellow indicating lower scores.

below the maxim of quantity. This suggests that capturing relevancy within the given context plays a significant role in a more precise interpretation of implied information, contributing to better overall performance.

4.2 Assessing the Stability of Pragmatic Inference

We further explore the stability of LLMs in pragmatic inference under two specific setups. First, we evaluate the models on a subset of each category of maxims, specifically designed where the test questions lack an appropriate answer. This subset is intended to be more challenging as it requires the models to identify incorrect interpretations and select the option ‘(E) None of the above’ without reference to a correct meaning. Secondly, we test the models on additional test units consisting of context, utterance, and question, structured similarly, but where the context is irrelevant to the utterance. This setup is designed to assess whether LLMs can accurately distinguish purely literal meanings from inappropriate interpretations.

Subset of No Correct Answer. Figure 1 illustrates that the scores on the subset without correct answers (Opt. None) generally align with the

overall scores, yet they reveal subtle differences in performance details. While Claude3-Opus consistently outperforms GPT-4 by a certain margin in overall scores across all languages, GPT-4 surpasses Claude3-Opus by approximately 5 points in both German and Korean. This result indicates that both models are comparably robust in the challenging setup of pragmatic consideration.

It is evident that models with lower overall scores exhibit significant declines when tested in the setup without a correct answer. Among proprietary LLMs, Claude3-Haiku, along with medium and small-sized models by Mistral, notably drop in scores, indicating their struggles with the task. Similarly, 7-billion parameter models such as Llama2, Gemma, and Qwen also show poor performance, underscoring the complexity of the task for models of this size.

Additional Set of Literal Meaning. The scores on the set asking literal meanings also demonstrate a general increase along with the overall scores. While the flagship models of GPT and Claude show performance close to perfect, GPT-4 demonstrates a slight edge over Claude-3-Opus for English, German, and Chinese. This may suggest a trade-off between pragmatic and literal focus in their infer-

Model	MultiPragEval (Eng.)	MMLU 5-shot	MATH 4-shot	Arena Elo*	ARC 25-shot	HumanEval 0-shot	GSM-8K 8-shot
GPT-4	75.28	86.4	52.9	1252	96.3	67.0	92.0
GPT-3.5	53.6	70.0	34.1	1110	85.2	48.1	57.1
Claude3-Opus	85.0	86.8	61.0	1246	96.4	84.9	95.0
Claude3-Sonnet	66.4	79.0	40.5	1199	93.2	73.0	92.3
Claude3-Haiku	56.7	75.2	40.9	1181	89.2	75.9	88.9
Llama3-8B	53.3	68.4	30.0	1154	60.7	62.2	79.6
Llama2-13B	26.9	47.8	6.7	1065	59.4	14.0	77.4
Llama2-7B	21.8	34.1	3.8	1042	53.1	7.9	25.7
Gemma-7B	35.0	66.0	24.3	1091	61.1	32.3	46.4
Qwen-14B	53.3	69.4	24.8	1119	56.6	32.3	61.3
Qwen-7B	50.3	61.7	11.6	1079	54.2	29.9	51.7
Kendall τ	1.00	0.95	0.92	0.84	0.81	0.80	0.73

Table 6: Performance scores of LLMs across multiple benchmarks and Kendall’s Tau correlation Coefficients Relative to MultiPragEval.

* The Arena Elo scores are as of May 17, 2024.

ences.

The Llama2 models, particularly Llama2-7B, show the lowest scores among the others, with 42.22, 45.56, and 49.44 for Korean, German, and English, respectively. These results generally correlate with lower overall scores in both the pragmatic and no-correct-answer subset questions. We interpret this to mean that these tasks are not independent of each other, but instead mutually influence one another, highlighting the importance of maintaining a good balance between the sub-tasks.

4.3 Comparison with Existing Benchmarks

To further delve into the implications of our findings, we compare the results from our English test suite with existing English-based benchmarks. This analysis encompasses scores from 11 models, for which other benchmark scores were publicly available. We consider seven popular benchmarks: MMLU (Hendrycks et al., 2020) and ARC (Clark et al., 2018) for general reasoning, HumanEval (Chen et al., 2021) for coding, GSM-8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) for mathematics, and Chatbot Arena (Chiang et al., 2024), a crowd-sourced evaluation. We opted to calculate the correlation coefficients using Kendall’s Tau (Kendall, 1938) due to its better handling of varying ranges and subtle differences between benchmarks.

The correlations of MultiPragEval with other benchmarks consistently show high values, indicating a general trend toward ‘good’ performance across different benchmarks. This suggests that improvements in a model’s performance on one task generally enhance its performance on other tasks

(Raffel et al., 2020).

MMLU and MATH exhibit the highest correlations among other benchmarks, suggesting that the abilities assessed by these benchmarks align closely with those required for pragmatic inference. It is anticipated that MMLU, which evaluates the general language understanding capabilities of LLMs across a broad spectrum of disciplines, reflects the ability to consider contextual information in language, which is a key requirement of MultiPragEval.

However, the high correlation observed with the MATH benchmark is surprising, given its primary focus on mathematical reasoning. Notably, the score gap between Claude3-Opus and GPT-4, which is around 10 points on MultiPragEval, is similarly reflected on MATH but not distinctively on MMLU. This pattern suggests that the sophisticated mathematical problem-solving required by MATH—which demands a higher level of logical reasoning compared to the basic mathematical problems in GSM-8K—may also tap into core capabilities essential for pragmatic inference. This connection between mathematical reasoning and high-level linguistic comprehension indicates an intricate relationship that requires deeper investigation.

5 Conclusion

In this work, we present the first multilingual study of LLMs’ capabilities of their pragmatic inference, particularly in the context of Grice’s theory of conversational implicature. Our findings demonstrate the usefulness of MultiPragEval test suite in distinguishing the levels of comprehension among various proprietary and open-source models.

The results reveal that among the models evaluated, Claude3-Opus and GPT-4 particularly stand out, with Claude3-Opus consistently outperforming GPT-4 by 6 to 10 points across all languages, affirming its state-of-the-art capability in pragmatic understanding. Top-performing open-source models like Solar-10.7B and Qwen-14B demonstrate superior or comparable performance to lite-size proprietary models such as GPT-3.5, Claude3-Haiku, and Mistral-Small. The performance gaps across languages within models and individual Grice’s maxims further highlight language biases and areas for improvement.

Our findings, with the highest correlations with MMLU and MATH, suggest that general language understanding and complex logical reasoning are intricately linked to pragmatic inference abilities. This insight guides us towards further research to empirically demonstrate how these abilities relate to pragmatic reasoning.

Limitations

While our study provides a comprehensive comparison of 15 proprietary and open-source models, it does not include a comparison with human performance. Including human performance would offer deeper insights into how closely LLMs approximate human abilities. Moreover, human performance can vary across languages, which would enrich our understanding of the LLMs’ multilingual pragmatic abilities. Recognizing this gap, we aim to incorporate human performance comparisons in our future research.

Another limitation of our study is its exclusive focus on implicature, despite pragmatics encompassing a broader range of phenomena such as speech acts, presupposition, and politeness. This focus was chosen due to the increasing role of LLMs as AI assistants, which often need to interpret human expressions that are frequently conveyed implicitly. The ability of LLMs to capture these subtle nuances directly influences human judgments about the quality of these systems. Furthermore, contextual awareness is critical not only for linguists but also for NLP engineers who aim to provide reliable services to users. We believe that our specific focus on implicature provides valuable insights into how effectively current LLMs manage the complexities inherent in interpreting implied meanings, a crucial aspect of human communication.

Our study set the temperature value to 0.5 to

achieve a moderate balance between consistency and creativity in responses. However, it is important to note that the optimal temperature may vary for each LLM, and the effect of temperature settings on pragmatic inference remains unclear. Recognizing the potential influence of temperature on LLMs’ pragmatic abilities, we suggest that future studies investigate the relationship between temperature and pragmatic reasoning to gain deeper insights into how LLMs handle nuanced language tasks.

Ethics Statement

In this work, we introduce a test suite designed to evaluate the pragmatic abilities of LLMs. We have ensured that all data created for this study does not infringe on any existing intellectual property rights, while also ensuring it contains no personally identifiable information. Linguistic experts were involved in the creation and translation of the test suite; all contributors were fully informed about the research’s purpose and the methods employed. We commit to making the dataset publicly available to foster transparency and further research in the field.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00274280). Additionally, this work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. *Gpt-4 technical report*. *arXiv preprint arXiv:2303.08774*.
- AI Anthropic. 2024. *The claude 3 model family: Opus, sonnet, haiku*. *Claude-3 Model Card*.
- Giorgio Arcara and Valentina Bambini. 2016. *A test for the assessment of pragmatic abilities and cognitive substrates (apacs): Normative data and psychometric properties*. *Frontiers in psychology*, 7:172889.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei

- Huang, et al. 2023. [Qwen technical report](#). *arXiv preprint arXiv:2309.16609*.
- Edward Beeching, Clémentine Fourier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
- Ljubisa Bojic, Predrag Kovacevic, and Milan Cabarkapa. 2023. [Gpt-4 surpassing human performance in linguistic pragmatics](#). *arXiv preprint arXiv:2312.09545*.
- Rishi Bommasani, Percy Liang, and Tony Lee. 2023. [Holistic evaluation of language models](#). *Annals of the New York Academy of Sciences*, 1525(1):140–146.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. [A survey on evaluation of large language models](#). *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: An open platform for evaluating llms by human preference](#).
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. [The pragmatic profile of chatgpt: assessing the pragmatic skills of a conversational agent](#).
- Arpad E Elo and Sam Sloan. 1978. [The rating of chess-players: Past and present](#).
- Herbert P Grice. 1975. [Logic and conversation](#). In *Speech acts*, pages 41–58. Brill.
- Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. [Evaluating large language models: A comprehensive survey](#). *arXiv preprint arXiv:2310.19736*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. [Measuring massive multitask language understanding](#). *arXiv preprint arXiv:2009.03300*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the math dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Ludger Hoffmann. 2010. *Sprachwissenschaft: ein Reader*. de Gruyter.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Maurice G Kendall. 1938. [A new measure of rank correlation](#). *Biometrika*, 30(1/2):81–93.
- Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. [Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling](#). *arXiv preprint arXiv:2312.15166*.
- Robert M Krauss and Susan R Fussell. 1996. [Social psychological models of interpersonal communication](#). *Social psychology: Handbook of basic principles*, pages 655–701.
- Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. [Beyond English: Evaluating LLMs for Arabic grammatical error correction](#). In *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid). Association for Computational Linguistics.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. [Cmmlu: Measuring massive multitask language understanding in chinese](#). *arXiv preprint arXiv:2306.09212*.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Annual Meeting of the Association for Computational Linguistics*.

- Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024. [Pragmatic competence evaluation of large language models for korean](#). *arXiv preprint arXiv:2403.12675*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of machine learning research*, 21(140):1–67.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. [Recipes for building an open-domain chatbot](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- Damien Sileo, Philippe Muller, Tim Van de Cruys, and Camille Pradel. 2022. [A pragmatics-centered evaluation framework for natural language understanding](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2382–2394, Marseille, France. European Language Resources Association.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. [Kmmmlu: Measuring massive multitask language understanding in korean](#). *arXiv preprint arXiv:2402.11548*.
- Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.
- Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. [Pub: A pragmatics understanding benchmark for assessing llms’ pragmatics capabilities](#). *arXiv preprint arXiv:2401.07078*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. [Gemma: Open models based on gemini research and technology](#). *arXiv preprint arXiv:2403.08295*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruiti Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. [Recipes for safety in open-domain chatbots](#). *arXiv preprint arXiv:2010.07079*.
- Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. [Hyperclova x technical report](#). *arXiv preprint arXiv:2404.01954*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Demonstration of Test Unit Example

Language	Context	Utterance	MCQ
English	A student asks their professor if they can extend the due date of an assignment just a little longer. The professor replies:	"Rules are rules."	Choose the most appropriate meaning of the above utterance from the following options. (A) The deadline can't be extended because rules must be followed. (B) Rules are rules. (C) Breaking the rules isn't a big deal, so I'll give the student a chance. (D) The professor discovered a new theory after doing research. (E) None of the above.
German	Ein Student fragt seinen Professor, ob er den Abgabetermin für eine Aufgabe noch ein wenig hinauszögern kann.	"Regeln sind Regeln."	Wählen Sie die passendste Bedeutung der obigen Äußerung aus den folgenden Aussagen aus. (A) Die Frist kann nicht verlängert werden, weil die Regeln eingehalten werden müssen. (B) Regeln sind Regeln. (C) Ein Verstoß gegen die Regeln ist keine große Sache, also gebe ich dem Studenten eine Chance. (D) Der Professor hat durch Nachforschungen eine neue Theorie entdeckt. (E) Keine der obigen Aussagen ist richtig.
Korean	학생이 교수에게 과제의 마감 기한을 조금만 더 늘려 주실 수 없냐고 부탁하자 교수가 말한다.	"규칙은 규칙일세."	다음 보기에서 위 발화가 갖는 가장 적절한 의미를 고르세요. (A) 규칙은 지켜져야만 하므로 마감 기한을 늘릴 수 없다. (B) 규칙은 규칙이다. (C) 규칙을 깨는 것은 큰 문제가 되지 않으므로 학생에게 기회를 주겠다. (D) 교수는 연구 끝에 새로운 이론을 발견했다. (E) 정답 없음.
Chinese	一名学生问教授可不可以将作业的截止日期再延长一点, 教授说:	"规则就是规则。"	请在以下选项中选择最恰当地表达上述话语含义的选项。 (A) 规则必须遵守, 因此不能延长截止日期。 (B) 规矩就是规矩。 (C) 违反规则没什么大不了的, 所以教授会给学生一个机会。 (D) 教授经过研究发现了一个新理论。 (E) 没有正确答案。

Table 7: Multilingual test unit example on the maxim of quantity. The utterance "Rules are rules" is not sufficiently informative because it provides less information than necessary. This under-informativeness constitutes a violation of Grice's maxim of quantity, which demands that enough information be given to be fully informative. In this context, "the rules" implicitly refer to the adherence to established guidelines, such as the due date for assignments. Therefore, the most appropriate interpretation of the professor's statement is option (A) "The deadline can't be extended because rules must be followed," which accurately captures the implied meaning behind the response.

Language	Context	Utterance	MCQ
English	When Emily, a PhD student, spoke at length about the theory she had studied yesterday, Charlie said:	"You're the professor."	Choose the most appropriate meaning of the above utterance from the following options. (A) Emily was hired as a professor. (B) Emily knows a lot, but she talks too much. (C) Emily is not good at graduate studies. (D) Emily lives in a dormitory. (E) None of the above.
German	Als Anna, eine Doktorandin, ausführlich über die Theorie sprach, die sie gestern untersucht hatte, sagte Felix:	"Du bist ja Professorin."	Wählen Sie die passendste Bedeutung der obigen Äußerung aus den folgenden Aussagen aus. (A) Anna wurde zur Professorin ernannt. (B) Anna weiß eine Menge, aber sie redet zu viel. (C) Anna ist nicht gut im Studium. (D) Anna wohnt in einem Studentenwohnheim. (E) Keine der obigen Aussagen ist richtig.
Korean	박사생인 영희가 어제 공부한 이론에 대해 길게 이야기하자 철수가 다음과 같이 말했다.	"네가 교수다."	다음 보기에서 위 발화가 갖는 가장 적절한 의미를 고르세요. (A) 영희는 교수로 임용되었다. (B) 영희는 아는 것이 많지만 말이 너무 많다. (C) 영희는 대학원 공부에 소질이 없다. (D) 영희는 기숙사에 살고 있다. (E) 정답 없음.
Chinese	当博士生王芳详细讲述她昨天学习的理论时，张伟说：	"你是教授吗？"	请在以下选项中选择最恰当地表达上述话语含义的选项。 (A) 王芳被任命为教授。 (B) 王芳知道很多，但她说得太多了。 (C) 王芳不适合读研。 (D) 王芳住在宿舍里。 (E) 没有正确答案。

Table 8: Multilingual test unit example on the maxim of quality. This example illustrates a violation of Grice's maxim of quality, which requires contributions to be true. Although Charlie refers to Emily as "the professor," he does not literally mean that she holds this academic position, as she is a PhD student. Instead, this utterance uses irony to comment on Emily's detailed and extensive explanation, typical of a professor's depth of knowledge. Therefore, the utterance "You're the professor" acknowledges Emily's thorough knowledge while subtly critiquing her for possibly providing more information than necessary in casual conversation. Thus, option (B) "Emily knows a lot, but she talks too much." best captures the implied meaning of Charlie's statement.

Language	Context	Utterance	MCQ
English	When Charlie confessed to Emily that he wanted to go out with her, she replied:	"I really like you as a friend, too, but I don't think I'm in the right frame of mind to meet someone right now."	Choose the most appropriate meaning of the above utterance from the following options. (A) Charlie and Emily have a good personality match. (B) Emily wants to date Charlie's brother. (C) Emily doesn't want to go out with Charlie. (D) There are no friends between men and women. (E) None of the above.
German	Als Felix Anna gestand, dass er mit ihr ausgehen wollte, sagte sie ihm:	"Ich mag dich sehr als Freund, aber ich glaube nicht, dass ich im Moment in der richtigen Stimmung bin, um mit jemandem in einer Beziehung sein."	Wählen Sie die passendste Bedeutung der obigen Äußerung aus den folgenden Aussagen aus. (A) Felix und Anna passen charakterlich gut zusammen. (B) Anna will mit Felix' Bruder ausgehen. (C) Anna will nicht mit Felix ausgehen. (D) Es gibt keine echte Freundschaft zwischen Männern und Frauen. (E) Keine der obigen Aussagen ist richtig.
Korean	철수가 영희에게 사귀자고 고백하자 영희가 다음과 같이 말했다.	"나도 너를 친구로서 정말 좋아하지만 내가 지금 사람을 만날 만한 마음의 여유가 없는 것 같아."	다음 보기에서 위 발화가 갖는 가장 적절한 의미를 고르세요. (A) 철수와 영희는 성격이 잘 맞는다. (B) 영희는 철수의 친오빠와 사귀고 싶다. (C) 철수와 사귀고 싶지 않다. (D) 남자와 여자 사이에 친구란 없다. (E) 정답 없음.
Chinese	当张伟向王芳表白, 王芳说:	"作为朋友我真的很喜欢你, 但是我现在状态不适合和别人在一起。"	请在以下选项中选择最恰当地表达上述话语含义的选项。 (A) 张伟和王芳性格很合得来。 (B) 王芳想和张伟的哥哥约会。 (C) 王芳不想和张伟谈恋爱。 (D) 男女之间没有朋友。 (E) 没有正确答案。

Table 9: Multilingual test unit example on the maxim of manner. Emily's response to Charlie's confession is a classic example of violating Grice's maxim of manner, which advocates for clarity and brevity in communication. Instead of a direct answer, Emily's reply is ambiguously structured, suggesting a rejection without explicitly stating one. This ambiguity is strategic, preserving social harmony while conveying her feelings indirectly. Given the content and context of the conversation, options (A), (B), and (D) do not align with the information provided. Emily emphasizes her current emotional state and her appreciation of their friendship as reasons for not pursuing a romantic relationship, which implicitly suggests she does not wish to date Charlie. Thus, option (C) "Emily doesn't want to go out with Charlie" captures the underlying implication of her response most accurately.

Language	Context	Utterance	MCQ
English	Emily and Charlie are working on a writing assignment from class. Emily asks Charlie when the writing assignment is due, and Charlie replies:	"It's due next Thursday."	Choose the most appropriate meaning of the above utterance from the following options. (A) Charlie is asking Emily for help. (B) Charlie is not confident in English and wants to postpone the writing assignment. (C) Charlie wants to finish the writing assignment today. (D) The writing assignment is due next Thursday. (E) None of the above.
German	Anna und Felix arbeiteten an einer schriftlichen Aufgabe aus ihrem Unterricht. Anna fragte Felix, wann die Schreibaufgabe fällig sei, und Felix antwortete:	"Sie ist nächsten Donnerstag fällig."	Wählen Sie die passendste Bedeutung der obigen Äußerung aus den folgenden Aussagen aus. (A) Er bittet Anna um Hilfe. (B) Felix ist unsicher in Englisch und möchte die Schreibaufgabe verschieben. (C) Er möchte die schriftliche Aufgabe sofort fertigstellen. (D) Die Schreibaufgabe soll bis zum nächsten Donnerstag fertig sein. (E) Keine der obigen Aussagen ist richtig.
Korean	영희와 철수는 수업에서 나온 글쓰기 과제를 하고 있다. 영희가 철수에게 글쓰기 과제 마감일이 언제인지 묻자, 철수가 다음과 같이 대답했다.	"다음주 목요일까지 제출해야 해."	다음 보기에서 위 발화가 갖는 가장 적절한 의미를 고르세요. (A) 철수는 영희에게 도움을 요청하는 중이다. (B) 철수는 영어에 자신이 없어서 글쓰기 과제를 미루고 싶다. (C) 철수는 오늘 글쓰기 과제를 끝내려고 한다. (D) 글쓰기 과제 마감일이 다음주 목요일이다. (E) 정답 없음.
Chinese	王芳和张伟正在完成课堂上的写作任务。王芳问张伟什么时候交写作业，张伟回答说：	"下周四前得交上去。"	请在以下选项中选择最恰当地表达上述话语含义的选项。 (A) 张伟在向王芳寻求帮助。 (B) 张伟对英语没有信心，想推迟写作任务。 (C) 张伟想在今天完成写作任务。 (D) 下周四之前要交写作业。 (E) 没有正确答案。

Table 10: Multilingual test unit example on the category of literal interpretation. Charlie's reply is a direct answer to Emily's question about the deadline. His utterance does not trigger any implications based on the violation of Grice's maxims. It straightforwardly indicates that the due date is next Thursday. Therefore, option (D) "The writing assignment is due next Thursday" is the most appropriate meaning.

Language	Context	Utterance	MCQ
English	Emily saw Charlie's brother in a family photo and asked Charlie how old his brother was, to which he replied:	"He's 28."	Choose the most appropriate meaning of the above utterance from the following options. (A) Charlie does not know his brother's age. (B) Charlie's brother is not in college. (C) Charlie doesn't have a brother. (D) Charlie's brother is unemployed. (E) None of the above.
German	Anna sah Felix' Bruder auf einem Familienfoto und fragte ihn, wie alt er sei, woraufhin Felix antwortete:	"Er ist 28."	Wählen Sie die passendste Bedeutung der obigen Äußerung aus den folgenden Aussagen aus. (A) Felix weiß nicht, wie alt sein Bruder ist. (B) Felix' Bruder geht nicht auf eine Universität. (C) Felix hat keinen Bruder. (D) Felix' Bruder ist arbeitslos. (E) Keine der obigen Aussagen ist richtig.
Korean	영희는 철수의 가족사진에서 그의 동생을 보았고, 동생의 나이를 물었다. 이에 철수는 다음과 같이 대답했다.	"28살이야."	다음 보기에서 위 발화가 갖는 가장 적절한 의미를 고르세요. (A) 철수는 동생 나이를 알지 못한다. (B) 철수의 동생은 대학생이 아니다. (C) 철수는 동생이 없다. (D) 철수의 동생은 무직이다. (E) 정답 없음.
Chinese	王芳在一张全家福照片上看到了张伟的弟弟, 并问他几岁了, 张伟回答说:	"他28岁。"	请在以下选项中选择最恰当地表达上述话语含义的选项。 (A) 张伟不知道他的弟弟是几岁。 (B) 张伟的弟弟不是大学生。 (C) 张伟没有弟弟。 (D) 张伟的弟弟失业了。 (E) 没有正确答案。

Table 11: Multilingual test unit example without correct answer. Charlie's reply to Emily's question about his brother's age is straightforward and direct, with no implications based on the violation of Grice's maxims. His response should thus be interpreted as literal meaning: Charlie's brother is 28 years old. Since none of the options (A) to (D) accurately reflect this literal expression, each introducing an unrelated assumption, the correct answer is (E) "None of the above."

B Prompt Demonstration and Inter-Rater Agreement Analysis

Prompt

While visiting Charlie’s house, Emily saw a large pile of oranges in the kitchen and asked why there were so many. Charlie responded: (*context*)

"My uncle lives in Florida." (*statement*)

Choose the most appropriate meaning of the above utterance from the following options. (*MCQ*)

- (A) Charlie’s uncle sent the oranges.
 - (B) Charlie’s uncle resides in Florida.
 - (C) People in Florida do not like oranges.
 - (D) Charlie’s uncle lives in a rural house.
 - (E) None of the above.
-

Table 12: Example of the prompt using a test unit from our suite. It illustrates how the actual prompt is structured into a context and a corresponding statement followed by an MCQ with options. The words with parentheses are for clarification and are not part of the actual prompt.

		English	German	Korean	Chinese
Proprietary	GPT-4	0.87	0.86	<u>0.70</u>	0.90
	GPT-3.5	0.86	0.85	0.86	0.88
	Claude3-Opus	0.92	0.96	0.94	0.86
	Claude3-Sonnet	0.93	0.96	0.85	0.90
	Claude3-Haiku	0.95	0.95	0.90	0.91
	Mistral-Large	0.91	0.95	0.88	0.89
	Mistral-Medium	0.90	0.90	0.94	0.94
	Mistral-Small	0.80	0.84	0.84	0.85
Open-Source	Llama3-8B	0.86	0.91	0.90	0.90
	Llama2-13B	0.86	0.89	<u>0.56</u>	0.81
	Llama2-7B	0.88	0.86	0.87	0.92
	Gemma-7B	0.97	0.99	0.96	0.97
	Solar-10.7B	0.94	0.92	0.94	0.94
	Qwen-14B	0.96	0.95	<u>0.69</u>	0.95
	Qwen-7B	0.96	0.97	0.95	0.91

Table 13: Fleiss’ Kappa values representing inter-rater agreement across three trials on the MultiPragEval test suite for four languages. Most models demonstrate high Kappa values (above 0.80), indicating strong agreement across trials. However, models such as GPT-4, Llama2-13B, and Qwen-14B exhibit moderate agreement in generating Korean responses (0.56 to 0.70), suggesting some variability in their performance across the different trials.

C Score Tables

		German				
		Quan.	Qual.	Rel.	Man.	Avg.
Proprietary	GPT-4	70.56	76.67	77.22	65.56	72.50
	GPT-3.5	58.89	51.67	53.89	47.22	52.92
	Claude-Opus	85.56	87.78	85.00	72.78	82.78
	Claude-Sonnet	53.89	70.00	66.11	50.56	60.14
	Claude-Haiku	36.67	51.67	52.78	39.44	45.14
	Mistral-Large	60.00	70.00	73.33	51.67	63.75
	Mistral-Medium	47.22	68.89	56.11	42.22	53.61
	Mistral-Small	50.56	53.33	58.89	41.67	51.11
Open-Source	Llama3-8B	35.56	40.00	46.67	37.78	40.00
	Llama2-13B	20.00	13.33	15.00	17.22	16.39
	Llama2-7B	5.56	3.89	3.33	5.00	4.44
	Gemma-7B	29.44	23.89	35.00	20.56	27.22
	Solar-10B	56.67	59.44	62.78	43.89	55.69
	Qwen-14B	53.89	38.89	45.56	33.89	43.06
	Qwen-7B	45.56	37.78	41.11	33.33	39.44

Table 14: Performance scores on the MultiPragEval test suite across four maxims with overall averages for German. While the maxim of manner generally shows the lowest scores, high scores are more evenly distributed across the other three maxims.

		Korean				
		Quan.	Qual.	Rel.	Man.	Avg.
Proprietary	GPT-4	81.67	86.67	85.56	71.11	81.25
	GPT-3.5	42.22	47.22	37.22	28.89	38.89
	Claude-Opus	86.67	87.78	93.33	80.56	87.08
	Claude-Sonnet	58.89	74.44	67.78	52.22	63.33
	Claude-Haiku	37.22	49.44	37.78	29.44	38.47
	Mistral-Large	67.78	68.33	74.44	51.67	65.56
	Mistral-Medium	59.44	51.11	53.89	47.22	52.92
	Mistral-Small	41.11	52.22	42.78	25.56	40.42
Open-Source	Llama3-8B	34.44	39.44	31.11	25.00	32.50
	Llama2-13B	45.00	61.11	42.22	41.67	47.50
	Llama2-7B	5.56	5.00	0.00	1.67	3.06
	Gemma-7B	30.56	15.00	25.00	12.78	20.83
	Solar-10B	52.78	52.22	57.22	33.89	49.03
	Qwen-14B	53.33	58.89	44.44	42.22	49.72
	Qwen-7B	36.67	35.56	38.33	30.00	35.14

Table 15: Performance scores on the MultiPragEval test suite across four maxims with overall averages for Korean. The maxim of quality typically achieves the highest rankings, while the maxim of manner consistently records the lowest scores, reflecting a similar pattern observed in English.

		Chinese				
		Quan.	Qual.	Rel.	Man.	Avg.
Proprietary	GPT-4	59.44	85.00	72.78	57.78	68.75
	GPT-3.5	47.22	42.22	43.89	41.11	43.61
	Claude-Opus	80.56	82.22	80.56	63.33	76.67
	Claude-Sonnet	46.11	63.89	48.33	36.11	48.61
	Claude-Haiku	40.00	52.78	40.56	30.00	40.83
	Mistral-Large	47.22	60.56	66.67	44.44	54.72
	Mistral-Medium	43.89	46.67	36.67	28.33	38.89
	Mistral-Small	35.56	41.11	39.44	18.33	33.61
Open-Source	Llama3-8B	45.56	49.44	53.33	38.89	46.81
	Llama2-13B	6.67	12.78	3.33	12.22	8.75
	Llama2-7B	7.78	3.33	0.56	5.00	4.17
	Gemma-7B	29.44	26.67	18.89	25.28	25.28
	Solar-10B	49.44	57.78	46.67	31.67	46.39
	Qwen-14B	51.67	47.22	58.89	42.22	50.00
	Qwen-7B	45.00	46.11	35.56	37.78	41.11

Table 16: Performance scores on the MultiPragEval test suite across four maxims with overall averages for Chinese. The pattern of rankings mirrors those observed in English and Korean.

D Example Responses

Id	7	Category	Maxim of Quantity	Correct Answer	A
Question	<p>Charlie and Emily, who are in the same club, got into an argument at a club meeting last week. It's almost time for today's meeting and Charlie hasn't shown up. Michael, a friend from the club, asks Emily why he hasn't shown up. Emily tells him as follows:</p> <p>"Either he comes or he doesn't."</p> <p>Choose the most appropriate meaning of the above utterance from the following options.</p> <p>(A) Emily doesn't care if Charlie comes or not. (B) Charlie may or may not come. (C) Emily decided to apologize to Charlie. (D) Emily is not a good speaker. (E) None of the above.</p>				
Answer by GPT-4	<p>The most appropriate meaning of Emily's statement "Either he comes or he doesn't" is: (B) Charlie may or may not come.</p> <p>This interpretation is based on the literal meaning of the statement, which indicates uncertainty or neutrality regarding Charlie's attendance. It simply states the possibilities without any additional emotional or situational context.</p>				
Answer by Claude3-Opus	<p>The most appropriate meaning of Emily's utterance "Either he comes or he doesn't" is: (A) Emily doesn't care if Charlie comes or not.</p> <p>Explanation: Emily's statement suggests indifference towards Charlie's attendance at the club meeting. By saying "Either he comes or he doesn't," she implies that she is not concerned about whether Charlie shows up or not, likely due to their recent argument.</p> <p>The other options are less fitting because: (B) While the statement does suggest uncertainty about Charlie's attendance, it doesn't capture Emily's indifference. (C) There is no indication in the given context that Emily has decided to apologize to Charlie. (D) The statement does not reflect on Emily's speaking abilities. (E) Option (A) is the most appropriate interpretation of the given utterance.</p>				

Table 17: Responses by GPT-4 and Claude3-Opus on a test unit evaluating the Maxim of Quantity in English. While Claude3-Opus captures the implied nuance of indifference in the utterance, GPT-4 focuses on a literal interpretation, leading to an incorrect answer.

Id	127	Category	Maxim of Relation	Correct Answer	A
Question	<p>At a company dinner, a coworker invited Charlie to have a drink, and Charlie said:</p> <p>"I drove here."</p> <p>Choose the most appropriate meaning of the above utterance from the following options.</p> <p>(A) Charlie will not drink alcohol. (B) The car is parked in the parking lot. (C) Charlie plans to drink. (D) Charlie is planning to move. (E) None of the above.</p>				
Answer by Llama2-13B	<p>Answer: (C) Charlie plans to drink.</p> <p>Explanation: The statement "I drove here" is often used as a way of saying that one plans to drink alcohol. The speaker is indicating that they have arrived at the destination (the company dinner) by driving, and that they are prepared to consume alcohol. Therefore, option (C) is the most appropriate meaning of the utterance.</p>				
Answer by Llama2-7B	<p>Answer: (C) Charlie plans to drink.</p> <p>Explanation: The sentence "I drove here" is a common way to indicate that the speaker plans to drink alcohol. It implies that the speaker arrived at the location by driving, and therefore, they are not planning to drink. Therefore, option (C) is the most appropriate meaning of the utterance.</p>				

Table 18: Responses by Llama2-13B and Llama2-7B on a test unit evaluating the Maxim of Relation in English. Both models fail to capture the implied meaning of the utterance. While Llama2-7B demonstrates some degree of adequate inference in its reasoning, it still selects the incorrect option.

Id	75	Category	Maxim of Quality	Correct Answer	A
Question	<p>Emily, who saw Charlie come to work on Sunday, said:</p> <p>"He's not even breathing."</p> <p>Choose the most appropriate meaning of the above utterance from the following options.</p> <p>(A) Charlie works hard and diligently. (B) Charlie doesn't breathe when he works. (C) Time passes slowly, as if Charlie is out of breath. (D) Charlie is a bad sleeper. (E) None of the above.</p>				
Korean Answer by Llama2-13B	<p>답: ① 철수는 바쁘게 부지런히 일한다.</p> <p>설명: 위 발화 "재는 숨도 안 쉬어"는 "철수는 바쁘게 부지런히 일한다"라는 의미를 갖고 있습니다. 이 발화는 철수가 일요일에 나와서 일하는 것을 본 영희가 철수의 열의와 힘든 일을 직감적으로 표현한 것입니다.</p>				
Translation of Korean Answer	<p>Answer: ① Charlie works hard and diligently.</p> <p>Explanation: The utterance "He's not even breathing." means "Charlie works hard and diligently." This utterance is Emily's instinctive expression of Charlie's enthusiasm and difficulty at work after seeing him come to work on Sunday.</p>				
Chinese Answer by Llama2-13B	<p>答: B</p>				
Translation of Chinese Answer	<p>Answer: B</p>				

Table 19: Responses by Llama2-13B on a test unit evaluating the Maxim of Quality in Korean and Chinese. The model demonstrates a significant performance gap between the two languages. While it selects the correct answer and provides a well-reasoned explanation in Korean, it chooses the literal interpretation without any explanation in Chinese.

Id	169	Category	Maxim of Relation	Correct Answer	E
Question	<p>Charlie and Emily walked into a cafe. As Emily was choosing a beverage from the menu, Charlie said: "I have a gift card."</p> <p>Choose the most appropriate meaning of the above utterance from the following options. (A) Charlie wants Emily to give him a gift card. (B) Charlie gave Emily a gift card. (C) Charlie will save the gift card instead of using it. (D) Charlie used to work part-time at a cafe. (E) None of the above.</p>				
Answer by GPT-4	<p>The most appropriate meaning of Charlie's utterance "I have a gift card" in the context provided is: (E) None of the above.</p> <p>The correct interpretation, which isn't listed among the options, is likely that Charlie is indicating he has a gift card to use at the cafe, possibly suggesting he can pay for the beverages using the gift card. This interpretation is not explicitly stated in the options provided.</p>				
Answer by Claude3-Opus	<p>The most appropriate meaning of Charlie's utterance "I have a gift card" is: (E) None of the above.</p> <p>In the given context, Charlie mentioning that he has a gift card most likely implies that he intends to use it to pay for their purchases at the cafe. This suggests that:</p> <ul style="list-style-type: none"> - Charlie is offering to treat Emily using his gift card, or - Charlie is informing Emily that they can use the gift card to save money on their purchases. <p>The other options (A, B, C, and D) do not accurately represent the meaning of Charlie's statement in the given context.</p>				

Table 20: Responses by GPT-4 and Claude3-Opus on a test unit where no correct answer is provided. Both models correctly identify the absence of a correct answer and successfully explain the intended meaning of the utterance in their responses.