

Do Large Language Models Understand Conversational Implicature – A case study with Chinese sitcom

Shisen Yue

Shanghai Jiao Tong University

lyw520@sjtu.edu.cn

Abstract

Understanding the non-literal meaning of an utterance is critical for large language models (LLMs) to become human-like social communicators. In this work, we introduce SwordsmanImp, the first Chinese multi-turn-dialogue-based dataset aimed at conversational implicature, sourced from dialogues in the Chinese sitcom *My Own Swordsman*. It includes 200 carefully handcrafted questions, all annotated on which Gricean maxims have been violated. We test eight close-source and open-source LLMs under two tasks: a multiple-choice question task and an implicature explanation task. Our results show that GPT-4 attains human-level accuracy (94%) on multiple-choice questions but fails to generate explanations of implicature that are fully aligned with human understanding. Other models, including earlier versions of GPT and several open-source models, demonstrate a low accuracy ranging from 20% to 60% on multiple-choice questions; they also struggle to generate sensible, logical and fluent explanations about conversational implicatures. Moreover, we find LLMs’ performance does not vary significantly by Gricean maxims. These results suggest that there is still much room for improvement for most of the current LLMs in pragmatic reasoning.

1 Introduction

The complexity of communication is largely epitomized by indirect, or non-literal utterances. A common instance is hinting at a busy schedule as a polite refusal to engage in an unwanted activity. This form of indirect expression is known as conversational implicature in the field of pragmatics. It encodes implied meanings within a situational context rather than in the linguistic message itself. As such, it allows for a range of interpretations, unlike the straightforward mapping of meaning from language in semantics. This ambiguity of meaning presents further challenges for disambiguation

白眉	你说得轻巧。这个人的一辈子，只能开三次天眼。我已经开了两次了，最后一次，我想给自己留着，数数寿数。
秀才：	喔，那好吧，那这样就算了吧。
白眉：	哎哎，你要真想算，不是没有办法。我可以给你一次机会。
秀才：	呵呵，谢谢你。（白眉不说话）哦，银子我有。
白眉：	你我都是文人，提钱多俗啊！你有多少？
Bai Mei:	You speak lightly. In a person’s lifetime, they can only open the Heavenly Eyes three times. I’ve already opened them twice. For the last time, I want to keep it for myself, to predict my lifespan.
Xiucui:	Oh, okay, let’s just leave it at that then.
Bai Mei:	Hey, hey, if you really want to do prediction, there is a way. I can give you a chance.
Xiucui:	Hmmm, thank you. (Bai Mei remains silent) Oh, by the way, I have money.
Bai Mei:	Both you and I are intellectuals; talking about money is so vulgar! But how much do you have?

Experiment 1 requires LLMs to choose the correct pragmatic interpretation of the target sentence from four choices (Section 4).
Experiment 2 requires LLMs to generate explanations of the target sentence (Section 5).

Table 1: An example dialogue from the SwordsmanImp dataset. The target sentence that contains conversational implicature is marked in bold. Multimodal information is supplemented in the bracket to better understand the conversational implicature.

and is a key subject of investigation in pragmatics research (Grice, 1975; Searle et al., 1980; Brown and Levinson, 1987; Wilson and Sperber, 2006).

This ambiguity also presents a challenge for large language models (LLMs) in attaining proficient pragmatic comprehension. Recent studies have evaluated LLMs’ pragmatic reasoning in

multiple aspects, including scalar inference (Hu et al., 2023b), discourse connectives (Pandia et al., 2021), gradable adjectives (Lipkin et al., 2023) and conversational implicatures (Qiu et al., 2023; Kim et al., 2023; Ruis et al., 2022; Hu et al., 2023a; Zheng et al., 2021). However, the prior test materials are primarily in English, leaving a gap for pragmatic understanding in other languages. Moreover, previous conversational reasoning datasets involved few turns of dialogue, while in daily communication, situational context is more abundant and usually embedded in multi-turn dialogues. This requires a more human-like capacity for extracting context messages to decode conversational implicature. In addition, the various forms of questions applied in prior evaluation include multiple-choice questions (Hu et al., 2023a) and binary-answer questions (Ruis et al., 2022). While subjective questions such as interpreting non-literal meaning typically provide a more comprehensive gauge of their pragmatic skills, they are barely involved in previous assessments.

From a linguistic perspective, conversational implicature can be classified according to the violated maxim in the cooperative principle. As stated by Grice (Grice, 1989), the violation of maxims varies by their tolerance. In other words, some conversational implicatures are more perceptible than others, according to the Gricean maxim they violate. The maxim of quantity is verified in empirical studies to be less perceptible (Engelhardt et al., 2006; Rubio-Fernandez, 2019; Okanda et al., 2015; Panzeri and Foppolo, 2021). By evaluating LLMs’ understanding of conversational implicatures that are evoked by the violation of different maxims, we can find out if LLMs’ tolerance to implicature varies by maxims, which serves as a clue to answer the question of whether LLMs resemble humans in processing such information.

To bridge the aforementioned gaps, we present our Chinese conversational dataset *Swordsman-Imp*, the first Chinese multi-turn-dialogue-based dataset aimed at conversational implicature. It consists of 200 dialogues carefully excerpted by linguistic experts from the Chinese sitcom *My Own Swordsman*. Table 1 shows an example dialogue in the dataset. Each dialogue contains one sentence that carries a non-literal meaning. We provide four well-organized interpretations of this sentence, including a pragmatic meaning, a literal understand-

ing, and two incorrect inferences that involve related information. The questions in our dataset cover four types of conversational implicature that correspond to four maxims in the cooperative principle.

Based on our dataset, we aim to answer the following questions with two experiments:

(1) How do state-of-the-art LLMs perform in interpreting implicatures written in Chinese, in multiple choice style? (Section 4.3)

(2) What are the error patterns of LLMs on multiple choice questions? (Section 4.3)

(3) How do human participants rate the interpretation of conversational implicature generated by LLMs? (Section 5.3)

(4) Do models differ from human in processing conversational implicature? (Section 6)

In Section 4, we measure the models’ accuracy on multiple-choice questions derived from our dataset. We test four models from GPT series (Radford et al., 2018), three open-source models and BLOOMZ-7.1B (Muennighoff et al., 2023). In Section 5, we use five models selected from Experiment 1 to generate pragmatic interpretations of 32 dialogues, and then we invite human participants to rate these interpretations in three dimensions.

We find that GPT-4 attains a human-level accuracy in multiple-choice questions while other models perform poorly in this task. We also find that OpenAI LLMs choose less frequently the two false inference than other open source models. Besides, we find little evidence that implicatures evoked by different maxims are processed differently by LLMs. Moreover, we reveal the different patterns of LLMs in explaining implicatures, and find that confused logic, redundant information and unexpected language tokens are the main reasons for human evaluators to underrate the generated explanations.

2 Related Work

Understanding non-literal meanings has long been considered a difficult task for language models. Previous studies have explored the capabilities of language models in recognizing metaphors (Wachowiak and Gromann, 2023; Neidlein et al., 2020), humor (Hessel et al., 2023; Jentzsch and Kersting, 2023; Chen et al., 2023) and social commonsense (Sap et al., 2019). Broad-scale datasets aimed at pragmatic reasoning collect data mainly through crowdsourcing or crawling from the in-

Dataset	Task	Context	Manually	Language
CoQa (Reddy et al., 2019)	conversational QA	Paragraph	✓	English
Narrative-QA (Kočíský et al., 2018)	Reading comprehension	Documents	✓	English
DREAM (Sun et al., 2019)	QA	Free-form, multi-turn dialogue	✓	English
MuTual (Cui et al., 2020)	Next utterance prediction	Free-form, multi-turn dialogue	✓	English
GRICE (Zheng et al., 2021)	Gricean Implicature Recovery, QA	Fixed form, multi-turn dialogue	✗	English
PragMega (Floyd et al., 2023)	ToM QA	Paragraph	✓	English
LUDWIG (Ruis et al., 2022)	Implicature QA	Single-turn dialogue	✓	English
Douban (Wu et al., 2017)	Next utterance prediction	Free-form multi-turn dialogue	✗	Chinese
SwordsmanImp (This work)	Gricean Implicature QA	Free-form, multi-turn dialogue	✓	Chinese

Table 2: Comparing our datasets and the existing datasets. “Manually” indicates whether the questions or answers are written partly or entirely by human.

ternet (Reddy et al., 2019; Kočíský et al., 2018). Some others transform the existing tests for humans to train and evaluate language models (Cui et al., 2020; Sun et al., 2019). The pragmatic knowledge included in these large-scale datasets is too comprehensive to focus on a specific aspect. In our work, we craft our dataset manually to keep each entry within the boundary of Gricean conversational implicature. This enables us to perform a fine-grained analysis of LLMs’ understanding of this pragmatic aspect. It also allows for comparison with human performance on similar tasks to better understand the underlying process of LLMs’ pragmatic reasoning.

Driven by a similar motivation, Zheng et al. (2021) construct GRICE with a well-defined grammar model. Each entry in this dataset is formulated through fixed syntactic structures. This grammar-based method ensures a precise control of the sentence structure to provide tenable insights within the linguistic framework. However, it deviates from the daily conversation which characterizes abundant context information involved in diverse sentence structures that are harder to learn for LLMs. While PragMega (Floyd et al., 2023) includes Gricean implicature as one of the seven phenomena and provides a rich context for each question, context information is narrated through paragraphs preceding the questions. This eliminates the difficulty of understanding the situation from dialogues, a process that happens commonly in human conversations. In this study, the pragmatic understanding of the target sentence is deduced from multi-turn dialogues that are similar to natural conversations.

Language differences should also influence LLMs’ dialogue understanding. Previous large-scale conversational datasets in Chinese include

Sina Weibo (Shang et al., 2015), Douban conversational corpus (Wu et al., 2017) and E-commerce Dialogue corpus (Zhang et al., 2018). However, no previous datasets in Chinese are specifically aimed at pragmatic reasoning. Given that the understanding of conversational implicature relies heavily on social norms and commonsense knowledge encoded in the spoken language, investigating cross-linguistic pragmatic reasoning in LLMs demonstrates considerable value. To address this gap, we select the Chinese sitcom *My Own Swordsman* as the source to formulate pragmatic questions.

3 Dataset Construction

Data Collection To the best of our knowledge, there’s no reliable method to automatically detect the implicature from a conversation and to precisely clip the dialogue that contains all necessary information for pragmatic reasoning. Also, the judgment of conversational implicature relies on linguistic knowledge that is hard to program. Thus we manually collect the dialogues that involve implicature from the transcripts of the Chinese sitcom *My Own Swordsman*. We select the segment of conversation by judging if any conversational maxim is violated on purpose (Grice, 1975). We also do the cancellation test² (Hirschberg, 1985) to ensure that a conversational implicature rather than a semantic entailment is involved.

Four Interpretations We construct the four interpretations of the sentence that carries implicature. We deduce the pragmatic meaning based on our understanding of the context and storyline. We

¹“sneak out” and “skate” translate to the same Chinese character “溜”.

²A test to diagnose the conversational implicature by encoding semantically the negation of the target meaning. If the result seems consistent, then the target meaning is likely an implicature.

Dialogue	
小郭:	知道这意味着什么吗？以后再敢胡来，就不光是挨顿打的事了。
薇:	烦死了，在家就是听爹娘罗嗦，好不容易溜出来，倒听你罗嗦。
小郭:	什么？你是溜出来的？
薇:	我说的那个溜，是溜冰的溜啊，现在京城都在下雪，满地都是冰，我是穿着冰刀，溜出来的，嘿嘿……
Xiaoguo:	Do you know what this implies? If you dare to mess around again, it won't just be about getting a beating.
Wei:	So annoying. At home, I have to listen to my parents nagging. Finally sneaking out, now I have to listen to you nagging.
Xiaoguo:	What? You sneaked out?
Wei:	When I said "sneak out," ¹ I meant ice skating. It's snowing in the capital now, the ground is covered in ice. I wore ice skates and skated out, hehe...
Choices	
Pragmatic	薇发现自己说溜出来说漏嘴了，就找补说是溜冰出来的。 Wei realized she accidentally slipped up by saying she sneaked out and tried to cover it up by saying she meant ice skating.
Literal	薇是从京城的家里溜冰出来的。 Wei sneaked out from her home in the capital to go ice skating.
Distractor#1	薇很喜欢溜冰。 Wei really enjoys ice skating.
Distractor#2	薇怕下雪冷，所以离开了京城。 Wei was afraid of the cold due to the snow, so she left the capital.
Maxim Check	
Maxim	Sub-maxims
Quality	✗ Do not say what you believe is false.
	✗ Do not say that for which you lack adequate evidence
Quantity	✓ Make your contribution as informative as is required.
	✗ Do not make your contribution more informative than is required.
Relation	✓ Be relevant
Manner	✓ Avoid obscurity of expression.
	✓ Avoid ambiguity
	✗ Be brief
	✓ Be orderly

Table 3: An example entry in our dataset. The tick (✓) and fork (✗) denotes if the target sentence, marked in red, comply with or violate the sub-maxim. This entry belongs to the classes of Quality, Quantity and Manner.

construct the literal meaning by paraphrasing the sentence and explaining the abstract concepts. We generate incorrect inferences as distractors, which are relevant to the the topic of the conversation. The character names referred to in the four interpretations are aligned with the mentions in the dialogue clip. These four interpretations are directly transformed into four choices in the multiple-choice questions.

Verification We hired three Linguistic-major graduate students to complete a multiple-choice question task, which is to choose the pragmatic understanding of the sentence from the shuffled four interpretations for all dialogues in our dataset. The students are required to complete the task independently without a time limit. After they finish, we invite them to discuss about their wrong answers and their reasoning process. This validation process guarantees that the provided pragmatic interpretation is closely aligned with the commonsense intuitive understanding and can be deduced from the limited context. Additional information that is necessary to pragmatic reasoning, such as relationships between characters, their personality, social background, and multi-modal information is supplemented in the bracket at the beginning of the dialogue. Besides, implicatures derived from vernacular, slang, and network-specific jokes have been filtered out from our dataset.

Classification We classify the dialogues in our dataset according to the conversational maxims they violate. The criteria are the description of these maxims by Grice (1975). To allow for a more fine-grained classification, we employ sub-maxims as criteria, assessing if the target sentence fulfills each requirement individually. An utterance is considered to violate a maxim if it infringes on any sub-maxim. Moreover, a dialogue might belong to multiple classes according to the sub-maxims the utterance violates. An illustrative dataset entry featuring a dialogue, four interpretations and a class is presented in Table 3.

4 Experiment 1: multiple-choice question for LLMs

In this experiment, we assess models with multiple-choice questions derived from our dataset. The requirement is to choose the pragmatic interpretation of the target sentence by distinguishing it from the literal meaning and the two semantically connected yet incorrect inferences.

	Total	Quality	Quantity	Relevance	Manner
# of questions	200	76	33	71	62
# of turns per dialogue	6.80	7.84	5.91	6.23	6.35
Avg. dialogue length	158.22	184.53	143.67	147.20	152.79
Avg. Utterance length	23.27	23.53	24.31	23.64	24.04
Avg. Answer length	15.08	14.47	14.85	15.82	14.86

Table 4: Question numbers, average number of Chinese characters contained in each dialogue and utterance, and the number of turns per dialogue in our dataset.

	Pragmatic	Literal	Distractor#1	Distractor#2
Total	19.20	16.38	12.54	12.21
Quality	18.32	15.88	11.96	11.72
Quantity	18.36	16.73	12.30	12.00
Relevance	20.77	16.62	13.11	12.77
Manner	18.10	17.18	12.13	12.06

Table 5: Average number of Chinese characters in each interpretation for all questions and for each type of questions.

4.1 Experimental setup

We experiment with eight models in this experiment. The first four models are from the OpenAI GPT family, which are tested through OpenAI API, with the default hyperparameters: text-davinci-002, text-davinci-003, GPT-3.5-turbo and GPT4. We also examine four open source models with the Transformers library (Wolf et al., 2020) from Huggingface³. We evaluate Chinese-Alpaca-2-13B (Cui et al., 2023), OpenBuddy-Llama2-13B⁴ (based on Llama2 (Touvron et al., 2023)), CausalLM-13B (based on Llama2 and Qwen (Bai et al., 2023)), as well as BLOOMZ-7.1B from the BLOOM series (Muennighoff et al., 2023). To draw direct comparisons between models, the answers are generated using the same script and parameters (See Appendix A).

4.2 Prompts Structure

We use zero-shot prompts to simulate the real-life situations where humans come across these non-literal expressions daily. The prompts used to test open-sourced models and GPT series models are the same. We embed the multiple-choice questions in a prompt template as follows:

你现在是一个中文母语者。对于以下对话，请识别特定人物的话语中的言外之意，在给

³<https://huggingface.co>

⁴<https://huggingface.co/OpenBuddy/openbuddy-llama2-13b-v8.1-fp16>

Subject	Acc (%)
Random baseline	25
BLOOMZ (7.1B)	35.5
OpenBuddy-Llama2 (13B)	21.5
Chinese-Alpaca-2 (13B)	42.0
CausalLM (14B)	31.5
text-davinci-002	24.0
text-davinci-003	56.5
GPT-3.5-Turbo	55.0
GPT-4	94.0
Human	93.1

Table 6: Accuracy of language models and humans in all multiple choice questions.

出的四个选项中选择一个你认为的正确答案。请在'Response:'后写你的答案。

< Dialogue >

< Four interpretations as choices >

请在'Response:'后写出你选择的答案。

(English translation for illustrative purposes:

You are now a native Chinese speaker. For the following dialogue, please identify the implied meaning in the specific character's speech, and choose one of the four given options that you think is the correct answer. Please write your answer after 'Response:'.

< Dialogue >

< Four interpretations as choices >

Please write your answer after "Response: ".)

4.3 Results

For comparative purposes, we ask 10 native speakers to work on 32 questions randomly sampled from the data set and they achieve an average accuracy of 93.1%. The overall accuracy of all models and human participants is displayed in Table 6. GPT-4 achieves the highest accuracy of 94% in all the questions, showing a performance on par with the human participants. Other models exhibit difficulty in identifying the pragmatic meaning of the line containing an implicature, with accuracy ranging from 20% to 60%. OpenBuddy-Llama2 (13B)

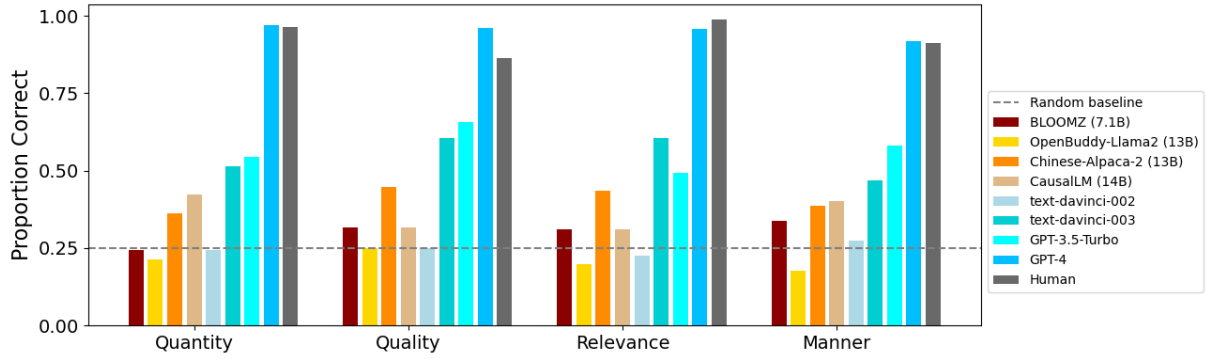


Figure 1: Performance of models across question types in the multiple choice scenario. Dashed line represents chance accuracy

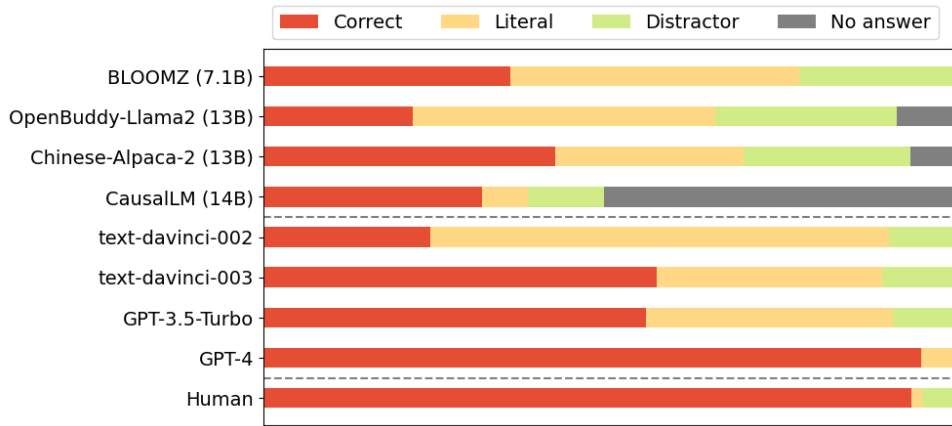


Figure 2: Frequency of each option in models' answers. the cases that models refute all of the choices or simply repeat the prompt are counted as "no answer".

and text-davinci-002 even fail to perform above chance (25%).

Model accuracy across different types of questions is shown in Table 1. The accuracy of OpenBuddy-Llama2 (13B) and text-davinci-002 are below chance in all types of questions. The accuracy of BLOOMZ (7.1B) manifests an incompetency in interpreting sentences that flout the maxim of Quantity. In questions related to Quality and Relevance infringements, the performance of Chinese-Alpaca-2 (13B) surpasses that of CausalLM (14B). Conversely, when comparing the results of the remaining two categories of implicatures, CausalLM (14B) exhibits superior performance. ChatGPT and text-davinci-003 display higher-level performance, attaining accuracy above 50%. GPT-3.5-Turbo excels particularly in dialogues that breach the maxims of Quality and Manner. Meanwhile, text-davinci-003 outperforms in the remaining areas. GPT-4 attains the highest accuracy, surpassing 90% across all categories of questions. Overall, certain models

demonstrate proficiency in answering questions related to a set of maxims, while others excel in different ones. We do not observe a uniform pattern indicating a particular strength or weakness in any of these maxims across the models. This lack of uniformity is also observed in the results obtained from human participants who worked on the 32 sampled questions.

The distribution of choices made by models is displayed in Figure 2. When compared to open-source models, GPT models seem to make fewer choices on distractors. This could possibly indicate a stronger emphasis that GPT models place on the selected sentence, rather than being easily sidetracked by irrelevant information in the context. The evolution of GPT models also witnesses a boost in distinguishing literal meaning from implied meaning, culminating in the considerably low ratio of literal understanding in GPT-4. The open-source Chinese LLMs seem to struggle much more. Chinese-Alpaca-2, Openbuddy-Llama (13B) and CausalLM (14B) all produce re-

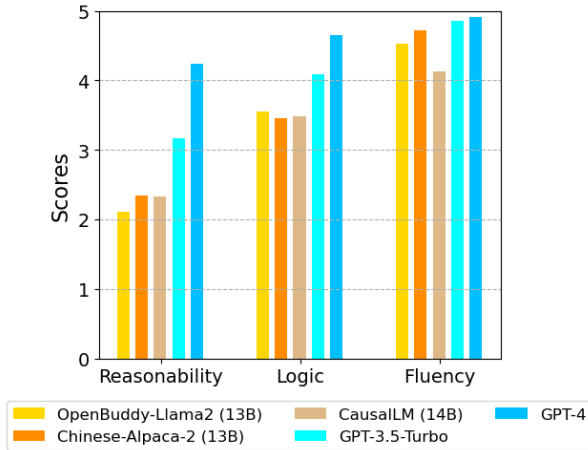


Figure 3: Scores of five models’ responses in three assessment dimensions, averaged across 32 questions and 5 human participants.

sponses that contain no answer. Notably, over half of the questions are null in the answers of CausalLM (14B), and the majority of these simply repeat the questions.

5 Experiment 2: evaluating the quality of explanations generated by LLMs

In the previous experiment, we ask LLMs to choose one answer from four choices. In this experiment, we design open-ended questions where the models are asked to generate explanations of the implicature, which will then be evaluated manually by native speakers of Chinese, based on the reasonability, logic and fluency of the generated explanations.

5.1 Experimental setup

In this experiment, we only use the models in Experiment 1 that demonstrate a capacity for generating complete sentences. We then carry out a pilot trial comprising 5 five questions. Among the eight models, BLOOMZ (7.1B), text-davinci-002 and text-davinci-003 produce brief, fragmented responses, despite explicit prompts for detailed explanation. We therefore select GPT-3.5-Turbo, GPT-4, Chinese-Alpaca-2 (13B), OpenBuddy-Llama2 (13B) and CausalLM (14B) for this experiment. To avoid incomplete answers, we increase the maximum number of newly generated tokens in calling open-sourced models, while GPT models perform well with the hyperparameters used in Experiment 1.

We use zero-shot prompts embedded within a similar structure to that used in Experiment 1.

These prompts are designed to instruct the models to formulate an explanation for the target sentence, grounded in the context of the dialogue. An example prompt can be found in Table 8.

5.2 Human evaluation

We invite 15 human participants to evaluate the answers generated by models based on their reasonability, logic and fluency. Reasonability is gauged by the degree to which the response is aligned with the interpretation provided by the human annotator. Logic is assessed based on the presence of a clear and consistent chain of reasoning in the response. Fluency, on the other hand, evaluates whether the generated text is devoid of incoherent elements such as foreign language insertions or unintelligible code fragments.

5.3 Results

Figure 3 depicts the ratings of five models’ responses in three dimensions and Table 7 presents the mean scores with standard deviations in each dimension, along with the average number of characters in each model’s responses. Reasonability as a direct indicator of the model’s pragmatic understanding reveals an identical order in their performance on multiple-choice questions. Responses from GPT-4 scores the highest in all three dimensions with the smallest variance. Responses from GPT-3.5-Turbo receive high but unstable scores throughout the 32 questions. Answers from the other three models are comparable, with Chinese-Alpaca-2 outperforms the other two models in an insignificant level ($F = .964, p = .549 > .05$).

Examining the generated answers in detail explains the distribution of human rating scores. Text generated by GPT-4 and GPT-3.5-Turbo are both identified with a minimal number of “code switching”, the phenomenon of adulterating non-Chinese tokens into their responses, with the interpretation from GPT-4 being more precise and coherent than that from GPT-3.5-Turbo. While Chinese-Alpaca-2 (13B), OpenBuddy-Llama2 (13B) and CausalLM (14B) have comparable performance, they all underperform the two GPT models. They reflect different patterns of generated text. Both Chinese-alpaca-2 (13B) and Openbuddy-Llama2 (13B) feature highly pure Chinese responses with

⁵Escort Agency, refers to historical security firms that were typically hired by trade merchants for protection during transportation of goods. They were known to use martial arts for defense, similar to bodyguards or security personnel.

	Reasonability	Logic	Fluency	Avg. response length
GPT-4	4.24 ± 0.68	4.65 ± 0.39	4.91 ± 0.13	114.44
GPT-3.5-Turbo	3.17 ± 1.30	4.09 ± 0.77	4.86 ± 0.21	125.41
Chinese-Alpaca-2 (13B)	2.34 ± 1.10	3.45 ± 0.82	4.72 ± 0.39	156.19
CausalLM (14B)	2.33 ± 1.03	3.48 ± 0.67	4.13 ± 1.01	147.41
Openbuddy-Llama2 (13B)	2.11 ± 0.99	3.55 ± 0.71	4.52 ± 0.65	153.56

Table 7: Models’ mean scores in three dimensions with standard deviation and the average number of Chinese characters in their responses.

Question	Model Responses
<p>你现在是一个中文母语者。对于以下对话，请识别特定人物的话语中的言外之意，并解释。</p> <p>李大嘴： 是啊，要么怎么说是龙门镖局的少东家呢，一般人哪有这气派是不是。啥叫家传，啥叫传统（刚说道这儿佟石头“哇”的一声吐了出来，众人挥手趋散气味）</p> <p>佟湘玉： 这可不是我家的传统。</p> <p>佟石头： 再来！</p> <p>佟湘玉： 佟石头你有完没完了？</p> <p>佟石头： 给洒家倒酒！</p> <p>佟湘玉： 不用倒了，捧着坛子一口气喝完那多爷们儿啊！</p> <p>佟石头： 姐，那一坛子二斤半呐！</p> <p>佟湘玉： 你不是一口气能喝五六十斤嘛？</p> <p>请根据以上情景判断佟湘玉说的“你不是一口气能喝五六十斤嘛？”有什么言外之意，并解释。</p> <p>You are now a native Chinese speaker. Regarding the following conversation, please identify the implicit meanings in the speech of specific individuals and explain them.</p> <p>Dazui : Yeah, otherwise, how could he be considered the young master of the Longmen Escort Agency⁵? Ordinary people don’t have this aura, right? What’s meant by family heritage, what’s tradition...(Just as he was saying this, Shitou, with a ”wow,” spat out, and everyone waved their hands, dispersing the smell.)</p> <p>Xiangyu : That’s not our family’s tradition.</p> <p>Shitou : Next!</p> <p>Xiangyu : Shitou, are you ever going to finish?</p> <p>Shitou : Pour me a drink!</p> <p>Xiangyu : No need to pour. If you hold the jug, you’ll finish it all, big guy!</p> <p>Shitou : Sis, that jug is forty four ounces!</p> <p>Xiangyu : Aren’t you capable of drinking a thousand ounces in one gulp?</p> <p>Please judge the implicit meaning of Xiangyu’s statement “Aren’t you capable of drinking a thousand ounces in one gulp?” based on the above context, and explain it.</p>	<p>GPT-4: 佟湘玉的话”你不是一口气能喝五六十斤嘛？”的言外之意是在讽刺和挑衅佟石头。她的话语表面上是在赞扬佟石头的饮酒能力，实际上是在质疑他的酒量，挑衅他是否真的能一口气喝完一整坛的酒。同时，这句话也表达了她对佟石头酒后行为的不满和无奈。 (114 characters)</p> <p>Xiangyu’s statement ”Aren’t you capable of drinking a thousand ounces in one gulp?” implies sarcasm and provocation towards Shitou. On the surface, her words seem to praise Shitou’s drinking ability, but in reality, she questions his capacity for alcohol, challenging whether he can really drink an entire jug in one go. Simultaneously, this sentence also conveys her dissatisfaction and helplessness toward Shitou’s behavior when he’s drunk.</p> <p>CausalLM (14B): 佟湘玉的言外之意是质疑佟石头的酒量，暗示他可能酒品NST。她可能认为，佟石头的言语和行为过于狂妄自大，让他在众人面前出丑。通过这句话，佟湘玉试图让佟石头意识到自己的错误，并提醒他不要 forgot his place. (109 characters)</p> <p>Xiangyu’s implicit meaning questions Shitou’s alcohol capacity, suggesting he might not handle alcohol well. She might perceive Shitou’s words and actions as overly arrogant, causing embarrassment in front of others. Through this statement, Xiangyu attempts to make Shitou aware of his mistake and reminds him not to forget his place.</p> <p>OpenBuddy-Llama2 (13B): 从这段对话中，我们可以看出佟湘玉说的话中存在一些言外之意。她说“你不是一辈子能喝五六八番呢？”其实这句话暗示着佟石头喝酒的行为可能已经到了了一定的程度，而她自己也已经接受了这个现实。这种暗示的意思是，佟石头喝酒的习惯已经形成，他已经不能像以前那样随意地停止喝酒了。因此，佟湘玉的话可以被解释为一种对佟石头喝酒行为的默许和接受。 (163 characters)</p> <p>From this conversation, we can see that Xiangyu’s words contain some implicit meanings. When she says, ”Aren’t you capable of drinking for a lifetime?” in fact, this sentence implies that Shitou’s drinking behavior may have reached a certain level, and she herself has accepted this reality. The implication here is that Shitou’s drinking habit has formed, and he can no longer stop drinking as freely as before. Therefore, Xiangyu’s words can be interpreted as a kind of tacit approval and acceptance of Shitou’s drinking behavior.</p>

Table 8: Models’ interpretations to a character line (highlighted with red color) that carries extra meanings in a dialogue with multiple turns. Explanations aberrant from the correct interpretation and non-Chinese characters are highlighted with yellow bottom color.

their scores in text quality above 4.5, but the latter model scores distinctively low in reasonability. CausalLM-14B, on the contrary, has a relatively

higher performance in providing correct interpretations while generating cross-linguistic tokens frequently.

Table 8 presents a typical example to illustrate how the different styles of generation influence readers’ acceptance levels. The explanation from the annotator for this sentence is that Xiangyu’s words are absurd by itself. She purposely makes such a ridiculous statement to notify Shitou of his absurd behaviors and exaggerated words. Her words express her irony and dissatisfaction towards Shitou. GPT-4 gives a concise interpretation that aligns closest to the human interpretation among the selected models. However, it mistakes the ironic tone as questioning Shitou’s capacity for liquor, as highlighted in the example. CausalLM (14B) actually reaches the correct interpretation, but the quality of its answer is dragged down by its poor fluency, that some English words and meaningless character sequences are inserted in its answer. It’s noteworthy that the “forgot his place” carries the correct meaning in an unexpected language. The response from Openbuddy-Llama2 (13B) exemplifies the low acceptability of a response that is both verbose and irrelevant.

6 Discussion and Future Work

Do models differ from human in processing conversational implicature?

The result in Experiment 1 doesn’t demonstrate a significant by-maxim variance in human accuracy. We attribute this balanced performance to the low difficulty of this task instead of the fact that humans have the same sensitivity to the four conversational maxims. However, for the models that don’t share the same level of accuracy with GPT-4 and human participants, their balanced accuracy shouldn’t be ascribed to the setting of this task. Rather, we believe there’s a significant difference in the way they process conversational implicature from that of humans.

We draw this conclusion from two pieces of evidence. First, based on the cooperative principle, human beings sense implicatures when some of the requirements for being a cooperative communicator are overtly infringed. This understanding is formulated by first assuming cooperation in standard scenarios, and then, when this norm is disrupted, the listener engages in inference to deduce the speaker’s underlying intentions. (Levinson, 1995). However, we find little evidence from LLMs’ explanations that they rely on the abnormality of utterance to infer a plausible intention. Secondly, LLMs balanced performance isn’t aligned

with the pattern described by Grice’s, that humans treat the conversational maxims differently, with more urgency imposed on some maxims than others (Grice, 1975, p. 46). In an empirical aspect, LLMs’ statistically insignificant variance in sensitivity to different maxims draws a direct comparison to the results from previous studies with human participants (Engelhardt et al., 2006; Rubio-Fernandez, 2019; Okanda et al., 2015; Panzeri and Foppolo, 2021).

Future directions

The multiple choice task in this dataset might be no longer be a hard task for future LLMs, while the results in Experiment 2 necessitates a further improvement for GPT-4 to produce interpretations fully aligned with the human interpretations. We thus call for more sophisticated design to better quantify their free-form explanations of conversational implicature.

We also foresee the possibility of creating a large-scale dataset composed of conversational implicature embedded in natural dialogues. Current datasets for pragmatic reasoning can attain a tremendous scale at the expense of dialogue naturalness, linguistic constraints or a focus on certain pragmatic aspects. We look forward to modeling the conversational implicature in a rule-based approach, particularly through classic theories in pragmatics, to extract precisely the target sentences from a broad scope of language data.

7 Conclusions

In this paper, we present SwordsmanImp, a fine-grained Chinese dataset to evaluate language models’ understanding of conversational implicature. In the two experiments, we evaluate the state-of-the-art language models’ pragmatic skills with two tasks. In Experiment 1, we reveal that GPT-4 attains a human-level accuracy in answering multiple choice questions, with other models lagging a considerable distance behind. We also find that no significant difference exists within both LLMs and human’s accuracy with respect to conversational implicatures that violate different maxims. The result from Experiment 2 reflects the different patterns of generated text across LLMs and to what extent they are accepted by human judges. Through this more difficult task we also reveal a space to improve for GPT-4 in its understanding of conversational implicature.

Acknowledgments

I would like to thank Siyuan Song and Xinyuan Cheng for their involvement in the data collection and annotation process. I would also thank all the anonymous participants in the study. I am grateful to Professor Hai Hu for his constructive suggestions on experimental design and paper writing.

References

- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.
- Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. 2023. Can pre-trained language models understand chinese humor? In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 465–480.
- Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. **MuTual: A dataset for multi-turn dialogue reasoning**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2023. **Efficient and effective text encoding for chinese llama and alpaca**.
- Paul E Engelhardt, Karl GD Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of memory and language*, 54(4):554–573.
- Sammy Floyd, Edward Gibson, and Evelina Fedorenko. 2023. Pragmeqa. Retrieved from <https://osf.io/dpge6>.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Herbert Paul Grice. 1989. Retrospective epilogue. *Studies in the Way of Words*, 339:386.
- Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. **Do androids laugh at electric sheep? humor “understanding” benchmarks from the new yorker caption contest**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.
- J.L.B. Hirschberg. 1985. *A Theory of Scalar Implicature*. University of Pennsylvania.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023a. **A fine-grained comparison of pragmatic language understanding in humans and language models**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.
- Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023b. Expectations over unspoken alternatives predict pragmatic inferences. *arXiv preprint arXiv:2304.04758*.
- Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.
- Zae Myung Kim, David E Taylor, and Dongyeop Kang. 2023. “is the pope catholic?” applying chain-of-thought reasoning to understanding conversational implicatures. *arXiv preprint arXiv:2305.13826*.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. **The NarrativeQA reading comprehension challenge**. *Transactions of the Association for Computational Linguistics*, 6:317–328.
- Stephen C Levinson. 1995. Interactional biases in human thinking. In *Social intelligence and interaction*, pages 221–260. Cambridge University Press.
- Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. **Crosslingual generalization through multitask finetuning**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

- Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736.
- Mako Okanda, Kosuke Asada, Yusuke Moriguchi, and Shoji Itakura. 2015. Understanding violations of gricean maxims in preschoolers and adults. *Frontiers in psychology*, 6:901.
- Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv preprint arXiv:2109.12951*.
- Francesca Panzeri and Francesca Foppolo. 2021. Children’s and adults’ sensitivity to gricean maxims and to the maximize presupposition principle. *Frontiers in Psychology*, 12:624628.
- Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. 2023. Does chatgpt resemble humans in processing implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pages 25–34.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Paula Rubio-Fernandez. 2019. Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity. *Cognitive science*, 43(11):e12797.
- Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. **Social IQa: Commonsense reasoning about social interactions**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.
- John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. 1980. *Speech act theory and pragmatics*, volume 10. Springer.
- Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. **Neural responding machine for short-text conversation**. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. **DREAM: A challenge data set and models for dialogue-based reading comprehension**. *Transactions of the Association for Computational Linguistics*, 7:217–231.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.
- Deirdre Wilson and Dan Sperber. 2006. Relevance theory. *The handbook of pragmatics*, pages 606–632.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. **Transformers: State-of-the-art natural language processing**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. **Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.
- Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. **Modeling multi-turn conversation with deep utterance aggregation**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- 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.

A Hyperparameter setting

Parameter	Value
max_new_tokens	50
temperature	0.9
top_k	3
top_p	0.1
repetition_penalty	1.0
num_return_sequence	1
do_sample	True

Table 9: Parameter setting for open-source models in Experiment 1

Parameter	Value
max_new_tokens	300
temperature	0.9
top_k	0
top_p	0.9
repetition_penalty	1.0
num_return_sequence	1
do_sample	True

Table 10: Parameter setting for open-source models in Experiment 2