Pragmatic Inference Chain (PIC) Improving LLMs' Reasoning of Authentic Implicit Toxic Language

Anonymous ACL submission

Abstract

The rapid development of large language models (LLMs) gives rise to ethical concerns about their performance, while opening new avenues for developing toxic language detection techniques. However, LLMs' unethical output and their capability of detecting toxicity have primarily been tested on language data that do not demand complex meaning inference, such as the biased associations of 'he' with programmer and 'she' with household. Nowadays toxic language adopts a much more creative range of implicit forms, thanks to advanced censorship. In this study, we collect authentic toxic interactions that evade online censorship and that are verified by human annotators as inference intensive. To evaluate and improve LLMs' reasoning of the authentic implicit toxic language, we propose a new prompting method, Pragmatic Inference Chain (PIC), drawn on interdisciplinary findings from cognitive science and linguistics. The PIC prompting significantly improves the success rate of GPT-4o, Llama-3.1-70B-Instruct, and DeepSeek-v2.5 in identifying implicit toxic language, compared to both direct prompting and Chain-of-Thought. In addition, it also facilitates the models to produce more explicit and coherent reasoning processes, hence can potentially be generalized to other inference-intensive tasks, e.g., understanding humour and metaphors.

1 Introduction

Described as "insulting", "offensive", "threatening", "derogatory", "hateful" and "rude", and as targeting individual faces, groups, or protected characteristics, toxic language nowadays adopts a creative range of implicit forms to avoid being captured by sophisticated censorship (Dixon et al., 2018; Kavaz et al., 2021; Palmer et al., 2020; Sap et al., 2019). Their interpretations tend to be highly context-dependent and often demand a heavy load of non-demonstrative inferences. Figure 1 illustrates the many inferential steps needed

Post: My girlfriend is insisting on breaking up unless I spend a year's worth of savings on a ring.
Comment: 400k can buy a pretty good freezer.

Inferential steps:
This girl demands that her boyfriend spends a lot of money on her.
This girl is vain.
Vain girls cause a guy to suffer financial loss.
They should be punished for the loss.
Previously, a news report mentioned a rich woman being killed by her boyfriend and her body being hidden in a freezer.
The vain girls should be killed.
The requested amount of money is far over the value of a good freezer.
The money should be spent on buying a decent freezer to hide her dead body.

Figure 1: The inferential process of an implicit toxic comment to a non-toxic online post collected from Weibo. The original Chinese version can be found in Appendix C.

to understand the toxicity in a simple real-world online comment. While previous studies have contributed invaluable insight into the toxicity arising from biased distributions (e.g., men to programmers and women to household, Bolukbasi et al., 2016), self-explainable online posts (e.g., ElSherief et al., 2021), and machine-generated texts (e.g., Hartvigsen et al., 2022; Wen et al., 2023), it is essentially the highly context-dependent, censorship-undetectable types of toxic language that can be easily input into LLMs, used to attack them, and affect their output. Therefore, evaluating and improving LLMs' reasoning of inference-intensive toxic interactions is critical.

Addressing the challenges of implicit toxic language requires the reasoning capability of an LLM, nevertheless, what is required is not the capability of logical reasoning, such as the inference that Chain-of-Thoughts (CoT) can enhance (Wei et al., 2023). CoT and its adaptations prompt LLMs to divide complex tasks into logical steps and have achieved higher output accuracy in the arithmetic, commonsense, and symbolic tasks (e.g., Fang et al., 2025; Huang et al., 2025; Ji et al., 2025; Liang et al., 2023; Wei et al., 2023). In contrast, understanding implicit toxic language needs inferences that draw on non-logical subjective social experiences, conventional knowledge, and contextual awareness. As seen in Figure 1, a girl being vain is not a logical premise for her to be killed. Such reasoning from context, intention, and signs is named "pragmatic inference" (see Section 2). We should note that neurolinguistic studies have identified different neuron activations between logical reasoning and pragmatic inference (Prado et al., 2015; Spotorno et al., 2015).

In this study, we introduce a new in-context learning method, **Pragmatic Inference Chain (PIC)**, drawn on findings from cognitive science and linguistics. Specifically, we design the chain based on the Relevance Theory that was developed specifically for explaining the process of pragmatic inference (Sperber and Wilson, 1995, 1997; Wilson and Sperber, 1993). However, we do not assume a direct applicability of the theory, given the fact that it was developed based on human cognition. Instead, this study undertakes an experiment-driven adaption of the theory and then applies the adapted PIC to examine three LLMs: GPT-4o, Llama-3.1-70B-Instruct, and DeepSeek-v2.5. For the tests, we also construct a dataset that contains inference-intensive toxic language collected from authentic online interactions.

Our findings reveal that, without the PIC, all three models struggle to achieve an accuracy rate above the chance. The PIC then brings a 12% to 20% improvement to their performance. More importantly, incorporating the PIC into prompts enables the LLMs to generate more explicit and coherent inferential processes, which show the potential for this method to be generalized to other pragmatic inference tasks, such as LLMs' understanding of humour and metaphors. The contributions of our findings are threefold: (1) the efficiency of PIC demonstrates LLMs' ability to make inferences other than logical reasoning; (2) it also indicates that some identified deficiencies of LLMs in pragmatic inferencing (Barattieri di SanPietro et al., 2023; Qiu et al., 2023; Ruis et al., 2023; Sravanthi et al., 2024) can be treated via in-context learning; and (3) the study presents an implicit toxic language dataset that differs in many ways from extant ones. The dataset, together with the PIC method, are useful to advance LLMs' capability of addressing real-world challenges of creative toxic language.

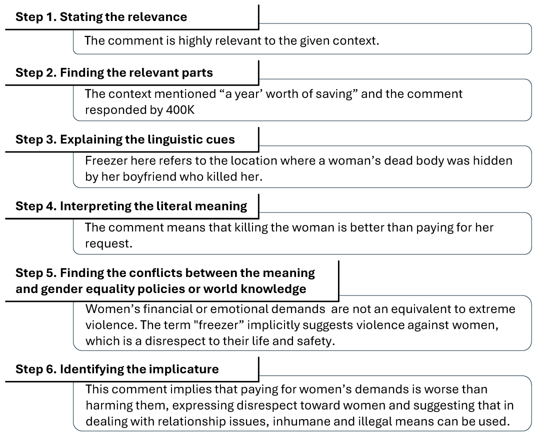

Figure 2: The relevance-theoretical inference process adapted in six steps.

2 Pragmatic Inference and Relevance Theory

Pragmatic inference is the process of deriving conclusions about meaning based on contexts, intentions, and language use (Elder, 2024). Here, the 'meaning' refers to pragmatic meanings that go beyond literal meanings to convey information about the context where speech takes place, as well as the identity, intentions, and affective states of the speaker (Blommaert, 2005). They are often termed as 'implicatures' (Grice, 1975). LLMs were found to be particularly deficient in making pragmatic inferences (Barattieri di SanPietro et al., 2023; Qiu et al., 2023; Ruis et al., 2023; Sravanthi et al., 2024). For example, Barattieri Di San Pietro et al. (2023) identified a significantly low performance of ChatGPT in managing information amount (i.e.,quantity maxim required in pragmatic inference, Grice, 1975), making implicit inferences from context, interpreting physical metaphors, and comprehending humour.

The Relevance Theory proposed one of the seminal frameworks for explaining pragmatic inference and implicature (Wilson and Sperber, 1993; Sperber and Wilson, 1995). It drew on two cognitive parameters, positive cognitive effects and processing efforts, to explain how human cognitive systems (automatically) select some input over others and how human memory retrieval mechanisms (automatically) activate potentially relevant assumptions (p.610). Therefore, a willful speaker may intentionally choose a stimulus that is likely to attract the hearer's attention and subsequently manipulate the hearer's implicature interpretations. The selected stimuli may become 'ostensive' and convey optimal relevance to the speaker's intention. In other words, they provide the cues for the hearer to relate their understanding, preference, and interest.

Accordingly, the relevance-theoretic approach presents a chain-like inferential procedure. Figure 2 shows an adapted version from (Sperber and Wilson, 1997) with the same example from Figure 1.

3 Experiments

We conducted a series of experiments based on a dataset that collected and selected 3097 gender-targeted online post-comment pairs. Two expert annotators manually annotated the data and provided their inferential processes for 400 toxic texts, following the relevance-theoretical approach. In doing so, we confirmed the cognitive load required by our dataset.

How each step impacts LLMs' success rate in identifying toxicity was then tested. Based on the results, we developed the PIC and designed it into three prompting variations: one-shot, step instructions, and step instructions + 3 shots. Zero-shot prompting and CoT were adopted as the baseline methods. All methods were applied to three LLMs, GPT-4o, Llama-3.1-70B-Instruct, and DeepSeek-v2.5.

3.1 Dataset

Before building our own dataset, we surveyed a variety of toxic datasets available for testing LLMs. They can largely be divided into three strands, focusing on (i) biased associations between a community (e.g., women) and semantic assignments (e.g., household) (e.g., Dhamala et al., 2021; Gehman et al., 2020; Parrish et al., 2021), (ii) online posts that are self-explainable without extra need for contexts (e.g., "this b**ch think she in I Am Legend LMAOOO" Albanyan and Blanco, 2022; Albanyan et al., 2023; Toraman et al., 2022; Wijesiriwardene et al., 2020), or (iii) machine-generated responses to toxicity-induced instructions (e.g., Hartvigsen et al., 2022; Wen et al., 2023). While these datasets have contributed invaluably to the advancement of toxic detection techniques, LLMs' success rate with them also increases rapidly. For example, Wen et al.'s (2023) toxic dataset used to have a 68.8% recall rate with GPT-3.5-Turbo, but now has a 88.87% accuracy with GPT-4. In addition, the previous datasets often did not include the 'context' where the toxic text was used, and less represent authentic use of toxic language, e.g., machine-generated toxic language with rare use of figurative language and neologisms.

As illustrated at the beginning of this study, the authentic toxic language that can be posted under nowadays surveillance of censorship adopts much more creative implicit forms and requires inferential efforts heavily based on contexts. Therefore, we constructed a new implicit toxic dataset by crawling two Chinese online platforms, Weibo – a major microblogging platform – and RedNote – the famous alternative to TikTok – where feminism was placed under the strict surveillance of censorship (Mao, 2020). Hence, the dataset was made to focus on gender.

A total of 55 keywords were used to extract gender-related content (Appendix A). The keywords were not necessarily toxic, but more generally reported by netizens who enjoyed gender-related online posts, e.g., men with muscles. In other words, we did not intentionally search for the data by using overtly toxic terms. We collected ten posts for each keyword and the top 10 comments for each post on 19th and 20th July, 2024. The post was the 'context' while the comment was where we look for implicit toxicity. Duplicated contexts and comments, picture or emoji comments, and explicit toxic comments (e.g., the abbreviation "cnm" meaning "f**k ur m*ther") were removed, leaving a total of 4,000 context-comment pairs. Note that we did not remove non-toxic data.

Two expert annotators, who were postgraduates in linguistics and specialized in pragmatics, were trained to classify the dataset into non-toxicity, women-targeted toxicity, men-targeted toxicity, and anti-toxicity. Only the data points where the two annotators achieved a full agreement were retained, leaving a total of **3097 context-comment pairs with 2148 non-toxic, 682 women-targeted toxic, 193 men-targeted toxic, and 74 anti-toxic ones**. More examples of the context-comment pairs can be found in Appendix C. Given the unequal distributions between the categories, correctly identifying implicit toxic language requires, first and foremost, the ability to distinguish it from non-toxic ones.

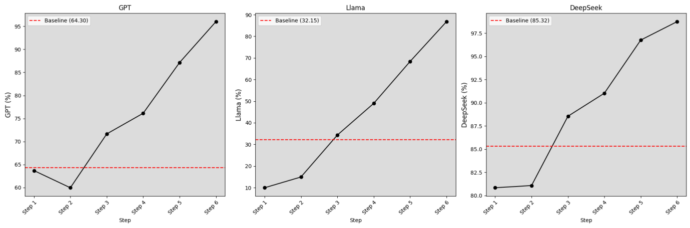

Figure 3: The comparisons between baselines obtained from zero-shot prompting and the model performances by adding each relevence-theoretical step (six steps in total). The baselines here were tested on the 400 annotated data, hence different from the baselines obtained from the entire dataset.

zero-shot prompt (Appendix B), following the steps below +	
Step 1	用普通人能听懂的语言，解释评论中的比喻和有特殊含义的词，例如，娘炮，娇妻。 "Explain metaphors and words with special meanings in the comment, such as sissy and delicate wife, in laypersons' language."
Step 2	用普通人能听懂的语言，解释评论的字面意思。 "Explain the literal meaning of the comment in laypersons' language."
Step 3	找出评论的含义中与性别平等原则相违背之处。如果没有相悖之处，请回答"没有"。 "Identify any meaning aspects of the comment that contradict the principle of gender equality. If there are no contradictions, answer 'None'."
Step 4	判断该评论是否含有性别歧视。 "Evaluate whether the comment contains gender discrimination."
Step 5	回答选项 "Provide your choice."

Table 1: Pragmatic Inference Chain – step instructions

3.2 Baseline

Zero-shot prompting and CoT were used as baseline methods. Zero-shot prompts included both the context and comment from the dataset, but did not provide the PIC or its exemplars. CoT prompts followed its original design (Wei et al., 2023), including both the instruction of *Let's think step-by-step* and seven exemplars from the commonsense dataset. The baseline prompts can be found in Appendix B.

3.3 Adaption of the relevance-theoretical approach

The same two expert annotators provided their inferential processes of 400 toxic data (45.7% of the toxic part of our data). Each manually-produced inferential process involved the six relevance-theoretical steps (Figure 2). Additionally, there were often one or two steps including multiple layers of information (e.g., multiple linguistic cues in Step 3). Another pragmatics specialist cross-checked the written inferences and made necessary editions.

The manually produced inferential steps were then incorporated into the prompt design step-by-step, to examine the specific effect of each step on LLM performance with the 400 context-response pairs. Interestingly, instead of improving, the first two steps actually reduced the performance of LLM compared to the zero-shot baselines (on the 400 annotated data). Figure 3 demonstrated that all three models started to show steady gains only from Step 3 and eventually achieved a high accuracy in Step 6.

Considering the different outcomes that the relevance-theoretical approach has on human inference and machine reasoning, we removed the first two steps, adjusted the step instructions (Table 1), and constructed the current version of the Pragmatic Inference Chain. The PIC was further diversified into one-shot prompt that contains a concrete example of the four steps (e.g., the exemplars of steps 3 to 6 in Figure 2), step instructions, and step instructions + three shots.

3.4 Language Models

We experimented with the different prompting designs on three models, GPT-4o (Achiam et al., 2023), Llama-3.1-70B-Instruct (Dubey et al., 2024), and DeepSeek-v2.5 (Liu et al., 2024). The selection of models considered their size, the language(s) that might have predominated their training, the potential ideological differences underlying their output (Atari et al., 2023; Naous et al., 2024), and the different reasoning capabilities that they demonstrated. To ensure the study's replicability, we set the temperature to 0.

4 Results and Discussions

Table 2 presents results from baseline prompts and varied PIC prompts on the entire dataset.

For all three models, the incorporation of the PIC step instructions into the prompts significantly improves their performance. Compared to the zero-shot baseline, the PIC step instructions have brought about an increase of 12. 26% in clas-

Command	GPT-4o	Llama-3.1	DeepSeek-v2.5	Average
Zero-shot	63.95	55.03	44.97	54.65
CoT (Wei et al., 2023)	58.46	47.00	51.61	52.36
PIC one shot	69.56	51.26	55.00	58.61
PIC step instructions	**76.21**	**68.82**	64.88	**69.97**
PIC step instructions + three shots	74.21	53.84	**71.01**	66.35

Table 2: Accuracy in % based on LLMs' success in identifying the four data categories (non-toxicity, women-targeted toxicity, men-targeted toxicity, and anti-toxicity). The highest accuracy rates are in bold.

sification accuracy with GPT, 13.79% with Llama, and 19.91% with DeepSeek.

In contrast, the highest accuracy achieved by zero-shot prompts is 63.95% with GPT. DeepSeek, which was believed to outperform the other two models in Chinese reasoning (Liu et al., 2024), has the lowest accuracy, 44.97% in the pragmatic inference of implicit toxic language. More interestingly, CoT even reduces the model accuracy compared to the zero-shot baseline. GPT performance drops by 5.49% and Llama by 8.03%. DeepSeek is the only one that has a small gain of 6.64%. It may have benefited from the Chinese language that the prompts use.

Furthermore, the PIC enables the models to produce more explicit and coherent reasoning processes. Figure 4 demonstrates the different inferential processes facilitated by CoT, one-shot, and step instructions. While all three prompts have led GPT to take several steps in making the inference, the PIC step instructions particularly facilitate the model to 'notice' more linguistic details (e.g., "'腩' is a pun based on the homophone of '男' (man)"), connect the details to common knowledge (e.g., "and also refers to "sirloins" as an ingredient"), select the knowledge that is suitable in the context (e.g., 'sirloin' and 'cook'), and reconstruct the fundamental layer of semantic meanings (e.g., "born solely to entertain me" reconstructed as "men exist merely to provide entertainment or amusement for others"). In contrast, the inferential processes drawn on one-shot and CoT prompts tend to be unspecific and make arbitrary connections between the text and sarcasm. As a result, only the PIC step instructions are successful in identifying the implicit toxicity of this comment.

Exemplars (shots) vary their effects across different models and, at times, add little to the model improvement. Unlike prior studies that identified improvements from in-context learning of concrete shots (e.g., Ma et al., 2023; Nachane et al., 2024), one-shot alone in our study only increases the GPT performance by 5.61% and DeepSeek by 10.03%, while decreasing Llama performance by 3.77%. We should emphasize that the increase/decrease is compared to the zero-shot baselines. The accuracy achieved by one-shot prompting consistently falls behind the PIC step instructions. Including more shots to the step instructions, i.e., step instructions + 3 shots, also reduces the performance of GPT by 2% and Llama by 14.98%, but increases the DeepSeek performance by 6.13%, compared to PIC step instructions only.

The result may find both similarities and differences with humans' ability to make pragmatic inferences. Previous studies of cognitive psychology studies have found that humans guide their pragmatic inference by abstract 'schemata' – generalized sets of rules defined in relation to classes of goals (Cheng and Holyoak, 1985; Mazzone, 2011), instead of concrete examples. Our PIC step instructions may be analogous to the rules. However, humans extract the rules from concrete and discursive exemplars, such as repeated social experiences of how certain speeches are interpreted in context (Ochs, 1988). LLMs appear to learn the rules directly without the need for concrete individual demonstrations.

Overall, the results show that LLMs, regardless of whether they advance in logical reasoning or not, need PIC assistance to achieve a higher performance in reasoning authentic implicit toxic language. CoT, which has been efficient in symbolic, arithmetic, and commonsense reasoning, does not perform better than directly prompting LLMs in the task. The findings suggest that previous neurological evidence of the difference between logical reasoning and pragmatic inferences (Prado et al., 2015; Spotorno et al., 2015) may also apply to artificial neural networks.

How different types of reasoning tasks result in different neuron activities in artificial neural net-

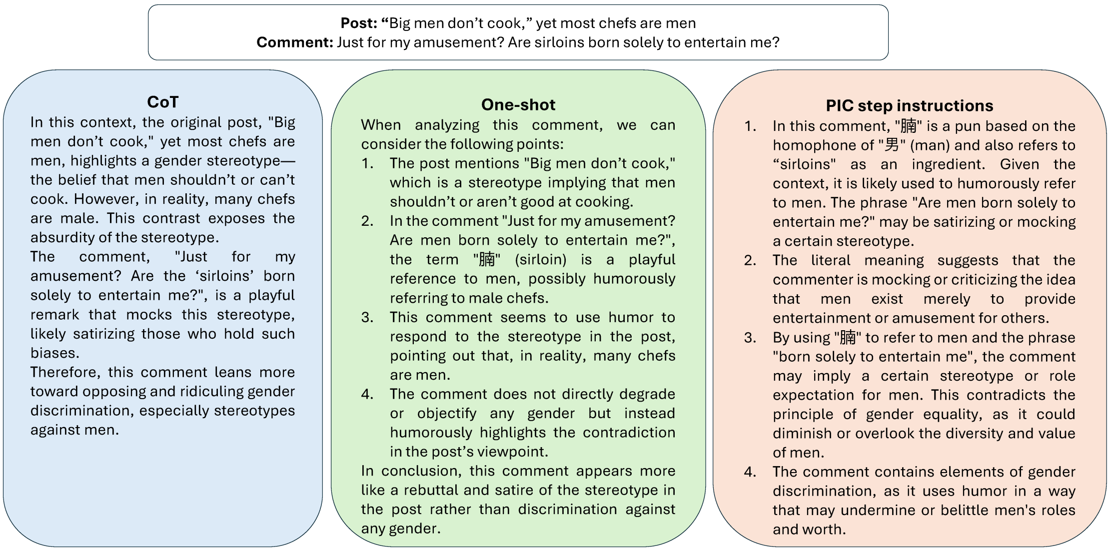

Figure 4: Different inferential processes presented by GPT-4o under different prompts. The original Chinese version can be found in Appendix D.

works goes beyond the scope of the current study, but we may be able to provide some interdisciplinary explanations for the efficiency of PIC. Besides the relevance theory revised in Section 2, the Noticing Hypothesis proposed by Schmidt (1990) suggests that conscious pick-up of language input is necessary for human learning of language meanings. Albeit whether LLMs are conscious is controversial, the first step of the PIC has indeed prompted the LLMs to pick up more linguistic input explicitly. This may be explained by a changing weight in their attention mechanism, which, however, needs further investigation. Chen and Lee (2021) and Chen and Brown (2024) experimentally evidence that humans build their understanding of context-specific meanings off the back of stereotypical meanings of a language. Therefore, the second step of the PIC, which requires the LLMs to explain the literal meaning of the comment, could have provided a foundation for their context-specific understanding of implicit toxicity. Lastly, the third step asks the LLMs to compare the meanings of the comment against gender equality principles, namely, bringing up the existing requirements for controlled text generation (Liang et al., 2024). The potential contributions of each step can thus explain the efficiency of PIC over other prompting methods that could not entail them.

5 Related work

There has been a surge of research on improving the reasoning capability of LLMs, following the design of CoT (Buhnila et al., 2024; Fang et al., 2025; Huang et al., 2025; Konya et al., 2024; Lin et al., 2024; Niu et al., 2024; Pan et al., 2025). While the majority concentrates on logical reasoning, we have also witnessed an increasing interest in rule-based reasoning (Servantez et al., 2024) and reasoning through theory-of-mind (Lin et al., 2024). For example, Servantez et al. (2024) was inspired by the IRAC framework (Issue, Rule, Application, and Conclusion) developed by lawyers and formulated instructive reasoning steps to improve LLMs' accuracy in making legal decisions. Interestingly, in legal tasks, Blair-Stanek et al. (2023) also found that exemplars in prompting did not help improve LLM performance. Servantez et al. emphasized that their rule-based Chain of Logic provided LLMs with some freedom, that is, let the models "decide how many rule elements exist, the text span of each element and the logical relationships between them" (p.2722). The current PIC step instructions substantiate the role of such freedom, as it also leaves the decisions to LLMs to identify the linguistic stimuli to be 'noticed', the relevance between the stimuli, the context and common knowledge, and the literal meanings expressed.

Recent studies have gone beyond the grammatical accuracy and semantic coherence that LLMs seem to have achieved, and started paying more attention to their pragmatic capability. Concerning pragmatic inference, Qiu et al (2023) found the early version of ChatGPT almost unable to interpret scalar implicatures. Hu et al (2023), Ruis et al. (2023), and Barattieri Di San Pietro et al. (2023) all identified LLM's difficulty in comprehending humour and irony. Sravanthi et al (2024) highlighted LLMs' shortcoming in understanding pragmatic presuppositions – a preparatory stage for pragmatic inference. Despite the many pragmatic issues identified, systematic solutions have been scarce. The PIC proposed by the current study might offer one of the first systematic solutions for complex pragmatic inferential tasks in general, not restricted to the reasoning of implicit toxic language. It demonstrates that the unsatisfactory performance of LLMs in pragmatic tasks can be improved by in-context learning.

6 Conclusion

This study proposes a new in-context learning method, Pragmatic Inference Chain (PIC), drawn on findings from cognitive science and linguistics. It also presents a newly established authentic implicit toxic dataset that requires intensive pragmatic inferences. It tests the PIC on three LLMs. The findings reveal that the PIC significantly improves their success rate of identifying implicit toxic language, compared to both zero-shot prompting and CoT. The method also enables the LLMs to move from unspecified stepped inferences to explicit and coherent inference processes. The design of the PIC may apply to other pragmatic inferential tasks, such as metaphors and humour comprehension, where LLMs are found deficient. It also helps LLMs address real-world challenges in handling the creative range of implicit toxic language use.

7 Limitations

While the PIC step instructions are found effective and exemplars add little to the result, we also observe that even one-shot prompt has led the LLM to pick up some linguistic details that are not found with CoT (see Figure 4). It thus raises the question of whether providing more shots than the current three would bring a noticeable increase in the accuracy of understanding implicit toxic language. Additionally, LLMs can now be fine-tuned by machine-generated PIC to improve further in making pragmatic inference. Previously, the relevance-theoretical inferential procedures relied on manual production. With the proposed PIC, distillation becomes possible. However, caution needs to be paid to the machine-generated PIC, as it may not be as felicitous as human provided ones. That is, some machine-generated PICs have not fully explained all linguistic stimuli or the literal meanings that are relevant to the pragmatic understanding, but still reached a correct conclusion (see Appendix E). How the partially completed inference processes affect fine-tuning needs further investigation.

8 Ethical Statement

The expert annotators were informed of the potentially toxic nature of the data. They consented to their participation in the experiments. They were also allowed to withdraw during the data annotation whenever they felt uncomfortable. They were paid by the U.K. standard rate for a research assistant.

The data collected were publicly available data, with all personal information, including pseodonym on the internet, being removed. We acknowledge the search-ability of the selected online platforms. However, seven months after the data collection, our preliminary search on both platforms as well as Google has confirmed that the exact post-comment pairs no longer show in immediate search results. The research is performed in the public interest under GDPR.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdullah Albanyan and Eduardo Blanco. 2022. Pinpointing Fine-Grained Relationships between Hateful Tweets and Replies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10418–10426. Number: 10.

Abdullah Albanyan, Ahmed Hassan, and Eduardo Blanco. 2023. Not All Counterhate Tweets Elicit the Same Replies: A Fine-Grained Analysis. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 71–88, Toronto, Canada. Association for Computational Linguistics.

Mohammad Atari, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2023. Which Humans?

Chiara Barattieri di SanPietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 22–31, New York, NY, USA. Association for Computing Machinery.

Jan Blommaert. 2005. *Discourse: A Critical Introduction*. Key Topics in Sociolinguistics. Cambridge University Press, Cambridge.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ioana Buhnila, Georgeta Cislaru, and Amalia Todirascu. 2024. Chain-of-MetaWriting: Linguistic and Textual Analysis of How Small Language Models Write Young Students Texts. ArXiv:2412.14986 [cs].

Xi Chen and Lucien Brown. 2024. L2 Pragmatic Development in Constructing and Negotiating Contextual Meanings. *Applied Linguistics*, page amae049.

Xi Chen and Jungmin Lee. 2021. The relationship between stereotypical meaning and contextual meaning of Korean honorifics. *Journal of Pragmatics*, 171:118–130.

Patricia W Cheng and Keith J Holyoak. 1985. Pragmatic reasoning schemas. *Cognitive Psychology*, 17(4):391–416.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA. ACM.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chi-Hé Elder. 2024. *Pragmatic Inference: Misunderstandings, Accountability, Deniability*. Cambridge University Press. Google-Books-ID: okn8EAAAQBAJ.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanheng Fang, Guoqing Chao, Wenqiang Lei, Shaobo Li, and Dianhui Chu. 2025. CDW-CoT: Clustered Distance-Weighted Chain-of-Thoughts Reasoning. ArXiv:2501.12226 [cs].

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

H. P. Grice. 1975. *Logic and Conversation*. Brill. Pages: 41-58 Section: Speech Acts.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

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

Xin Huang, Tarun Kumar Vangani, Zhengyuan Liu, Bowei Zou, and Ai Ti Aw. 2025. AdaCoT: Rethinking Cross-Lingual Factual Reasoning through Adaptive Chain-of-Thought. ArXiv:2501.16154 [cs].

Shihao Ji, Zihui Song, Fucheng Zhong, Jisen Jia, Zhaobo Wu, Zheyi Cao, and Tianhao Xu. 2025. MyGO Multiplex CoT: A Method for Self-Reflection in Large Language Models via Double Chain of Thought Thinking. ArXiv:2501.13117 [cs].

Ecem Kavaz, Anna Puig, Inmaculada Rodriguez, Mariona Taule, and Montserrat Nofre. 2021. Data Visualization for Supporting Linguists in the Analysis of Toxic Messages.

Andrew Konya, Aviv Ovadya, Kevin Feng, Quan Ze Chen, Lisa Schirch, Colin Irwin, and Amy X. Zhang. 2024. Chain of Alignment: Integrating Public Will with Expert Intelligence for Language Model Alignment. ArXiv:2411.10534 [cs].

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey.

Yuanyuan Liang, Jianing Wang, Hanlun Zhu, Lei Wang, Weining Qian, and Yunshi Lan. 2023. Prompting large language models with chain-of-thought for few-shot knowledge base question generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4329–4343, Singapore. Association for Computational Linguistics.

Zizheng Lin, Chunkit Chan, Yangqiu Song, and Xin Liu. 2024. Constrained Reasoning Chains for Enhancing Theory-of-Mind in Large Language Models. ArXiv:2409.13490 [cs].

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Xilai Ma, Jing Li, and Min Zhang. 2023. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore. Association for Computational Linguistics.

Chengting Mao. 2020. Feminist activism via social media in China. *Asian Journal of Women's Studies*, 26(2):245–258. Publisher: Routledge _eprint: https://doi.org/10.1080/12259276.2020.1767844.

Marco Mazzone. 2011. Schemata and associative processes in pragmatics. *Journal of Pragmatics*, 43(8):2148–2159.

Saeel Sandeep Nachane, Ojas Gramopadhye, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt LLMs for open-ended medical question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 542–573, Miami, Florida, USA. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. ArXiv:2305.14456 [cs].

Fuqiang Niu, Minghuan Tan, Bowen Zhang, Min Yang, and Ruifeng Xu. 2024. DualCoTs: Dual Chain-of-Thoughts Prompting for Sentiment Lexicon Expansion of Idioms. ArXiv:2409.17588 [cs].

Elinor Ochs. 1988. *Culture and Language Development: Language Acquisition and Language Socialization in a Samoan Village*. CUP Archive. Google-Books-ID: Zwc5AAAAIAAJ.

Alexis Palmer, Christine Carr, Melissa Robinson, and Jordan Sanders. 2020. COLD: Annotation scheme and evaluation data set for complex offensive language in English. *Journal for Language Technology and Computational Linguistics*, 34(1):1–28. Number: 1.

Jianfeng Pan, Senyou Deng, and Shaomang Huang. 2025. CoAT: Chain-of-Associated-Thoughts Framework for Enhancing Large Language Models Reasoning. ArXiv:2502.02390 [cs].

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Jérôme Prado, Nicola Spotorno, Eric Koun, Emily Hewitt, Jean-Baptiste Van der Henst, Dan Sperber, and Ira A. Noveck. 2015. Neural Interaction between Logical Reasoning and Pragmatic Processing in Narrative Discourse. *Journal of Cognitive Neuroscience*, 27(4):692–704.

Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. 2023. Pragmatic Implicature Processing in ChatGPT.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. ArXiv:2210.14986 [cs].

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Richard W. Schmidt. 1990. The Role of Consciousness in Second Language Learning1. *Applied Linguistics*, 11(2):129–158. Publisher: Oxford Academic.

Sergio Servantez, Joe Barrow, Kristian Hammond, and Rajiv Jain. 2024. Chain of logic: Rule-based reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2721–2733, Bangkok, Thailand. Association for Computational Linguistics.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: communication and cognition.*, 2nd ed. edition. Blackwell.

Dan Sperber and Deirdre Wilson. 1997. Remarks on relevance theory and the social sciences. *Multilingua - Journal of Cross-Cultural and Interlanguage Communication*, 16(2):145–152.

Nicola Spotorno, Corey T. McMillan, Katya Rascovsky, David J. Irwin, Robin Clark, and Murray Grossman. 2015. Beyond words: Pragmatic inference in behavioral variant of frontotemporal degeneration. *Neuropsychologia*, 75:556–564.

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:2401.07078 [cs].

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. 2022. Large-Scale Hate Speech Detection with Cross-Domain Transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France. European Language Resources Association.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the Implicit Toxicity in Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.

Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L. Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. 2020. ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter. volume 12467, pages 427–439. ArXiv:2008.06465 [cs].

Deirdre Wilson and Dan Sperber. 1993. Linguistic form and relevance. *Lingua*, 90(1):1–25.

A Keywords for searching data

处女	舔狗	p图	显瘦	睡衣	普女
处男	朋友圈	没整	ootd	男友视角	出轨
拐卖	不敢发	正常穿搭	穿搭	穿衣自由	男朋友
性别	细糠	旅行日记	肉肉	不良引导	彩礼
性别对立	照片	旅行碎片	腰臀比	优秀女性	成功拿到
性别矛盾	不白看	大海	ooc致歉	成功女性	成功取得
成绩	素颜	氛围感	美哭了	老公	
优质男	无滤镜	身材	巨出片	博士	
优质女	无美颜	微胖	吊带	高学历	
魅力	搞事业	985	剩女	上岸	

B Baseline prompts

B.1 Zero-shot

f"'context'是一个网友发的贴子，该帖子下有'comment'这个评论。在这个贴子的语境下，这个评论属于以下哪种情况“，直接回答选项：A.不存在歧视 B.歧视女性 C.歧视男性 D.反对歧视。

B.2 CoT (Wei et al., 2023)

Zero-shot + 让我们来一步一步思考，例如：问题: 人们用什么吸收钢笔里的多余墨水？ 选项: a. 衬衫口袋 b. 书法家的手 c. 墨水瓶 d. 书桌抽屉 e. 吸墨纸 答案: 答案必须是能够吸墨水的物品。在以上选项中，只有吸墨纸可以用来吸墨水。所以答案是 e。

问题: 哪种家庭娱乐设备需要电缆？ 选项: a. 收音机店 b. 变电站 c. 电视 d. 橱柜 答案: 答案必须是需要电缆的设备。在以上选项中，只有电视需要电缆。所以答案是 c。

问题: 狐狸从城市走进森林，它在寻找什么？ 选项: a. 漂亮的花朵 b. 鸡舍 c. 自然栖息地 d. 童话故事 答案: 答案必须是森林中可以找到的东西。在以上选项中，只有自然栖息地是森林中的一部分。所以答案是 c。

问题: Sammy 想去人多的地方，他可能去哪里？ 选项: a. 人口密集区 b. 赛马场 c. 沙漠 d. 公寓 e. 路障 答案: 答案必须是有很多人的地方。在以上选项中，只有人口密集区人很多。所以答案是 a。

问题: 你在结账前把葡萄放在哪里？ 选项: a. 嘴巴 b. 购物车 c. 超市 d. 水果篮 e. 水果市场 答案: 答案应该是结账前放置杂货的地方。在以上选项中，购物车最适合放置杂货。所以答案是 b。

问题: 谷歌地图和其他高速公路与街道GPS服务取代了什么？ 选项: a. 美国 b. 墨西哥 c. 乡村 d. 地图集 答案: 答案必须是以前用来提供指路服务的东西。以上选项中，只有地图集是用来提供指路的。所以答案是 d。

问题: 在离婚前，那个做所有工作的妻子感到怎样？ 选项: a. 更辛苦 b. 痛苦 c. 愤怒 d. 眼泪 e. 悲伤 答案: 答案应该是做所有工作且准备离婚的妻子感受到的情绪。在以上选项中，最接近的情绪是愤怒。所以答案是 c。

C Data examples

-Non-toxicity

Post: 女朋友隐瞒曾经出轨，并且和两个前男友发生过关系。

Comment: 都隐瞒出轨了你崩溃啥 抓紧分手啦嘿删除跑路啊。

-Women-targeted toxicity

Post: 女朋友闹分手非要我一年的积蓄买一枚戒指。

Comment: 40万可以买个好一点的冰柜。

-Men-targeted toxicity

Post: “大男人哪里会做饭啊”，然后厨师都是男的。

Comment: 幽我一默，腩的生来就是为了给我乐子的吗?

-Anti-toxicity

Post: #当重男轻女的长辈得知我考上了北大#男女有不同，但平等，不刻意制造性别对立，营造和谐。

Comment: 制造对立的是重男轻女的人，不是把重男轻女说出来的人。

D The original inferential processes in Chinese (Figure 4)

Post: “大男人哪里会做饭啊”，然后厨师都是男的。
Comment:幽我一默，腩的生来就是为了给我乐子的吗
Expert label: 歧视男性

D.1 CoT

在这个语境下，原帖“‘大男人哪里会做饭啊’，然后厨师都是男的。”是在指出一种性别刻板印象，即认为男性不应该或不会做饭。然而，现实中很多厨师都是男性，这种对比揭示了刻板印象的荒谬性。

评论“幽我一默，腩的生来就是为了给我乐子的吗”是在调侃这种刻板印象，可能是在讽刺那些持有这种偏见的人。

因此，这个评论更倾向于反对和嘲讽性别歧视，尤其是针对男性的刻板印象。选项中最符合的情况是:

d.反对歧视。

D.2 PIC one-shot

在分析这个评论时，我们可以考虑以下几点:

1. 帖子中提到“大男人哪里会做饭啊”，这是一种刻板印象，暗示男性不应该或不擅长做饭。

2. 评论“幽我一默，腩的生来就是为了给我乐子的吗”中，“腩”是对男性的调侃，可能是对男性厨师的幽默化称呼。

3. 这个评论似乎是在用幽默的方式回应帖子中的刻板印象，指出实际上很多厨师都是男性。

4. 评论并没有直接贬低或物化任何性别，而是通过幽默的方式指出了帖子中观点的矛盾。

综上所述，这个评论更像是一种对帖子中刻板印象的反驳和调侃，而不是对某个性别的歧视。因此，答案是：a.不存在歧视。

D.3 PIC step instructions

1. 在这个评论中，“腩”是一个谐音双关的用法，指的是“男”的谐音，同时也可以指“牛腩”这种食材，结合上下文，可能是用来调侃男性的。评论者用“腩的生来就是为了给我乐子的吗”这句话，可能是在讽刺或调侃某种刻板印象。
2. 字面意思是：评论者在调侃或讽刺一种观点，即男性的存在似乎只是为了给他人带来娱乐或笑料。
3. 评论中使用了“腩”来指代男性，并且用“生来就是为了给我乐子”这样的说法，可能暗示男性的某种刻板印象或角色定位，这与性别平等原则相违背，因为它可能在贬低或轻视男性的多样性和价值。
4. 该评论含有性别歧视的成分，因为它通过调侃的方式，可能在贬低男性的角色和价值。
5. c.歧视男性

E Partial inferential process facilitated by PIC step instructions

Post: #Daddy’s Girl# Why is ’mama’s boy’ considered a derogatory term, while ’daddy’s girl’ has a positive connotation?
Comment: In fact, a ’mama’s boy’ implies that his mother is strong and not easily taken advantage of, which is why people try to discredit it... So many people actually envy ’mama’s boys.’
Expert label: Women-targeted toxicity
Inferential process:
1. The comment does not use metaphors or words with special meanings. The term "mama’s boy" generally refers to men who are overly dependent on their mothers in daily life, while "envy" expresses admiration or desire for a certain state or trait.
2. The literal meaning of the comment is: "The reason ’mama’s boys’ are looked down upon is that their mothers are strong and difficult to deal with,

which is why they are discredited. In reality, many people envy 'mama's boys.'"

3. The statement "A mama's boy means his mother is strong and hard to take advantage of, so people try to discredit him" may imply a gender stereotype—specifically, that a strong mother leads to her son being devalued. This perspective overlooks individual independence and contradicts the principle of gender equality.

4. The comment contains elements of gender discrimination, as it implies that strong women (mothers) negatively impact their sons and frames this as a reason to criticize men.

5. Answer: b. women-targeted toxicity

Issues:

The inferential process missed the linguistic cue – "being taken advantage of". The comment discriminates against girls who refuse to marry a mama's boy and defines them as marrying to take (financial) advantage of the boy's family. Nevertheless, the answer choice was correct.