

# Pragmatic Metacognitive Prompting Improves LLM Performance on Sarcasm Detection

Joshua Lee<sup>†</sup>, Wyatt Fong<sup>†</sup>, Alexander Le<sup>†</sup>, Sur Shah<sup>†</sup>, Kevin Han<sup>†</sup>, Kevin Zhu<sup>†\*</sup>

<sup>†</sup>Algorverse AI Research  
{zhukevin, kevin.x.han}@berkeley.edu

## Abstract

Sarcasm detection is a significant challenge in sentiment analysis due to the nuanced and context-dependent nature of verbiage. We introduce **Pragmatic Metacognitive Prompting (PMP)** to improve the performance of Large Language Models (LLMs) in sarcasm detection, which leverages principles from pragmatics and reflection helping LLMs interpret implied meanings, consider contextual cues, and reflect on discrepancies to identify sarcasm. Using state-of-the-art LLMs such as LLaMA-3-8B, GPT-4o, and Claude 3.5 Sonnet, PMP achieves state-of-the-art performance on GPT-4o on MUStARD and SemEval2018. This study demonstrates that integrating pragmatic reasoning and metacognitive strategies into prompting significantly enhances LLMs' ability to detect sarcasm, offering a promising direction for future research in sentiment analysis.

## 1 Introduction

Within the field of sentiment analysis, various approaches exist to improve emotion classification, from bidirectional transformers to prompt tuning for aspect-based sentiment analysis (Ataei et al., 2020; Ouyang et al., 2015; Devlin et al., 2019; Li et al., 2021; Zadeh et al., 2017; Kanakaraj and Gudeti, 2015). Yet one present limitation sentiment analysis models face is in determining sarcasm (Tan et al., 2023a).

Recent discoveries found that LLMs underperform compared to specially trained transformer encoder models in both sarcasm detection and sentiment analysis. The speculated cause of poor LLM performance is that LLMs are built on logical pipelines, which may contradict sarcasm's non-sequential nature. Regardless, studies believe improving prompting methods is a step towards the solution (Zhang et al., 2024, 2023; Tan et al., 2023b;

Liu et al., 2023; Yao et al., 2024; Wei et al., 2022; Besta et al., 2024; Yao et al., 2023).

This work presents PMP<sup>1</sup> based on Wei et al.'s Metacognitive prompting (MP). PMP is a new approach to improving LLM sarcasm detection. Our approach incorporates linguistic principles to mimic how humans reason through emotionally complex text as well as reflection strategies commonly found in LLM reasoning and planning agents (Shinn et al., 2023). This paper presents a novel prompting approach through the use of pragmatics and reflection to improve sarcasm detection, runs its prompting method on sarcasm benchmarks, and at times exceeds the prompt results of the current state-of-the-art (SoTA) prompt for LLM sarcasm detection.

## 2 Background

### 2.1 Pragmatics

Pragmatics is a field of linguistics that goes beyond the literal meaning of a conversation. It's the social context of a statement that is needed to comprehend the subtleties of human language. (Grice, 1975; Clark, 1996; Horn and Ward, 2004). Various studies in linguistics have been conducted on the pragmatics of sarcasm. One pragmatic theory called Grice's Maxims of Conversation, poses the 4 different factors that a conversation must have to be a meaningful conversation. One study, in the field of pragmatics, analyzed Grice's Maxims. It concluded that if Grice's Maxims were exceeded, like with sarcasm in TV shows, it could be a determining factor as to whether dialogue is sarcastic (Al Anssari and Hadi, 2021).

Our method, PMP, incorporates proposed pragmatic theories on how to detect sarcasm from the field of linguistics into LLM prompting. Our

<sup>1</sup>Our code can be found at: <https://github.com/wyatt-fong/Pragmatic-Metacognitive-Prompting-Improves-LLM-Performance-on-Sarcasm-Detection>

\*Corresponding Author

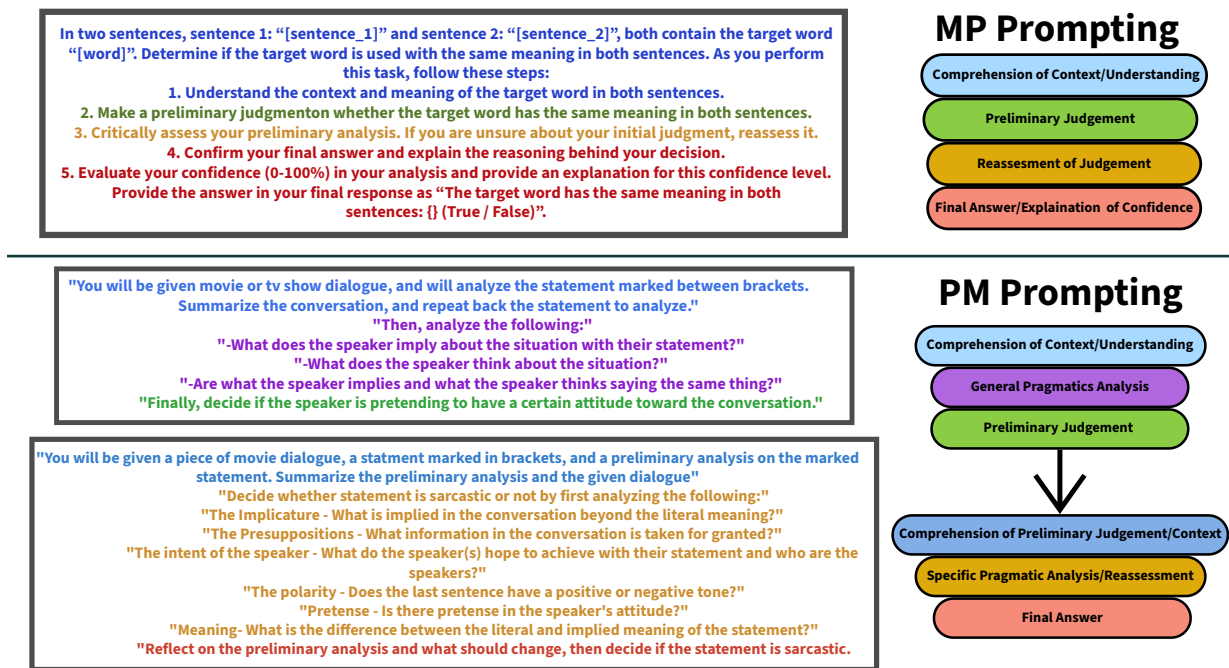


Figure 1: Metacognitive Prompt structure compared to proposed Pragmatic Metacognitive Prompt.

method encourages the LLM to analyze multiple pragmatic theories analyzing sarcasm before reaching its conclusion. A simplified explanation of the theories used in PMP is provided below.

**The Standard Pragmatic Model (TSPM):** Building upon the foundations of the TSPM, Gibbs and Colston, refined the understanding of sarcasm detection. His version of TSPM emphasized a process of contacting literal and non-literal meanings alongside context to determine sarcasm.

**The Pretense Theory of Irony:** When a speaker is ironic or sarcastic, they set up a facade to what they actually believe. (Clark and Gerrig, 1984). For instance, if someone says “Your jacket looks soooo nice” in a sarcastic tone, they are presenting an attitude that they do like the look of your jacket when they actually do not.

**The Echoic Reminder Theory of Verbal Irony:** This method is characterized by positive and neutral statements that ironically reference a past statement. It is often used as a critique of a scenario. An example would be the phrase “What a great idea!” which typically has a positive connotation. However, if it was used to describe a terrible plan it would take a negative connotation, conveying sarcasm. (Kreuz and Glucksberg, 1989).

**Implicature:** Implicatures are implicit inferences drawn automatically from the information provided in a sentence, relying on the shared con-

text between the speaker and listener. For example, if Bob says “Do you want any cake for lunch?” and Joey responds “I don’t want to get fat”, the implied meaning is that Joey is declining the offer. However, inferences like these can sometimes be incorrect. For instance, Joey may have been making a completely unrelated comment. As the interpreters, we may assume that Joey’s statement was relevant to the conversation, which causes us to infer his refusal.

**Presupposition:** The presupposition is that the information is automatically accepted as true in order for a statement to make sense. For example, in the statement “The king of France is bald”, the presupposition is that there exists a King of France. This statement assumes that a king exists in France, even though there might not be one, for the sake of making sense of the statement.

### 3 Method

Our prompting method builds on top of Wei et al.’s Metacognitive prompting (MP). MP consists of prompting an LLM to repeat the given information, create a preliminary analysis, reflect on their preliminary analysis, and then create a final judgment (Wei et al., 2023). See Figure 1 for more details on MP. In our method, PMP, the LLM is encouraged to analyze simplified elements of pragmatic theories in the preliminary analysis and reflection stages.

Model	SemEval 2018		MUSTARD	
	Acc.	Ma-F1	Acc.	Ma-F1
GPT-4o (IO)	64.03	63.17	67.24	65.79
GPT-4o (CoT)	58.92	51.99	58.11	55.76
GPT-4o (ToT)	63.90	63.02	69.00	68.27
GPT-4o (CoC)	70.79	70.60	69.42	68.48
GPT-4o (GoC)	74.03	74.02	70.69	69.91
GPT-4o (BoC)	62.12	61.85	69.42	68.45
GPT-4o (PMP)	<b>86.68</b>	<b>83.18</b>	<b>79.42</b>	<b>77.65</b>
GPT-4o-mini (PMP)	81.88	79.85	65.79	62.29
Claude 3.5 Sonnet (IO)	75.13	75.11	<b>74.78</b>	<b>74.78</b>
Claude 3.5 Sonnet (CoT)	71.56	71.47	73.62	73.53
Claude 3.5 Sonnet (ToT)	68.62	68.61	58.84	54.46
Claude 3.5 Sonnet (CoC)	<b>82.27</b>	<b>82.23</b>	74.20	74.16
Claude 3.5 Sonnet (GoC)	57.33	57.24	52.77	52.67
Claude 3.5 Sonnet (BoC)	65.94	65.50	59.71	56.70
Claude 3.5 Sonnet (PMP)	81.50	76.72	72.60	71.66
LLaMA-3-70B (PMP)	80.86	78.15	72.73	73.06
LLaMA-3-8B (IO)	49.36	44.47	54.64	44.99
LLaMA-3-8B (CoT)	49.36	44.55	54.20	44.86
LLaMA-3-8B (ToT)	50.64	48.63	54.35	50.56
LLaMA-3-8B (CoC)	49.23	44.36	54.93	45.66
LLaMA-3-8B (GoC)	57.33	57.24	52.7	52.67
LLaMA-3-8B (BoC)	65.94	65.50	59.71	56.70
LLaMA-3-8B (ToC)	68.88	68.21	<b>61.26</b>	<b>58.03</b>
LLaMA-3-8B (PMP)	<b>78.21</b>	<b>77.65</b>	53.48	54.69

Table 1: Comparison of PMP with Claude 3.5 Sonnet, GPT4o, GPT4o-mini, LLaMa-3-70B and LLaMA-3-8B to prompting methods. The best results are bolded.

We establish two separate LLM calls, one which analyzes the prompt from the lens of each pragmatic factor: implicature, presuppositions, intent, polarity, pretense, and potential meanings individually, and a second LLM call that reflects on the analysis and outputs a final prediction. A detailed explanation of PMP is provided in Figure 1.

## 4 Experimental Design

### 4.1 Benchmarks

We evaluated our sarcasm detection method on the same benchmarks as (Yao et al., 2024): MUSTARD (Castro et al., 2019), which consists of sarcastic and non-sarcastic comments in TV and movie dialogue paired with context; and SemEval 2018 Task 3 (Van Hee et al., 2018) consisting of sarcastic and non-sarcastic twitter statements.

### 4.2 Models

We tested our method using models also utilized in SarcasmCue (Yao et al., 2024). The models are: GPT-4o, LLaMA 3-8B and Claude 3.5 Sonnet (Anthropic, 2024). Furthermore, we additionally tested on GPT-4o mini (OpenAI, 2023) and LLaMA 3-70B (Touvron et al., 2023).

#### 4.2.1 SarcasmCue

Our method achieves the new SoTA in comparison to SarcasmCue. The method SarcasmCue modifies popular SoTA prompts to analyze a “cue”, which is a coherent language sequence that serves as an indicator towards identifying sarcasm, from either linguistic (rhetorical devices, punctuation), contextual (topic, common knowledge), or emotional (emotional words, emojis) parts of a sentence.

SarcasmCue<sup>2</sup> introduces four sarcasm detection methods; three prompting techniques: Bag of Cues (BoC), Chain of Cues (CoC), and Graph of Cues (GoC), and one that requires explicit model training, Tensor of Cues (ToC). BoC removes sequential bias by treating cues independently. CoC arranges cues in a sequential order to capture the step-by-step reasoning process of sarcasm detection. GoC analyzes the relationships between cues without imposing a fixed sequence. ToC adds encoded indications through explicit training to leverage higher-order interactions among cues. For the exact prompts, please see Appendix A.

## 5 Results

The accuracy and Macro-F1 scores comparing PMP with prompting method baselines are compared in Table 1. The accuracy and Macro-F1 scores comparing PMP with SarcasmCue’s BoC, CoC, GoC, and ToC strategies are reported in Table 1.

### Comparison to Popular Prompting methods:

PMP surpasses popular prompting methods such as Zero Shot, Chain of Thought, and Tree of Thought in both SemEval 2018 Task 3 and MUsTARD. Across both performing well on LLaMA-3-8B and GPT-4o, with the exception of Claude 3.5 Sonnet. Zero-shot prompting still works well with Claude 3.5 Sonnet in 2 benchmarks, aligning with Yao et al.’s results. PMP’s performance with LLaMA-3-70B is significantly higher than with LLaMA-3-8B.

### Comparison to State of the Art (SoTA):

PMP is competitive with and exceeds SarcasmCue’s performance on all datasets with GPT-4o while performing well on LLaMA-3-8B on SemEval 2018. As shown in Table 2, zero-shot Claude 3.5 Sonnet achieves the highest accuracy on the MUsTARD datasets, outperforming it in SarcasmCue and PMP.

**Datasets:** PMP performs the best on SemEval 2018 Task 3, although it falls slightly short of SarcasmCue on Claude. PMP struggles on Sarcasm Corpus V1 the most, with current SoTA and Tree of Thought outperforming it across certain models.

**State of The Art and PMP:** Between PMP and SarcasmCue, neither consistently achieves higher accuracies than the other across all models and datasets, excluding Claude. However, one notable

factor is that for both datasets, GPT-4o utilizing PMP performs best in comparison with all other models and prompting methods. GPT-4o outperforming other LLMs aligns with previous studies such as Zhang et al.’s work, suggesting GPT-4o’s performance is a common factor in sarcasm detection. Another inconsistency is that SarcasmCue underperforms some prompts in SemEval 2018 Task 3 across all models except Claude 3.5 Sonnet, while PMP outperforms prompts in SemEval 2018 Task 3 across all models but underperforms SoTA in Claude 3.5 Sonnet. Analyzing SemEval 2018 as a dataset could help explain these performance patterns.

## 6 Conclusion

Pragmatic Metacognitive Prompting is a novel approach for enhancing sarcasm detection in LLMs. PMP is competitive with or beats the current state-of-the-art methods for sarcasm detection with pre trained LLMs such as GPT4o and LLaMA-3-8B. It introduces various pragmatic theories into the prompt design, fosters a deeper contextual understanding that improves sarcasm identification, and incorporates a human-like reflection step for final verification and sarcasm reasoning. After testing across models like GPT-4o and LLaMA-3-8B, PMP underscores the potential of pragmatic-informed methods to outperform traditional prompting methods and points to a continued focus on linguistic theories to bridge performance gaps in sentiment analysis.

## 7 Limitations

While PMP represents an approach to implementing pragmatic reflection, prompting is only one implementation of pragmatics and reflection in natural language processing. A key limitation to using zero-shot prompting is that PMP does not guarantee high performance in sarcasm detection that deviates from general linguistics norms and in domain-specific contexts. Due to PMP’s reliance on LLM’s pretraining with data, underrepresented cultural or linguistic norms are also not accounted for with prompting. These limitations suggest PMP is a step towards improving sarcasm detection, but does not represent a comprehensive solution.

## References

R. S. Al Anssari and H. A. N. Hadi. 2021. *A pragmatic study of sarcasm in selected tv shows*. *International*

<sup>2</sup>At the time of the writing of this paper, Yao et al. has not published their source code. Therefore, we compare our results with the reported results from their paper.



- Journal of Linguistics, Literature and Translation*, 4(7):148–153.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- T. S. Ataei, S. Javdan, and B. Minaei-Bidgoli. 2020. [Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection](#). In *Proceedings of the Second Workshop on Figurative Language Processing*.
- M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, and T. Hoefler. 2024. [Graph of thoughts: Solving elaborate problems with large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.
- Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. [Reliability-aware dynamic feature composition for name tagging](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1021, Florence, Italy. Association for Computational Linguistics.
- Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.
- Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- Raymond W. Gibbs and Herbert L. Colston. 2007. *Irony in Language and Thought*. Taylor & Francis Group, New York.
- H. Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press.
- N.A. Helal, A. Hassan, N.L. Badr, et al. 2024. [A contextual-based approach for sarcasm detection](#). *Scientific Reports*, 14:15415.
- Laurence R. Horn and Gregory L. Ward. 2004. *The Handbook of Pragmatics*. Wiley Online Library.
- M. Kanakaraj and R. M. R. Guddeti. 2015. [Performance analysis of ensemble methods on twitter sentiment analysis using nlp techniques](#). In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 169–170, Anaheim, CA, USA. IEEE.
- Roger J. Kreuz and Sam Glucksberg. 1989. How to be sarcastic: The echoic reminder theory of irony. *Journal of Experimental Psychology: General*, 118(4):374–386.
- Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, and Zhi Yu. 2021. [Sentiprompt: Sentiment knowledge enhanced prompting for aspect-based sentiment analysis](#). *Preprint*, arXiv:2109.08306.
- Y. Liu, R. Zhang, Y. Fan, and J. Guo. 2023. [Prompt tuning with contradictory intentions for sarcasm recognition](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 328–339.
- OpenAI. 2023. [Gpt-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- X. Ouyang, P. Zhou, C. Li, and L. Liu. 2015. [Sentiment analysis using convolutional neural networks](#). In *Proceedings of the IEEE International Conference on Computer and Information Technology*.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#). *Preprint*, arXiv:2303.11366.
- K. L. Tan, C. P. Lee, and K. M. Lim. 2023a. [A survey of sentiment analysis: Approaches, datasets, and future research](#). *Applied Sciences*, 13(7).
- Y. Y. Tan, C. Chow, J. Kanesan, J. H. Chuah, and Y. Lim. 2023b. [Sentiment analysis and sarcasm detection using deep multi-task learning](#). *Wireless Personal Communications*, 129(3):2213–2237.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Armand Joulin, Edouard Grave, Max Chatalain, and Hervé Jegou. 2023. [Llama: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2018. [Semeval-2018 task 3: Irony detection in english tweets](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.
- J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). arXiv.org.
- Jason Wei, Yi Tay, Shixiang Shane Gu, William Fedus, Xuezhi Wang, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2023. [Metacognitive prompting improves understanding in large language models](#). *Preprint*, arXiv:2308.05342.
- B. Yao, Y. Zhang, Q. Li, and J. Qin. 2024. [Is sarcasm detection a step-by-step reasoning process in large language models?](#) arXiv.org.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). arXiv.org.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. [Tensor fusion network for multimodal sentiment analysis](#). Preprint, arXiv:1707.07250.

W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing. 2023. [Sentiment analysis in the era of large language models: A reality check](#). arXiv.org.

Y. Zhang, C. Zou, Z. Lian, P. Tiwari, and J. Qin. 2024. [Towards evaluating large language models on sarcasm understanding](#). arXiv.org.

## A Prompts

The prompting method we utilized in our approach guides the model through a structured reasoning process before reaching a conclusion. The prompt instructs the model to analyze a statement checking its comprehension of the given information before asking the LLM to generate an accompanying preliminary analysis analyzing basic pragmatic factors. After completing the preliminary analysis, the model then passes its generated analysis to another LLM call, where the model has a chance to reflect and comprehend the preliminary analysis originally generated. It then directs the model to refine the analysis by systematically addressing specific pragmatic aspects, including implicature, presuppositions, speaker intent, polarity, pretense, and the relationship between literal and implied meanings. The wording of our initial prompt varies per dataset to ensure that all information about the benchmark is given for the LLM to generate a proper analysis. An example of this would be with MUStARD’s dataset:

“You will be given movie or TV show dialogue and will analyze the statement marked between brackets. Summarize the conversation, and repeat back the statement to analyze. Then, analyze the following:”

Decide whether the statement is sarcastic or not by first analyzing the following:

1. The Implicature – What is implied in the conversation beyond the literal meaning?
2. The Presuppositions – What information in the conversation is taken for granted?

3. The Intent of the Speaker – What do the speaker(s) hope to achieve with their statement, and who are the speakers?

4. The Polarity – Does the last sentence have a positive or negative tone?

5. Pretense – Is there pretense in the speaker’s attitude?

6. Meaning – What is the difference between the literal and implied meaning of the statement?

Reflect on the preliminary analysis and what should change, then decide if the statement is sarcastic.”

### A.1 Cues

#### A.1.1 Bag of Cues

The Bag of Cues method evaluates sarcasm by treating cues and without order.

**Prompt Example:** “Identify if the given statement is sarcastic based on the presence of the following cues:

Rhetorical devices (e.g., irony, hyperbole, or understatement) Emotional language (e.g., frustration, happiness, or sarcasm-laden phrases) Contextual inconsistencies (e.g., contradictory meanings or unexpected word choices). Does the statement exhibit any of these cues?”

**Example Application:** Input: “Oh, great! Another meeting that could have been an email.” Rhetorical Device: Irony Emotional Cue: Frustration Contextual Cue: Work-related sarcasm Detection: Likely sarcastic

#### A.1.2 Chain of Cues

The Chain of Cues method evaluates sarcasm by analyzing cues sequentially. It simulates logical reasoning to then assess the overall sarcastic nature of a statement.

**Prompt Example:** “Analyze the statement step-by-step: Identify any rhetorical device (e.g., hyperbole, irony). Determine if emotional cues such as frustration or humor are present. Check for contextual markers that may suggest sarcasm. Does the progression or order of these cues indicate sarcasm?”

**Example Application:** Input: “Thanks for breaking the printer. Really helpful.” Rhetorical Device: Irony detected in “Really helpful.” Emotional Cue: Frustration in the context of the statement.

Contextual Marker: Complaints about a broken printer. Detection: Sarcastic

### A.1.3 Graph of Cues

The Graph of Cues method evaluates sarcasm by analyzing the relationships between cues. This method leverages interdependencies between linguistic, emotional, and contextual features.

**Prompt Example:** “Construct a graph where: Nodes represent sarcasm cues (e.g., rhetorical devices, emotional cues, contextual features). Edges represent relationships between these cues (e.g., reinforcement, contrast). Based on the interconnected cues, does the statement appear sarcastic?”

**Example Application:** Input: “Wow, you’re so good at driving (said during a near accident).” Nodes: Rhetorical Device: Sarcastic praise (“so good”). Emotional Cue: Anxiety/frustration. Contextual Cue: Near accident. Edges: Reinforcement between rhetorical device and emotional cue. Contextual cue amplifies sarcasm. Detection: Sarcastic

### A.1.4 Tensor of Cues

The Tensor of Cues method uses a structured, multi-dimensional representation of sarcasm cues to train a model explicitly. This approach captures interactions between cues in a numerical format. The implementation details include implementation details, as well as linguistic, emotional, and contextual cues to be encoded as tensors. The model is then trained to revolve around optimizing the model to learn patterns across these dimensions.

**Example Tensor Encoding:** Input: “Nice job ignoring me all day!” Linguistic Cue: Irony (tensor dimension 1). Emotional Cue: Frustration (tensor dimension 2). Contextual Cue: Social neglect (tensor dimension 3). Combined Tensor: Captures interrelations of cues for sarcasm prediction.

**Performance Highlights:** This method achieves higher accuracy by explicitly modeling multi-cue interactions compared to the prompting methods. This structured prompt ensures that the model’s reasoning aligns with pragmatic analysis principles, fostering a more nuanced understanding of sarcasm detection.

## B LLM Pragmatic Reasoning

Figure 2 illustrates the application of PMP and the reasoning process demonstrated by the model. As depicted, the LLM leverages various elements of the pragmatic framework to arrive at a well-considered conclusion. In the appendix, we include

a detailed PMP analysis of the phrase, “*Lots of people tweeting pictures from their cars of their snowy commutes to work, whilst saying ‘stay safe’ Oh, the #irony!*”. In this example, PMP successfully identifies the nuanced contrast between literal and implied meanings, allowing the model to detect sarcasm effectively by contextualizing the speaker’s intent, polarity, and presuppositions.

## C Transformer Encoder Models

Model	SemEval 2018	
	Acc.	Ma-F1
BERT	69.60	68.40
RoBERTa	75.00	72.00
XLNet	66.00	66.00
DC-Net-RoBERTa	70.90	68.70
GPT-4o (PMP)	86.68	83.18
GPT-4o-mini (PMP)	81.88	79.85
Claude 3.5 Sonnet (PMP)	81.50	76.72
LLaMA-3-70B (PMP)	80.86	78.15
LLaMA-3-8B (PMP)	78.21	77.65

Table 2: Comparison of PMP among LLMs with reported results of Zhang et al. with transformer encoder models on SemEval 2018 Task 3

Previous works have studied the performance of transformer encoder models in sarcasm detection with MUSTARD and SemEval 2018 Task 3. The models listed in Table 2 and Table 3 are reported from Helal et al. and Zhang et al. which show their performance compared to PMP across tested LLMs. A brief explanation of each model is provided:

**BERT** a model built for sarcasm detection which produces context-based word embeddings.

**RoBERTa** based off BERT, which removes the next-sentence predictions, utilizes larger datasets and can have larger inputs of text in training.

**XLNet** Uses permutation language modeling for training, which predicts the order of tokens in a random order, helping the model learn the dependencies of words better.

**DC-Net-RoBERTa** processes two separate inputs (DC-Net) analyzing both literal and implied sentiments of a text separately, using RoBERTa as a base.

**DistilBERT** Uses knowledge distillation where DistilBERT is trained on BERT data as well as mimicking the outputs of BERT while having reduced parameters.

**"Lots of people tweeting pictures from their cars of their snowy commutes to work, whilst saying 'stay safe' Oh, the #irony!"**

- The speaker implies that there is a contradiction in the actions of people who are driving in potentially dangerous snowy conditions while also advising others to stay safe. This irony arises from the fact that they are engaging in an activity that may compromise safety (using a phone to tweet while driving) while simultaneously promoting safety.
- The speaker likely thinks that there is hypocrisy or a lack of self-awareness in these people's actions. They find it ironic or somewhat ridiculous that the same individuals advising others to be cautious are engaging in risky behavior themselves.
- Yes, what the speaker implies and what they think are largely saying the same thing: the behavior they observe is contradictory or ironic.

Finally, the speaker is not pretending to have a certain attitude toward the conversation. Their statement expresses genuine sarcasm or criticism about the contrast between the act of tweeting while driving and the advice to "stay safe."

-----

The preliminary analysis identifies the irony in people tweeting pictures of snowy commutes while advising others to "stay safe." Here's a summary:

- **Implicature**: By calling out the irony, the speaker highlights the contradiction in urging safety while potentially engaging in unsafe behavior (tweeting while driving).
- **Presuppositions**: It is assumed that driving in snow is risky and using a phone while driving is unsafe.
- **Intent of the Speaker**: The speaker aims to point out and criticize the perceived hypocrisy or lack of self-awareness of those tweeting from their cars, by highlighting the irony.
- **Polarity**: The tone of the last sentence is negative, as it critiques the contradiction in behavior.
- **Pretense**: There's no pretense in the speaker's attitude; they openly express criticism through sarcasm.
- **Meaning**: The literal act of tweeting "stay safe" contradicts the implied riskiness of the situation, emphasizing the irony and making the tone sarcastic.

Reflecting on the analysis, the statement is indeed sarcastic. The speaker uses sarcasm to draw attention to the irony and criticize those who are tweeting about safety while potentially compromising it themselves. No major changes are needed in the preliminary analysis.

Figure 2: PMP analysis on the phrase “Lots of people tweeting pictures from their cars of their snowy commutes to work, whilst saying 'stay safe' Oh, the #irony!”.

MUSARD		
Model	Acc.	Ma-F1
RoBERTa	86.80	87.70
DistilBERT	87.00	87.70
GPT-4o (PMP)	79.42	77.65
GPT-4o-mini (PMP)	65.79	62.29
Claude 3.5 Sonnet (PMP)	72.60	71.66
LLaMA-3-70B (PMP)	72.73	73.06
LLaMA-3-8B (PMP)	53.48	54.69

Table 3: Comparison of PMP among LLMs with reported results of Helal et al. with transformer encoder models on MUSARD