

Nek Minit: Harnessing Pragmatic Metacognitive Prompting for Explainable Sarcasm Detection of Australian and Indian English

Ishmanbir Singh* Dipankar Srirag* Aditya Joshi

University of New South Wales, Sydney, Australia

ishman.singh@student.unsw.edu.au, {d.srirag, aditya.joshi}@unsw.edu.au

* Equal contribution

Abstract

Sarcasm is a challenge to sentiment analysis because of the incongruity between stated and implied sentiment. The challenge is exacerbated when the implication may be relevant to a specific country or geographical region. Pragmatic metacognitive prompting (PMP) is a cognition-inspired technique that has been used for pragmatic reasoning. In this paper, we harness PMP for explainable sarcasm detection for Australian and Indian English, alongside a benchmark dataset for standard American English. We manually add sarcasm explanations to an existing sarcasm-labeled dataset for Australian and Indian English called BESSTIE, and compare the performance for explainable sarcasm detection for them with FLUTE, a standard American English dataset containing sarcasm explanations. Our approach utilising PMP when evaluated on two open-weight LLMs (GEMMA and LLAMA) achieves statistically significant performance improvement across all tasks and datasets when compared with four alternative prompting strategies. We also find that alternative techniques such as agentic prompting mitigate context-related failures by enabling external knowledge retrieval. The focused contribution of our work is utilising PMP in generating sarcasm explanations for varieties of English.

1 Introduction

Sarcasm is a form of verbal irony used to express contempt or ridicule, often by saying the opposite of what one means (Joshi et al., 2017). This paper focuses on **explainable sarcasm detection as a generation task where, given a text, the LLM must predict if it contains sarcasm, and, if it does, it must generate a textual explanation** (Kumar et al., 2022). However, sarcasm is socio-culturally situated, and its interpretation often depends on local conventions (Oprea and Magdy, 2020). This may be evidenced via phrases

(such as ‘eshay’ or ‘nek minit¹ in the case of Australian English) or statements (such as ‘The sun is out and I am at work yay’ may not be understood as sarcastic by Indian English speakers). Recent work shows that large language models (LLMs) may misinterpret sarcasm in non-Western cultural contexts (Atari et al., 2023; Tao et al., 2024). Therefore, we focus on explainable sarcasm detection for two varieties of English: Australian (native variety) and Indian English (non-native variety), alongside standard American English. The extended discussion on the related works is provided in Appendix A.

We harness pragmatic metacognitive prompting (PMP; Lee et al., 2025), an extension of metacognitive prompting (Wang and Zhao, 2024), to incorporate contextual understanding with respect to the geographical region for the task of explainable sarcasm detection. We compare PMP alongside four prompting baselines for a standard American English dataset (FLUTE; Chakrabarty et al., 2022) and corresponding Australian and Indian English subsets from BESSTIE (Srirag et al., 2025a) using two open-weight LLMs. Our proposed approach based on PMP significantly ($p \leq 0.001$) improves the performance on both tasks across all datasets². The novelty of this paper is three-fold: (a) We are the first to explore sarcasm explanation generation in the context of varieties of English; (b) we are the first to explore PMP and KG in the context of sarcasm explanation generation.; (c) we release a manually created dataset of sarcasm explanations for Australian and Indian English. Our findings corroborate past work that shows degradation in performance for language varieties other than standard American English for other tasks (Blodgett et al., 2016; Ziems et al., 2023; Joshi et al., 2025; Srirag et al., 2025b).

¹https://en.wikipedia.org/wiki/Nek_minnit; Accessed on 20th May 2025.

²We provide the prompts used for our experiments in Appendix B.

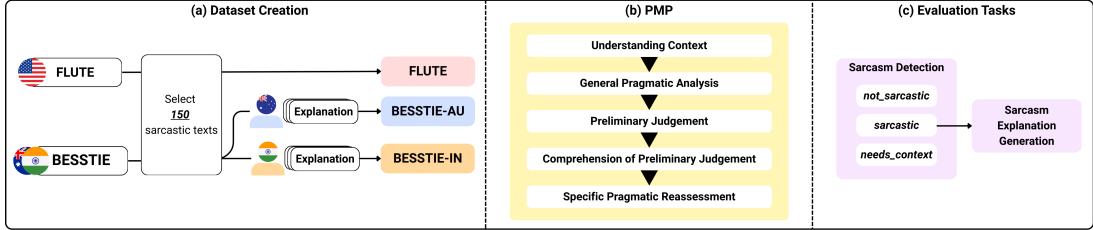


Figure 1: Evaluation methodology; (a) Dataset Creation in Section 2.1; (b) Prompting strategies in Section 2.2; (c) Evaluation tasks in Section 2.

Subset	Samples	Text	Expl.
FLUTE	150	19.1	28.2
BESSTIE-AU	150	47.7	25.7
BESSTIE-IN	150	14.3	29.6

Table 1: Dataset statistics for the three evaluation subsets. The average length of the sarcastic text (Text), and the average length of the corresponding ground truth explanation (Expl.), are measured in words.

2 Proposed Approach

Given an input text, our objective is explainable sarcasm detection. This spans two tasks: (a) sarcasm detection, and (b) the generation of a textual explanation when sarcasm is detected. Additionally, the model may output two auxiliary labels: *not_sarcastic*, if no sarcasm is found; and *needs_context*, if the model is uncertain and requires additional information to make a judgment. As the texts sampled are all labeled to have sarcasm, the expected output is an explanation for sarcasm for all test samples. Our evaluation methodology for the task is outlined in Figure 1.

2.1 Extending BESSTIE

As shown in Figure 1(a), we **annotate 150 sarcastic examples each from Australian and Indian English subsets of BESSTIE with text-based explanations**. Two authors of this paper, one native to Australia and one to India, served as annotators to independently write these sarcasm explanations for texts from their corresponding regional subset. Annotators were instructed to provide semi-structured explanations in a format stylistically aligned with FLUTE. This enables consistency in explanation structure across all samples. We refer to these subsets as BESSTIE-AU and BESSTIE-IN respectively. We also randomly sample **150 sarcastic instances and their corresponding explanations**

from FLUTE as representative of standard American English. Table 1 reports basic statistics of the evaluation datasets, including the average length (in whitespace-tokenised words) of the sarcastic text and its explanation across the three subsets.

2.2 PMP for Explainable Sarcasm Detection

As shown in Figure 1(b), we structure the prompt as a *five-step* reasoning scaffold based on pragmatic linguistic theory (Lee et al., 2025)³. While Lee et al. (2025) utilise PMP for sarcasm detection, we adapt PMP to sarcasm explanation generation for language varieties (Australian English and Indian English) as follows.

- 1. Comprehension of Context/Understanding:** The LLM identifies key situational elements that may trigger sarcastic meanings, enabling the deduction of intended meanings even for unfamiliar words or expressions.
- 2. General Pragmatic Analysis:** In this step, the LLM clarifies the true intent in the text by separating genuine beliefs from exaggerated or pretended attitudes. This can further clarify the meaning of unusual or unclear words or expressions.
- 3. Preliminary Judgment:** An initial hypothesis explanation for sarcasm is generated as a focused seed to guide interpretation.
- 4. Meta-Comprehension:** The LLM validates the hypothesis explanation by ensuring all relevant contextual cues (both explicit and implicit) are correctly interpreted.
- 5. Specific Pragmatic Reassessment:** Finally, the LLM systematically examines detailed pragmatic elements, improving and justifying the final explanation, further resolving any residual ambiguity.

³We provide an example output in the Appendix C.

Prompt	FLUTE			BESSTIE-AU			BESSTIE-IN		
	accuracy	similarity	judge	accuracy	similarity	judge	accuracy	similarity	judge
(a) GEMMA									
ZERO	0.97	0.51	4.7	0.70	0.4	2.92	0.59	0.49	2.62
FEW	0.98	0.63	4.74	0.73	0.42	2.99	0.59	0.49	2.55
ORIGIN	-	-	-	0.72	0.42	3.12	0.61	0.52	2.78
KG	0.93	0.45	4.47	0.74	0.4	2.99	0.85	0.59	3.31
PMP (Ours)	1.0	0.5	4.92	0.94	0.49	3.98	0.91	0.63	3.76
(b) LLAMA									
ZERO	0.61	0.35	2.94	0.55	0.33	2.28	0.49	0.41	1.62
FEW	0.69	0.43	3.22	0.35	0.26	1.66	0.42	0.39	1.43
ORIGIN	-	-	-	0.57	0.34	2.15	0.55	0.41	1.69
KG	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PMP (Ours)	0.9	0.38	3.97	0.92	0.41	3.2	0.79	0.52	2.47

Table 2: Performance comparison between baseline prompting methods and PMP, for tasks: sarcasm detection and sarcasm explanation generation. We test (a) GEMMA and (b) LLAMA on the three datasets. We use *accuracy* to measure the performance on sarcasm detection task, while *similarity* and *judge* are used for sarcasm explanation generation. N/A in (b) indicates KG was not applicable because it is not compatible with LLAMA. The best task performances are denoted by numbers in **bold**.

3 Experiment Setup

We experiment with *two* open-weight decoder-only model optimised for instruction following, namely, Gemma3-12B-Instruct (GEMMA; [Gemma Team et al., 2025](#)) and Llama-3.2-3B-Instruct (LLAMA; [Grattafiori et al., 2024](#))⁴. All experiments are performed using *one* A100 GPU. We compare PMP with *four* baseline methods:

1. **ZERO**: We prompt the model directly with a task instruction for sarcasm detection and explanation, without any examples.
2. **FEW**: We include five manually curated examples from FLUTE to the ZERO prompt. Each example contains a sarcastic sentence paired with an explanation, and these examples are fixed across all evaluations.
3. **ORIGIN**: We append the geographical origin to the ZERO prompt to help the model incorporate cultural and linguistic cues specific to the source variety.
4. **KG**: We implement an agentic prompting strategy inspired by ReAct ([Yao et al., 2023](#)). The model is prompted to identify knowledge gaps, integrate information retrieved through search queries, and then respond. The model uses the DuckDuckGo tool⁵ to issue search queries.

⁴Text Generation Settings: max_new_tokens=1024, temperature=1.0, top_p=0.95, top_k=64

⁵https://github.com/deedy5/duckduckgo_search;

This method of prompting minimises the gap in context and knowledge from local culture, idioms, or events⁶.

We measure performance across three metrics: (a) *accuracy* (the proportion of correctly predicted labels over all instances); (b) *similarity* (average cosine similarity between the Sentence-BERT embeddings of the reference explanation and generated explanation; [Reimers and Gurevych, 2019](#)); and *judge*. For *judge*, we employ GPT-4o ([OpenAI et al., 2024](#)) with default parameters as an evaluator to assess explanation quality. Given a sarcastic text sample, the corresponding ground truth explanation, and the generated explanation, GPT-4o assigns a score from 0 to 5 based on an arbitrary scoring criterion defined in Appendix E. While *accuracy* measures the sarcasm detection performance, *similarity* and *judge* evaluate the generated explanations.

4 Results

Our results address three questions: (a) How do baseline prompting strategies perform on sarcasm detection and explanation generation?; (b) Does PMP improve over these strategies, particularly for sarcasm explanation task?; and (c) Under what settings does PMP yield the best performance?

Accessed on 19 May 2025

⁶An example output is provided in the Appendix D.

Prompt	FLUTE		BESSTIE-AU		BESSTIE-IN	
	NS	NC	NS	NC	NS	NC
ZERO	4	0	36	9	18	43
FEW	2	0	31	10	15	46
ORIGIN	-	-	30	10	21	36
KG	8	0	36	1	18	4
PMP (Ours)	0	0	4	2	1	13

Table 3: Counts of sarcastic instances, across datasets and prompting strategies, flagged by GEMMA to require more context (NC) or incorrectly identified to not present sarcasm (NS).

Table 2 reports results across the *three* datasets for the *two* tasks. For both models, ZERO and FEW yield high results on FLUTE for sarcasm detection, with GEMMA achieving 0.97 and 0.98 *accuracy* respectively, and LLAMA yielding 0.61 and 0.69. However, these methods degrade substantially on the BESSTIE subsets. Particularly, the performance of GEMMA when tested on BESSTIE-IN remains at 0.59 for both methods, while LLAMA shows a slight degradation (ZERO: 0.49; FEW: 0.42). Appending geographic cues in ORIGIN prompt yields inconsistent improvements when compared to ZERO. For example, GEMMA reports a minor improvement when tested on BESSTIE-AU (ZERO: 0.7; ORIGIN: 0.72), LLAMA reports a higher performance gain when tested on BESSTIE-IN (ZERO: 0.49; ORIGIN: 0.55). These findings indicate that appending geographical cues alone is insufficient to capture cultural nuances. For explanation generation, the trends among baselines remain consistent. GEMMA, when prompted with KG, demonstrates improvements in explanation quality on BESSTIE-IN (*similarity*: 0.59; *judge*: 3.31), suggesting some benefit from agentic reasoning. However, these gains do not consistently transfer across datasets. Moreover, KG is not compatible with LLAMA, limiting its applicability.

PMP significantly improves over all baselines across both tasks and models. For GEMMA, PMP achieves the highest detection accuracy (1.00) and explanation quality (*judge*: 4.92) on FLUTE, and similarly strong performance on BESSTIE-AU (*accuracy*: 0.94; *judge*: 3.98) and BESSTIE-IN (*accuracy*: 0.91; *judge*: 3.76). For LLAMA, PMP significantly improves over baseline performances: on BESSTIE-IN, it improves *similarity* and *judge* to 0.52 and 2.47 compared to ORIGIN(0.41 and 1.69), and *accuracy* of 0.79 from 0.55. We also note that

while each row of the table reports lower values for Australian and Indian English as compared to standard American English present in FLUTE, our results on PMP method highlight the potential to bridge the performance gap between varieties. The improvements across all datasets are statistically significant ($p \leq 0.001$) relative to the ZERO baseline.

Table 3 presents an error analysis of two types of explanation errors: instances incorrectly classified as NS, and those requiring NC. On both BESSTIE datasets, PMP substantially reduces these errors. On BESSTIE-IN, the count of NS cases drops from 18 under ZERO to 1 with PMP; similarly, NC cases reduce from 43 to 13. These results are consistent for BESSTIE-AU, confirming that PMP is particularly effective in cases requiring nuanced cultural or pragmatic interpretation. However, for BESSTIE-IN, models more often output NC, especially for ZERO and FEW. This indicates that the model often cannot interpret the sarcasm from surface-level language alone, likely due to cultural or regional differences in how sarcasm is expressed. Here, KG proves effective by supplying external context that helps the model bridge this gap, reducing the NS errors to 4, when compared to 43 from ZERO.

5 Conclusion

We demonstrated how pragmatic metacognitive prompting (PMP) can be harnessed to generate sarcasm explanations for text written in varieties. We evaluate our approach on three language varieties of English: standard American English, Australian and Indian English, and two open-source LLMs: GEMMA and LLAMA. We annotated an existing dataset for the latter two with sarcasm explanations to assess model performance for explainable sarcasm detection. Standard prompting methods, such as zero-shot and few-shot, perform well on FLUTE, but failed to generalise to Australian and Indian English subsets of BESSTIE. **PMP significantly ($p \leq 0.001$) improved the performance across both tasks, both Australian and Indian English, and both the models.** Agentic prompting methods like KG also reduced context-related failures by enabling dynamic knowledge integration. Our results and error analysis demonstrated the limitations of generic prompts and the importance of pragmatic scaffolding for figurative language understanding. Our findings suggest that reasoning-aware prompting offers a viable pathway to improve explanation generation for sarcasm.

Limitations

We perform sarcasm detection only on positive text samples, i.e. sarcastic text, as this study primarily explores sarcasm explanation as a task. There may be multiple possible explanations for sarcasm. Also, we acknowledge that there are varieties within a country related to the state, native language, and so on. Note that we perform sarcasm detection only on positive text samples, i.e. sarcastic text, as this study primarily explores sarcasm explanation as a task. However, as a first study on sarcasm explanation for language varieties, our observations are a good starting point for the future. The proposed PMP approach helps BESSTIE-AU and BESSTIE-IN to retrieve appropriate context which is not necessary in the case of FLUTE based on past findings in the bias of LLMs towards a western-centric context.

Ethical Considerations

The research was approved by the ethics board of the host institution. All annotators were native speakers of the respective English varieties (Australian and Indian) and participated voluntarily in the annotation process. Given that sarcasm is often context-dependent and culturally embedded, we acknowledge the sensitivity involved in interpreting or misclassifying user-generated content. No personally identifiable information was used in this study.

Acknowledgment

This paper is the outcome of a Taste-of-Research scholarship awarded to Ishmanbir Singh by the Faculty of Engineering at UNSW Sydney. The paper is dedicated to the memory of Pushpak Bhattacharyya from Indian Institute of Technology Bombay who served as the ACL President (2016), and was considered a pioneer of NLP in India.

References

- Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. [Which humans?](#)
- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. [Demographic dialectal variation in social media: A case study of African-American English](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.
- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. [FLUTE: Figurative language understanding through textual explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Keanealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Patterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathi-halli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotiinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivantha, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhannagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotrutz,

Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. *Gemma 3 technical report*. Preprint, arXiv:2503.19786.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. *The role of conversation context for sarcasm detection in online interactions*. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Alionsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junting Jia, Kalyan Vasudevan Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew

Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Couder, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenber, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant

- Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanaz-
eri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry As-
pegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,
Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher,
Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan Mc-
Phie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khan-
delwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madien Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaoqian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. *The llama 3 herd of models*. Preprint, arXiv:2407.21783.
- Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. *Multi-source semantic graph-based multimodal sarcasm explanation generation*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11349–11361, Toronto, Canada. Association for Computational Linguistics.
- Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. *Automatic sarcasm detection: A survey*. *ACM Comput. Surv.*, 50(5).
- Aditya Joshi, Raj Dabre, Diptesh Kanodia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dipold. 2025. *Natural language processing for dialects of a language: A survey*. *ACM Comput. Surv.* Just Accepted.
- Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. *When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.
- Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. 2025. *Pragmatic metacognitive prompting improves LLM performance on sarcasm detection*. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 63–70, Online. Association for Computational Linguistics.
- Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. *Multimodal sarcasm detection via cross-modal graph convolutional network*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.
- Yiyi Liu, Ruqing Zhang, Yixing Fan, Jiafeng Guo, and Xueqi Cheng. 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, Dubrovnik, Croatia. Association for Computational Linguistics.
- Chenwei Lou, Bin Liang, Lin Gui, Yulan He, Yixue Dang, and Ruifeng Xu. 2021. *Affective dependency graph for sarcasm detection*. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’21*, page 1844–1849, New York, NY, USA. Association for Computing Machinery.

- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Sham, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Silviu Vlad Oprea and Walid Magdy. 2020. [The effect of sociocultural variables on sarcasm communication online](#). *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).
- Soujanya Poria, Erik Cambria, Devamanyu Hazarika, and Prateek Vir. 2016. [A deeper look into sarcastic tweets using deep convolutional neural networks](#). In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1601–1612, Osaka, Japan. The COLING 2016 Organizing Committee.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. [Sarcasm as contrast between a positive sentiment and negative situation](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.
- Dipankar Srirag, Aditya Joshi, Jordan Painter, and Diptesh Kanodia. 2025a. [Besstie: A benchmark for sentiment and sarcasm classification for varieties of english](#). *Preprint*, arXiv:2412.04726.
- Dipankar Srirag, Nihar Ranjan Sahoo, and Aditya Joshi. 2025b. [Evaluating dialect robustness of language](#)

models via conversation understanding. In *Proceedings of the Second Workshop on Scaling Up Multilingual & Multi-Cultural Evaluation*, pages 24–38, Abu Dhabi. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico. Association for Computational Linguistics.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2024. Is sarcasm detection a step-by-step reasoning process in large language models? *Preprint*, arXiv:2407.12725.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2025. Is sarcasm detection a step-by-step reasoning process in large language models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25651–25659.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-value: A framework for cross-dialectal english nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

A Related Work

Sarcasm detection has employed statistical models (Joshi et al., 2017), sentiment-incongruity heuristics (Riloff et al., 2013), and traditional neural architectures such as LSTMs (Ghosh et al., 2017), CNNs (Poria et al., 2016), GNNs (Lou et al., 2021; Liang et al., 2022) and Transformer (Yao et al., 2024). **Prompting LLMs** has shown promise in guiding models to reason about sarcasm effectively (Liu et al., 2023; Yao et al., 2025). Specifically, Pragmatic Metacognitive Prompting (PMP) (Lee et al., 2025) introduces a structured approach that mirrors human pragmatic reasoning by incorporating reflection and analysis of implied meanings, contextual cues, and speaker intent. This method has demonstrated improved performance in sarcasm detection tasks but **has not been used for sarcasm explanation generation**. Additionally, agentic prompting methods like ReAct (Yao et al., 2023) enable models to actively retrieve and integrate external knowledge, facilitating context-aware reasoning in sarcasm detection. Beyond sarcasm detection, **sarcasm explanation** generation has been investigated using datasets like FLUTE (Chakrabarty et al., 2022) which provide figurative language instances with explanations. Models such as TEAM (Jing et al., 2023) employ multi-source semantic graphs to generate multimodal sarcasm explanations, integrating visual and textual cues. Ours is the first work that examines the different prompting strategies for the explanation generation of sarcasm for language varieties.

B Prompts

In this paper, we evaluate models on the task of explainable sarcasm detection, by prompting models to first identify sarcasm and then provide an explanation for classifying the text to be sarcastic. Below are the prompts that we used.

B.1 ZERO

For the provided text, perform one of the tasks. If the text is not sarcastic, output ‘not_sarcastic’. If the text is sarcastic, provide an explanation in one or two sentences. Output ‘need_context’ if you cannot explain the sarcasm.

B.2 FEW

For the provided text, perform one of the tasks. If the text is not sarcastic, output ‘not_sarcastic’. If the text is sarcastic, provide an explanation in one or two sentences. Output ‘need_context’ if you cannot explain the sarcasm. Following are some examples of sarcastic texts and their explanations.

1. **Text:** I got the greatest news today when I found out my company is getting rid of me after 10 years of faithful service
Explanation: sarcastic. Being fired from a job after 10 years of service is not something to be happy about and hence receiving the greatest news is sarcastic.
2. **Text:** I feel great that I am single and have not accomplished anything significant inspite of being 30 years old
Explanation: sarcastic. Someone who is 30 years old and single has not achieved anything significant in their life so far and so should not feel honored by that fact.
3. **Text:** The idea of the genius Republican party blatantly rejecting climate change and its effects on our planet makes me feel safe for the future
Explanation: sarcastic. The Republican party’s denial of climate change will have catastrophic effects on our planet in the future and so to feel confident about the future is absurd.
4. **Text:** I was so impressed when my hair dresser tried a new product on my hair but the end result was me getting to shave all my hair off

Explanation: sarcastic. Shaving all your hair off is not an impressive hairstyle and is often seen as a last resort when someone has messed up your hair.

5. **Text:** I am on seventh heaven because I don't have enough money to pay my important bill.

Explanation: sarcastic. Having not enough money to pay an important bill creates a lot of stress and anxiety and is definitely not a heavenly feeling.

B.3 ORIGIN

This text is from [Australian/Indian] subreddit post or comment. If the text is not sarcastic, output ‘not_sarcastic’. If the text is sarcastic, provide an explanation in one or two sentences. Output ‘need_context’ if you cannot explain the sarcasm.

B.4 KG

Think step-by-step and feel free to use tools whenever they would be helpful. Once you have finished reasoning and using tools:

- If the text is not sarcastic:
{"action": "Final Answer", "action_input": "not_sarcastic"}
- If the text is sarcastic:
{"action": "Final Answer", "action_input": "sarcastic. Explanation: concise explanation (1-2 sentences)"}
- If you cannot explain the sarcasm:
{"action": "Final Answer", "action_input": "need_context"}

All outputs must be valid JSON, with no unescaped double quotes; use single quotes. You have access to the following tools:

- **Search:** Tool for getting up to date answers to current or historical events and word/phrase definitions., args: {'tool_input': {'type': 'string'}}

Use a json blob to specify a tool by providing an action key (tool name) and an action_input key (tool input). Valid “action” values are: “Final Answer” or “Search”.

Provide only ONE action per \$JSON_BLOB, as shown:

```
{"action": $TOOL_NAME, "action_input": $INPUT}
```

Follow this format:

Question: input question to answer

Thought: consider previous and subsequent steps

Action: \$JSON_BLOB

Observation: action result

... (repeat Thought/Action/Observation N times)

Thought: I know what to respond

Action:

```
{"action": "Final Answer", "action_input": "Final response to human"}
```

Begin! Reminder to ALWAYS respond with a valid json blob of a single action.

Use tools if necessary. Respond directly if appropriate. Format is

Action:\$JSON_BLOB then Observation:.

Thought:

Human:

Additional Information

Once the model outputs a thought and action, the action is executed and the thought, action and output are all appended back into the prompt. This repeats until the "Final Answer" action is output by the model.

B.5 PMP

For the provided text, perform one of the tasks. If the text is not sarcastic, output 'not_sarcastic'. If the text is sarcastic, follow these steps in your analysis:

1. Comprehension of Context/Understanding:

- Repeat the text and summarize it to confirm your understanding
- Identify the context, topic, and key elements of the situation described

2. General Pragmatic Analysis:

- What does the author imply about the situation with their statement?
- What does the author think about the situation?
- Are what the author implies and what they think saying the same thing?
- Is the author pretending to have a certain attitude toward the conversation?

3. Preliminary Judgment:

- Based on your analysis, formulate an initial explanation of why the text is sarcastic

4. Comprehension of Preliminary Judgment/Context:

- Summarize your preliminary analysis
- Ensure you've captured all relevant contextual elements

5. Specific Pragmatic Analysis/Reassessment:

- Implicature: What is implied beyond the literal meaning?
- Presuppositions: What information is taken for granted?
- Intent: What does the author hope to achieve with their statement?
- Polarity: Is there a mismatch between apparent and intended tone?
- Pretense: Is there pretense in the author's attitude?
- Meaning gap: What is the difference between literal and implied meaning?
- Context: Consider slang, cultural references, or communication patterns

6. After completing your analysis, perform one of the tasks.

- If the text is sarcastic, provide an explanation in one or two sentences.
- Output 'need_context' if you cannot explain the sarcasm.

C PMP Example Output

PMP Output Analysis

Input Text

Text: Whitlam wanted to shut it down did he not? Nek Minit

Step 1: Comprehension of Context/Understanding

Summary: The text poses a rhetorical question about Whitlam's desire to shut something down, followed by the phrase *Nek Minit*, which is Australian slang meaning *in a short time or very soon*.

Context: The context appears to be Australian politics, referencing Gough Whitlam, a former Prime Minister of Australia. The "it" refers to something he allegedly wanted to shut down. The inclusion of *Nek Minit* suggests a playful or dismissive response to that allegation.

Step 2: General Pragmatic Analysis

Author's Implied Meaning: The author likely implies that the idea of Whitlam wanting to shut something down is overstated or a common, possibly exaggerated, claim.

Author's Attitude: The author's attitude is likely skeptical or humorous. They are not taking the claim seriously.

Alignment of Implied and Stated: The implied meaning and the stated question don't perfectly align. The question *seems* to solicit confirmation, but the *Nek Minit* undermines that, suggesting the matter is trivial or quickly resolved.

Pretense of Attitude: The author is pretending to be earnestly seeking confirmation, but the *Nek Minit* reveals a dismissive or playful attitude.

Pretense: Appears serious but sets up mockery.

Step 3: Preliminary Judgment

The text is likely sarcastic because the rhetorical question is followed by a phrase (*Nek Minit*) that minimizes the significance of the issue being questioned, creating a mismatch between the apparent seriousness of the question and the implied triviality.

Step 4: Comprehension of Preliminary Judgment/Context

The text employs a rhetorical question about Whitlam's actions within the context of Australian politics. The combination of the question and *Nek Minit* creates a sarcastic effect by downplaying the significance of Whitlam's alleged desire to shut something down, suggesting the issue is either minor or quickly resolved.

Step 5: Specific Pragmatic Analysis/Reassessment

Implicature: The implicature is that the claim about Whitlam's actions is not a major concern or deserves lengthy consideration.

Presuppositions: It presupposes that the audience is familiar with Gough Whitlam and potentially has some knowledge of the issue being referenced

(whatever ""it"" is). It also presupposes an understanding of *Nek Minit* as slang indicating immediacy or triviality.

Intent: The author intends to express skepticism or amusement about the claim in a playful, culturally specific way.

Polarity: There's a significant polarity mismatch. The question has a formal, interrogative structure (suggesting seriousness), while *Nek Minit* introduces an informal, dismissive tone.

Pretense: The author is feigning serious inquiry while actually expressing a dismissive attitude.

Meaning Gap: The literal meaning of the question is a request for confirmation. The implied meaning is a casual rejection of the claim's importance.

Context: The use of Australian slang (*Nek Minit*) is crucial to understanding the sarcasm. It's a culturally specific expression that signals informality and downplaying.

Step 6: Final Explanation

The rhetorical question regarding Whitlam's alleged action is undermined by the casual Australian slang *Nek Minit*, creating a sarcastic effect by suggesting the issue is trivial and doesn't warrant serious consideration.

Ground Truth Explanation

The comment seems supportive of Prime Minister Whitlam but the *Nek Minit* (suggesting an ironic twist of events) pokes fun at how he was dismissed from government before being able to do anything.

D KG Example Output

KG Reasoning Chain

Input Text

Text: This case seriously is now sounding like a badly written *Crime Patrol* episode.

Thought 1

The text uses a comparison to a low-quality, fictional crime show (*Crime Patrol*) to express disbelief or mockery of a real-life case. The phrase “badly written” implies that the case is illogical, predictable, or absurd, but instead of stating this directly, it uses the humorous comparison to a poorly written TV show. This indirect, mocking comparison is characteristic of sarcasm. I need to check what *Crime Patrol* is to ensure the comparison is understood.

Action

Action: Search

Action Input: "What is Crime Patrol?"

Action Output

Crime Patrol, the iconic true-crime anthology series hosted by Anup Soni, is making its way to Netflix on March 17, 2025. The series, which premiered in 2003, is the longest-running reality crime television series in India. *Crime Patrol* not only tells such gripping stories but also teaches important lessons about staying safe and preventing crime. By showing the efforts of law enforcement, the show helps audiences see the importance of being aware and cautious in today's world.

Thought 2

The search result confirms that *Crime Patrol* is a well-known Indian crime show, known for its dramatized recreations of real crimes. Given that the original statement compares the case to “a badly written *Crime Patrol* episode,” it suggests the case being described is convoluted, unrealistic, or predictable—much like a poorly constructed fictional crime narrative. The sarcasm lies in the unexpected and humorous comparison to a show known for its sometimes sensationalized and dramatic portrayal of crime, implying the real-life case is far less sophisticated than it should be.

Final Output Explanation

The statement uses a negative comparison to *Crime Patrol* to mock the quality of the case, implying it is unrealistic or poorly handled.

Ground Truth Explanation

The user sarcastically compares the case to a poorly written episode of *Crime Patrol*, a popular Indian crime reenactment show—implying that the developments in the case are so bizarre, dramatic, or unbelievable that they feel fictional and poorly scripted, rather than real-life events.

E Scoring Criterion

Scoring Criterion

0: Irrelevant or incoherent explanation

1: Barely related, vague or generic statement

2: Somewhat related but incomplete or unclear reasoning

3: Reasonable explanation, covers core sarcastic cue

4: Strong explanation with appropriate contextual grounding

5: Excellent explanation, highly aligned with human interpretation