Sarc7: Evaluating Sarcasm Detection and Generation with Seven Types and Emotion-Informed Techniques

Anonymous authors
Paper under double-blind review

Abstract

Sarcasm is a complex linguistic phenomenon where expressions convey meanings that contrast with their literal interpretations. Accurately classifying and generating sarcasm is critical for improving large language model (LLM) understanding of human intent. We introduce **Sarc7**, a benchmark for fine-grained sarcasm evaluation based on the MUStARD dataset, annotated with seven distinct sarcasm types: self-deprecating, brooding, deadpan, polite, obnoxious, raging, and manic. These categories are adapted from prior linguistic work and used to create a structured dataset suitable for LLM evaluation. For classification, we evaluate multiple prompting strategies—zero-shot, few-shot, chain-of-thought (CoT), and a novel emotion-based technique—across five major LLMs. Emotion-based prompting yields the highest macro-averaged F1 score of 0.3664 (Gemini 2.5), outperforming CoT for several models and demonstrating its effectiveness in sarcasm type recognition. For generation, we design structured prompts using fixed values across four sarcasm-relevant dimensions: incongruity, shock value, context dependency, and emotion. Using Claude 3.5 Sonnet, this approach produces more subtype-aligned outputs, with human evaluators preferring emotion-based generations 38.46% more often than zero-shot baselines. Sarc7 offers a foundation for evaluating nuanced sarcasm understanding and controllable generation in LLMs, pushing beyond binary classification toward interpretable, emotion-informed language modeling.

1 Introduction

Sarcasm is defined as the use of remarks that convey the opposite of their literal meaning. Understanding sarcasm requires an intuitive grasp of humor and social cues, posing a challenge for natural language processing (NLP) tasks such as human-like conversation Yao et al. (2024); Gole et al. (2024). Large language models (LLMs) generally perform poorly on sarcasm classification and generation tasks due to the subtlety and context dependence of sarcastic language Yao et al. (2024). Traditional sentiment analysis and machine learning techniques also struggle with these challenges. This work introduces a novel sarcasm benchmark grounded in the seven recognized types of sarcasm and proposes an emotion-based approach for both classification and generation. In contrast to prior rule-based and template-driven methods, which often produced rigid outputs Zhang et al. (2024), and even more recent deep learning models that still fall short in capturing subtlety and social nuance Gole et al. (2024), our technique aims to improve contextual relevance and expressive range in sarcastic generation.

2 Related Work

Previously, SarcasmBench Zhang et al. (2024) established benchmarks for binary sarcasm classification by evaluating state-of-the-art (SOTA) large language models (LLMs) and pretrained language models (PLMs). Current benchmarks do not address specific sarcasm type classification or generation, or emotion as a controlled factor. Emotion and sarcasm

are directly correlated, as sarcasm is emotionally fueled and reflects the speaker's emotion, both intentionally and unintentionally Leggitt & Gibbs (2000); Biswas et al. (2019). Linguists have identified seven basic types of sarcasm, which we use to build our Sarc7 benchmark Qasim (2021).

Sarcasm Classification: Recent advances have focused on structured prompting techniques that use pragmatic reasoning to enhance sarcasm detection Lee et al. (2024). Approaches such as pragmatic metacognitive prompting method (PMP) have improved model performance by making sarcasm inference more explicit Yao et al. (2024); Lee et al. (2024). Furthermore, recent studies have shown that integrating commonsense, knowledge, and attention mechanisms help models identify subtleties in sarcastic statements Zhuang et al. (2025). These methods show that guiding LLMs with structured signals can help them better understand the nuances of sarcastic statements.

Sarcasm Generation: Recent studies have introduced controlled generation methods to guide LLMs toward producing sarcastic statements using contradiction strategies and dialogue cues Zhang et al. (2024); Helal et al. (2024). Structured prompting and contradiction-based strategies have shown to improve sarcasm generation. Some methods guide LLMs by introducing contrast between expected and actual meanings or using contextual dialogue cues for coherence Zhang et al. (2024); Helal et al. (2024); Skalicky & Crossley (2018). However, existing techniques struggle with controlling sarcasm levels and aligning them with contextual incongruence, shock value, and prior context dependency.

3 Methods

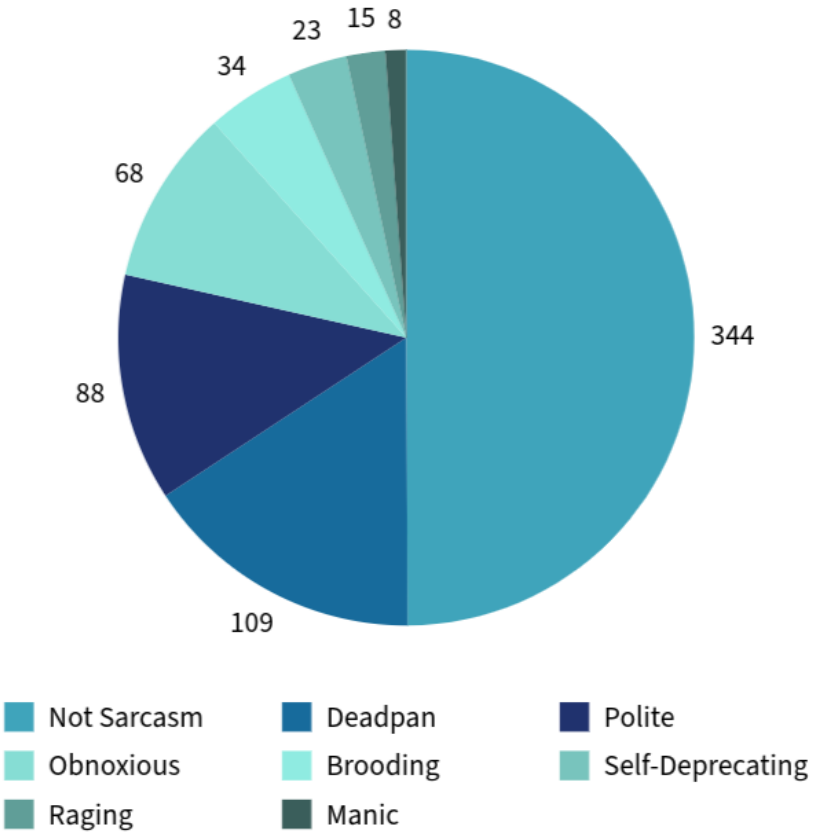

Figure 1: Distribution of annotation labels in the dataset.

3.1 Benchmark Construction

We introduce **Sarc7**, a novel benchmark for fine-grained sarcasm classification and generation. Building on the MUStARD dataset (Castro et al., 2019), which provides binary sarcasm annotations for short dialogue segments, we manually annotated each sarcastic utterance with one of seven distinct sarcasm types: *self-deprecating*, *brooding*, *deadpan*, *polite*, *obnoxious*, *raging*, and *manic*.

These seven categories are inspired by the linguistic taxonomy proposed in Qasim (2021), which identified common sarcasm types based on pragmatic and affective features. Our contribution lies in implementing these types of sarcasm for computational annotation. We defined each type using precise, example-grounded criteria suitable for large language model evaluation, and we applied this schema to build the first sarcasm benchmark that captures this level of granularity.

Each sarcastic utterance in the MUStARD dataset was independently labeled by four trained annotators using the seven sarcasm subtypes defined in Sarc7. Annotators received detailed definitions and examples of each category (see Table 1) to ensure consistent interpretation. The annotation process is illustrated in Figure 2.

- Each utterance was first labeled independently by all four annotators.
- If at least three annotators agreed on the same label, that label was accepted as the final annotation.
- In cases with no 3-out-of-4 agreement, a consensus discussion was held between annotators, with a final decision made by majority vote.

Figure 1 shows the distribution of the seven annotated sarcasm types. The resulting Sarc7 benchmark supports two tasks: (1) multi-class sarcasm classification, and (2) sarcasm-type-conditioned generation. These tasks allow for more fine-grained evaluation of sarcasm understanding in large language models.

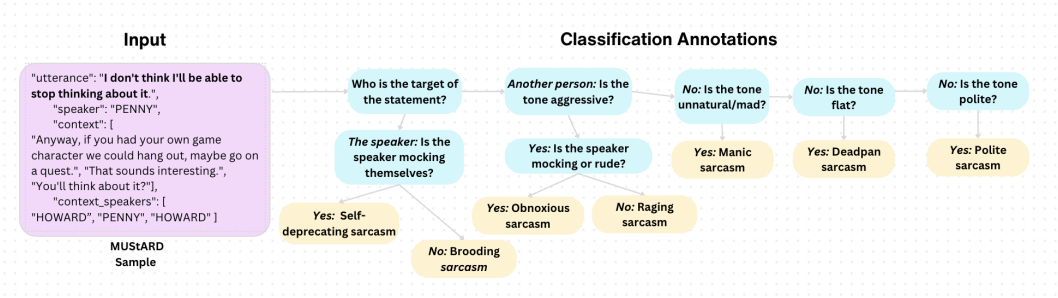

Figure 2: Flowchart of the Step-by-Step Process for Sarcasm Classification Annotation

3.2 Task Definition

We define two primary evaluation tasks:

- **Sarcasm Classification**: Given a sarcastic utterance and its dialogue context, correctly predict the dominant sarcasm type from among the seven annotated categories.
- **Sarcasm Generation**: Generate a sarcastic utterance consistent with one of the 7 types of sarcasm. Table 1 outlines definitions for each sarcasm category in the Sarc7 benchmark.

3.3 Baseline Classification

Our baseline testing focused on zero-shot, few-shot, and CoT prompting. For generations, baseline outputs were produced using a zero-shot prompt, without structured control over

Type	Definition	Example
Self-deprecating	Mocking oneself in a humorous or critical way.	"Oh yeah, I'm a genius — I only failed twice!"
Brooding	Passive-aggressive frustration masked by politeness.	"Sure, I'd love to stay late again — who needs weekends?"
Deadpan	Sarcasm delivered in a flat, emotionless tone.	"That's just the best news I've heard all day."
Polite	Insincere compliments or overly courteous remarks.	"Wow, what an *interesting* outfit you've chosen."
Obnoxious	Rude or provocative sarcasm aimed at others.	"Nice driving! Did you get your license in a cereal box?"
Raging	Intense, exaggerated sarcasm expressing anger.	"Of course! I *love* being yelled at in meetings!"
Manic	Overenthusiastic, erratic sarcasm with chaotic tone.	"This is AMAZING! Who needs food or sleep anyway?!"

Table 1: Operational definitions and examples of the seven sarcasm types used in Sarc7

dimensions. These baselines were evaluated by a human grader based on accuracy of sarcasm type and emotion.

3.4 Emotion-Based Prompting

Emotion-based prompting refers to the assistance of emotion categories per sarcasm type. American psychologist Paul Ekman identified six basic emotions: happiness, sadness, anger, fear, disgust, and surprise Ekman (1992). Our emotion-based prompting technique consists of three main steps: 1) Categorize the emotion of the context. 2) Classify the emotion of the utterance. 3) Identify the sarcasm based on the incongruity of the emotional situation.

3.5 Generation Dimensions

Our approach moves beyond general sarcasm generation by conditioning the model on four controllable dimensions intended to guide the tone, intensity, and context of the output:

- **Incongruity**: Degree of semantic mismatch.
- **Shock Value**: Intensity of sarcasm.
- **Context Dependency**: Reliance on conversational history.
- **Emotion**: One of Ekman's six basic emotions (e.g., anger, sadness).

Rather than tuning these dimensions dynamically, we assigned fixed values for each when prompting the model to generate a sarcastic utterance. These values were manually selected to reflect characteristics associated with specific sarcasm subtypes. For example, a "raging" sarcasm prompt might use high shock value, high incongruity, low context dependency, and anger as the target emotion. By anchoring each generation to these abstract but interpretable cues, we observed improved alignment between the generated outputs and their intended sarcasm type.

This structured prompting approach helps control for variation in tone and emotional affect, resulting in more consistent and subtype-specific sarcasm generation. A sample output from this technique is shown in Figure 3.

4 Experiments

4.1 Model Selection

We evaluate several state-of-the-art language models on our proposed sarcasm benchmark, including GPT-4o OpenAI (2024), Claude 3.5 Sonnet Anthropic (2024), Gemini 2.5 DeepMind et al. (2023), Qwen 2.5 Team (2024), and Llama 4 Maverick.

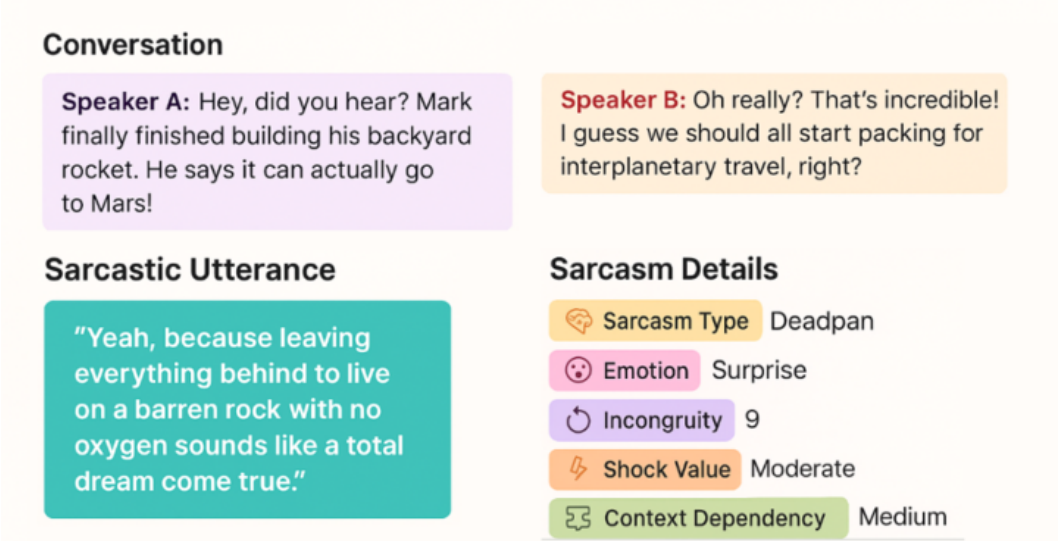

Figure 3: Sample output using emotion-based method

4.2 Evaluation

We evaluated classification by comparing model predictions to human-annotated labels across seven sarcasm types. For generation, Claude 3.5 Sonnet produced 100 sarcastic statements per prompting method, each rated by a human for type accuracy.

5 Results and Discussion

Model	0-shot	Few-shot	CoT	Emotion-based
GPT-4o	47.73%	50.29%	**55.07%**	48.94%
Claude 3.5 Sonnet	51.16%	52.61%	**57.10%**	52.32%
Qwen 2.5	41.45%	**46.96%**	46.09%	45.94%
Llama-4 Maverick	34.20%	35.51%	**50.29%**	49.86%
Gemini 2.5	46.81%	47.97%	**53.04%**	52.03%

Table 2: Classification Accuracy Across Models and Prompting Techniques

5.1 Classification Results

Across all evaluated prompting techniques, CoT prompting consistently outperformed zero-shot, few-shot, and emotion-based approaches in sarcasm classification. Table 2 shows its superior results compared to other methods in almost every model.

In terms of macro-averaged F1 score, emotion-based prompting outperformed zero-shot, few-shot, and chain-of-thought (CoT) prompting. As shown in Table 3, Gemini 2.5 achieved the highest F1 score overall under emotion-based prompting, with Claude 3.5 Sonnet, Llama-4 Maverick, and Qwen 2.5 also seeing gains relative to their CoT performance. While CoT prompting remains strong in absolute accuracy and reasoning through ambiguous cases, emotion-based prompting demonstrated greater ability to generalize across sarcasm types, especially those associated with emotional signals.

This improvement is particularly important given the dataset's class imbalance. Since types like "Deadpan" appear more frequently than others such as "Manic" or "Polite," raw accuracy metrics may disproportionately reflect dominant class performance. Macro-averaged F1 provides a more balanced evaluation by weighting each class equally. The

higher F1 scores observed under emotion-based prompting suggest that emotional cues may help LLMs better distinguish between low-frequency categories, even in the absence of detailed reasoning steps.

Model	0-shot F1	Few-shot F1	CoT F1	Emotion-based F1
GPT-4o	0.2089	**0.3255**	0.2674	0.2233
Claude 3.5 Sonnet	0.2964	0.3487	0.2471	**0.3487**
Qwen 2.5	0.2116	0.2075	0.2052	**0.2124**
Llama-4 Maverick	0.2184	0.2340	0.2040	**0.2841**
Gemini 2.5	0.2760	0.3274	0.3141	**0.3664**

Table 3: Macro-averaged F1 scores of models across prompting techniques.

5.2 Classification Confusion Analysis

While models showed moderate success identifying sarcastic utterances, they struggled to accurately categorize specific sarcasm types. Figure 4 shows that most models, including GPT4o, Claude 3.5 Sonnet, and Gemini 2.5, frequently defaulted to labeling content as either "not sarcastic" or "deadpan sarcasm" when uncertain. Deadpan emerged as the most frequent misclassification across all sarcasm types, underscoring its role as a default or fallback label in ambiguous cases.

This trend reveals a key limitation: although LLMs can sometimes detect cues associated with sarcastic tone, they often conflate subtle, flat, or ambiguous language with sarcasm—even when none is present. The frequent misclassification of non-sarcastic utterances as "deadpan" indicates that models are over-reliant on surface-level features such as flat affect or contrastive phrasing, rather than grounded pragmatic reasoning. As a result, fine-grained differentiation among sarcasm subtypes remains a substantial challenge. Improving model sensitivity to context and disambiguation of neutral tone from intentional sarcasm is critical for more accurate multi-class sarcasm detection.

5.3 Prompt Technique Analysis

Despite lower overall accuracy, emotion-based prompting achieved higher F1 scores across several models, suggesting better performance on minority classes. Emotion cues likely helped models interpret speaker intent and tone, key to recognizing nuanced sarcasm. CoT prompting had the highest accuracy by supporting pragmatic reasoning, while emotion-based methods suffered from label ambiguity, as multiple sarcasm types can share similar emotional signals. This highlights the trade-off between emotional intuition and structured reasoning in multi-class sarcasm classification.

5.4 Qualitative Error Analysis

Despite strong binary performance, models often misclassify playful language as sarcasm. Consider the following example:

> **Utterance:** `A lane frequented by liars. Like you, you big liar!`
> **Context:** `HOWARD: I just Googled "foo-foo little dogs."`
> `HOWARD: (Skype ringing) It's Raj. Stay quiet.`
> `HOWARD: (chuckles): Hey!`
> `Bad timing.`
> `Bernadette just took Cinnamon out for a walk.`
> `RAJ: Hmm. Interesting.`
> `Did they take a walk down Liars' Lane?`
> `HOWARD: What?`

The true label is *not sarcastic*, yet all models predicted *obnoxious sarcasm*. The CoT prompt overemphasized surface-level markers such as exaggeration and contradiction, failing to

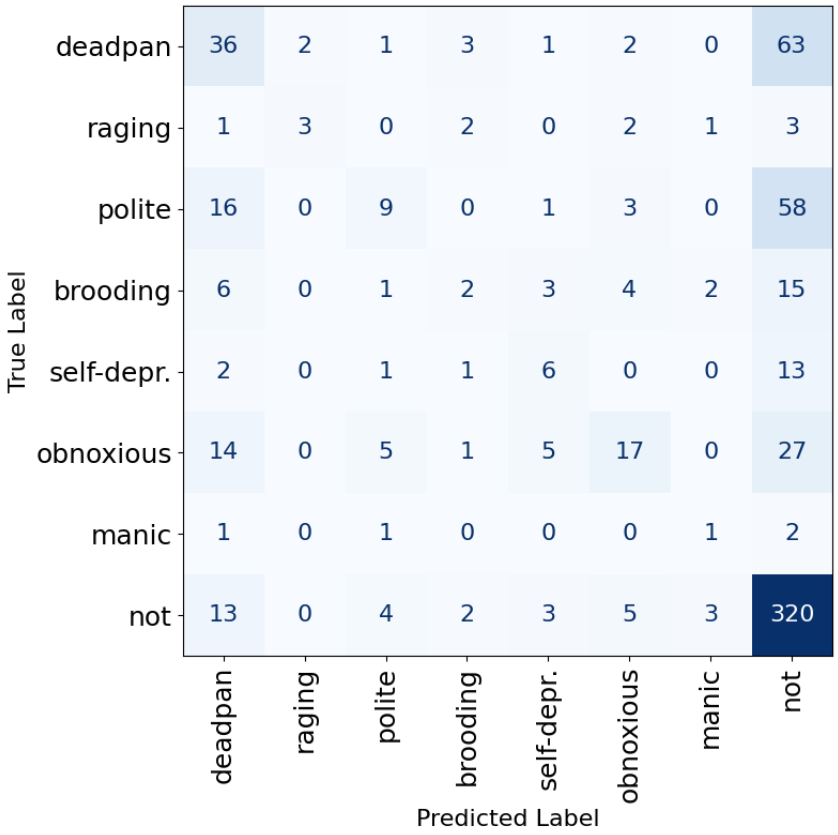

Figure 4: Confusion matrix for Claude 3.5 Sonnet using CoT.

consider the light tone of the exchange. Similarly, the emotion-based prompt misclassified the utterance by identifying "disgust" due to literal wording, despite the playful social context. These errors highlight a broader limitation: while structured prompting improves reasoning, both CoT and emotion-based methods lack sensitivity to pragmatic cues and interpersonal intent in conversational sarcasm.

5.5 Generation Results and Analysis

Emotion-based prompting generated more accurate sarcasm types. Table 4 shows a 38.42% increase in accuracy using the emotion-based structure compared to the baseline model.

Prompt	**Successful Generation**
Zero-shot	52/100
Emotion-based	**72/100**

Table 4: Generation Evaluation Scores

For example, when prompted for raging sarcasm zero-shot prompting produced a neutral response:

"Oh, absolutely! I only stayed up until 3 AM because sleep is just so overrated, right?"

The emotion-based prompt with angry context and high shock value generated:

"Isn't that just fantastic? I mean, who wouldn't want to spend an entire day writing reports on how well we walk from our desks to the restroom? It's a dream come true!"

The baseline prompt's neutral context made it difficult to generate raging sarcasm, likely confusing it with deadpan due to the absence of anger cues. However, our emotion-based prompt was able to identify the anger in the statement and appropriately express it in its response. Explicit emotional cues helped generate more distinct sarcasm types. See Appendix B for examples' context.

While multiple models were evaluated for the classification task, we selected Claude 3.5 Sonnet for generation due to its consistently strong performance in classification accuracy and F1 score (see Table 2 and 3). Our primary goal in this benchmark was not to compare model-specific generation capabilities, but to explore how structured prompting techniques—particularly emotion-based prompting—affect the quality and controllability of sarcasm generation. By holding the model constant, we isolate the impact of the prompting strategy itself. Future work may extend this evaluation to other models such as GPT-4o and Gemini 2.5 to assess cross-model generalization.

5.6 Dataset Imbalance and Cultural Considerations

Our evaluation revealed two key considerations for improving sarcasm modeling: label imbalance and cultural-linguistic scope. Sarc7 has a skewed distribution, with "Deadpan" far more frequent than types like "Manic," leading models to default to dominant labels under uncertainty. Future work could address this with balancing techniques like weighted loss or data augmentation.

Sarcasm also varies by culture and language. Since MUStARD consists of English dialogues from Western media, Sarc7 reflects only English-speaking norms. Expanding to other languages would reveal how sarcasm differs across cultures and test whether English-trained LLMs can generalize across diverse sarcastic styles.

6 Conclusions

In this work, we propose a new multi-class sarcasm benchmark, Sarc7, by categorizing sarcasm into seven types and establishing a benchmark for classification and generation. Sarc7 enhances the understanding of LLM's ability to identify subtle cues in sarcastic statements. We developed an emotion-based prompting generation technique that showed a higher sarcasm detection accuracy than traditional methods. The structured generation approach guided by incongruity, shock value, context, and emotion produced more accurate sarcastic statements than baseline methods. Our benchmark advances sarcasm research by enabling fine-grained, multi-class evaluation—crucial for modeling nuanced, human-like intent in language.

Limitations

The proposed Sarc7 method has several limitations. (1) While the process for annotating the MUStARD dataset had a rigorous structure, and annotations were peer-reviewed for consistency, there is still room for inconsistencies. (2) The MUStARD dataset had an imbalance of sarcasm types, which may have introduced bias during classification. For example, deadpan sarcasm was common in the dataset while manic sarcasm was rare. (3) Our emotion-based prompting technique relies on Ekman's six basic emotions, which assumes these categories are universally applicable. However, emotions like "anger" and "surprise" may be expressed or perceived differently across languages or cultures. Since MUStARD is based on English-language media, future research should explore the cross-lingual generalizability of our emotion-based approach. (4) Besides the cues introduced—incongruity, shock value, context dependency, and emotion—other cues could have potentially been useful. (5) Single-category classification of the dataset is limiting and may have skewed results. For example, a data point may have been brooding and manic, but was only classified as brooding. Future research may explore multi-category classification for a more comprehensive overview of LLMs' understanding of sarcasm.

References

Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Anthropic Report*, 2024.

Prasanna Biswas, Anupama Ray, and Pushpak Bhattacharyya. Computational model for understanding emotions in sarcasm: A survey. *CFILT Technical Report, Indian Institute of Technology Bombay*, 2019.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1455. URL https://aclanthology.org/P19-1455/.

Google DeepMind, Rohan Anil, Stefano Arolfo, Igor Babuschkin, Lucas Beyer, Maarten Bosma, and ... Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Paul Ekman. Are there basic emotions? *Psychological Review*, 99(3), 1992.

Montgomery Gole, Williams-Paul Nwadiugwu, and Andriy Miranskyy. On sarcasm detection with openai gpt-based models. In *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, pp. 1–6. IEEE, 2024.

Nivin A Helal, Ahmed Hassan, Nagwa L Badr, and Yasmine M Afify. A contextual-based approach for sarcasm detection. *Scientific Reports*, 14(1):15415, 2024.

Joshua Lee, Wyatt Fong, Alexander Le, Sur Shah, Kevin Han, and Kevin Zhu. Pragmatic metacognitive prompting improves llm performance on sarcasm detection. *arXiv preprint arXiv:2412.04509*, 2024.

John S Leggitt and Raymond W Gibbs. Emotional reactions to verbal irony. *Discourse processes*, 29(1):1–24, 2000.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Sawsan Abdul-Muneim Qasim. A critical pragmatic study of sarcasm in american and british social interviews. 2021. URL https://www.researchgate.net/publication/363925404_A_Critical_Pragmatic_Study_of_Sarcasms_in_American_and_British_Interviews.

Stephen Skalicky and Scott Crossley. Linguistic features of sarcasm and metaphor production quality. *Proceedings of the Workshop on Figurative Language Processing*, 2018.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *arXiv preprint arXiv:2408.11319*, 2024.

Xingjie Zhuang, Fengling Zhou, and Zhixin Li. Multi-modal sarcasm detection via knowledge-aware focused graph convolutional networks. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

A Classification Statistics

Below are the macro-averaged precision, recall, and F1 scores for all prompting techniques.

Model	Precision	Recall	F1 Score
GPT-4o	0.2104	0.2073	0.2089
Claude 3.5 Sonnet	**0.2982**	**0.2960**	**0.2964**
Gemini 2.5	0.2703	0.2824	0.2760
Llama-4 Maverick	0.2173	0.2196	0.2184
Qwen 2.5	0.2217	0.2025	0.2116

Table 5: Macro-Averaged Precision, Recall, and F1 Scores for Each Model under Zero-shot Prompting.

Model	Precision	Recall	F1 Score
GPT-4o	0.3067	0.3469	0.3255
Claude 3.5 Sonnet	**0.3322**	**0.3669**	**0.3487**
Gemini 2.5	0.3233	0.3314	0.3274
Llama-4 Maverick	0.2314	0.2361	0.2340
Qwen 2.5	0.2461	0.1794	0.2075

Table 6: Macro-Averaged Precision, Recall, and F1 Scores for Each Model under few-shot Prompting.

Model	Precision	Recall	F1 Score
GPT-4o	0.2682	0.2668	0.2674
Claude 3.5 Sonnet	0.2903	0.2148	0.2471
Gemini 2.5	**0.3178**	**0.3106**	**0.3141**
Llama-4 Maverick	0.2116	0.1970	0.2040
Qwen 2.5	0.2063	0.2038	0.2052

Table 7: Macro-Averaged Precision, Recall, and F1 Scores for Each Model under CoT Prompting.

Model	Precision	Recall	F1 Score
GPT-4o	0.2140	0.2331	0.2233
Claude 3.5 Sonnet	0.3322	0.3669	0.3487
Gemini 2.5	**0.3388**	**0.3990**	**0.3664**
Llama-4 Maverick	0.2936	0.2753	0.2841
Qwen 2.5	0.2352	0.1933	0.2124

Table 8: Macro-Averaged Precision, Recall, and F1 Scores for Each Model under Emotion Prompting.

B Generation Output

Below is an example of zero-shot and emotion-based generation results.

Sarcasm Generation Example

Emotion-based prompting was able to generate more targeted sarcasm types. For example, in the case of a contextually neutral statement, the baseline model produced a generic sarcastic response.

Zero-Shot Conversation:

- Speaker A: Did you finish the presentation for tomorrow's big meeting?
- Speaker B: Oh, absolutely! I only stayed up until 3 AM because sleep is just so overrated, right?
- Speaker A: Wow, sounds like you're thrilled about your life choices.

Zero-Shot Sarcastic Utterance:

- Speaker B: *Oh, absolutely! I only stayed up until 3 AM because sleep is just so overrated, right?*

Emotion-Based Context:

- Speaker A: Hey, did you see those new management rules they rolled out today?
- Speaker B: Oh yes, they're really something else. Now, we're going to document every minute of our bathroom breaks.
- Speaker A: Isn't that just fantastic? I mean, who wouldn't want to spend an entire day writing reports on how well we walk from our desks to the restroom? It's a dream come true!

Emotion-Based Sarcastic Utterance:

- Speaker A: *Isn't that just fantastic? I mean, who wouldn't want to spend an entire day writing reports on how well we walk from our desks to the restroom? It's a dream come true!*

C Prompts

Below are the zero-shot, few-shot, sarcasm analysis, and emotion-based prompts.

Zero-shot Prompt

You are tasked with determining the sarcasm type in a given statement. Read the statement carefully and classify the sarcasm type based on the context of the statement. Use one of the following categories:

- Self-deprecating sarcasm – mocking oneself
- Brooding sarcasm – passive-aggressive or emotionally repressed
- Deadpan sarcasm – flat or emotionless tone
- Polite sarcasm – fake politeness or ironic compliments
- Obnoxious sarcasm – mocking, mean-spirited, or rude
- Raging sarcasm – angry, exaggerated, or harsh
- Manic sarcasm – unnaturally cheerful, overly enthusiastic

If the statement is **not sarcastic**, **Output**: `[not sarcasm]`
If the statement is **sarcastic**, **Output**: `[Type of Sarcasm]`

Sarcasm Type Classification Prompt (Few-Shot)

You are tasked with determining the sarcasm type in a given statement. Read the statement carefully and classify the sarcasm type based on the context of the statement. Use one of the following categories:

- Self-deprecating sarcasm – mocking oneself
- Brooding sarcasm – passive-aggressive or emotionally repressed
- Deadpan sarcasm – flat or emotionless tone
- Polite sarcasm – fake politeness or ironic compliments
- Obnoxious sarcasm – mocking, mean-spirited, or rude
- Raging sarcasm – angry, exaggerated, or harsh
- Manic sarcasm – unnaturally cheerful, overly enthusiastic

If the statement is **not sarcastic**, **Output**: `[not sarcasm]`
If the statement is **sarcastic**, **Output**: `[Type of Sarcasm]`

Examples:

A person might say, "Your new shoes are just fantastic," to indicate that the person finds a friend's shoes distasteful.
Output: `[Polite sarcasm]`

A socially awkward person might say, "I'm a genius when it comes to chatting up new acquaintances."
Output: `[Self-deprecating sarcasm]`

A person who is asked to work overtime at one's job might respond, "I'd be happy to miss my tennis match and put in the extra hours."
Output: `[Brooding sarcasm]`

A person who is stressed out about a work project might say, "The project is moving along perfectly, as planned. It'll be a winner."
Output: `[Manic sarcasm]`

When asked to mow the lawn, a person might respond by yelling, "Why don't I weed the gardens and trim the hedges too? I already do all of the work around the house."
Output: `[Raging sarcasm]`

A person might say, "I'd love to attend your party, but I'm headlining in Vegas that evening," with a straight face, causing others to question whether they might be serious.
Output: `[Deadpan sarcasm]`

A person's friend may offer a ride to a party, prompting the person to callously answer, "Sure. I'd love to ride in your stinky rust bucket."
Output: `[Obnoxious sarcasm]`

Sarcasm Analysis Prompt

You are a sarcasm analyst. Your task is to determine whether a speaker's utterance is sarcastic or sincere. Only if you are reasonably confident the speaker is being sarcastic based on tone, behavior, and contradiction between words and context, classify it into a type.

—

Step 1: Contextual Emotion Analysis
Analyze the emotional tone of the surrounding context or situation (i.e., what is happening before or around the statement). Consider what emotion would be appropriate or expected in that situation.
Select one dominant contextual emotion from this fixed list:

- Happiness
- Sadness

- Anger
- Fear
- Surprise
- Disgust
- Neutral (use only if no strong emotion applies)

—

Step 2: Utterance Emotion Analysis
Analyze the emotional tone of the bracketed statement itself based on word choice, delivery cues (e.g., exaggeration, flatness, enthusiasm), and stylistic tone.
Select one dominant utterance emotion from the same list:

- Happiness
- Sadness
- Anger
- Fear
- Surprise
- Disgust
- Neutral

Use only one label for each step. Do not guess outside this list.

—

Step 3: Emotional Comparison and Incongruity Detection
Compare the contextual emotion and the utterance emotion. If there is a mismatch (e.g., the situation is sad but the speaker sounds happy), explain whether this emotional contrast suggests mockery, irony, insincerity, passive aggression, or theatrical overreaction.
If no such contrast or ironic delivery is present, conclude that the statement is not sarcastic.

—

Step 4: Sarcasm Type Classification
If the statement is sarcastic, classify it using the emotional cues, delivery style, and social function into one of the following types:

- Self-deprecating sarcasm – mocking oneself
- Brooding sarcasm – passive-aggressive or emotionally repressed
- Deadpan sarcasm – flat or emotionless tone
- Polite sarcasm – fake politeness or ironic compliments
- Obnoxious sarcasm – mocking, mean-spirited, or rude
- Raging sarcasm – angry, exaggerated, or harsh
- Manic sarcasm – unnaturally cheerful, overly enthusiastic

—

Step 5: Final Output
Clearly output the final classification on a new line in this exact format:

- If sarcastic: `[Type of Sarcasm]`
- If not sarcastic: `[Not Sarcasm]`

Emotion-based Prompt

You are an expert sarcasm and emotion analyst. For every input statement, follow the steps below in order, using the context and speaker's delivery to reason carefully.

—

Step 1: Contextual Emotion Analysis
Analyze the emotional tone of the surrounding context or situation (i.e., what is happening before or around the statement). Consider what emotion would be appropriate or expected in that situation.

Select one dominant contextual emotion from this fixed list:

- Happiness
- Sadness
- Anger
- Fear
- Surprise
- Disgust
- Neutral (use only if no strong emotion applies)

—

Step 2: Utterance Emotion Analysis
Analyze the emotional tone of the bracketed statement itself based on word choice, delivery cues (e.g., exaggeration, flatness, enthusiasm), and stylistic tone.
Select one dominant utterance emotion from the same list:

- Happiness
- Sadness
- Anger
- Fear
- Surprise
- Disgust
- Neutral

Use only one label for each step. Do not guess outside this list.

—

Step 3: Emotional Comparison and Incongruity Detection
Compare the contextual emotion and the utterance emotion. If there is a mismatch (e.g., the situation is sad but the speaker sounds happy), explain whether this emotional contrast suggests mockery, irony, insincerity, passive aggression, or theatrical overreaction.
If no such contrast or ironic delivery is present, conclude that the statement is not sarcastic.

—

Step 4: Sarcasm Type Classification
If the statement is sarcastic, classify it using the emotional cues, delivery style, and social function into one of the following types:

- Self-deprecating sarcasm – mocking oneself
- Brooding sarcasm – passive-aggressive or emotionally repressed
- Deadpan sarcasm – flat or emotionless tone
- Polite sarcasm – fake politeness or ironic compliments
- Obnoxious sarcasm – mocking, mean-spirited, or rude
- Raging sarcasm – angry, exaggerated, or harsh
- Manic sarcasm – unnaturally cheerful, overly enthusiastic

—

Step 5: Final Output
Clearly output the final classification on a new line in this exact format:

- If sarcastic: `[Type of Sarcasm]`
- If not sarcastic: `[Not Sarcasm]`

D Misclassification

Below are tables indicating the most misclassified sarcasm type for each sarcasm type for each of the prompting techniques.

Table 9: Most Frequent Misclassifications per Type using Zero-Shot Prompting

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Obnoxious	Polite	Not Sarcastic
Obnoxious	Not Sarcastic	Deadpan	Deadpan	Deadpan	Deadpan
Brooding	Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan
Polite	Not Sarcastic	Deadpan	Deadpan	Deadpan	Not Sarcastic
Raging	Obnoxious	Deadpan	Obnoxious	Obnoxious	Obnoxious
Manic	Not Sarcastic	Deadpan	Obnoxious	Deadpan	Not Sarcastic
Self-deprecating	Not Sarcastic	Deadpan	Deadpan	Deadpan	Deadpan
Not Sarcastic	Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan

Table 10: Most Frequent Misclassifications per Type using Few-Shot Prompting

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Obnoxious	Polite	Not Sarcastic
Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan	Deadpan
Brooding	Deadpan	Deadpan	Deadpan	Deadpan	Deadpan
Polite	Not Sarcastic	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Raging	Obnoxious	Deadpan	Obnoxious	Obnoxious	Obnoxious
Manic	Raging	Self-deprecating	Obnoxious	Obnoxious	Not Sarcastic
Self-deprecating	Deadpan	Not Sarcastic	Deadpan	Deadpan	Deadpan
Not Sarcastic	Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan

Table 11: Most Frequent Misclassifications per Type using CoT Prompting

Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Not Sarcastic	Not Sarcastic	Not Sarcastic
Obnoxious	Deadpan	Not Sarcastic	Deadpan	Deadpan	Deadpan
Brooding	Deadpan	Not Sarcastic	Deadpan	Deadpan	Deadpan
Polite	Not Sarcastic	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Raging	Deadpan	Not Sarcastic	Obnoxious	Deadpan	Obnoxious
Manic	Brooding	Not Sarcastic	Not Sarcastic	Deadpan	Brooding
Self-deprecating	Not Sarcastic	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Not Sarcastic	Deadpan	Deadpan	Deadpan	Deadpan	Deadpan

Table 12: Most Frequent Misclassifications per Type using Emotion-Based Prompting

Sarcasm Type	GPT-4o	Claude 3.5 Sonnet	Gemini 2.5	Llama-4 Maverick	Qwen 2.5
Deadpan	Not Sarcastic	Not Sarcastic	Not Sarcastic	Obnoxious	Not Sarcastic
Obnoxious	Deadpan	Deadpan	Deadpan	Deadpan	Not Sarcastic
Brooding	Deadpan	Deadpan	Deadpan	Obnoxious	Not Sarcastic
Polite	Deadpan	Deadpan	Not Sarcastic	Not Sarcastic	Not Sarcastic
Raging	Brooding	Deadpan	Obnoxious	Obnoxious	Not Sarcastic
Manic	Polite	Not Sarcastic	Self-deprecating	Obnoxious	Not Sarcastic
Self-deprecating	Deadpan	Not Sarcastic	Not Sarcastic	Deadpan	Not Sarcastic
Not Sarcastic	Deadpan	Deadpan	Deadpan	Obnoxious	Deadpan