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

Lang Xiong
langlglang@email.com

Raina Gao
rainatg9@gmail.com

Alyssa Jeong
alyssa.y.jeong@gmail.com

Abstract

Sarcasm is a complex linguistic and pragmatic phenomenon where expressions convey meanings that contrast with their literal interpretations, requiring sensitivity to the speaker's intent and context. Misinterpreting sarcasm in collaborative human–AI settings can lead to under- or overreliance on LLM outputs, with consequences ranging from breakdowns in communication to critical safety failures. We introduce **Sarc7**, a benchmark for fine-grained sarcasm evaluation based on the MUStARD dataset, annotated with seven pragmatically defined 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 sarcasm 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). Sarcasm is a pragmatic act, where meaning depends not only on words but also on speaker intent, emotional tone, and shared context. 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. We examine whether LLMs can demonstrate pragmatic reasoning. 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

While prior benchmarks (Zhang et al., 2024) focus on binary detection by evaluating state-of-the-art (SOTA) large language models (LLMs) and pretrained language models (PLMs), (Leggitt and Gibbs, 2000; Biswas et al., 2019) real-world agents require subtype sensitivity. According to (Qasim, 2021), Lamb (2011) first introduced a seven-type classification of sarcasm based on observational studies of classroom discourse. (Qasim, 2021) then refined these categories into operational definitions tailored for social-interview data, providing clear examples and criteria. (Zuhri and Sagala, 2022) subsequently applied this refined taxonomy in an irony and sarcasm detection system for public-figure speech.

Proceedings of the 9th Widening NLP Workshop, pages 157–166
November 8, 2025 ©2025 Association for Computational Linguistics

Sarcasm Classification: Research has progressed from early sentiment-contrast frameworks (Riloff et al., 2013) to modern techniques that guide LLM inference. Recent advances leverage structured prompting for pragmatic reasoning (Lee et al., 2024; Yao et al., 2024) and integrate external knowledge to help models identify subtleties (Zhuang et al., 2025), confirming that structured signals improve nuance detection.

Sarcasm Generation: Current generation methods use controlled techniques like structured prompting and contradiction strategies to guide LLM outputs (Zhang et al., 2024; Helal et al., 2024; Skalicky and Crossley, 2018). Despite these advances, existing approaches lack fine-grained control over sarcasm levels and key dimensions like contextual incongruity or shock value.

3 Methods

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.

3.2 Annotation Methodology

Each of the 690 sarcastic utterances from MUStARD was labeled by four native-english speaking annotators using our seven-type schema (see Table 3), guided by pragmatic definitions and examples. Labels with at least three annotator agreements were accepted; remaining cases were resolved via majority-vote discussion. A fifth annotator then re-labeled all examples, yielding Cohen's κ = 0.6694 (substantial agreement) and human macro-averaged precision/recall/F1 of 0.6586/0.6847/0.6663. Brooding, deadpan, and polite subtypes were hardest even for humans, setting realistic performance ceilings for models.

Figure 2 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.

3.3 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 3 outlines definitions for each sarcasm category in the Sarc7 benchmark.

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

3.5 Emotion-Based Prompting

Our emotion-based prompting goes beyond traditional sentiment analysis by leveraging the six basic emotions identified by American psychologist Paul Ekman: 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. By comparing these two emotion labels, we capture nuanced contrasts that a simple positive/negative split cannot distinguish.

3.6 Generation Dimensions

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

- **Incongruity**: Degree of semantic mismatch (1-10).
- **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 subtype based on our intuitive understanding (see Table 9). 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.

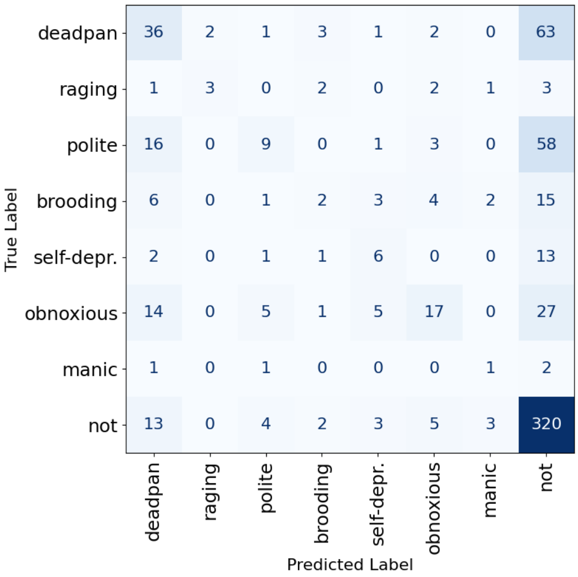
4 Experiments

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 (Meta AI, 2024).

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 sarcasm type accuracy.



	deadpan	raging	polite	brooding	self-depr.	obnoxious	manic	not
deadpan	36	2	1	3	1	2	0	63
raging	1	3	0	2	0	2	1	3
polite	16	0	9	0	1	3	0	58
brooding	6	0	1	2	3	4	2	15
self-depr.	2	0	1	1	6	0	0	13
obnoxious	14	0	5	1	5	17	0	27
manic	1	0	1	0	0	0	1	2
not	13	0	4	2	3	5	3	320

Figure 1: Confusion Matrix for Claude 3.5 Sonnet using CoT.

Subtype	**CoT**	**Emotion-based**	**Human**
Brooding	6.06%	9.09%	39.39%
Deadpan	33.03%	50.46%	55.45%
Polite	10.34%	33.33%	57.30%
Manic	20.00%	20.00%	75.00%
Obnoxious	24.64%	39.13%	67.14%
Raging	25.00%	41.67%	71.43%
Self-deprecating	26.09%	34.78%	86.96%
Not sarcasm	91.17%	66.38%	95.04%

Table 1: Per-class Accuracy for Claude 3.5 using CoT vs. Emotion-based Prompting, Alongside Human Agreement.

5 Results and Discussion

5.1 Classification Results and Analysis

Our results highlight a key trade-off between prompting methods. While Chain-of-Thought (CoT) prompting achieves the highest raw accuracy (57.10%), emotion-based prompting yields a superior macro-averaged F1-score (0.3664). This is because emotion-based prompts significantly improve the detection of low-frequency sarcasm subtypes like "Polite" (+23.0%) and "Raging" (+16.7%). Given the Sarc7 dataset's class imbalance, the macro-F1 score provides a fairer assessment of performance.

However, a significant drawback emerges: emotion-based prompts decrease accuracy on non-sarcastic inputs by 24.8%. This suggests the models become "trigger-happy," creating a critical precision-recall trade-off where false positives increase. This behavior stems from a general model bias to default to "Deadpan" or "Not sarcasm" when uncertain, relying on surface cues over genuine pragmatic inference. While emotion-informed prompting is a vital step toward more context-aware detection, this trade-off reveals a key robustness and alignment challenge for real-world applications where misclassifying neutral text is problematic.

5.2 Prompt Technique Analysis

Our analysis reveals a trade-off between prompting techniques. Emotion-based prompting yields a higher macro-F1 score by using discrete emotional cues to help models identify low-frequency sarcasm subtypes, especially when context is limited. In contrast, Chain-of-Thought (CoT) prompting achieves higher overall accuracy through its structured reasoning but can overlook these subtle emotional distinctions. This also explains why

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 2: Macro-averaged F1 scores of Models Across Prompting Techniques.

few-shot prompting surpasses CoT in macro-F1; its concrete examples provide a stronger signal for rare classes, whereas CoT's abstract reasoning may default to more common labels like 'deadpan' or 'not sarcastic'.

5.3 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 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.

5.4 Generation Results and Analysis

Emotion-based prompting generated more accurate sarcasm types. Table 10 shows a 38.42% increase in accuracy using the emotion-based structure compared to the baseline model.

By explicitly specifying dimensions like shock value and target emotion, our generation technique makes the model's choices transparent—each sarcastic output can be traced back to the intended setting—thereby improving interpretability. For raging sarcasm, the zero-shot prompt yielded a bland reply—"Oh, absolutely! I only stayed up until 3 AM because sleep is just so overrated, right?"—whereas our emotion-based prompt (high shock value, anger) produced a clearly enraged quip: "Isn't that just fantastic? Who wouldn't want to track every restroom trip all day? Dream come true!" directly reflecting the selected parameters. This structured control also mitigates bias toward the most frequent "deadpan" or overly neutral styles: by anchoring each subtype in distinct emotional and intensity cues, we prevent the model from defaulting to bland or stereotyped responses and ensure more equitable coverage of underrepresented sarcasm types (e.g., brooding, manic).

We selected Claude 3.5 Sonnet for generation due to its consistently strong performance in classification accuracy and F1 score (see Table 4 and 2). 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.

6 Conclusion

We present **Sarc7**, the first benchmark to evaluate both the detection and controlled generation of seven nuanced sarcasm subtypes, framing the task as a test of an LLM's pragmatic competence. Our classification experiments show that while chain-of-thought prompting yields the highest accuracy, emotion-based prompts achieve a superior macro-averaged F1 score (0.3664 with Gemini 2.5). A human baseline ($\kappa = 0.6694$) confirms the inherent difficulty of subtypes like brooding and deadpan. For generation, structured prompts specifying dimensions like incongruity and emotion improved subtype alignment by 38% over zero-shot baselines with Claude 3.5 Sonnet. By benchmarking fine-grained performance, Sarc7 moves beyond binary detection and lays the groundwork for more natural, context-sensitive dialogue agents with potential for future multimodal and cross-lingual extensions.

References

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

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

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

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

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

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

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

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

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

Meta AI. 2024. Llama-4-maverick-17b-128e-original. Hugging Face Model Hub: `https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Original`. Accessed: 2025-06-27.

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

Sawsan Abdul-Muneim Qasim. 2021. A critical pragmatic study of sarcasm in american and british social interviews. *Journal of Strategic Research in Social Science*.

Ellen Riloff, Aditya Qadir, Prajakta Surve, Lakshika De Silva, Nisheeth 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. ACL.

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

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

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

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

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

Ari Tantra Zuhri and Rakhmat Wahyudin Sagala. 2022. Irony and sarcasm detection on public figure speech. *Journal of Elementary School Education*, 1(1):41–45.

A Limitations and Safety

Our evaluation revealed several areas for improvement. Although our peer-reviewed annotation process was rigorous, some disagreement remains under a forced single-label scheme, and the heavy class imbalance (e.g. many deadpan but few manic examples) introduces bias—future work could use multi-label annotations and data balancing. Relying on Ekman's six basic emotions also misses subtler affects like irony or embarrassment and may not generalize across languages or cultures, so richer emotion taxonomies and cross-lingual validation are needed. Finally, prosody, discourse structure, and dialogue history are untapped sources of pragmatic nuance, and expanding Sarc7 with multilingual and multimodal data will help ensure equitable sarcasm detection across diverse communities. Transparent rationales are also crucial for safe deployment: mis-interpreting sarcasm in mission-critical dialogues (e.g. negotiations, medical advice) risks harmful actions. Our emotion-based prompts surface whether the model truly identified an anger or disgust signal before labeling an utterance sarcastic, substantially reducing the model's bias toward the dominant "not sarcasm" label. This improved true-positive rates on genuine sarcastic subtypes by up to 23 percent—thereby avoiding safety hazards where an agent might otherwise fail to detect critical ironic intent.

B Reproducibility Statement

All data and code required to reproduce the findings of this study are publicly available at: https://github.com/langlglang/sarc7 under an apache 2.0 license. All prompts are included in the appendix.

C Classification Definition and Statistics

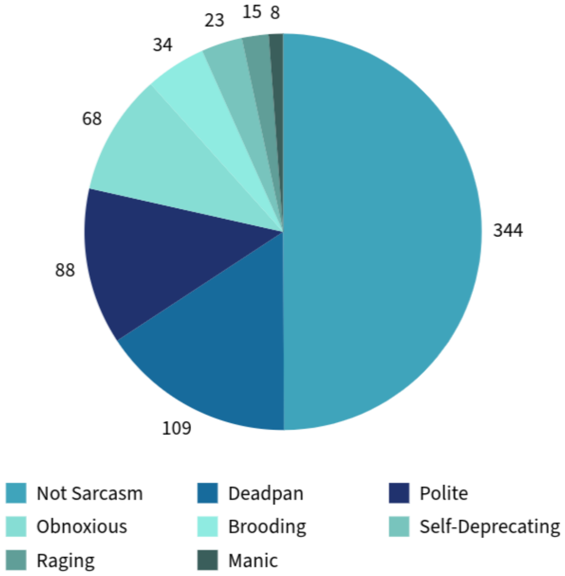


Figure 2: Distribution of Annotation Labels in the Dataset.

Type	Definition
Self-deprecating	Mocking oneself in a humorous or critical way.
Brooding	Passive-aggressive frustration masked by politeness.
Deadpan	Sarcasm delivered in a flat, emotionless tone.
Polite	Insincere compliments or overly courteous remarks.
Obnoxious	Rude or provocative sarcasm aimed at others.
Raging	Intense, exaggerated sarcasm expressing anger
Manic	Overenthusiastic, erratic sarcasm with chaotic tone.

Table 3: Operational Definitions and Examples of the Seven Sarcasm Types used in Sarc7

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

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. Gemini 2.5 produces the highest precision, recall, and F1 score.

D Generation Settings and 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*

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 4: Classification Accuracy Across Models and 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. Claude 3.5 Sonnet produces the highest precision, recall, and F1 score.

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	0.075

Table 6: Macro-Averaged Precision, Recall, and F1 Scores for Each Model under few-shot Prompting. 3.5 Sonnet produces the highest precision and recall score, while GPT-4o produces the highest F1 score.

> *well we walk from our desks to the restroom? It's a dream come true!*

E 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**
- **Brooding sarcasm**
- **Deadpan sarcasm**
- **Polite sarcasm**
- **Obnoxious sarcasm**

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. 3.5 Sonnet produces the highest precision and recall score, while GPT-4o produces the highest F1 score.

- **Raging sarcasm**
- **Manic sarcasm**

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**
- **Brooding sarcasm**
- **Deadpan sarcasm**
- **Polite sarcasm**
- **Obnoxious sarcasm**
- **Raging sarcasm**
- **Manic sarcasm**

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

If the statement is **sarcastic**, **Output**: `[Type of Sarcasm]`

Examples:

Subtype	**Incongruity (1–10)**	**Shock Value**	**Context Dependency**	**Emotion**
Self-deprecating	3–5	low	medium	sadness
Brooding	5–7	medium	medium	anger
Deadpan	4–6	low	high	neutral
Polite	3–5	low	medium	happiness
Obnoxious	6–9	high	low	disgust
Raging	7–9	high	low	anger
Manic	5–7	high	medium	surprise

Table 9: Dimension Settings and Target Emotion for Each Sarcasm Subtype used in our Emotion-based Prompting.

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

Table 10: Generation Evaluation Scores

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 subtype. If there is no strong evidence of sarcasm (no exaggeration, no mismatch, no insincere tone), assume the speaker is genuine.

Think step by step:

1. Analyze speaker delivery and tone.
2. Check whether their words contradict the situation.
3. Ask: "Could a sincere person say this the same way?"
   - If yes: **Output**: [not sarcasm]
   - Otherwise: proceed to step 4.
4. Match to one of the following subtypes:
   - Self-deprecating sarcasm
   - Brooding sarcasm
   - Deadpan sarcasm
   - Polite sarcasm
   - Obnoxious sarcasm
   - Raging sarcasm
   - Manic sarcasm

Format your answer like this:

```
Utterance: <the target utterance>
Context:   <brief dialogue or situation>
Reasoning:
- <first reasoning bullet>
- <second reasoning bullet>
- . . .
Output: [Type of Sarcasm]
```

Example: *Utterance: "Oh yeah, I love getting stuck in traffic for hours." Context: (Someone is running late and stuck in traffic.) Reasoning:*

- Uses exaggeration ("love") about a negative event.
- Clear mismatch between words and reality.
- Tone is bitter and frustrated.

Output: [Brooding 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]`

Sarcasm Generation Prompt

You are a sarcasm simulation system. Create a short fictional dialogue that includes a clearly sarcastic utterance. Use the inputs below to guide the tone and structure.

Parameters:

- Incongruity Rating (1–10): incongruity
- Shock Value: shock_value
- Context Dependency: context_dependency
- Emotion of Sarcastic Utterance: emotion

Output format:

```
Conversation:
Speaker A: ...
Speaker B: ...
Speaker A: ...
(At least 3 turns)

Sarcastic Utterance: (copy the
sarcastic utterance exactly
here)

Sarcasm Type: (Self-deprecating,
Brooding, Deadpan, Polite,
Obnoxious, Raging, or Manic)

Emotion: {emotion}

Incongruity Rating:
{incongruity}

Shock Value: {shock_value}

Context Dependency:
{context_dependency}
```

F Misclassification

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

Table 11: Most Frequent Misclassifications per Type using Zero-Shot Prompting

Type	GPT-4o	Claude 3.5	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 12: Most Frequent Misclassifications per Type using Few-Shot Prompting

Type	GPT-4o	Claude 3.5	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 13: Most Frequent Misclassifications per Type using CoT Prompting

Type	GPT-4o	Claude 3.5	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 14: Most Frequent Misclassifications per Sarcasm Type using Emotion-Based Prompting

Sarcasm Type	GPT-4o	Claude 3.5	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