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

Anonymous authors
Paper under double-blind review

Abstract

In interactive systems, misreading sarcasm can undermine safety and robustness by causing models to interpret ironic remarks literally or generate unintended hostility. Sarc7 addresses this with a fine-grained, pragmatically grounded benchmark—built on MUStARD and annotated with seven distinct subtypes (self-deprecating, brooding, deadpan, polite, obnoxious, raging, manic)—that measures both classification and controlled generation performance. For classification, we compare zero-shot, few-shot, chain-of-thought (CoT), and a novel emotion-based prompting across five LLMs and find that emotion prompts boost macro-F1 compared to CoT prompting, reaching a highest of 0.3664 (Gemini 2.5).On generation, structured prompts defined by incongruity, shock value, context dependency, and emotion improve subtype alignment by 38.5% over zero-shot (Claude 3.5 Sonnet), enhancing interpretability and alignment with user intent. A human baseline (Cohen's κ = 0.6694, macro-F1 = 0.6663) further highlights persistent error modes in brooding, deadpan, and polite sarcasm. By quantifying model versus human performance and exposing alignment failures—bias toward "not sarcasm" or "deadpan"—Sarc7 advances transparency, explainability, and the safe deployment of LLMs where pragmatic understanding is critical.

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. Misunderstanding sarcasm can have real-world safety and alignment implications: an assistant that mistakes a genuine request for sarcastic mockery (or vice versa) can erode user trust, propagate harmful stereotypes, or even give unsafe advice under the wrong tone. By evaluating LLMs on fine-grained subtypes, Sarc7 not only probes robustness to subtle pragmatic cues but also surfaces alignment failures when models misinterpret intent.

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 & Gibbs (2000); Biswas et al. (2019) real-world agents require subtype sensitivity. 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 & Sagala (2022) subsequently applied this refined taxonomy in an irony and sarcasm detection system for public-figure speech.

Sarcasm Classification: Riloff et al. (2013) introduced a sentiment-contrast framework for binary sarcasm detection, flagging instances where positive wording clashes with negatively described contexts. 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

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. 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 trained annotators using our seven-type schema (see Table 4), 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 $\kappa = 0.6694$ (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 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.

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 4 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.

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 chose state-of-the-art LLM models, 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.

5 Results and Discussion

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

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

Table 2: Per-class Accuracy for Claude 3.5 using CoT vs. Emotion-based Prompting, Alongside Human Agreement.

5.1 Classification Results and Analysis

Table 10 confirms that chain-of-thought prompting yields the highest raw accuracy across models, while emotion-based prompting achieves superior macro-averaged F1 scores (Table 1), peaking at 0.3664 with Gemini 2.5. This gap reflects emotion-based prompting's strength on low-frequency classes—e.g. "Manic" and "Polite"—whereas CoT excels at overall reasoning. Given Sarc7's label imbalance (e.g. "Deadpan" dominates), macro-F1 offers a fairer assessment by equally weighting each subtype.

However, Figure 2 and Table 2 reveal a strong bias toward labeling uncertain cases as "Not sarcasm" or "Deadpan," underscoring models' reliance on surface cues rather than genuine pragmatic inference. Although emotion-based prompts boost performance on subtler categories (+3.0 % on Brooding, +17.5 % on Deadpan, +23.0 % on Polite, +16.7 % on Raging), they reduce accuracy on clear non-sarcastic inputs by 24.8 %, indicating a trade-off in robustness. In safety-critical agent applications, misclassifying neutral user language as "deadpan" could lead to inappropriate tone shifts or misunderstandings. Sarc7 therefore not only exposes these robustness and alignment failure modes but also demonstrates that emotion-informed prompting is a vital first step toward more resilient, context- and intent-aware sarcasm detection.

5.2 Prompt Technique Analysis

Emotion-based prompting, which explicitly models the listener's pragmatic hypothesis—"What emotion is intended here?"—yields higher macro-F1, demonstrating better performance on low-frequency sarcasm subtypes, indicating that discrete emotional cues guide LLMs toward the correct implicature when literal context is sparse. In contrast, CoT prompting excels at overall accuracy by simulating pragmatic inference, but can overlook subtler emotional distinctions; this trade-off underscores the need to balance structured reasoning with direct emotion signals when modeling conversational implicature in multi-class sarcasm. Sarc7's emotion-based prompting dramatically cuts misclassification of sarcastic input as non-sarcastic, critical for avoiding scenarios where an LLM fails to take urgent requests seriously or, worse, humorously "jokes" about emergencies.

5.3 Generation Results and Analysis

Emotion-based prompting generated more accurate sarcasm types. Table 3 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 3: Generation Evaluation Scores

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).

5.4 Safety, Robustness, and Bias

Sarcasm misclassification poses safety and alignment risks in downstream systems. For example, an AI agent that misconstrues "Oh, great, another meeting" as sincere approval could schedule further unwanted meetings, frustrating users or triggering policy violations in sensitive contexts (e.g., medical or legal advice). Robustness to tonal nuance is therefore a prerequisite for safe deployment.

Moreover, our reliance on English, Western-sourced dialogues may embed cultural biases: subtypes like "polite sarcasm" or "brooding sarcasm" may not map cleanly onto other languages or sociolects, leading to exclusion of non-Western speech patterns. Future work must both diversify training data and audit models for alignment failures across demographic and cultural groups, ensuring that sarcasm detection does not systematically misinterpret minority voices.

6 Conclusions

We present Sarc7, the first benchmark to distinguish seven nuanced sarcasm subtypes and to evaluate both detection and controlled generation. Sarc7 frames sarcasm understanding as a test of LLMs' pragmatic competence and their ability to reason about speaker goals and context-sensitive meaning. In classification experiments, emotion-based prompts raised macro-averaged F1 scores—reaching 0.3664 with Gemini 2.5—while chain-of-thought prompting achieved the highest overall accuracy. A human baseline (Cohen's $\kappa = 0.6694$) reveals that brooding, deadpan, and polite sarcasm remain the toughest subtypes to identify. For generations, structured prompts that specify incongruity, shock value, context dependency, and emotion improved subtype alignment by 38% over zero-shot prompts with Claude 3.5 Sonnet. By standardizing fine-grained sarcasm detection and controlled generation, Sarc7 can directly plug into AI agent stacks—serving both as a diagnostic for understanding user intent and as a parameterized engine for socially aware agent responses.

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.

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

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.

Ellen Riloff, Aditya Qadir, Prajakta Surve, Lakshika De Silva, Nisheeth Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714. ACL, 2013.

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.

Ari Tantra Zuhri and Rakhmat Wahyudin Sagala. Irony and sarcasm detection on public figure speech. *Journal of Elementary School Education*, 1(1):41–45, 2022. doi: 10.1234/joese.v1i1.13. URL https://journal.berpusi.co.id/index.php/joese/article/view/13.

A Annotations

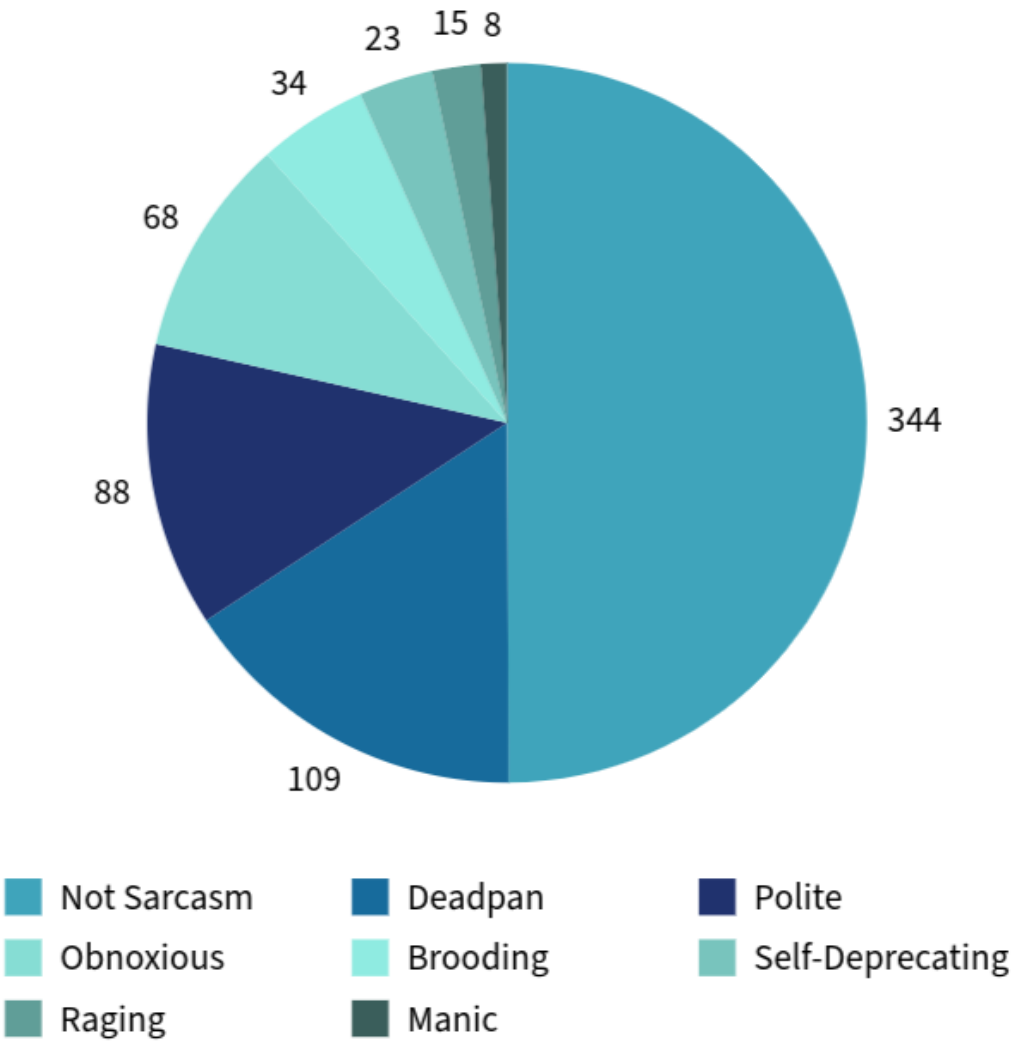

Figure 1: Distribution of Annotation Labels in the Dataset.

B Classification Statistics

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

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 under Xero-shot Prompting.

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

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 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 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 under Emotion-based Prompting.

C Generation Input

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.

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 10: Classification Accuracy Across Models and Prompting Techniques

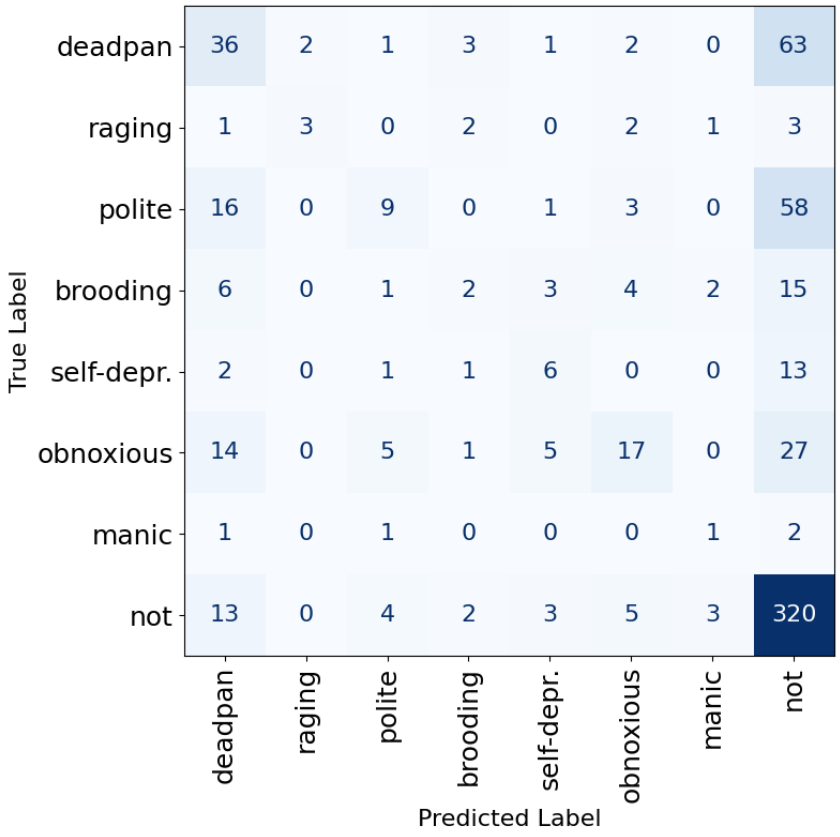

Figure 2: Confusion Matrix for Claude 3.5 Sonnet using CoT.

D 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!*

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 – 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]`

F Misclassification

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

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 11: 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	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 12: 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	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 13: 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	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

Table 14: Most Frequent Misclassifications per Type using Emotion-Based Prompting