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

Sarcasm Detection and Generation; Explainable AI (XAI); Controllable Generation; AI Safety; AI Alignment; Commonsense Reasoning

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. Accurately classifying and generating sarcasm is critical for improving large language models' (LLM) understanding of human intent. We introduce Sarc7, a benchmark for fine-grained sarcasm evaluation that grounds LLM analysis in linguistic knowledge by operationalizing 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 generation, we introduce a method for controllable generation along four pragmatic dimensions: incongruity, shock value, context dependency, and emotion. Sarc7 offers a foundation for evaluating nuanced sarcasm understanding, pushing beyond binary classification toward interpretable, knowledge-informed, and trustworthy language modeling. 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 is a fundamental challenge in commonsense reasoning, requiring an intuitive grasp of humor, social cues, and speaker intent (Yao et al., 2024; Gole, Nwadiugwu, and Miranskyy, 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, Nwadiugwu, and Miranskyy (2024), our technique aims to improve contextual relevance and expressive range in sarcastic generation. While binary sarcasm detection can flag an utterance for review, it cannot distinguish between playful banter and hostile mockery. This distinction is not merely academic; it is critical for AI safety and building trustworthy systems. An agent that misinterprets hostile sarcasm as a joke, or vice-versa, can erode user trust and lead to harmful interaction dynamics. Our multi-class approach provides the necessary granularity for an AI to navigate these social complexities safely. Our multi-class approach provides crucial insight into the speaker's underlying intent, which is essential for any system aiming for deep pragmatic understanding.

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). Leggitt and Gibbs (2000); Biswas, Ray, and Bhattacharyya (2019). 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. Building on this lineage, we translate those high-level categories into concrete, example-driven definitions and detailed annotation guidelines to construct and evaluate our Sarc7 benchmark for LLMs.

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.

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, Zhou, and Li (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 and Crossley (2018). However, existing techniques struggle with controlling sarcasm levels and aligning them with contextual incongruence, shock value, and prior context dependency.

3 Methods

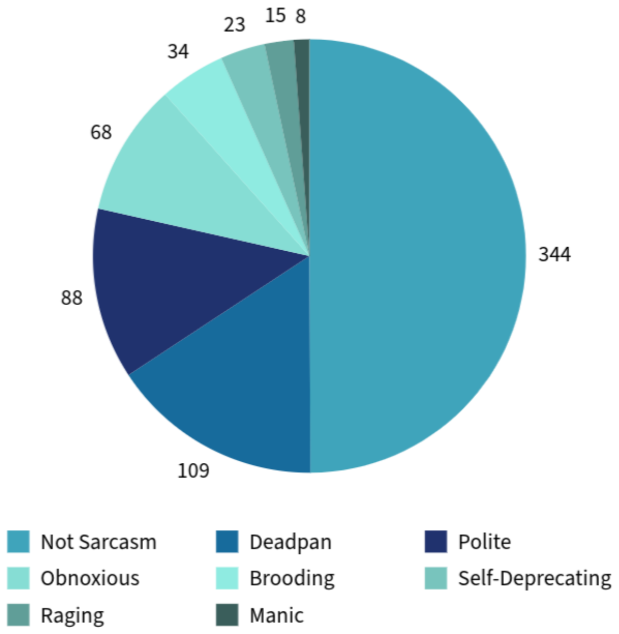

Figure 1: Distribution of Annotation Labels in the Dataset.

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.

3.2 Annotation Methodology

Each sarcastic utterance in the MUStARD dataset (n=690) was independently labeled by four trained annotators using the seven sarcasm subtypes defined in Sarc7. Annotators were instructed to consider pragmatic cues and 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.

To quantify the reliability of our 3-of-4 consensus labels, we recruited a fifth trained annotator to re-label all utterances independently. We then computed Cohen's kappa between the majority vote (from the original four annotators) and this fifth annotator's labels. The resulting Cohen's $\kappa = 0.6694$ indicates substantial agreement according to Landis and Koch (1977) scale. The macro-averaged precision, recall, and F1 for this human comparison were 0.6586, 0.6847, and 0.6663, respectively. This provides further evidence that our annotation schema is both consistent and replicable.

Even for trained readers, **brooding**, **deadpan**, and **polite** sarcasm proved the most challenging to label consistently, establishing realistic upper bounds for model performance on these subtypes.

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 1 outlines

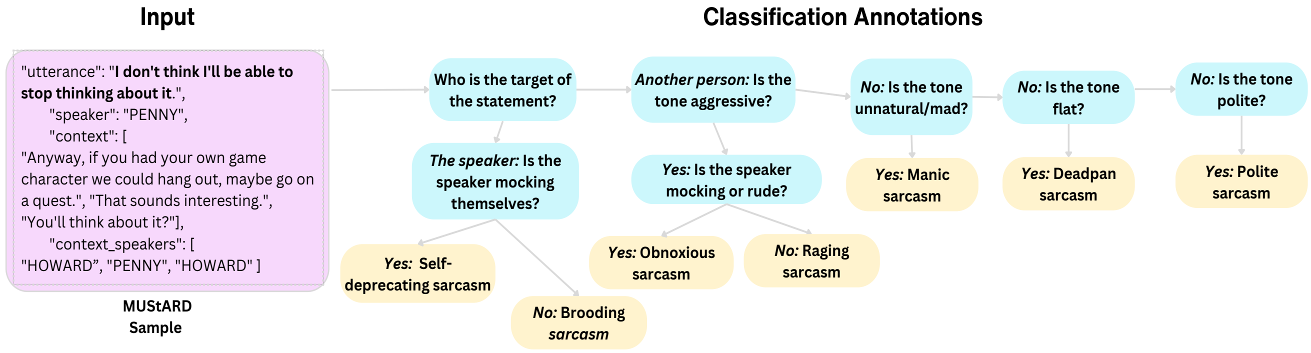
Input
Classification Annotations
"utterance": "I don't think I'll be able to stop thinking about it.",
"speaker": "PENNY",
"context": [
"Anyway, if you had your own game character we could hang out, maybe go on a quest.", "That sounds interesting.", "You'll think about it?"],
"context_speakers": [
"HOWARD", "PENNY", "HOWARD"]
MUStARD Sample
Who is the target of the statement?
The speaker: Is the speaker mocking themselves?
Yes: Self-deprecating sarcasm
No: Brooding sarcasm
Another person: Is the tone aggressive?
Yes: Is the speaker mocking or rude?
Yes: Obnoxious sarcasm
No: Raging sarcasm
No: Is the tone unnatural/mad?
Yes: Manic sarcasm
No: Is the tone flat?
Yes: Deadpan sarcasm
No: Is the tone polite?
Yes: Polite sarcasm

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

definitions for each sarcasm category in the Sarc7 benchmark.

3.4 Baseline Classification

Our baseline testing focused on zero-shot, few-shot, and chain-of-thought (CoT) prompting.

- Zero-shot: The model classifies the utterance with only a definition of the sarcasm types and no examples.
- Few-shot The model is provided with the defintions of the sarcasm types and a few examples of correct classifications within the prompt to guide its response.
- CoT: The model is provided with the definitions of the sarcasm types and is prompted to break down its reasoning into steps, with examples that also show the reasoning process.

Our novel emotion-based prompting method is detailed separately in Section 3.5, as it introduces a unique reasoning framework based on affective incongruity.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

To make the model's pragmatic reasoning more explicit and explainable, our emotion-based prompting method operationalizes the detection of emotional incongruity. This can be viewed as a pragmatic consistency check, where the model must reason about the expected emotion of a context versus the expressed emotion of an utterance Our emotion-based prompting goes beyond traditional sentiment analysis by leveraging discrete emotion categories rather than coarse positive/negative polarity. This method captures pragmatic incongruity through emotional mismatches, approximating listener inference. Whereas sentiment classifiers typically flag a mismatch between overall sentiment and context Riloff et al. (2013), our approach leverages 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—such as polite sarcasm pairing happiness with a neutral situation or obnoxious sarcasm pairing neutral context with a superficially disgusting utterance—that a simple positive/negative split cannot distinguish. This fine-grained emotional reasoning provides a clear advantage for multi-class sarcasm classification: it supplies subtype-specific cues (e.g., "raging" sarcasm requires anger, "manic" requires surprise or happiness) and thus helps disambiguate among several closely related sarcasm types rather than collapsing them all into a single sarcastic category.

3.6 Generation Dimensions

A key pillar of explainability and controllability in LLMs is the ability to steer their outputs in a predictable manner. 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 2). We opted for fixed values for each subtype to create a controlled and interpretable baseline for generation. This approach allows us to directly test a model's ability to adhere to explicit pragmatic instructions, whereas a data-driven approach would conflate feature extraction with generation quality. 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.

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

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

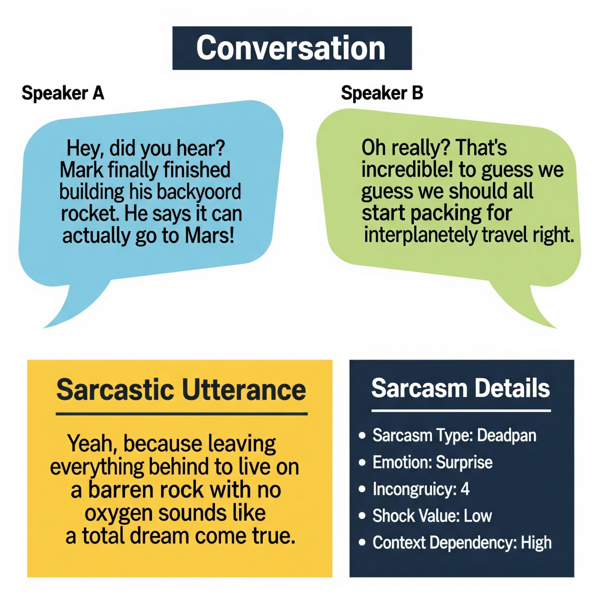
Conversation
Speaker A
Speaker B
Hey, did you hear? Mark finally finished building his backyoord rocket. He says it can actually go to Mars!
Oh really? That's incredible! to guess we guess we should all start packing for interplanetely travel right.
Sarcastic Utterance
Yeah, because leaving everything behind to live on a barren rock with no oxygen sounds like a total dream come true.
Sarcasm Details
• Sarcasm Type: Deadpan
• Emotion: Surprise
• Incongruity: 4
• Shock Value: Low
• Context Dependency: High

Figure 3: Sample Output Using Emotion-based Generation Method

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.

5 Results and Discussion

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 3 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 4, 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

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 3: Classification Accuracy Across Models and Prompting Techniques

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

of detailed reasoning steps.

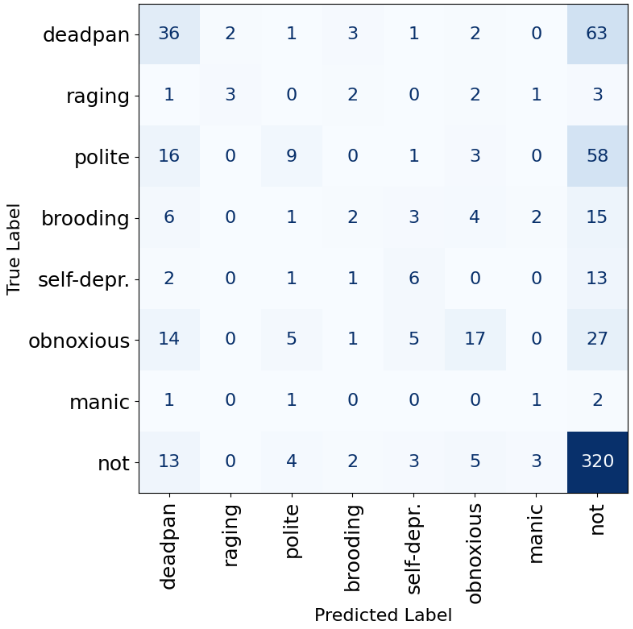

Figure 4: Confusion Matrix for Claude 3.5 Sonnet using CoT.

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.

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

Table 5 shows that emotion-based prompting yields consistent relative improvements over CoT prompting, though absolute accuracy remains below the human ceiling. In particular, brooding gains +3.04%, polite +23.0 %, deadpan +17.47 %, and raging +16.67 %, demonstrating that emotion cues help disambiguate more subtle tones. Conversely, “not sarcasm” drops by –24.82 %, indicating that adding emotion information can sometimes introduce noise for clear non-sarcastic cases. These shifts confirm that emotion-based prompts move the model closer to human-level nuance on mid-difficulty classes, but the largest remaining gaps still align with the hardest human distinctions—especially brooding, deadpan, and polite sarcasm—suggesting the need for richer contextual and pragmatic reasoning beyond fixed emotion settings.

While emotion-based prompting significantly boosts the macro-averaged F1 score by improving performance on rare subtypes, this comes at the cost of misclassifying non-sarcastic text more often. This suggests that adding emotional cues makes the models more ’trigger-happy’ in their sarcasm detection, highlighting a critical precision-recall trade-off that must be considered in real-world applications where false positives can be problematic.

From a pragmatic standpoint, these patterns show that fixed emotion cues can help LLMs avoid the default “deadpan” trap in nuanced cases, but true conversational implicature often depends on richer context and iterative hypothesis testing. The persistent gaps on brooding, deadpan, and polite highlight subtypes whose disambiguation relies heavily on prosodic and interpersonal cues—elements our current text-only prompting cannot capture. This trend reveals the model’s high uncertainty when faced with ambiguous inputs. This highlights the need for models that can not only classify sarcasm but also express when they are uncertain. Developing such capabilities is a crucial step toward the automated verification of an LLM’s pragmatic understanding. Future work should integrate dialogue history, world knowledge, or multimodal signals to approximate the full pragmatic reasoning humans employ.

5.3 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.

One possible explanation for few-shot prompting achieving a higher macro-F1 score than CoT, despite CoT’s higher overall accuracy, is its directness. The concrete examples in few-shot prompts may provide a stronger signal for rare classes, which a macro-F1 score weights heavily. In contrast, the abstract reasoning steps of CoT may inadvertently bias the model towards the more frequent ’deadpan’ or ’not sarcastic’ labels.

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 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 6 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 6: 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. By structuring generation through pragmatic dimensions like context dependency and incongruity, our method implicitly guides the model to replicate speaker goals. See Appendix **??** for examples’ context. Notably, brooding and manic sarcasm were the toughest for LLMs to generate. Brooding depends on a courteous veneer masking genuine frustration, a nuance carried by tone and pacing, not keywords, so single-turn prompts either over-polish or slip into blunt reproach. Manic sarcasm requires sustained, erratic enthusiasm that signals insincerity through vocal intensity; without prosody, models fall back on generic hyperbole. In both cases, missing nonverbal and contextual cues hinder authentic reproduction. Future work might integrate audio–text alignment or fine-tune on prosody-annotated dialogues to better capture these complex styles.

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 3 and 4). Our primary goal in this benchmark was 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 Real-World and Agent Applications

The ability to distinguish nuanced sarcasm subtypes has significant real-world applications, particularly in enhancing the social and pragmatic competence of AI systems. Our Sarc7 benchmark and emotion-informed methods can be directly integrated into several domains:

- **Advanced Conversational Agents:** Beyond simple chatbots, sophisticated AI agents could leverage Sarc7 for more natural human-computer interaction. The classification framework can help an agent understand a user's true intent behind a sarcastic remark, preventing literal misinterpretations. For instance, recognizing *self-deprecating* sarcasm could allow an agent to respond with appropriate humor, while identifying *raging* sarcasm could trigger de-escalation protocols. The generation engine enables an agent to use sarcasm in a controlled manner, such as employing *polite* sarcasm to gently refuse a request or using playful banter to build rapport. In multi-agent AI systems, robust communication is key. An agent equipped with our framework could interpret sarcastic cues from other agents or humans, preventing misunderstandings that could derail cooperative tasks. This capability for nuanced pragmatic understanding is essential for more effective and human-like collaboration in multi-agent AI frameworks.
- **Nuanced Content Moderation and Toxicity Detection:** Sarcasm is deeply intertwined with toxicity, humor, and playfulness. Automated moderation systems often struggle to differentiate between malicious insults and sarcastic teasing (e.g., *obnoxious sarcasm*) between friends. By classifying the type of sarcasm, a system could make more informed decisions, reducing false positives that erode user trust and better identifying genuinely harmful content. This increases the overall trustworthiness and fairness of automated moderation systems.
- **Reliable Sentiment Analysis:** Standard sentiment analysis tools are notoriously brittle in the face of sarcasm, often misinterpreting a negative opinion cloaked in positive words. For example, a product review stating, "This battery life is absolutely amazing, it lasts a whole 20 minutes!" would be incorrectly flagged as positive. Our framework allows for a more granular analysis; identifying the utterance as sarcastic immediately flips the polarity, leading to more accurate metrics for market research, brand monitoring, and public opinion tracking.

By providing a foundation for both detecting and generating subtype-specific sarcasm, Sarc7 enables the systematic benchmarking and development of more socially intelligent AI systems capable of navigating complex pragmatic phenomena.

6 Conclusions

We present Sarc7, the first benchmark to distinguish seven nuanced sarcasm subtypes and to evaluate both detection and controlled generation. Sarcasm, as a fundamentally pragmatic act, depends on interpreting intent, emotional incongruity, and social context beyond surface form. 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 benchmarking both model and human performance, Sarc7 demonstrates LLMs' ability to handle intentional, socially informed sarcasm and lays the groundwork for deeper pragmatic reasoning. Moving beyond binary detection to fine-grained, context-sensitive inference and generation, it enables more natural, emotion-aware dialogue agents and supports future multimodal and cross-lingual extensions.

6.1 Future Work

Our analysis opens several avenues for future research. Future work should explore hybrid prompting strategies that combine the structured reasoning of CoT with the targeted cues of emotion-based prompting to potentially achieve both high accuracy and a strong F1 score. To improve alignment with human social norms, the preference ratings from our generation evaluation could be used to fine-tune models via techniques like Direct Preference Optimization (DPO). This would directly leverage human feedback to create agents that generate more appropriate and context-aware sarcasm. Furthermore, a deeper error analysis is needed to address persistent misclassifications, especially the confusion between polite, brooding, and deadpan sarcasm. Developing robust mechanisms to reduce the misclassification of non-sarcastic text is particularly critical for improving the reliability and safety of these models in real-world applications.

7 Limitations

Our evaluation also surfaced key limitations to guide future work. First, while the process for annotating the MUStARD dataset had a rigorous structure, and annotations were peer-reviewed for consistency, there is still room for annotator disagreement. Second, our forced single-label scheme and skewed class distribution (e.g. abundant deadpan vs. scarce manic examples) bias both annotation and model defaults; multi-label annotations and data balancing (e.g. weighted loss, augmentation) could mitigate this. Third, relying on Ekman's six basic emotions overlooks finer affective states (irony, embarrassment) and may not transfer across languages or cultures—MUStARD's English-only dialogues underscore the need for richer emotion taxonomies and cross-lingual validation. Finally, and most critically, our evaluation is constrained by its reliance on purely textual data. Sarcasm is a fundamentally multimodal phenomenon, where meaning is often conveyed through non-textual cues like prosody (tone, pitch, and pacing), facial expressions, and gestures. The difficulty in generating authentic brooding or manic sarcasm, for instance, stems directly from the absence of vocal intensity and tonal nuance in text-only prompts. The persistent confusion between sincere statements and deadpan sarcasm further underscores this limitation, as the flat affective tone that defines this subtype is primarily an audio-visual cue. The absence of this multimodal context imposes a natural ceiling on the performance of any text-based system and is a key factor behind the modest classification accuracies observed. Future work must move towards integrating multimodal signals to capture the full pragmatic richness of human communication.

8 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.

A 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]`

CoT 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]`

References

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

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

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4619–4629. Florence, Italy: Association for Computational Linguistics.

DeepMind, G.; Anil, R.; Arolfo, S.; Babuschkin, I.; Beyer, L.; Bosma, M.; and ... 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

Gole, M.; Nwadiugwu, W.-P.; and Miranskyy, A. 2024. On sarcasm detection with openai gpt-based models. In *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, 1–6. IEEE.

Helal, N. A.; Hassan, A.; Badr, N. L.; and Afify, Y. M. 2024. A contextual-based approach for sarcasm detection. *Scientific Reports* 14(1):15415.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1):159–174.

Lee, J.; Fong, W.; Le, A.; Shah, S.; Han, K.; and Zhu, K. 2024. Pragmatic metacognitive prompting improves llm performance on sarcasm detection. *arXiv preprint arXiv:2412.04509*.

Leggitt, J. S., and Gibbs, R. W. 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*.

Qasim, S. A.-M. 2021. A critical pragmatic study of sarcasm in american and british social interviews. *Journal of Strategic Research in Social Science*.

Riloff, E.; Qadir, A.; Surve, P.; De Silva, L.; Gilbert, N.; and Huang, R. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 704–714. ACL.

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

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

Yao, B.; Zhang, Y.; Li, Q.; and Qin, J. 2024. Is sarcasm detection a step-by-step reasoning process in large language models? *arXiv preprint arXiv:2407.12725*.

Zhang, Y.; Zou, C.; Lian, Z.; Tiwari, P.; and Qin, J. 2024. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *arXiv preprint arXiv:2408.11319*.

Zhuang, X.; Zhou, F.; and Li, Z. 2025. Multi-modal sarcasm detection via knowledge-aware focused graph convolutional networks. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Zuhri, A. T., and Sagala, R. W. 2022. Irony and sarcasm detection on public figure speech. *Journal of Elementary School Education* 1(1):41–45.