Can LLMs Get Sarcasm?

Group 33
Jeffrey **Kim** (37364585)
Miso **Lee** (36594455)
Yiou **Li** (89673826)
Ina **Lee** (72308307)

Introduction

When someone says, "Oh, great, another Monday!" we don't interpret that as meaning they're having a wonderful Monday. The phrase subtly provokes annoyance, discontent, or even laughter. But can LLMs (Large-Scale Language Models) or artificial intelligence truly understand this sarcasm?

Sarcasm isn't simply a choice of words. It is a sophisticated linguistic skill that combines human social intelligence, emotions, and cultural context. While LLM can generate sarcastic sentences based on extensive data and patterns, it can't actually feel, express, or fully understand the intent and emotion behind them.

However, the question of whether LLMs can truly understand sarcasm remains unanswered. LLMs can learn from vast amounts of text data to generate and, to a certain extent, categorise sentences that look like sarcasm, but they lack the ability to actually sense or infer the emotional or social implications behind them. In other words, while human sarcasm comprehension requires social intelligence, emotional processing, and understanding of relational context, LL

Ms rely on statistical pattern recognition to mimic output.

Because sarcasm often differs from literal language, the ability to comprehend it is crucial, as it can pose serious risks beyond simple misunderstanding. The Adam Raine incident in 2025 exemplifies this extreme. In this case, GPT-4o failed to interpret the user's emotional distress and danger signals, leading to misguided empathy or inappropriate advice that worsened the situation.

While not directly related to sarcasm itself, this case highlights the serious problems that can arise when LLMs misinterpret hidden human intentions, emotions, and context. This vulnerability goes beyond mere technical limitations and leads to social conflicts that directly impact stability, reliability, and ethics.

Therefore, this study analyses how LLMs handle sarcasm and how they differ from human language and social reasoning. By doing so, it aims to identify areas where LLM can and cannot understand human intentions and emotions. This comparison clearly reveals the structural limitations of LLM and is necessary to redefine what it actually means to say that "LLM understands sarcasm." Understanding sarcasm is a key indicator of LLMs' social and linguistic literacy, and research on it is essential for AI stability and risk reduction in real-world environments.

Background Research

Researchers in both linguistics and computer science have studied sarcasm for a long time, but most of this work has focused on detecting sarcasm rather than understanding it. Early computational approaches mostly looked for surface cues like specific words, polarity shifts, or

even hashtags. For example, Joshi et al. (2017) reviewed many sarcasm detection systems and found that most models were trained on simple signals such as contrast between positive and negative words. These systems could label something as sarcastic, but they didn't capture why it was sarcastic or what the speaker actually meant.

Later research started using more context. Ghosh et al. (2017) showed that including conversation history improved sarcasm detection in social media. Their LSTM and attention-based models looked at previous turns in the dialogue to catch mismatches between literal wording and the overall tone. This was an improvement, but the system was still relying on statistical patterns rather than real understanding of speaker intention.

Oprea and Magdy (2020) introduced the iSarcasm dataset, where people labeled their own sarcastic tweets. This was meant to fix problems with hashtag-based datasets, which often mislabeled tweets or failed to capture true intent. However, even this newer dataset still supported models that predicted sarcasm labels but did not explain human-like interpretation.

On the linguistics and cognitive science side, researchers emphasize that human sarcasm depends on pragmatic inference, intention-reading, and multimodal processing. Classic studies such as the McGurk effect (1976) and Gick & Derrick (2009) show that humans naturally combine auditory, visual, and even tactile cues when interpreting language. These findings suggest that human understanding is embodied and context-sensitive, while LLMs only work with text and lack access to sensory or intentional information. This creates a theoretical gap between how humans understand sarcasm and how AI systems process it.

Overall, the existing literature shows two clear limitations:

1. computational models focus on detection rather than true comprehension, and

2. human sarcasm relies on multimodal and social cues that LLMs currently cannot use.

These gaps motivate our project, which compares human sarcasm judgments with LLM outputs across different levels of context to see whether modern language models show any signs of pragmatic reasoning or whether they still rely purely on pattern-based detection.

Overview of Study Design

The primary focus of our study is to test whether LLMs have the ability to correctly interpret sarcasm under different prompting strategies. The design of our experiment will be a 3 x 3 factorial implementing 3 prompting styles and 3 different models. The prompting styles will consist of: either zero-shot, few-shot, or chain-of-thought prompting. Additionally, the models we will be testing are OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, each at their free tier, which are amongst the best in the artificial intelligence market. The design allows us to measure how instructional scaffolding and model architecture affect LLMs' interpretation of sarcastic text.

Materials

A dataset will consist of 50-80 “items,” these items will be the foundation to the prompts being delivered. By having a large variety of items, this ensured sufficient variability in cue types and context while remaining manageable for manual data collection and analysis. The items will include: a target utterance (e.g., “Oh great, another Monday”), a context manipulation, a ground-truth sarcasm label (i.e., whether it was sarcastic or not), and finally a cue type representing known sarcasm mechanisms (e.g., echoic mention, hyperbole, understatement, etc.). The context manipulations will be the amount of contextual information offered in the prompt and can be the prompt alone, light context, or a full thread context (a short conversation excerpt). Using this structure, we can reflect on common linguistic distinctions between sentence meaning vs. speaker meaning which allows us to test how much context LLMs need before sarcasm becomes interpretable. All models will be accessed through their free-tier public interfaces, offering both a practical resource constraint and an ecologically valid choice. This constraint ensures that there is equal accessibility across group members, lower funding needs, and allows this experiment to be easily replicable for future researchers. Moreover, the free-tier will be the model that is most generally used by the majority of the public, contributing to a consistent baseline performance.

Procedure

The prompting conditions, as mentioned before, will consist of three prompting styles: zero-shot prompting is when the model only receives the target item, few-shot prompting is when the model is shown 3 example items with correct sarcasm labels before the target item, and chain-of-thought prompting is when the model is asked to provide reasoning steps before giving its final label. Chain-of-thought prompting, in addition to its given advantage, is essential to this study as the AI may label an item correctly, yet still have an incorrect thought process. The usage of multiple prompting styles allows us to test whether structured guidance improves the model’s ability to recover speaker intention. There will be a total of 9 experimental conditions (3 prompting styles x 3 models) and the fully crossed design allows us to compare differences due to model architecture, prompting structure, and their interaction, while maintaining consistent input. To present further consistency and avoid contamination from previous interactions, each prompt will be entered into a fresh chat session. This prevents memory carryover, the practice effect, and ensures that all items are evaluated independently. Overall, this procedure improves internal validity and helps isolate the effect of the prompt itself.

Analysis

For every response, we will collect: a sarcasm classification, the explanation quality (for chain-of-though prompting styles), a confidence estimate, and notes on irrelevant reasoning. These components will require both quantitative accuracy analysis and qualitative evaluation of pragmatic reasoning. A trained group of 5 human raters will judge these results and evaluate based on a rubric. This criterion will be based on accuracy of the model’s classification, appropriateness of the explanation (e.g., does it use the correct pragmatic cues?), and calibration to assess whether the model’s confidence is appropriate for the difficulty of the item. By using human evaluation, it offers a baseline analytical level and is crucial for identifying whether the model’s reasoning is genuinely pragmatic or simply pattern matching.

Funding and Ethical Considerations

Although models are accessed through their free-tier interfaces, the study still requires funding to support the development of high-quality stimuli, compensate training for human raters, and hire research assistants to manually collect, organize, and evaluate model outputs. Ethically, the study requires informed participation from all staff and fair compensation, as well as secure handling of all collected data.

Discussion

Our choice of method is motivated by both the pragmatic properties of sarcasm and the prompt-driven nature of LLMs.

Although the interpretation of sarcasm relies heavily on context, intention and cues, humans can understand several isolated sarcasm since we have preconceived judgement about them. For instance, humans consider "Oh, great, another Monday" as sarcasm since Mondays represent the end of rest and the start of work or study and we can reason out the intention behind this sentence. Given that LLMs lack such shared experience, our method aims to figure out to what extent they can understand isolated sarcasm, and how their performance in identifying sarcasm varies with different amounts of contextual information.

In addition, we seek to learn in which way LLMs identify sarcasm, and whether their interpretation of sarcasm aligns with humans'.Our method is to test them with three prompting styles. The variation of their performance regarding different styles indicates whether they understand sarcasm or they replicate online answers. Among the three styles, zero-shot prompting reflects the most common usage. Performance in this condition indicates LLMs' capacity to identify sarcasm in everyday conversations. By contrast, chain-of-thought prompting guides LLMs to reason step by step, allowing us to learn the upper bound of their capacity to interpret sarcasm.

We predict three main patterns in results. First, we expect LLMs' performance to improve as the amount of contextual information increases, while the control group maintains high accuracy regardless of the amount of contextual information. This would suggest that LLMs rely more heavily on context than humans do. Second, we expect LLMs' performance to improve as prompting becomes more specific. However, their reasoning process might not be the same as humans'. Third, we expect different models to adopt different strategies for identifying sarcasm. One model might aggressively identify a sentence as sarcasm, which would lead to higher false-positive rate, while another model might take a more conservative approach.

This study contributes to linguistics by adapting different amounts of contextual information and different cues to semi-quantify the contextual threshold, and how these factors affect the interpretation of sarcasm.

It contributes to computer science by systematically comparing the performance of humans and LLMs on understanding sarcasm, and pointing out that LLMs might benefit from additional training focusing on the difference between literal meaning and intended meaning, which strengthen their capacity in natural-language processing.

The limitation of this study is located in the format. Our experiment is text-only and could not examine LLMs' capacity of identifying sarcasm by reading users' facial expression and tone. In the future, we could carry out an experiment using the voice mode and video mode

if applicable.

Conclusion

Through this project, we began to think more seriously about what it means for LLMs to identify and interpret language. Sarcasm is so common in human society that we often overlook its pragmatic complexity. By predicting LLMs' performance under different contextual conditions, we became more aware of the significance of context in interpreting sarcasm. In addition, the reasoning behind sarcasm, which in everyday life seems to happen instantaneously, was broken down and planned to be examined in our study, allowing us to see the inherent process regarding sarcasm more clearly. Our findings would suggest that LLMs' understanding of sarcasm relies heavily on context, and that their reasoning about sarcasm might not align with humans' Overall, LLMs can understand sarcasm to some extent, but there is still a substantial gap between them and humans in both accuracy and reasoning process.

Moreover, we learned how to design primary research. In the earliest version of our design, we did not include a human control group and did not specify evaluation measures. These issues were addressed after revising according to other groups' feedback. As a result, we became more familiar with the process of designing primary research. At the same time, this project changed the way we think about LLMs. Even though we did not actually carry out the experiment, the design process pointed out several potential areas where LLMs could be improved.

Bibliography

Ghosh, D., Fabbri, A. R., & Muresan, S. (2017), The Role of Conversation Context for Sarcasm Detection in Online Interactions. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.48550/arxiv.1707.06226

Gick, B., & Derrick, D. (2009), Visual-tactile integration in speech perception: Evidence for modality-neutral speech primitives. *The Journal of the Acoustical Society of America*, 126(4), 2054–2062. https://doi.org/10.1121/1.3212921

Joshi, A., Bhattacharyya, P., & Carman, M. (2017), Automatic Sarcasm Detection: A Survey. *ACM Computing Surveys*, 50(5), 1–22. https://doi.org/10.1145/3124420

McGurk, H., & MacDonald, J. (1976), Hearing lips and seeing voices. *Nature*, 264, 746–748. https://doi.org/10.1038/264746a0

Oprea, S. V., & Magdy, W. (2020), iSarcasm: A Dataset of Intended Sarcasm. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 1279–1289. https://doi.org/10.48550/arxiv.1911.03123