Language models achieve human-level sarcasm detection

Lili Tamás
Károli Gáspár University
Budapest, Hungary
lili.tamas.contact@gmail.com

Mariann Lengyel
Language Technology Research Group
ELTE Research Centre for Linguistics
Budapest, Hungary
lengyel.mariann@nytud.elte.hu

Noémi Ligeti-Nagy, PhD
Language Technology Research Group
ELTE Research Centre for Linguistics
Budapest, Hungary
ligeti-nagy.noemi@nytud.elte.hu

Abstract

We report a comparative evaluation of two advanced language models, GPT-4 and OpenAI o3, alongside a human baseline on a sarcasm comprehension benchmark. The corpus consists of several hundred paired narratives rendered in neutral and sarcastic forms, which are assessed under both closed (yes/no) and open (wh-question) prompts. Model responses were elicited via zero-temperature API calls with uniform system and user prompts; human annotations were gathered through randomized surveys and inter-annotator agreement was quantified. Evaluation focuses on overall accuracy and a sarcasm-specific F1 metric. GPT-4 demonstrates strong narrative understanding but shows a notable decrease in performance on the sarcastic condition compared to neutral context questions. o3 attains performance on par with, or exceeding, the human baseline across both prompt types, with particularly high effectiveness in identifying non-literal intent. Humans maintain consistent accuracy across formats, suggesting that constrained answer options pose greater difficulty for models than for people. Further analysis reveals that closed-form prompts tend to amplify literal interpretations, especially for GPT-4, whereas open-ended prompts afford greater flexibility in capturing subtle intent. The gap between context comprehension and sarcasm detection underscores the challenge of pragmatic inference in language models. These findings indicate that instruction-tuned architectures can approach human-level sarcasm understanding under controlled conditions, yet targeted strategies such as chain-of-thought prompting, few-shot examples, and richer contextual cues, may be necessary to fully close remaining performance gaps in non-literal language comprehension.

1 Introduction

Large generative language models (LLMs) have seen a meteoric rise in recent years, revolutionizing natural language processing (NLP) and human-computer interaction. Models such as OpenAI's GPT-4 (OpenAI, 2023), Anthropic's Claude, Google's Gemini, and Meta's LLaMA (Touvron et al., 2023) can produce fluent, contextually relevant text, enabling applications from conversational agents to writing assistants. Their deployment in widely used chatbots (e.g., ChatGPT) has raised new questions about the depth of their language understanding capabilities. In particular, as these systems increasingly mediate everyday communication, there is growing interest in whether they grasp the nuances of non-literal language in human dialogue.

One prominent challenge in this realm is sarcasm detection. Sarcasm is a form of verbal irony in which the intended meaning of an utterance is opposite to its literal wording to mock or convey contempt. Detecting sarcasm is a well-known linguistic and cognitive challenge: it requires the interpreter to integrate contextual cues, tone, and real-world knowledge to infer the speaker's true intent. Humans

The code, the full dataset with the models' answers, and the human answers will be publicly released upon acceptance.

naturally leverage prosody, facial expressions, and shared context to recognize sarcasm, and even then misunderstandings can occur in text-only settings. In cognitive studies, understanding sarcasm has been shown to engage theory-of-mind reasoning and social cognition abilities (Capelli et al., 1990; Shamay-Tsoory & Aharon-Peretz, 2005). For instance, children and even adults often rely on context or intonation to identify sarcasm (Capelli et al., 1990), and developmental research indicates that sarcasm comprehension improves with linguistic and social maturity (Fanari et al., 2023).

Despite the remarkable success of LLMs on many tasks, interpreting non-literal language like sarcasm remains a challenge. LLMs primarily learn from literal text patterns; they do not directly perceive tone or facial cues, and they lack genuine social awareness. As a result, state-of-the-art models often struggle to recognize sarcastic intent, especially in the absence of explicit markers. In fact, current LLMs underperform earlier specialized sarcasm classifiers on standard benchmarks (Zhang et al., 2024), suggesting that these large models have not fully solved the nuance of sarcasm understanding. For example, a comprehensive study by Zhang et al. (Zhang et al., 2024) found that supervised sarcasm detection models (fine-tuned on task-specific data) outperformed general-purpose LLMs across multiple sarcasm datasets. Similarly, other work noted that prompting GPT-3.5 or GPT-4 to identify sarcasm yields inconsistent results unless carefully tuned, and that chain-of-thought reasoning approaches (so effective in math or logic tasks) provide little benefit for sarcasm comprehension (Tan et al., 2023; Zhang et al., 2024). This gap is unsurprising – recognizing sarcasm often demands pragmatic inference beyond what large-scale language modeling alone can capture. Given how prevalent sarcasm is in human communication, a failure to handle it can lead to miscommunication in human–AI interaction.

Generative LLMs like GPT-3.5 and GPT-4 represent a fundamentally different paradigm compared to traditional NLP approaches. Rather than being trained explicitly for sarcasm identification, they learn from vast generic text corpora and are later aligned via instruction-tuning or reinforcement learning from human feedback (Ouyang et al., 2022). They can be asked to detect or explain sarcasm in a zero-shot manner (via prompting) without additional fine-tuning. This raises several open questions: Do LLMs implicitly know how to detect sarcasm from their pretraining? How does their sarcasm comprehension compare to that of earlier, dedicated sarcasm detectors or to humans? Early evidence suggests that while LLMs have some latent ability (likely from encountering sarcastic dialogues in training data), they may not reliably apply it. For instance, one study reported that a fine-tuned GPT-3 model achieved over 80% F_1 on a sarcasm corpus, outperforming prior systems, whereas GPT-4 in zero-shot mode reached a lower 75% F_1 on the same task (Gole et al., 2024). This indicates that targeted training can still surpass general intelligence in this task, and that different LLMs vary in their aptitude: indeed, GPT-4 tends to outperform GPT-3.5 on irony understanding (Zhang et al., 2024), and newer models like Claude or Bard (Gemini) have shown mixed results in preliminary comparisons.

In this paper, we present a comparative study of sarcasm comprehension in two advanced LLMs and a human baseline. We focus on two generations of GPT models – a GPT-4 (representative of the ChatGPT family) and o3 – and evaluate their ability to detect and understand sarcastic utterances. To ensure a fair and meaningful comparison, we adopt an updated evaluation methodology inspired by research in human sarcasm understanding. Rather than relying solely on brief, out-of-context social media snippets, we use context-rich scenarios and dialogues derived from psychology and linguistics studies of sarcasm (Capelli et al., 1990; Fanari et al., 2023). Each model is presented with short narratives or conversational contexts in which a critical utterance may or may not be sarcastic, and we prompt the model with questions analogous to those given in human experiments (e.g., asking what the speaker really meant, or whether an utterance was sincere). We also include a controlled set of simpler sarcasm benchmark sentences from prior NLP datasets for completeness. A group of human annotators serve as a baseline to benchmark model performance against human-level comprehension.

The contributions of our work are threefold. First, we provide a systematic evaluation of GPT-4 and o3 on the task of sarcasm detection and comprehension, using both standard sarcasm datasets and novel context-driven tests. We quantify how well each model recognizes sarcastic intent, highlighting performance gaps between the two model generations. Second, we incorporate human performance as a reference point, comparing the models' outputs to human judgments to assess how close the models come

to human-like understanding. This human baseline, along with our context-rich evaluation, offers new insights beyond the traditional accuracy metrics. Third, we analyze the results to identify where and why these LLMs fail or succeed in interpreting sarcasm. Through qualitative examples and error analysis, we uncover common pitfalls (such as models taking sarcastic statements at face value or misreading subtle cues) and discuss what these reveal about the models' internal representations of language.

2 Methods

To systematically evaluate GPT-4 and OpenAI o3 in sarcasm detection, we tested both LLMs on 448 tasks. These tasks utilized 56 unique base stories: context-rich short narratives or conversational contexts. Each base story consisted of a social situation and a short ending remark from one character, either as the only utterance or as part of a brief dialogue. Two variants of each base story were included in our task sheet. Both versions of the same story shared the characters, setting, and ending remark, but differed in whether the ending remark was literal or sarcastic, depending on the specific events of the variant. We called these the neutral and sarcastic versions, belonging in the same group. The neutral and sarcastic versions of the same story differ only to the degree necessary to make the remark either literal or sarcastic. For many tasks, this was achieved by changing only one word, e.g., slowly to quickly (tasks in Group 1). In other cases, more significant change was required and achieved by rewriting some of the descriptive sentences while keeping the characters, the setting, and the ending remark the same.

Using both a neutral and a sarcastic version of the same story, we could check the models' ability to detect sarcasm more precisely. We also included tasks (questions) testing context comprehension, which allowed us to determine whether an incorrect answer to a sarcasm detection task resulted from failure to detect sarcasm or another deficiency, such as a lack of world knowledge about a specific domain. Additionally, we used both closed and open interrogatives, alternative and yes-no questions and wh-questions, respectively. Closed interrogatives indicate a set number of possible replies, which implies the possibility of a non-literal answer in this specific case. Open interrogatives "expect a reply from an open range of replies" and "contain an information gap" (Balogné Bérces, 2016).

3 Evaluation and discussion

Our results show that both GPT-4 and OpenAI o3 have a strong ability to detect and understand sarcastic utterances. Overall, GPT-4 achieves 87.05% accuracy, notably below o3 (94.41%) and the human baseline (93.48%). This gap highlights GPT-4's relative difficulty with pragmatic inference, especially when interpreting non-literal language.

When focusing on context comprehension alone, GPT-4 reaches 89.28%, while o3 and humans attain 95.55% and 94.20%, respectively. In sarcasm detection, GPT-4's accuracy falls to 84.82%, compared with 93.30% for o3 and 91.51% for humans. The sharper decline for GPT-4 indicates that sarcasm, requiring sensitivity to implied intent, poses a greater challenge than mere narrative understanding. OpenAI o3, by contrast, exhibits a smaller drop in performance, suggesting enhanced pragmatic reasoning.

The format of the interrogative also affects outcomes. Under closed (yes/no) questions, all systems perform more poorly than with open (wh-) questions, with GPT-4 at 84.82% versus 89.28% on open prompts. This contrast implies that constrained answer choices deepen misinterpretations of sarcasm. In open-ended settings, the broader answer space appears to afford both models and humans more flexibility to capture meaning.

OpenAI o3 consistently matches or exceeds human performance across nearly all conditions, outperforming humans by up to 2.8 percentage points in closed sarcasm tasks. GPT-4's weaker showing on sarcasm underscores the need for targeted fine-tuning or prompt engineering to improve its non-literal comprehension. Future work will include detailed error analysis of false positives and false negatives, exploration of different prompting strategies (e.g. chain-of-thought), and evaluation of additional models with differences in the language distribution of the training data.

The human baseline was defined based on a limited number of participants, and it is possibly insufficient for evaluating human sarcasm comprehension. Nonetheless, our focus was to provide information on the current state of two advanced LLMs, and the results encourage future research with a human

baseline of a bigger sample size. GPT-4 could not surpass the human baseline, except in the topic-related assessment, namely "work"-related tasks, where GPT-4 scored 89.29% and the human baseline was 71.43%. The o3 scored higher in most aspects than the human baseline. Naturally, due to the small sample size of the human baseline, comprehensive conclusions could and should not be drawn. We cannot state that o3 outperformed humans in sarcasm comprehension due to the limitations of this research. Possible explanations include that human participants might not have known the rules or lingo of certain sports, while the o3 knew about those. Additionally, while the human participants had a high level of language proficiency, they were not native speakers. Tiredness or rushing could have also affected the human participants' performance.

References

Balogné Bérces, K. (2016). *The structure of english*. Akadémiai Kiadó.

Capelli, C. A., Nakagawa, N., & Madden, C. M. (1990). How children understand sarcasm: The role of context and intonation. *Child Development*, *61*(6), 1824–1838. https://doi.org/10.2307/1130840

Fanari, R., Melogno, S., & Fadda, R. (2023). An experimental study on sarcasm comprehension in school children: The possible role of contextual, linguistic and meta-representative factors. *Brain Sciences*, *13*(6), 863. https://doi.org/10.3390/brainsci13060863

Gole, M., Nwadiugwu, W.-P., & Miranskyy, A. (2024). On Sarcasm Detection with OpenAI GPT-Based Models. *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, 1–6. https://doi.org/10.1109/CASCON62161.2024.10837875

OpenAI. (2023). *Gpt-4 technical report* (tech. rep.). OpenAI. https://cdn.openai.com/papers/gpt-4.pdf

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelcey, F., Pfister, G., Radford, A., Schulman, I., & Amodei, D. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2005). Understanding witty remarks: The role of theory of mind and executive functions in processing verbal irony. *Brain and Language*, *94*(1), 36–44. https://doi.org/10.1016/j.bandl.2004.12.003

Tan, Y., Chow, C. O., Kanesan, J., Chuah, J. H., & Lim, Y. (2023). Sentiment analysis and sarcasm detection using deep multi-task learning. *Wireless Personal Communications*, *129*, 2213–2237. https://doi.org/10.1007/s11277-023-10235-4

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, A., Rodriguez, A. H., Joulin, A., & Grave, E. (2023). LLaMA: Open and efficient foundation language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 5517–5533.

Zhang, W., Li, X., Chen, Y., & Sun, M. (2024). Evaluating sarcasm detection capabilities of large language models. *Transactions of the Association for Computational Linguistics*, *12*, 123–139. https://doi.org/10.1162/tacl_a_00345