

Irony in Emojis: A Comparative Study of Human and LLM Interpretation

Yawen Zheng¹, Hanjia Lyu², Jiebo Luo²

¹The Chinese University of Hong Kong, Shenzhen

²University of Rochester

121090840@link.cuhk.edu.cn, hlyu5@ur.rochester.edu, jluo@cs.rochester.edu

Abstract

Emojis have become a universal language in online communication, often carrying nuanced and context-dependent meanings. Among these, irony poses a significant challenge for Large Language Models (LLMs) due to its inherent incongruity between appearance and intent. This study examines the ability of GPT-4○ to interpret irony in emojis. By prompting GPT-4○ to evaluate the likelihood of specific emojis being used to express irony on social media and comparing its interpretations with human perceptions, we aim to bridge the gap between machine and human understanding. Our findings reveal nuanced insights into GPT-4○'s interpretive capabilities, highlighting areas of alignment with and divergence from human behavior. Additionally, this research underscores the importance of demographic factors, such as age and gender, in shaping emoji interpretation and evaluates how these factors influence GPT-4○'s performance.

Introduction

The emergence and advancement of Large Language Models (LLMs) have enabled more sophisticated simulations of human behavior, including the actions of social media users, through generative agents (Park et al. 2023). Among the diverse forms of online communication, emojis have evolved from simple pictorial representations to a universal language transcending cultural boundaries (Herring and Dainas 2017; Ai et al. 2017). Despite their widespread use, Lyu et al. (2024b) highlighted notable discrepancies between the interpretation of emojis by GPT-4V and human behavior, underscoring a critical gap in understanding.

Irony, a pervasive literary and communicative technique on social media, is defined by an inherent incongruity (Zhang et al. 2019). Weissman and Tanner (2018) identified two primary carriers of irony on social media: verbal content and emojis. Emojis often embody irony through the contrast between their outward appearance and intended meaning, complicating their interpretation. Their euphemistic, humorous, and context-dependent uses further challenge the ability of LLMs to accurately discern sentiment (Lyu et al. 2024b). Addressing this challenge is essential, as accurate detection of irony in emojis could significantly enhance applications such as virtual assistants, chatbots, and sentiment analysis tools (Lyu et al. 2024a).

This study investigates the following research question:

- How does GPT-4○'s interpretation of ironic emojis compare to that of humans?

To answer this, we prompt GPT-4○ to assess how likely it is to choose a specific emoji to express irony on social media and compare its responses to human perceptions.

By focusing on irony in emojis, this research aims to evaluate LLMs' ability to understand subtle, sentiment-rich elements of this emerging universal language. Beyond technical insights, the study underscores the broader implications of AI in enhancing communication and human behavior simulation.

Related Work

The use of emojis on social media has attracted significant research attention, particularly in the context of sentiment analysis (Hu et al. 2017). The advent of large language models (LLMs) has introduced innovative approaches for analyzing sentiment involving emojis (Wankhade, Rao, and Kulkarni 2022; Weissman and Tanner 2018). However, a notable gap remains between LLMs' interpretation of emojis and human understanding. For instance, Lyu et al. (2024b) identified significant discrepancies in behavior between humans and LLMs, which can be attributed to the subjective nature of emoji interpretation and the limitations imposed by cultural biases and insufficient representation of non-English cultures. Similarly, Zhou et al. (2024) demonstrated that while ChatGPT exhibits extensive knowledge of emojis, it also perpetuates stereotypes across communities. Furthermore, Qiu et al. (2024) highlighted that LLMs face challenges in suggesting emojis that align with the semantic meaning of social media posts. Building on this foundation, our study focuses on investigating how LLMs interpret emojis within a specific category: ironic emojis.

Several studies have explored factors influencing human interpretation of emojis, including personality traits (Li et al. 2018) and demographics. Gender and age are frequently cited as having a profound impact on emoji comprehension, with additional associations identified between cultural background, religious beliefs, and emoji interpretation (Guntuku et al. 2019; Wang 2022). These demographic influences, particularly age and gender, are central to our investigation.

Gender-based differences in emoji comprehension have

been well-documented. Differences are observed in accuracy in interpreting sentiments conveyed by emojis—such as joy, sorrow, fear, and anger—though no significant gender differences have been observed in the interpretation of surprise or disgust (Chen et al. 2024). Additionally, preferences for emojis that express positive sentiments differ across gender groups (Chen et al. 2018). Age also plays a critical role in emoji usage and comprehension. Younger individuals often exhibit more advanced emoji usage skills, characterized by greater diversity, more nuanced applications, and higher accuracy in interpreting sarcasm and irony (Chen et al. 2018; Garcia et al. 2022; Chen et al. 2024).

Our study contributes to this body of work by conducting a fine-grained analysis of emoji interpretation through the lens of demographic factors. By incorporating age and gender information directly into prompts for LLMs, we aim to understand better how these models account for demographic variations in interpreting ironic emojis.

Method

This section describes the experimental procedure and its results. A quantitative analysis is conducted to evaluate the ability of the GPT-4○ variant to interpret irony.

Human Perception of Emoji Irony

We quantify how humans perceive the irony of emojis by analyzing their usage patterns. Specifically, we measure the frequency with which an emoji is used to convey irony in real-world social media posts and calculate its relative proportion of ironic usage as an **irony score**. We follow Xiang et al. (2020) and define “irony” as instances where an emoji conveys a meaning opposite to its literal interpretation, resulting in a reversal of understanding.

We use the Ciron dataset compiled by Xiang et al. (2020), which consists of over 8,700 posts (including around 3,000 posts containing emojis) from Weibo, a Chinese social media platform. We select Ciron over the SemEval 2018 Irony Detection in English Tweets dataset (Van Hee, Lefever, and Hoste 2018), which, despite being an established English dataset, includes only 494 tweets containing emojis.

Each post is independently annotated by five postgraduate students, all native Chinese speakers, with an irony rating on a scale from 1 to 5, where 1 indicates “not ironic,” 2 “unlikely ironic,” 3 “insufficient evidence of irony,” 4 “weakly ironic,” and 5 “strongly ironic.” Table 1 presents the distribution of irony scores across the collected posts. See Xiang et al. (2020) for further annotation details.

| Category | Count | Percentage |
|--------------------------------|-------|------------|
| Not ironic | 4,342 | 49.5% |
| Unlikely ironic | 3,391 | 38.7% |
| Insufficient evidence of irony | 64 | 0.7% |
| Weakly ironic | 837 | 9.6% |
| Strongly ironic | 129 | 1.5% |

Table 1: Post irony distribution of Ciron (Xiang et al. 2020).

We use the irony rating of a post to represent the irony level of the emojis found within it. For instance, if a post is labeled as “5 (strongly ironic),” the emojis used in that post are considered strongly ironic. To compute the irony score $S(e)$ for a specific emoji e , we take the average of the irony scores of all posts P_e that contain that emoji:

$$S(e) = \frac{\sum_{p \in P_e} (R(p))}{|P_e|} \quad (1)$$

where $R(p)$ represents the irony rating of post p , and $|P_e|$ denotes the total number of posts containing the emoji e . For instance, consider the “kiss” emoji , which appears in 27 posts. Among these, 23 posts have an irony rating of 1, two posts have a rating of 2, and two posts have a rating of 3. Using the formula, the irony score of the “kiss” emoji is calculated as:

$$S(\text{😘}) = \frac{(23 \times 1) + (2 \times 2) + (2 \times 3)}{27} = 1.22 \quad (2)$$

GPT-4○’s Classification of Emoji Irony

To assess GPT-4○’s classification of irony, we prompt it to evaluate how likely it would use a given emoji in a social media post to express irony.

Prompt Design The exact prompt reads: “*Imagine you are a social media user; rate your likelihood of using this emoji if your intention is to express irony on an 11-point scale (11=very likely, 1=very unlikely). The rating may depend on the context. You need to give the most likely rating and your explanation. Do not give multiple ratings in terms of scenarios. Only one rating is required.*” Following the approach of Czestochowska et al. (2022) and Lyu et al. (2024b), we provide the model with emojis in image format.

We further investigate whether GPT-4○’s classification changes when demographic information is included in the prompt by revising the first sentence to “*Imagine you are a [gender] social media user aged [age] ...*”

The study considers five age groups: “under 20,” “20-34,” “35-49,” “50-64,” and “over 65.” Following previous studies (Kotek, Dockum, and Sun 2023; Wan et al. 2023), gender groups include male and female, and experiments are conducted with all combinations of these demographic categories. We recognize that gender identity is diverse and encompasses a wide spectrum, including non-binary and other gender non-conforming identities that are frequently underrepresented. The binary gender framework applied in this study does not capture the complexity of human gender expression. Given these limitations, we approach the results with care, fully aware of the constraints inherent in the current dataset and methodology. We encourage future research to adopt more inclusive and representative frameworks.

Experiment Setting To ensure the robustness and consistency of the results, the model GPT-4○ is queried three times for each emoji used in the study to assess irony scores, with its temperature set to 0.5.

Results

The collected social media posts contain a total of 82 unique emojis. Table 2 presents the descriptive statistics of human-perceived and model-classified irony scores for these emojis. To enable comparison, the model’s ratings are rescaled from a range of 1–11 to align with the 1–5 scale. Using the Wilcoxon signed-rank test, we find that the median irony score assigned by GPT-4○ is significantly higher than the scores perceived by humans ($W = 918.5, p < .001$). This indicates that, on average, GPT-4○ considers the same emoji more likely to be used for expressing irony compared to human perception. This discrepancy may stem from GPT-4○ being training on data with a disproportionate representation of ironic emoji usage. We explore this issue further in the following section. The irony scores rated by GPT-4○ also exhibit greater variability.

| | Mean | Std | Min | Median | Max |
|------------------|------|------|------|--------|------|
| Human-perceived | 1.73 | 0.58 | 1.00 | 1.65 | 4.00 |
| Model-classified | 2.21 | 1.58 | 1.00 | 1.80 | 4.73 |

Table 2: Descriptive statistics of human-perceived and normalized model-classified irony scores for emojis.

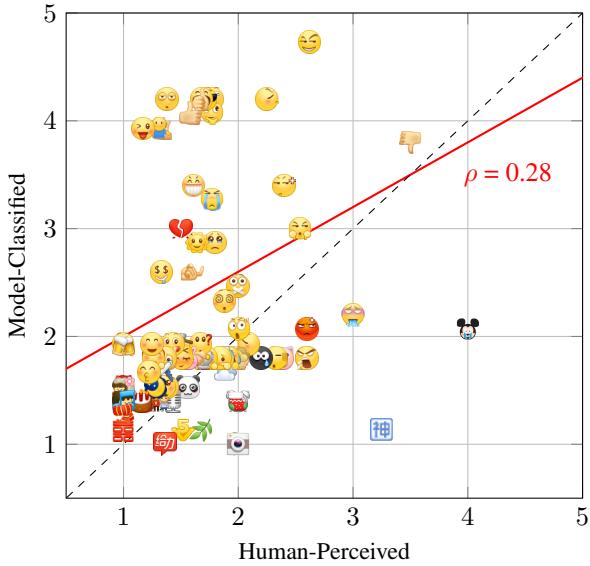


Figure 1: While GPT-4○ generally rates the same emoji as more likely to be used for expressing irony compared to human perception, the irony scores assigned by GPT-4○ and those perceived by humans show a significant correlation. Emojis positioned closer to the dashed line indicate greater alignment between GPT-4○’s classification of their use for expressing irony and human perception.

Moreover, as shown in Figure 1, by using the Spearman correlation test, we find that the irony scores assigned by GPT-4○ and those perceived by humans show a significant positive correlation ($\rho = 0.28, p < .05$). Emojis located

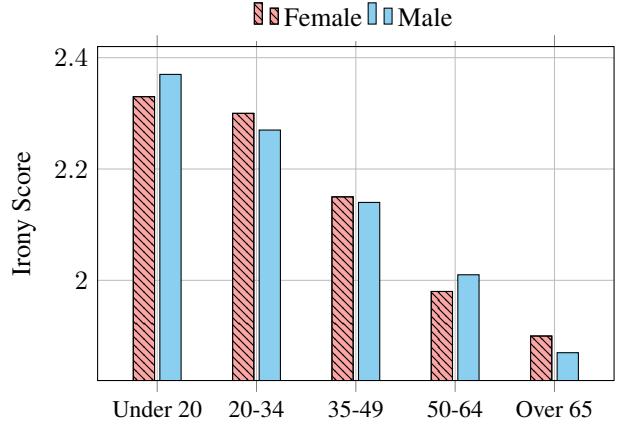


Figure 2: When the prompt includes demographic information, no significant differences in irony scores are observed between prompts specifying female or male gender. However, the irony scores tend to decrease on average as the specified age in the prompt increases.

closer to the dashed line reflect a higher degree of alignment between GPT-4○’s classification of their use for expressing irony and human perception.

To explore this alignment further, we prompt GPT-4○ to interpret emojis. For example, GPT-4○ explains that the emoji “” can convey irony due to its nuanced facial expression. The smirk’s inherent ambiguity makes it well-suited for ironic statements, where the intended meaning diverges from the literal interpretation. This emoji can suggest sentiments like “I know something you don’t” or “I’m not being entirely serious,” which are consistent with the subtle and indirect nature of irony.

However, in the Ciron dataset, we observe that the usage of this emoji differs. For instance, one post using this emoji states, “A Weibo post that challenges your psychological limits—dare to try?” This example highlights a distinct context where the emoji may not primarily signify irony.

Another emoji, “,” is rated highly by GPT-4○ for its association with irony but receives lower ratings from human perception. In Ciron, this emoji is frequently linked to positive and playful sentiment. When prompted, GPT-4○ explains that the emoji may also imply a teasing or mocking tone, suggesting an alternate interpretation that sometimes aligns with an ironic context.

Prompts with Demographic Information

This section explores the classification of emoji irony by GPT-4○ when demographic information, including age and gender, is incorporated into the prompt.

Figure 2 presents the average irony scores assigned by GPT-4○ to each emoji across prompts specifying different demographic attributes. The results suggest minimal differences in irony scores between prompts indicating female or male gender. However, a notable pattern emerges with age: irony scores tend to decrease as the specified age in the prompt increases.

The observed differences across age groups may be influenced by generational variations in communication habits, cultural norms, and digital literacy. Men and women of the same age group might exhibit similar interpretations of emoji irony due to shared digital environments, where gender differences in language use can be less pronounced. Shared cultural exposure to digital norms, memes, and emoji trends may contribute to this similarity.

In contrast, age appears to play a more substantial role in shaping perceptions of emoji irony. Younger individuals, having grown up with emojis as an integral part of their communication toolkit, may use and interpret emojis with greater fluidity and creativity. They are more likely to assign nuanced or ironic meanings to emojis compared to older individuals, who might favor more literal interpretations.

The meanings and contexts of emojis have evolved significantly over time, often driven by younger generations. For instance, emojis like “😊” or “👍” might be perceived as straightforward by older users, while younger users may attribute layered, ironic, or sarcastic meanings to the same symbols.

Additionally, younger users may gravitate toward platforms like TikTok or Instagram, where ironic or sarcastic emoji use is more prevalent, whereas older users might primarily engage with platforms such as email or Facebook, where emoji use tends to be more literal.

These insights align with findings from Garcia et al. (2022) and Chen et al. (2024), which indicate that younger individuals generally demonstrate a heightened ability to detect irony in emoji usage compared to older cohorts. While these results provide a compelling perspective, further research is necessary to better understand the interplay of demographic factors in shaping emoji interpretation.

Discussions and Conclusions

In this study, we investigate how GPT-4o evaluates the level of irony in emojis compared to human perceptions.

Our findings reveal that, on average, GPT-4o assigns higher irony scores to emojis than humans do for the same emoji. This observation has several implications. First, GPT-4o may have been trained on data with a disproportionate representation of ironic emoji usage, leading to an inflated assessment of their ironic potential. Second, the model might overgeneralize irony based on patterns in its training data, potentially lacking the contextual nuance humans rely on when interpreting emojis in specific scenarios. Lastly, this overestimation could have practical implications, such as misinterpretations in applications where accurate understanding of human communication patterns is critical—examples include sentiment analysis, chatbot interactions, and social media analysis.

We also observe minimal differences in irony scores between prompts indicating female or male genders. However, a notable trend emerged with age: irony scores tend to decrease as the specified age in the prompt increases. This age-related pattern aligns with findings from prior studies.

An additional factor contributing to the discrepancy between model-classified and human-perceived scores could

be the dataset used for evaluation. We conduct the study using a Chinese social media post dataset, while GPT-4o is predominantly trained on English-language data. This language orientation, combined with cultural differences in the use and interpretation of emojis, may partially explain the gap. However, due to the unavailability of an English-language irony dataset, we are unable to evaluate GPT-4o’s performance on English-language emojis. Addressing this limitation is a key focus of our future work.

Another limitation of our study is that it focuses exclusively on one model, GPT-4o. To gain a more comprehensive understanding of irony evaluation by large language models, future work will include a broader range of LLMs, especially those not primarily oriented toward English. This will allow for a more robust analysis of how linguistic and cultural differences impact irony perception.

References

- Ai, W.; Lu, X.; Liu, X.; Wang, N.; Huang, G.; and Mei, Q. 2017. Untangling emoji popularity through semantic embeddings. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 2–11.
- Chen, Y.; Yang, X.; Howman, H.; and Filik, R. 2024. Individual differences in emoji comprehension: Gender, age, and culture. *Plos one*, 19(2): e0297379.
- Chen, Z.; Lu, X.; Ai, W.; Li, H.; Mei, Q.; and Liu, X. 2018. Through a gender lens: Learning usage patterns of emojis from large-scale android users. In *Proceedings of the 2018 world wide web conference*, 763–772.
- Częstochowska, J.; Gligorić, K.; Peyrard, M.; Mentha, Y.; Bień, M.; Grüter, A.; Auer, A.; Xanthos, A.; and West, R. 2022. On the Context-Free Ambiguity of Emoji. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1): 1388–1392.
- Garcia, C.; Turcan, A.; Howman, H.; and Filik, R. 2022. Emoji as a tool to aid the comprehension of written sarcasm: Evidence from younger and older adults. *Computers in Human Behavior*, 126: 106971.
- Guntuku, S. C.; Li, M.; Tay, L.; and Ungar, L. H. 2019. Studying cultural differences in emoji usage across the east and the west. In *Proceedings of the international AAAI conference on web and social media*, volume 13, 226–235.
- Herring, S.; and Dainas, A. 2017. “Nice picture comment!” Graphic icons in Facebook comment threads.
- Hu, T.; Guo, H.; Sun, H.; Nguyen, T.-v.; and Luo, J. 2017. Spice up your chat: the intentions and sentiment effects of using emojis. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 102–111.
- Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, 12–24.
- Li, W.; Chen, Y.; Hu, T.; and Luo, J. 2018. Mining the relationship between emoji usage patterns and personality. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

- Lyu, H.; Huang, J.; Zhang, D.; Yu, Y.; Mou, X.; Pan, J.; Yang, Z.; Wei, Z.; and Luo, J. 2024a. GPT-4V(ision) as A Social Media Analysis Engine. *ACM Trans. Intell. Syst. Technol.* Just Accepted.
- Lyu, H.; Qi, W.; Wei, Z.; and Luo, J. 2024b. Human vs. LMMs: Exploring the Discrepancy in Emoji Interpretation and Usage in Digital Communication. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 2104–2110.
- Park, J. S.; O’Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.
- Qiu, Z.; Qiu, K.; Lyu, H.; Xiong, W.; and Luo, J. 2024. Semantics Preserving Emoji Recommendation with Large Language Models. *arXiv preprint arXiv:2409.10760*.
- Van Hee, C.; Lefever, E.; and Hoste, V. 2018. SemEval-2018 Task 3: Irony Detection in English Tweets. In Apidianaki, M.; Mohammad, S. M.; May, J.; Shutova, E.; Bethard, S.; and Carpuat, M., eds., *Proceedings of the 12th International Workshop on Semantic Evaluation*, 39–50. New Orleans, Louisiana: Association for Computational Linguistics.
- Wan, Y.; Pu, G.; Sun, J.; Garimella, A.; Chang, K.-W.; and Peng, N. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*.
- Wang, S. 2022. Sarcastic meaning of the slightly smiling face emoji from Chinese Twitter users: When a smiling face does not show friendliness. *International Journal of Languages, Literature and Linguistics*, 8(2): 65–73.
- Wankhade, M.; Rao, A. C. S.; and Kulkarni, C. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7): 5731–5780.
- Weissman, B.; and Tanner, D. 2018. A strong wink between verbal and emoji-based irony: How the brain processes ironic emojis during language comprehension. *PloS one*, 13(8): e0201727.
- Xiang, R.; Gao, X.; Long, Y.; Li, A.; Chersoni, E.; Lu, Q.; Huang, C.-R.; et al. 2020. Ciron: a new benchmark dataset for Chinese irony detection.
- Zhang, S.; Zhang, X.; Chan, J.; and Rosso, P. 2019. Irony detection via sentiment-based transfer learning. *Information Processing & Management*, 56(5): 1633–1644.
- Zhou, Y.; Xu, P.; Wang, X.; Lu, X.; Gao, G.; and Ai, W. 2024. Emojis decoded: Leveraging chatgpt for enhanced understanding in social media communications. *arXiv preprint arXiv:2402.01681*.

analysis systems. These improvements can lead to more effective and context-aware communication tools. Incorporating demographic information (e.g., age, gender) into AI models raises concerns about perpetuating or amplifying stereotypes. For example, demographic-based prompts might unintentionally reinforce assumptions about how certain groups use emojis. Care must be taken to ensure that the models remain equitable and do not propagate biases.

Appendix

Further Discussion on Potential Broader Impact and Ethical Considerations

By improving LLMs' understanding of ironic emojis, our work can enhance human-machine interactions in applications such as virtual assistants, chatbots, and sentiment