Bridging the Creativity Understanding Gap: Small-Scale Human Alignment Enables Expert-Level Humor Ranking in LLMs

Kuan Lok Zhou[*,1], **Jiayi Chen**[*,1], **Siddharth Suresh**[1], **Reuben Narad**[2], **Timothy T. Rogers**[1], **Lalit K Jain**[2], **Robert D Nowak**[1], **Bob Mankoff**[3], **Jifan Zhang**[1]

[1]University of Wisconsin-Madison, [2]University of Washington, Seattle,
[3]Air Mail and Cartoon Collections

Abstract

Large Language Models (LLMs) have shown significant limitations in understanding creative content, as demonstrated by Hessel et al. (2023)'s influential work on the New Yorker Cartoon Caption Contest. Their study exposed a substantial gap between LLMs and humans in humor evaluation, establishing that understanding and evaluating creative content is key challenge in AI development. We revisit this challenge by decomposing humor ranking into three components and systematically improve each: enhancing visual understanding through improved annotation, utilizing LLM-generated humor reasoning and explanations, and implementing targeted alignment with human preference data. Our refined approach achieves 84.7% accuracy in caption ranking, significantly improving upon the previous 67% benchmark and matching the performance of world-renowned human experts in this domain. Notably, while attempts to mimic subgroup preferences through various persona prompts showed minimal impact, model finetuning with crowd preferences proved remarkably effective. These findings reveal that LLM limitations in creative judgment can be effectively addressed through focused alignment to specific subgroups and individuals. Lastly, we advocate that truly improving LLM's creative understanding abilities necessitates systematic collection of human preference data across creative domains.

1 Introduction

Warning: this paper contains potentially offensive content due to the nature of humor.

The emergence of Large Language Models has revolutionized many domains of artificial intelligence, yet their ability to understand and evaluate creative content remains notably limited. This limitation is particularly evident in humor ranking, as demonstrated by several studies on the New Yorker Cartoon Caption Contest (Hessel et al., 2023; Zhang et al., 2024; Kazemi et al., 2025). These studies exposed a substantial gap between LLMs and human performance in ranking humorous captions. As shown by our test of the latest LLMs on the caption ranking task in Figure 1, various scaling efforts have not yielded significant improvement. In this paper, we successfully close the gap between LLMs and human experts.

Our work revisits this challenge by decomposing the ranking task into three components: visual understanding, cartoon-caption comprehension, and alignment with audience preferences (see Figure 2). Human experts demonstrate extraordinary ability in all three components. Previous research by Hessel et al. (2023) suggests that LLMs struggle with the second component—they are unable to comprehend the humor in cartoon-caption pairs. However, our study reveals that current LLMs, even with sophisticated prompting, struggle most with the third component—understanding the preferred jokes of specific audiences. While audience preferences are subjective in nature, it is a necessary and crucial ability that enables human experts to choose the best jokes tailored to their audiences.

Our initial attempts to bridge the gap between LLMs and human experts through various persona-based prompting techniques showed minimal impact, suggesting a fundamental limitation in how LLMs understand human preferences. The breakthrough came through explicit finetuning on human preference data from the caption contest crowd. Combined with improvements in the other components, we dramatically increased ranking performance from 67% to 84.7% accuracy, matching or exceeding human expert performance. This success extends to an even more challenging variant of the task where caption pairs are substantially closer in crowd preferences.

Lastly, our results highlight a broader challenge in the ability of LLM to understand the subgroup and individual preferences for subjective and cre-

[*] Equal contribution, random draw.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16273–16287
November 4-9, 2025

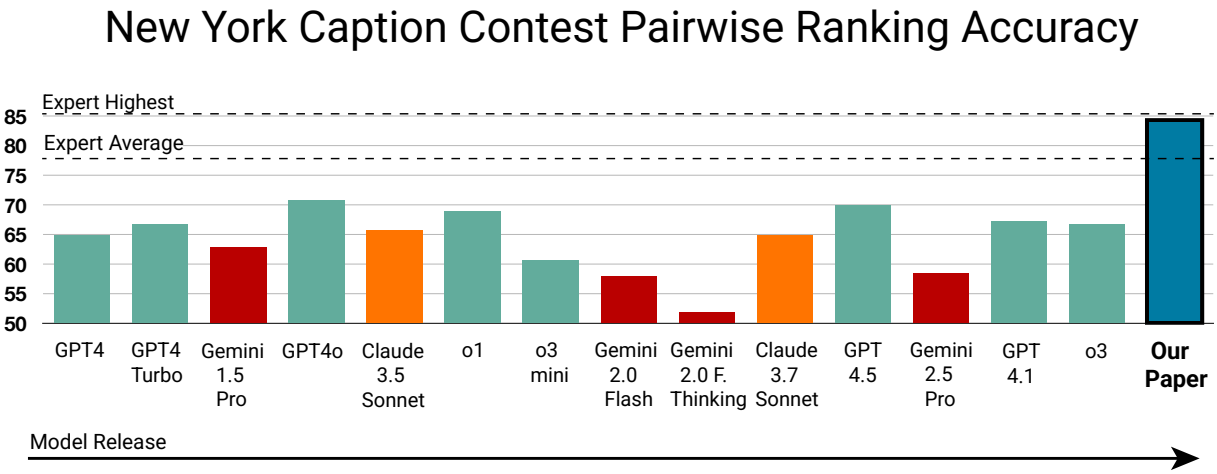

Figure 1: We evaluate the accuracies of latest model on the pairwise caption ranking task, which compares the ranking between a top 10 ranking caption versus a caption ranked around 1000. Over the past years, LLM development has not improved on this task despite various scaling efforts. Our paper closes this gap.

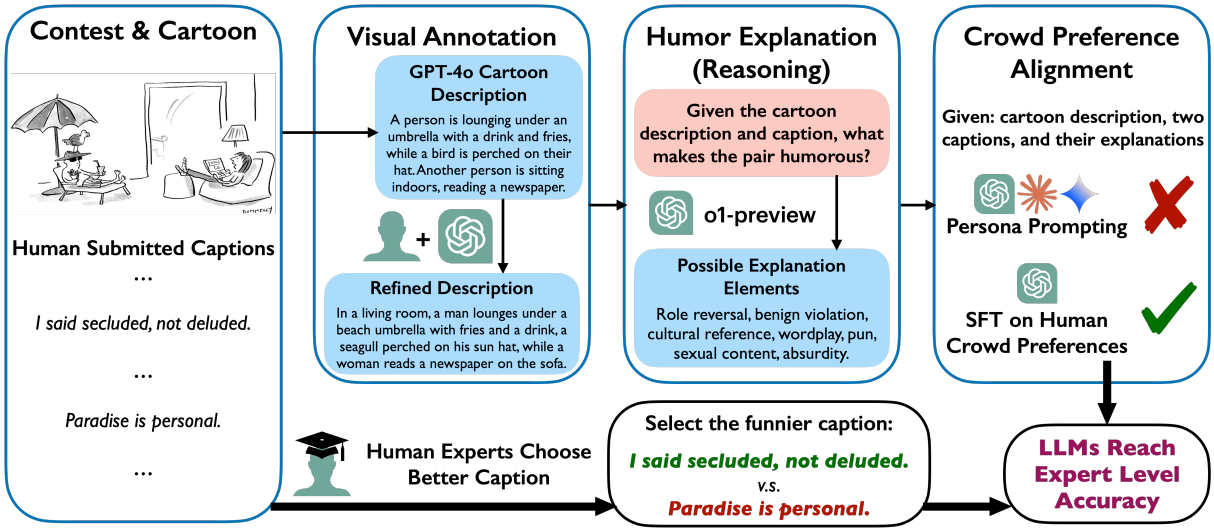

Figure 2: Our work improve over state-of-art caption ranking through a three-stage process. With multimodel LLM assistance, we manually fix visual understanding and cartoon description flaws. Our framework also incorporates o1 reasoning capabilities in explaining a joke, before utilizing two different alignment methods to align an LLM preferences with the human preferences from the NYYCC. Our experiments demonstrate that we are achieving human expert level accuracy in this caption ranking task.

ative tasks. In Section 5, we argue that the focus of the research community on problems with verifiable rewards, in domains such as mathematics and coding, may be insufficient to achieve a sophisticated understanding and abilities for creative works.

2 Related Work

Humor and LLMs. Research on computational humor has evolved significantly—from early rule-based, template-driven systems that generated puns via fixed linguistic rules (Binsted et al., 2006; Apte, 1988) to modern large language models that capture the nuances of human wit. Fine-tuned transformer encoders like BERT have yielded near–human performance in humor detection and rating (Weller and Seppi, 2019; Hossain et al., 2020), while very large generative LLMs such as GPT-3 and PaLM demonstrate emergent reasoning capabilities, explaining semantic incongruity in anti-jokes via chain-of-thought prompting (Chowdhery et al., 2022; Jentzsch and Kersting, 2023). To enhance creative generation, paradigms like Creative Leap-of-Thought (CLoT) prompting encourage unexpected conceptual associations (Zhong et al., 2023), and multimodal approaches incor-

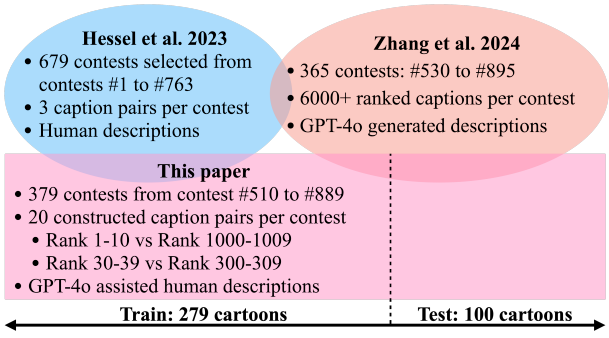

Figure 3: Composition of cartoon caption contest datasets across Hessel et al. (2023), Zhang et al. (2024) and our paper. In our paper, we examine 20 pairs of captions selected from 379 contests (#510-#889). The dataset is further split into 279 contest for training and 100 for testing.

Figure 4: Example voting page for the caption contest.

porating auditory cues have improved the modeling of phonetic ambiguities critical for pun understanding (Baluja, 2025). Curated humor datasets have exposed LLM limitations and driven progress in generation tasks (Horvitz et al., 2024), while real-world evaluations by stand-up comedians reveal that, despite fluent outputs, LLM-generated humor can appear generic compared to human creativity (Mirowski et al., 2024). Notably, under controlled evaluations, AI-generated jokes can rival human-produced humor (New York Post, 2024), and LLMs have been used as humor judges in the Crowd Score framework, producing funniness scores correlating with human judgments (Goes et al., 2022). Challenges remain in producing contextually rich, culturally sensitive, and genuinely surprising humor, motivating ongoing research into more sophisticated modeling and training paradigms (Robison, 2024). Humor detection still remains a hard problem, especially in the New Yorker Caption Contest ranking task as described below.

New Yorker Cartoon Caption Contest. Recent advances in computational humor have been bolstered by the availability of large, well-curated datasets derived from The New Yorker Cartoon Caption Contest. Previous works used this dataset to analyze the complex interplay between visual cues and linguistic humor, shedding light on the mechanisms that make captions amusing (Zhang et al., 2024). The seminal work of Bob Mankoff, whose editorial work shaped the contest's creative process, provides essential context and insight into what constitutes successful humor in this setting (Mankoff, 2008). However, recent studies have demonstrated that state-of-the-art AI models struggle to fully capture the nuanced judgment required to select and explain winning captions (Hessel et al., 2023). Together, these works underscore the utility of the New Yorker dataset as a powerful benchmark for advancing our understanding of humor in both human and machine-generated contexts.

LLM Post-training/Alignment Post-training alignment of large language models (LLMs) can be achieved through a blend of supervised fine-tuning (SFT) on instruction datasets and/or lightweight persona- or role-based prompting at inference time. Early instruction-tuning work showed that finetuning on a modest set of natural-language instructions yields large zero-shot gains on unseen tasks (Wei et al., 2021). Subsequent scaling along task diversity and model size continued this trend (Chung et al., 2022; Wang et al., 2022; Taori et al., 2023). Complementing SFT, persona prompting steers generation by prefacing the user query with a concise role description; Reynolds and McDonell (2021) formalize this "prompt programming" paradigm among others (Zheng et al., 2024; Chuang et al., 2023, 2024). Moral-reasoning systems such as *Delphi* demonstrate that persona-style conditioning can also be baked into the training data to yield consistent ethical judgments (Jiang et al., 2021). In this paper, we find persona-based prompting to be insufficient for current LLMs to understand subgroup human preferences. As preference learning is a binary output task, we adopt supervised finetuning for state-of-art LLMs to obtain expert level accuracy.

Notably, by training a better preference model, our method can also serve as the critical first step in training LLMs to generate funnier captions through Reinforcement Learning (Ouyang et al., 2022).

3 Cartoon Caption Ranking Task

The New Yorker Cartoon Caption Contest is a long-standing weekly feature hosted by The New Yorker magazine, in which a captionless cartoon is published and readers are invited to submit humorous captions. Each week, over 6,000 captions are submitted. From contest #530 to contest #895, a bandit-based crowdsource rating system (Jamieson et al., 2015) has been employed, allowing users to score captions as "funny", "somewhat funny", or "unfunny" (see Figure 4). At the end of each contest, a complete crowdsourced ranking of captions is obtained based on their perceived humor. Over the past eight years, the dataset of cartoons, captions and their rankings (Zhang et al., 2024) has proven invaluable for computational humor research. Notably, prior work by Hessel et al. (2023) and Zhang et al. (2024) has leveraged the caption contest dataset to benchmark both humor understanding and generation, two essential domains of humor reasoning.

To evaluate caption understanding, we employ the pairwise ranking task, a method widely used to study humor (Shahaf et al., 2015; Radev et al., 2016; King et al., 2013; Hessel et al., 2023; Zhang et al., 2024). We adopt the variant described by Hessel et al. (2023). In this task, given a cartoon description[1], evaluators (models or humans) compare two captions at a time, each randomly sampled from distinct ranking tiers. Specifically, one caption is drawn from a high-ranked group (ranks #1–10) and the other from a lower-ranked group (ranks #1000–1009) (see Figure 3). This sampling strategy allows us to directly measure an evaluator's ability to discern differences in humor quality while controlling task difficulty through the selection of ranking tiers. Additionally, we conduct a more challenging variant by asking models to distinguish between captions sampled from mid-ranked tiers (ranks #30–39 versus ranks #300–309). The results highlighted in Figure 1 underscore the persistent gap between state-of-the-art models and human expertise in humor evaluation, motivating our investigation into novel approaches to enhance model performance on this task.

[1]A cartoon image was used in place of the cartoon description when humor experts performed the same task.

4 Experiments

We break the ranking challenge into three components – visual understanding, humor reasoning, and targeted alignment to human crowd preferences. This parallels the human experts' abilities: visually perceiving cartoons, understanding and creating humorous content, and deliberately tailoring jokes to specific audiences.

Generating cartoon description is a crucial first step in understanding the humor correctly. However, we found 23.5% of the GPT-4o generated cartoon descriptions in Zhang et al. (2024) have erroneous descriptions. Therefore, in Section 4.1, we employ an AI-assisted annotation with human-in-the-loop assistance to fix cartoon descriptions. In Section 4.2, we find that the o1-preview model can explain captions correctly and demonstrates extensive humor reasoning more than 85% of the time. We therefore generate such explanations, which serve as intermediate reasoning steps that inform the final pairwise comparison of captions. To better align our system with human crowd preferences, we implement two different strategies in Section 4.3. First, we conducted extensive persona-based system prompting, which does not exhibit significant improvements. Our second, more sophisticated approach directly finetunes the model based on a set of ground truth rankings collected in the crowdsource ranking. This second approach significantly improves the ranking accuracy, and closes the performance gap between LLMs and human experts.

Throughout this section, we use a random train/test split (see Figure 3) with 279 cartoons for training and 100 for testing. The training set is also used for sampling 5-shot in-context learning, with five meaningfully sampled caption pairs per cartoon. All reported results are evaluated on the test set.

4.1 Improved Visual Annotation

Our dataset comprises 379 cartoons from the caption contest, including a subset from the annotated dataset introduced by Hessel et al. (2023). For cartoons lacking human annotations, we extend the description generation approach of Zhang et al. (2024). Through an LLM-assisted annotation process, we refine and improve the existing cartoon

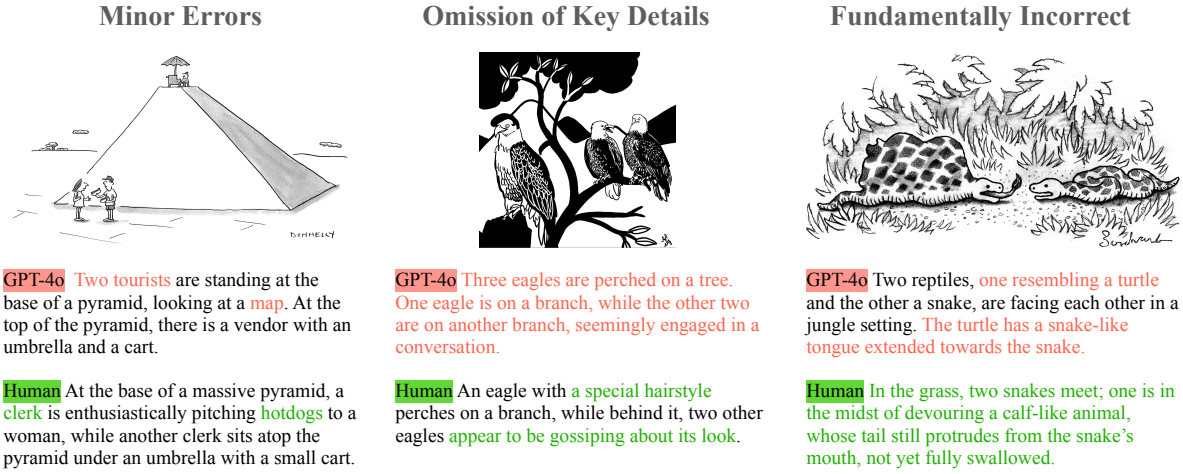

Figure 5: Examples of three types of errors in machine-generated cartoon descriptions and their human-annotated corrections. Left: Minor errors in word choice ("tourists" vs. "clerk", "map" vs. "hotdogs"). Center: Omission of key narrative details (missing the humorous implication of eagles gossiping about another eagle's appearance). Right: Fundamentally incorrect scene interpretation (misidentifying two snakes as a turtle and snake).

descriptions to build a comprehensive dataset.

Our visual annotation aims to generate both canny and uncanny descriptions. The canny descriptions accurately capture the literal contents of a cartoon, while the uncanny descriptions highlight its unusual or unexpected elements.

Our quality assessment reveals that 23.5% (89/379) of the machine-generated descriptions contain inaccuracies of varying severity, ranging from minor semantic errors and missing contextual elements to fundamental misinterpretations of the scene (see Figure 5). To address these issues, we develop a two-phase annotation refinement process. In the first phase, human reviewers iteratively improve the canny descriptions by identifying and correcting incorrect or omitted details. Based on their feedback, the descriptions undergo targeted revisions until they achieve comprehensive accuracy. In the second phase, these validated canny descriptions are used to generate corresponding uncanny elements, ensuring analytical consistency throughout the annotation process. Further details on this process can be found in Appendix A.1.

Comparative experiments between using the original and refined descriptions show an accuracy improvement from 70% to 73% with GPT-4o prompting. With the refined descriptions, finetuned models (more details in Section 4.3.2) obtain a performance gain from 81.3% to 84.7%.

4.2 Does reasoning through a joke improve humor ranking?

The New Yorker style humor is particularly sophisticated, characterized by layered cultural references, subtle wordplay, and implied backstories. To truly appreciate this humor, one must often draw connections between the explicit content and implicit knowledge, requiring a form of meta-understanding that bridges the cartoon-caption pairing with broader contextual associations (as illustrated in Figure 6).

This complex nature of humor suggests that explicit reasoning about jokes may facilitate better understanding, similar to how humans often explain jokes to clarify their humor elements. We hypothesize that incorporating explanations as an intermediate step in the ranking process could improve model performance by making implicit humor relationships more explicit.

Recent reasoning-focused language models like o1 (OpenAI, 2024b) appear particularly promising for this task. Our humor expert found that over 85% of o1-preview explanations effectively captured a cartoon's humor. To test this approach systematically, we generated explanations using two language models, GPT-4o and o1-preview (see Appendix A.2).

As shown in Table 1, incorporating o1-preview explanations increased ranking accuracy to 76%, compared to 73% for the baseline without explanation and 71% for explanations generated by GPT-4o. This improvement underscores the importance of explanation quality, with o1-preview better capturing humor elements such as the wordplay demonstrated in Figure 6.

However, despite models now being able to recognize and explain humor elements 85% of the time, we still observe persistently lower accuracy

Explanation Model	Accuracy
none (baseline)	73%
GPT-4o	71%
o1-preview	**76%**

Table 1: GPT-4o pairwise caption ranking accuracy of top 10 vs 1000-1009 captions. We compare explanations generated by different models. The experiment is conducted with five in-context examples (detailed prompts in Appendix A.4).

in ranking captions compared to human experts. This gap suggests that while understanding structural humor components is necessary, it may not be sufficient for fully aligning with human humor preferences. In the next section, we explore the subjective preferences unique to the New Yorker caption contest audience and implement various alignment strategies to help our model better learn and internalize this specific style of humor.

4.3 Alignment to Crowd Preference

Given the persistent gap above, we hypothesize that this discrepancy arises from a fundamental misalignment: GPT-4o's inherent subjective preferences does not match the specific taste and evaluative criteria of New Yorker crowd. Many subjective factors could contribute to this discrepancy of preferences, such as perceptions towards different types of culture references, or simply the New Yorker crowd may have a harder time recognizing certain types of jokes.

To address this misalignment, we build on our previous findings by exploring two alignment strategies. The first approach employs persona-based prompting to simulate the subjective evaluative criteria of New Yorker caption contest participants, conditioning the model's preference toward the target audiences' distinctive preferences. Our second strategy takes a more direct route through supervised finetuning (SFT) on a large corpus of New Yorker caption contest caption ranking data released by Zhang et al. (2024).

In the following sections, we detail these two alignment strategies and evaluate their effectiveness in bridging the gap between our model's performance and human experts' ranking accuracy.

4.3.1 Persona-Based Prompting

Persona-based prompting embeds persona information within system prompts to steer language model outputs toward reflecting target audience preferences, emerging as a promising method for aligning model behavior. Prior research has demonstrated the effectiveness of this approach in various tasks (Park et al., 2023; Chuang et al., 2023; Chen et al., 2024; Chuang et al., 2024). We design eleven distinct prompts that simulate personas representative of the New Yorker Cartoon Contest audience and evaluate their impact on humor preference alignment using four language models—Claude-3.5-sonnet, Gemini, GPT-4o, and o3-mini (Anthropic, 2024; Google, 2024; OpenAI, 2024a, 2025).

In all of our persona-based experiments, we incorporate five random in-context learning examples, and also the o1-preview explanations. See Appendix A.4 for detailed prompts. Results in Table 2 show that persona-based prompts yield only modest improvements in aligning the model with the intended audience. Moreover, we also observe a high agreement rate between drastically different persona prompts. For example, the CS PhD persona prompt yields an 88% agreement rate with the other personas, suggesting a fundamental defect in LLMs' understanding towards individual- and subgroup-level preferences (more discussions in Section 5). These results also indicate that persona-based alignment is not strong enough to capture the preferences of the New Yorker crowd.

4.3.2 GPT-4o Finetuning

Despite the limited accuracy gain achieved through persona-based prompting, aligning a model to specific crowd preferences should not be an inherently difficult task. When accurate humor explanations are available, learning subjective preferences becomes primarily a statistical exercise in identifying types of jokes more New Yorker audience will find amusing. Machine learning has proven remarkably effective at capturing such preference patterns, as evidenced by today's powerful recommendation systems. Indeed, our experiments confirm this intuition—once we finetune GPT-4o on even a small set of human preference data, we observe a dramatic improvement in ranking accuracy. This significant performance gain not only validates our approach but also reveals an interesting gap in LLMs' ability to understand subgroup and individual preferences for subjective tasks without explicit preference-based training (further discussed in Section 5).

To construct the training set, for each of the 279 training cartoons, we randomly form 10 pairwise comparisons between captions ranked 1-10 and those ranked 1000-1009, and another ten pairs be-

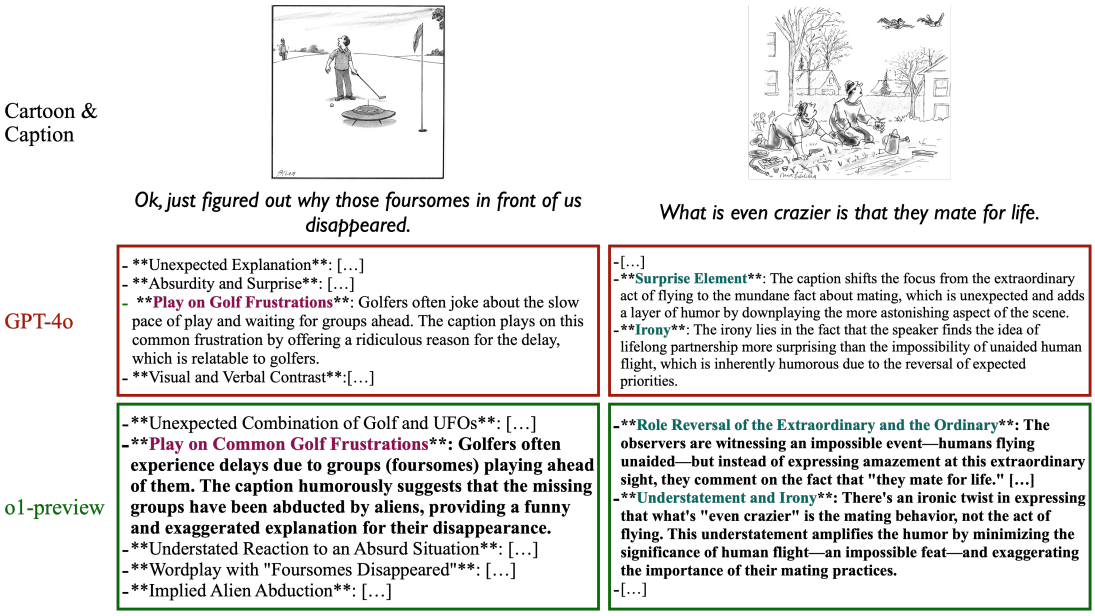

Figure 6: Comparison of humor explanation quality between GPT-4o and o1-preview, illustrated through two cartoon-caption pairs and their respective AI-generated humor explanation. o1-preview demonstrates a deeper comprehension of the humor, and its explanations are highlighted in bold text.

tween 30-39 and 300-309. This in total results in 5580 pairs of captions. In our experiments, we performed a simple supervised finetuning of GPT-4o, for it to choose between the two candidate captions. The model is given the cartoon description, the two captions and their corresponding explanations generated by o1-preview.

Note that pairwise comparisons of captions ranked 30-39 versus those ranked 300-309 have a much narrower gap, and are thus much more challenging. Results on both the easier (1-10 vs 1000-1009) and the the more challenging (30-39 vs 300-309) sets of comparisons are reported. As shown in Table 2, finetuned GPT-4o models can significantly improve upon all prompting-based baseline before. When incorporating o1-preview generated explanations, the finetuned GPT-4o can even achieve slightly higher accuracy than the majority vote of human experts. Below, we give some details on the human expert experiments.

4.3.3 Human Expert Accuracy

To evaluate the performance of our model, we conducted a study with five highly renowned human experts in the New Yorker Cartoon Caption Contest world, including famous cartoonists, editors and podcast hosts in this area. In our experiments, these human experts were presented with both the original cartoon image and paired captions, tasked with selecting the more humorous option from each pair (demonstrated in Figure 2). Due to busy schedules of these experts, the evaluation corpus consisted of 50 contests selected from our testing set, with two distinct caption pairs each. The two pairs are consisted of one for comparison between the top-ranked caption (rank 1) and a lower-ranked caption (rank 1000), and a comparison between a mid-ranked caption (rank 30) and a lower-ranked caption (rank 300).

The result in Table 2 shows the accuracy of the majority vote among the five experts as well as the average of their individual performances. In addition, the best individual performance was from Bob Mankoff, the former chief cartoon editor for the New Yorker, who created this contest more than 25 years ago. Our model still slightly underperforms the performance Mankoff, leaving space for further improvement. More details about our expert experiments can be found in Appendix C, where we do see strong inter-rater agreement among human experts.

5 Position: LLMs Need More Human Preference Data for Creative Domains

Our empirical findings on humor ranking highlight a significant challenge in developing more capable language models. While recent advances in LLMs have demonstrated remarkable capabilities in analytical reasoning and structured problem-solving, our results suggest that creative domains present a unique hurdle that requires special attention. We argue that this challenge stems from two fundamen-

Methods	10vs1000	30vs300
Expert Majority Vote	$84.0_{\pm 5.2}$	$66.0_{\pm 6.8}$
Expert Average Accuracy	$78.0_{\pm 2.6}$	$61.6_{\pm 3.0}$
Best Expert Accuracy	$85.3_{\pm 2.9}$	$68.0_{\pm 3.8}$
Gemini 2.0 Flash Thinking	$51.8_{\pm 1.6}$	$50.6_{\pm 1.6}$
o1	$69.0_{\pm 1.5}$	$58.0_{\pm 1.6}$
o3-mini	$60.7_{\pm 1.5}$	$53.2_{\pm 1.6}$
Claude 3.5 Sonnet	$65.8_{\pm 1.5}$	$54.0_{\pm 1.6}$
Claude 3.7 Sonnet	$66.4_{\pm 1.5}$	$53.7_{\pm 1.6}$
GPT-4.1	$67.5_{\pm 1.5}$	$56.0_{\pm 1.6}$
o3	$67.0_{\pm 1.5}$	$55.9_{\pm 1.6}$
Gemini 2.5 Pro	$58.4_{\pm 1.6}$	$51.1_{\pm 1.6}$
GPT-4o Prompting	$71.3_{\pm 1.4}$	$56.2_{\pm 1.6}$
GPT-4o SFT w/o Expl. (Ours)	$79.4_{\pm 1.3}$	$59.7_{\pm 1.6}$
GPT-4o SFT w/ Expl. (Ours)	$\mathbf{84.7}_{\pm \mathbf{1.1}}$	$\mathbf{63.0}_{\pm \mathbf{1.5}}$

Table 2: Accuracy(%) Comparison of Different Methods. All models use prompting techniques including best persona, 5-shot in-context learning, and o1-preview generated explanations. Our finetuned GPT-4o model with explanations even outperforms several human experts. Note that human expert accuracies are evaluated based on a random subset of 50 cartoons due to limited availabilities.

System Prompt	GPT-4o	Claude	Gemini	o3-mini
New Yorker Audiences	71.3	70.5	61.5	60.0
New Yorker Editors	70.2	69.0	61.5	61.0
Empty (Baseline)	73.5	68.0	47.0	70.0
Judge	73.0	67.5	55.0	58.0
Male Lawyer	75.0	74.0	51.0	59.0
Female Lawyer	76.5	69.0	59.0	67.0
CS PhD	73.5	68.0	49.0	65.0
Sociologist & Psychologist	73.5	67.0	61.0	60.0
Literature Student	73.5	72.0	51.0	58.0
Bob Mankoff	73.5	66.0	50.0	62.0
Larry Wood	73.0	68.0	57.0	61.0
Cartoon Author	71.5	61.0	46.0	62.0

Table 3: Accuracy(%) of using different persona-based system prompts on 10 vs 1000 pairwise caption ranking task across four language models: GPT-4o, Claude-3.5-Sonnet, Gemini-2.0-Thinking-Experiment, and o3-mini. Each number is measured on a size 200 subset of the test set, except the GPT-4o performance of audiences and editors prompts are measured on the full test set. Each row represents a distinct persona-based prompt. See Appendix B for system prompts and Appendix A.4 for the task prompt.

tal characteristics of creative tasks that are often overlooked in current AI research.

First, creative tasks inherently lack verifiable rewards. Unlike mathematical proofs or programming challenges where correctness can be definitively verified, creative success often depends on subjective human judgment. Our experiments with the New Yorker Caption Contest illustrate this clearly: while our models can now generate sophisticated explanations of why a caption might be humorous, these explanations alone do not translate to accurate predictions of human preferences.

Second, creative excellence requires understanding and internalizing group-specific preferences and cultural contexts. This is particularly evidenced by creative experts, who excel at tailoring their content towards specific audiences. In our experiments, persona-based prompting failed to improve caption ranking, while direct preference learning proved effective. This highlights a crucial gap in current LLM capabilities in understanding the individual- and subgroup- level preferences. While the New Yorker Caption Contest provides us with extensive ranking data from a specific audience, collecting similarly comprehensive preference data for every creative domain, cultural group, and individual taste remains prohibitively difficult. For instance, how might we gather equivalent preference data for domains like musical composition, architectural design, or scientific research, where expert judgment is highly specialized? How do we gather the same type of preference data across different groups of people with distinctive tastes?

These observations lead us to propose that improving LLMs on creative domains fundamentally requires solving the challenge of preference understanding. Current reinforcement learning based techniques can improve performance once we have reliable judgment models, but the path forward requires models that can develop generalizable insights about preferences across different contexts, domains and audiences.

6 Conclusion

Our work demonstrates that by decomposing humor evaluation into visual comprehension, reasoning, and preference alignment components, LLMs can achieve expert-level performance in humor evaluation. While persona-based prompting showed limited success, direct finetuning on crowd preferences yielded dramatic improvements, suggesting that systematic collection of human preference data across creative domains may be essential for achieving true creative understanding in AI systems. Looking ahead, our high-performing model can serve as a reliable verifier for humor generation, enabling inference-time scaling techniques (Zelikman et al., 2022; Snell et al., 2024) to improve creative output. This creates a promising pathway for advancing both humor understanding

and generation capabilities in AI systems.

Limitations

Our work has several limitations that we acknowledge below:

- **Domain Specificity.** Our study is based solely on the New Yorker Cartoon Caption Contest dataset. Although this dataset provides a rich benchmark for humor evaluation, its narrow focus may limit the generalizability of our findings to other forms of humor and creative tasks, especially for other cultures. However, we believe the shortcoming of audience understanding of current LLMs will likely persist under other cultures and creative domains. We also expect our proposed fix would work to address the audience preference understanding issue.
- **Evaluation Focus.** We primarily evaluate caption understanding using a pairwise ranking task. While this approach is effective for assessing relative humor quality, it may not fully capture the broader nuances of humor understanding or the challenges involved in humor generation.
- **Subjectivity and Bias in Preference Data.** The human preference data employed for finetuning and evaluation is inherently subjective and reflects the tastes of a specific audience (e.g., New Yorker readers). This limitation, however, reinforces our position that systematic collection of diverse human preference data is crucial for improving model performance on creative tasks.
- **Scalability of Human Alignment.** While our results demonstrate that aligning models with human preferences can substantially enhance performance, the process of gathering high-quality, curated human data is resource-intensive and may not scale easily to other creative domains. This challenge underlines our broader argument that advancing creative AI requires scalable methods for collecting and integrating human interaction data.
- **Humorous Content May Be Offensive.** Humor often walks a fine line between eliciting laughter and being potentially offensive. While our focus on the New Yorker dataset biases our work towards a certain style of humor, we acknowledge that humorous content can sometimes be culturally insensitive or derogatory. Our current framework does not explicitly address the detection or mitigation of offensive content, highlighting the need for future research to incorporate robust ethical safeguards alongside creative performance.

References

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Surender Apte. 1988. *Humor and Laughter: An Anthropological Approach*. Transaction Publishers.

Ashwin Baluja. 2025. Text is not all you need: Multimodal prompting helps llms understand humor. https://aclanthology.org/2025.chum-1.2.pdf.

Kim Binsted et al. 2006. Computational humor: An implemented model of puns. *IEEE Intelligent Systems*, 21(2):59–69.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, and et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2024. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, and et al. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Fabricio Goes, Zisen Zhou, Piotr Sawicki, Marek Grzes, and Daniel G. Brown. 2022. Crowd Score: A method for the evaluation of jokes using large language model ai voters as judges. In *arXiv preprint arXiv:2212.11214*.

Google. 2024. Gemini 2.0 flash thinkingg. https://deepmind.google/technologies/gemini/flash-thinking.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Zachary Horvitz et al. 2024. Getting serious about humor: Crafting humor datasets with unfunny large language models. https://arxiv.org/abs/2403.00794.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758. International Committee for Computational Linguistics.

Kevin G Jamieson, Lalit Jain, Chris Fernandez, Nicholas J Glattard, and Rob Nowak. 2015. Next: A system for real-world development, evaluation, and application of active learning. *Advances in neural information processing systems*, 28.

Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. 2025. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*.

Ben King, Rahul Jha, Dragomir Radev, and Robert Mankoff. 2013. Random walk factoid annotation for collective discourse. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 249–254.

Robert Mankoff. 2008. *The New Yorker Cartoon Caption Contest*, first edition edition. Andrews McMeel Publishing, Kansas City.

Piotr W. Mirowski et al. 2024. A robot walks into a bar: Can language models serve as creativity support tools for comedy? https://doi.org/10.48550/arxiv.2405.20956.

New York Post. 2024. Is ai funnier than humans? this study says so — but you be the judge.

OpenAI. 2024a. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

OpenAI. 2024b. Introducing openai o1-preview. https://openai.com/index/introducing-openai-o1-preview/.

OpenAI. 2025. Openai o3-mini. https://openai.com/index/openai-o3-mini/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. arxiv. *arXiv preprint ArXiv:2304.03442*.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in the New Yorker cartoon caption contest. In *LREC*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7. ACM.

Greg Robison. 2024. The last laugh: Exploring the role of humor as a benchmark for large language models.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *KDD*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. `https://crfm.stanford.edu/2023/03/13/alpaca.html`. Stanford CRFM blog post, accessed 19 May 2025.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *CoRR*, abs/2212.10560.

Jason Wei, Maarten Bosma, Vincent Zhao, et al. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Orion Weller and Kevin Seppi. 2019. Humor detection: A transformer gets the last laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, et al. 2024. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *arXiv preprint arXiv:2406.10522*.

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. 2024. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA. Association for Computational Linguistics.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2023. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. `https://arxiv.org/abs/2312.02439`.

A Language Model Prompt

A.1 Updating description

We conduct a comprehensive quality assessment of the cartoon descriptions generated by GPT-4o across our dataset of 379 images. Initial evaluation reveals that 76.5% of the generated descriptions meet our quality criteria for reasonableness and completeness. The remaining 23.5% exhibits various types of deficiencies that required remediation.

To address these quality issues, we implement a systematic two-phase refinement process:

In the first phase, for the identified problematic descriptions, we provide GPT-4o with specific feedback detailing the observed errors and request regeneration of these descriptions. This iterative process continues until the descriptions achieve the required level of accuracy and completeness.

In the second phase, following the establishment of a clean description set, we employ a 5-shot learning approach to generate corresponding uncanny descriptions for those updated canny descriptions. The following is a detailed prompt of the second phase.

User: In this task, you will see a cartoon image and a canny description written about the image. You need to write an uncanny description. I'm going to give you five examples first. Write an uncanny description for the last set.
User: <Insert Cartoon Image>
User: The canny description is <Insert canny description>
Assistant: The uncanny description is <Insert uncanny description>
.....Repeat user/assistant for four more examples.....
User: <Insert Cartoon Image>
User: The canny description is <Insert canny description>
User: The uncanny description is

A.2 Explanation Generation

We employ both GPT-4o and o1-preview to generate explanations for the humorous elements in the captions. We implement a zero-shot approach, providing each model with both the uncanny and cannny descriptions alongside the caption in question. The prompt structure utilized in our experiments is illustrated below.

User: I will give you a description of the cartoon and the winning caption. Explain to me why the caption is funny.
User: The descriptions for the images are <Insert canny description> and <Insert uncanny description> The winning captions is: <Insert cartoon captions>
User: There may or may not be multiple reasons for the caption being funny. Put them into bullet point(s).

A.3 Baseline Caption Evaluation

For our baseline evaluation, we employ a 5-shot prompting approach. In this setup, we provide the model with cartoon descriptions and the corresponding pair of captions. The prompt follows a structured format where the model is first assigned the role of a judge for the New Yorker cartoon caption contest. We then present five examples of caption ranking, allowing the model to observe the evaluation pattern. For the final test case, the model is tasked with selecting the funnier caption between two options, as the examples. The prompt structure is illustrated below.

System: You are a judge for the New Yorker cartoon caption contest.
User: In this task, you will see two descriptions for a cartoon. Then, you will see two captions that were written about the cartoon. Then you will choose which caption is funnier. I am going to give you five examples first and you answer the last question with either A or B
User: For example, the descriptions for the images are <Insert canny description> and <Insert uncanny description>. The two captions are A: <Insert Caption A>. B: <Insert Caption B>
Assistant: The caption that is funnier is <Insert Answer>
.....Repeat user/assistant for four more examples.....
User: The descriptions for the images are <Insert canny description> and <Insert uncanny description>. The two captions are A: <Insert Caption A>. B: <Insert Caption B>
User: Choose the caption that is funnier. Answer with either A or B and nothing else.

A.4 ICL Explanation Caption Evaluation

Building on the baseline evaluation, we incorperate o1-preview generated the model an explanation that is generated by o1. We changed the system prompts to test different persona. The detailed persona prompts is in Appendix B.

System: You are a judge for the new yorker cartoon caption contest.
User: In this task, you will see two descriptions for a cartoon. Then, you will see two captions about the cartoon and an explanation for why each caption is funny. I am going to first give you five examples where I will tell you which one is funnier then you answer the last one with either A or B and nothing else.
User: For example, the descriptions for the images are <Insert canny description> and <Insert uncanny description>. Captions A: <Insert Caption A>, and why the caption is funny is <Insert explanation for Caption A >. Caption B: <Insert Caption B>, and why the caption is funny is <Insert explanation for Caption B>,
Assistant: The caption that is funnier is <Insert Answer>
.....Repeat user/assistant for four more examples.....
User: Last one, the descriptions for the images are <Insert canny description> and <Insert uncanny description>. Caption A: <Insert caption A>, and why the caption is funny is <Insert explanation for Caption A>. Caption B: <Insert Caption B>, and why the caption is funny is <Insert explanation for Caption B>.
User: The caption that is funnier is

B Persona Prompt

We develop different system prompts, trying to represent different demographic groups of the New Yorker Cartoon Contest audience. See Table 4 for details.

Prompt Name	System Prompt
New Yorker Audiences	Judge the following cartoon based on the preference of the entire group of New Yorker audiences.
New Yorker Editors	Judge the following cartoon based on the preference of the entire group of New Yorker editors.
Judge	You are a judge for the New Yorker cartoon caption contest
Male Lawyer	Imagine you are a white male lawyer in your 50s. You grew up in New York City and have been reading the New Yorker Magazine ever since.
Female Lawyer	Imagine you are a white female lawyer in your 50s. You grew up in New York City and have been reading the New Yorker Magazine ever since.
CS PhD	Imagine you are a computer science PhD student. You have been submitting captions for every New Yorker cartoon caption contest for the past three years.
Sociologist & Psychologist	Imagine you are a sociology and psychology researcher that studies the New Yorker humor.
Literature Student	Imagine you are an English literature student that loves the New Yorker Magazine and its humor.
Bob Mankoff	Imagine you are Bob Mankoff, the editor of the New Yorker Cartoon Contest.
Larry Wood	Imagine you are Larry Wood, the 8-time New Yorker Cartoon Contest winner.
Cartoon Author	Imagine you are a cartoon author who often reads the New Yorker Cartoon Contest for inspiration.

Table 4: Persona prompt names and their corresponding text.

C Human Expert Annotation

To evaluate human performance, we collect assessments from five expert annotators and compute three different accuracy metrics. Each expert is given the following instruction at the beginning of the task.

> In each trial of this task, you will see one cartoon and two captions: the cartoon is on top, and the two caption choices are beneath the cartoon.
> For each trial, **please select the caption that is the funniest for the cartoon.**
> There will be around 100 trials. You will have opportunities to take a break throughout. There are attention checks during the experiment. Please chose the same image as the one on top for these trials.
> Click 'Continue' to begin the test.

After the instruction page, the participants complete 100 trials, each of which looks like the following.

Metrics	10vs1000	30vs300
Average Accuracy	78.00	61.60
Highest Accuracy	85.33	68.00
Majority Accuracy	84.00	66.00
Fleis Kappa	0.3641	0.2304
Agreement Rate	77.28	67.68

Table 5: Human expert performance. There are total of 5 human expert in this group.

Unless the apocalypse comes, I'll be back for dinner.

This was easier to carry when it was flat.

As shown in Table 5, the average accuracy represents the mean performance across all five experts. For the highest accuracy metric, we independently identify the best-performing expert for each of our two ranking tasks (Rank 10 vs 1000 and Rank 30 vs 300 pairs). The majority vote accuracy reflects the performance of collective human judgment. For each test instance, we aggregate the five individual expert annotations through majority voting to determine the final prediction, then calculate the accuracy of these consensus-based decisions. To assess inter-annotator agreement, we employ two complementary metrics, Fleis Kappa and agreement rate. The Fleiss Kappa values indicate fair to moderate agreement, accounting for the chance agreement. The agreement rate measure if randomly selected two annotators' judgments for a random caption pair, they would agree 77.28% of the time for the ranking tasks of Rank 10 vs 1000 and 67.68% of the time for the ranking tasks of Rank 30 vs 300.

D Additional Paper Details

We used OpenAI, Anthropic and Google APIs for all experiments. Overall, our experiments cost around $4,000 USD. In addition, LLMs have been used to rephrase some parts of this paper.

This paper is for research purpose only, and complies with the CC-BY-4.0 license for the dataset from Hessel et al. (2023) and the CC-BY-NC-4.0 license from Zhang et al. (2024).