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

Anonymous ACL submission

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 (NYCCC). Their study exposed a substantial gap between LLMs and humans in humor comprehension, establishing that understanding and evaluating creative content is key challenge in AI development. We revisit this challenge by decomposing humor understanding 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 82.4% 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 propose the position that achieving artificial general intelligence necessitates systematic collection of human preference data across creative domains. We advocate that just as human creativity is deeply influenced by individual and cultural preferences, training LLMs with diverse human preference data may be essential for developing true creative understanding.

1 Introduction

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

The emergence of Large Language Models (LLMs) 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 comprehension, as demonstrated by Hessel et al. (2023)'s seminal work on the New Yorker Cartoon Caption Contest (NYCCC). Their study, which earned the best paper award at ACL 2023, exposed a substantial gap between LLMs and human performance in ranking humorous captions, establishing creative understanding as a key challenge in AI systems.

We revisit this challenge by decomposing humor understanding into three components: visual understanding, cartoon-caption reasoning, and alignment with human preferences as demonstrated in Figure 1. Through improved visual annotations and LLM-generated explanations, we significantly enhanced both visual understanding and cartoon-caption reasoning. However, the most critical and challenging component proved to be alignment with human preferences.

Our work reveals an intriguing paradoxical finding in this alignment challenge: while LLMs can now generate sophisticated and accurate explanations about why captions are humorous, they still struggle with the seemingly easier task of ranking pairs of captions. Our attempts to bridge this gap 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 the other improvements mentioned above, we dramatically increased our ranking performance from 67% to 82.4% accuracy, matching or exceeding the performance of human experts. This success extends to an even more challenging variant of the task where the crowd-averaged preference differences between caption pairs are substantially smaller.

Our results highlight a broader challenge in AI capacity to understand subgroup and individual
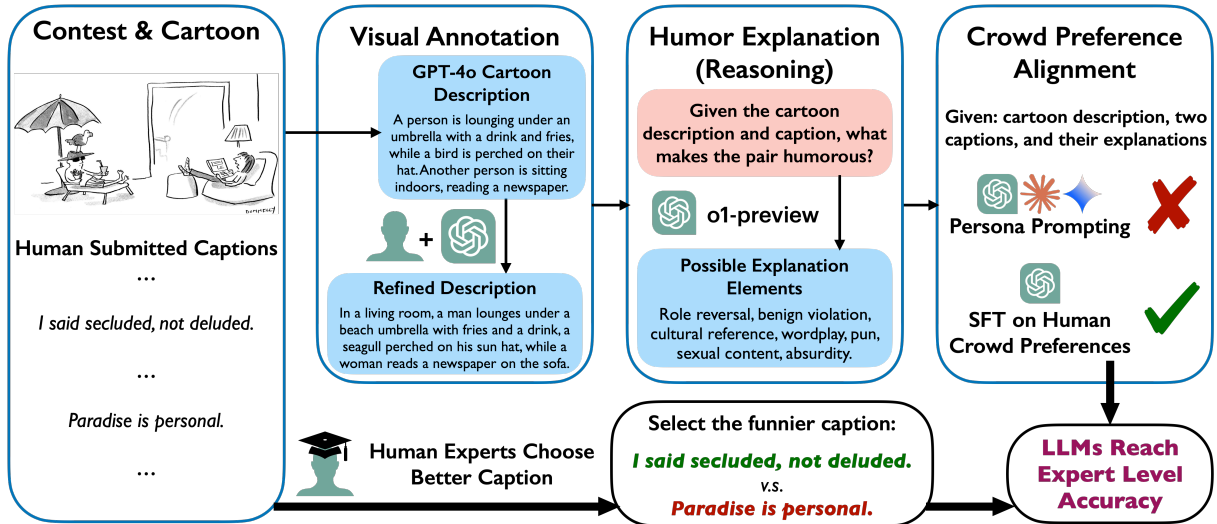
Figure 1: Our work improve over state-of-art caption ranking through a three-stage process. With multimodel LLM assistance, we manually fix visual understanding and cartoon description flaws. Our framework also incorporates o1 reasoning capabilities in explaining a joke, before utilizing two different 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.

preferences for subjective and creative tasks. In Section 5, we argue that the AI research community's focus on problems with verifiable rewards, in domains such as mathematics and coding, may be insufficient for achieving AGI. We propose that mastering creative domains – which lack objective metrics and require deep understanding of audience preferences – represents a crucial yet under-explored challenge on the path to AGI.

The contributions of this work are as follows.

1. We decompose LLM capability in humorous caption ranking into three fundamental components – visual understanding, humor reasoning and subgroup preference alignment.

2. By improving upon all of the three components, especially on the preference alignment, we obtain caption ranking models that achieve accuracy on par with human experts.

3. Our experiments reveal that extensive persona-based alignment significantly under-performs relative to improvements based on finetuning, revealing current LLM limitations in understanding subgroup and individual preferences.

4. We propose that systematic collection and integration of human preference data across creative domains may be essential for achieving AGI in creative tasks.

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 (LLMs) that strive to capture the nuances of human wit. Recent studies reveal that while models like Chat-GPT can produce coherent and seemingly humorous outputs, they often rely on a limited repertoire of pre-learned jokes rather than inventing truly original humor (Jentzsch and Kersting, 2023). To overcome these limitations, innovative prompting strategies such as the Creative Leap-of-Thought (CLoT) paradigm have been proposed, encouraging LLMs to make unexpected conceptual associations and thereby enhancing creative humor generation (Zhong et al., 2023). Complementing these approaches, multimodal techniques that integrate auditory cues have shown promise in capturing the phonetic ambiguities (essential for understanding puns) that text-only systems often miss (Baluja, 2025). Furthermore, research on curated humor datasets demonstrates how targeted data can expose LLM limitations and spur advances in humor generation (Horvitz et al., 2024), while real-world evaluations by stand-up comedians underscore that, despite impressive fluency, LLM outputs frequently appear generic or bland compared to human creativity (Mirowski et al., 2024). On the other hand, recent studies suggest that under controlled conditions AI-generated humor can rival human-produced jokes (New York Post, 2024). Challenges remain, however, in producing humor
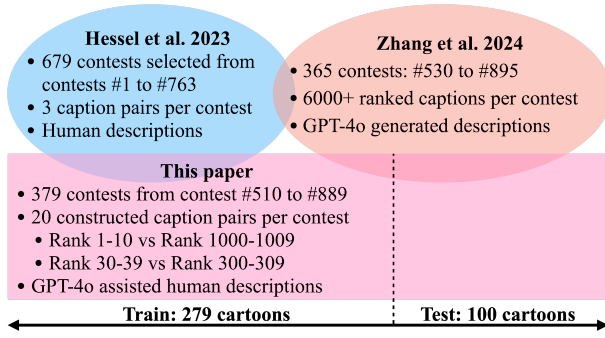
2

Figure 2: 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.

"I'm sorry, but it NOT "alimentary," my dear Watson!"

| UNFUNNY | SOMEWHAT FUNNY | FUNNY |

Figure 3: Example voting page for the caption contest.

that is contextually rich, culturally sensitive, and genuinely surprising, highlighting the need for continued research into more sophisticated models and training paradigms (Robison, 2024).

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 Recent advancements in post-training alignment techniques for LLMs have progressed through several distinct stages. Initially, supervised fine-tuning (SFT) was employed to adapt pre-trained models to specific tasks using high-quality, instruction-based datasets, demonstrating that even modest amounts of curated data can substantially improve downstream performance (Wei et al., 2021). Building on this, researchers introduced Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al.,

2022)to further align model outputs with human preferences. In this framework, Proximal Policy Optimization (PPO) is widely used to adjust the model's behavior based on human-provided preference comparisons (Schulman et al., 2017). However, the inherent complexity and instability of PPO-based RLHF motivated the development of simpler alternatives. Direct Preference Optimization (DPO) recasts the alignment objective as a supervised learning problem by directly contrasting the log-probabilities of preferred and non-preferred responses, thereby eliminating the need for an explicit reward model (Rafailov et al., 2023). More recently, extensions such as Group Relative Policy Optimization (GRPO) have been proposed, which incorporate group-level comparisons that further enhance training stability and mitigate issues like catastrophic forgetting (Guo et al., 2024). This evolution – from SFT through PPO-based RLHF to DPO and GRPO – reflects the field's ongoing efforts to develop robust, efficient, and reliable post-training alignment methods for LLMs.

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 3). At the end of each contest, a complete crowdsourced ranking of cap-

tions 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 2). 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). Previous work by Hessel et al. (2023) and Zhang et al. (2024) shows that state-of-the-art models, including variants of GPT-4, achieve only around 67% accuracy on the easy version of the pairwise task, whereas human experts significantly outperform the LLMs. These findings underscore the persistent gap between current state-of-the-art models and human expertise in humor understanding, motivating our investigation into novel approaches to enhance model performance on this task.

4 Experiments

We break the ranking challenge into three components – visual understanding, humor reasoning, and targeted alignment to human crowd preferences.

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 fine-tunes 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 2) 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 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 4). 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 accu-

---

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

4

Minor Errors	**Omission of Key Details**	**Fundamentally Incorrect**
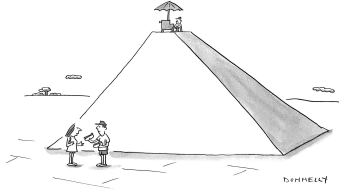
GPT-4o Two tourists are standing at the base of a pyramid, looking at a map. At the top of the pyramid, there is a vendor with an umbrella and a cart.

Human At the base of a massive pyramid, a clerk is enthusiastically pitching hotdogs to a woman, while another clerk sits atop the pyramid under an umbrella with a small cart.
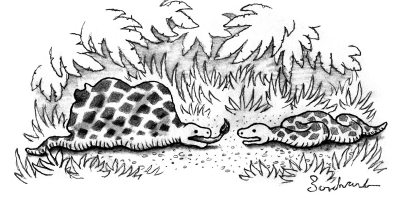
GPT-4o Three eagles are perched on a tree. One eagle is on a branch, while the other two are on another branch, seemingly engaged in a conversation.

Human An eagle with a special hairstyle perches on a branch, while behind it, two other eagles appear to be gossiping about its look.

GPT-4o Two reptiles, one resembling a turtle and the other a snake, are facing each other in a jungle setting. The turtle has a snake-like tongue extended towards the snake.

Human In the grass, two snakes meet; one is in the midst of devouring a calf-like animal, whose tail still protrudes from the snake's mouth, not yet fully swallowed.

Figure 4: 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).

racy. 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 82.4%.

4.2 Does reasoning through a joke improve humor understanding?

Humor assessment is challenging because it fuses objective cues with subjective preferences. While enhanced caption descriptions capture objective elements, they often miss the figurative aspects that make a caption truly funny.

We propose that enriching model inputs with explicit explanations can improve performance by highlighting both objective cues (e.g., wordplay) and subjective nuances (e.g., cultural context). Prior work (Hessel et al., 2023) shows that models detect objective features well but struggle with subjectivity. Encouragingly, recent reasoning models like o1 and DeepSeek (OpenAI, 2024b; DeepSeek, 2024) appear promising – our humor expert found that over 85% of o1-preview explanations effectively captured a cartoon's humor.

We generate explanations using two language models, GPT-4o and o1-preview (see Appendix A.2). As shown in Table 1, o1-preview explanations boost ranking accuracy to 76%, com-

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).

pared to 73% for the baseline and 71% for GPT-4o-generated explanations. This underscores the importance of explanation quality, with o1-preview better capturing humor nuances (see Figure 5).

Our findings indicate that equipping the ranking model with explicit explanations bridges the gap between objective cues and subjective humor elements. However, accuracy still falls short of expert human performance (90%, Table 3), possibly due to the unique subjective biases of the Caption Contest ranking crowd. In the next section, we explore aligning the verifier's preference with that of the New Yorker Caption Contest crowd to further narrow this gap.

4.3 Alignment to Crowd Preference

Despite our success in incorporating both objective structural cues and subjective reasoning-based explanations in previous sections, a significant performance gap remains between our models and humor experts. Our experiments in the pairwise caption ranking task demonstrate that while adding these elements enhances humor understanding, the

Cartoon & Caption		
	Ok, just figured out why those foursomes in front of us disappeared.	*What is even crazier is that they mate for life.*

GPT-4o

- **Unexpected Explanation**: […]
- **Absurdity and Surprise**: […]
- **Play on Golf Frustrations**: Golfers often joke about the slow pace of play and waiting for groups ahead. The caption plays on this common frustration by offering a ridiculous reason for the delay, which is relatable to golfers.
- **Visual and Verbal Contrast**:[…]

- […]
- **Surprise Element**: The caption shifts the focus from the extraordinary act of flying to the mundane fact about mating, which is unexpected and adds a layer of humor by downplaying the more astonishing aspect of the scene.
- **Irony**: The irony lies in the fact that the speaker finds the idea of lifelong partnership more surprising than the impossibility of unaided human flight, which is inherently humorous due to the reversal of expected priorities.

o1-preview

- **Unexpected Combination of Golf and UFOs**: […]
- **Play on Common Golf Frustrations**: Golfers often experience delays due to groups (foursomes) playing ahead of them. The caption humorously suggests that the missing groups have been abducted by aliens, providing a funny and exaggerated explanation for their disappearance.
- **Understated Reaction to an Absurd Situation**: […]
- **Wordplay with "Foursomes Disappeared"**: […]
- **Implied Alien Abduction**: […]

- **Role Reversal of the Extraordinary and the Ordinary**: The observers are witnessing an impossible event—humans flying unaided—but instead of expressing amazement at this extraordinary sight, they comment on the fact that "they mate for life." […]
- **Understatement and Irony**: There's an ironic twist in expressing that what's "even crazier" is the mating behavior, not the act of flying. This understatement amplifies the humor by minimizing the significance of human flight—an impossible feat—and exaggerating the importance of their mating practices.
- […]

Figure 5: 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.

verifier model (GPT-4o) still does not fully capture the nuanced humor preferences of the New Yorker caption contest crowd. We hypothesize that this discrepancy arises from a fundamental misalignment: GPT-4o's inherent subjective reasoning does not match the specific taste and evaluative criteria of New Yorker voters.

To address this misalignment, we build on our previous findings (that integrating objective and subjective elements enhances humor understanding) by exploring two alignment strategies. The first approach employs persona-based prompting to simulate the subjective evaluative criteria of active New Yorker caption contest participants, subtly steering the model's preference toward the target audience's distinctive preferences. Our second strategy takes a more direct route through supervised fine-tuning (SFT) on a large corpus of New Yorker caption contest caption ranking data released by Zhang et al. (2024). We hypothesize that direct adaptation to the target domain will better align the model's humor judgments with those of New Yorker voters, thereby narrowing the performance gap with human experts.

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

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 nine 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 3 show that persona-based prompts yield only modest improvements in aligning the model with the intended audience, with the highest accuracy of 76.5% achieved using the Female Lawyer prompt—a mere 3% gain over the baseline without any persona. These results indicate that persona-based alignment is not strong enough to capture the preferences of the New Yorker crowd. Instead, more powerful alignment strategies, such as supervised fine-tuning (SFT), are required. In the next

System Prompt	GPT-4o (%)	Claude (%)	Gemini (%)	o3-mini (%)
Empty (Baseline)	73.5	68	47	70
Judge	73	67.5	55	58
Male Lawyer	75	74	51	59
Female Lawyer	**76.5**	69	59	67
CS Phd	73.5	68	49	65
Sociologist & Psychologist	73.5	67	61	60
Literature Student	73.5	72	51	58
Bob Mankoff	73.5	66	50	62
Larry Wood	73	68	57	61
Cartoon Author	71.5	61	46	62

Table 2: Performance 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. Each row represents a distinct persona-based prompt. See Appendix B for system prompts and Appendix A.4 for the task prompt.

Methods	10vs1000	30vs300
Expert Majority Vote	$84_{\pm5.2}$	$66_{\pm6.8}$
Expert Average Accuracy	$78_{\pm2.6}$	$61.6_{\pm3.0}$
Best Expert Accuracy	$85.4_{\pm2.9}$	$68_{\pm3.8}$
Gemini 2.0 Flash Thinking	$61_{\pm3.5}$	$58_{\pm3.5}$
o1	$69_{\pm3.3}$	$58_{\pm3.5}$
o3-mini	$60.7_{\pm1.5}$	$53.2_{\pm1.6}$
Claude 3.5 Sonnet	$74_{\pm3.1}$	$61_{\pm3.5}$
GPT-4o Prompting	$67.3_{\pm1.5}$	$53.9_{\pm1.6}$
GPT-4o SFT w/o Expl. (Ours)	$79.4_{\pm1.3}$	$59.7_{\pm1.6}$
GPT-4o SFT w/ Expl. (Ours)	$\textbf{82}.4_{\pm\textbf{1.2}}$	$\textbf{63.2}_{\pm\textbf{1.5}}$

Table 3: Accuracy(%) Comparison of Different Methods. All models (Gemini, o1, o3-mini, Claude and GPT-4o) use prompting techniques including best persona, 5-shot in-context learning, and o1-preview generated explanations. Our fine-tuned GPT-4o model with humor explanations even outperforms the best human expert.

section, we detail our application of SFT to better align the model with human preferences and bridge the performance gap.

4.3.2 GPT-4o Finetuning

Despite the limited accuracy gain by persona-based prompting, aligning to a specific group of audience should not be a hard task especially when the model is given access to the correct understanding and explanations of the captions. Indeed, once we finetune SOTA LLMs on a small set of human preferences, we recover a significantly higher ranking accuracy. This also reveals an interesting failure case where LLMs fail in understanding subgroup and individual human preferences for subjective tasks (more discussions 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 between 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. To the best of our knowledge, we are the first to evaluate performance on this task. 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 3, fine-tuned GPT-4o models can significantly improve upon all prompting-based baseline before. When incorporating o1-preview generated explanations, the finetuned GPT-4o can achieve slightly higher accuracy than the average over 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 1). 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 3 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: To achieve AGI, LLMs require much more human interaction data to acquire the understanding of individual/subgroup level preferences.

Our empirical findings on humor comprehension point to a broader challenge in artificial intelligence: the development of true creative understanding. While recent advances in LLMs have demonstrated remarkable capabilities in analytical reasoning and structured problem-solving, our results suggest that creative domains may present a unique and potentially final hurdle in achieving AGI. We argue that this challenge stems from two fundamental 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. This suggests that current approaches to AI alignment, which often focus on optimizing for verifiable metrics, may be insufficient for creative domains.

Second, and perhaps most challengingly, cre-

ative excellence requires understanding and internalizing group-specific preferences and cultural contexts. Our finding that persona-based prompting failed to improve caption ranking, while direct preference learning proved effective, highlights a crucial gap in current LLM capabilities. 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 and preferences can vary dramatically across different communities?

These observations lead us to propose that achieving AGI may fundamentally require solving the challenge of preference understanding. While we can use reinforcement learning and inference-time scaling techniques to improve creative generation once we have reliable judgment models, the path to AGI requires models that can develop generalizable insights about how preferences function across different contexts and domains. This suggests that just as human creativity is deeply influenced by understanding others' perspectives and preferences, AGI systems will need to develop a fundamental grasp of how preferences work—not just in individual domains, but as a generalizable concept—to achieve truly intelligent creative behavior.

6 Conclusion

Our work demonstrates that by decomposing humor understanding 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 fine-tuning 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.

8

7 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.

- **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 fine-tuning 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.

9

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*.

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.

DeepSeek. 2024. Deepseek. https://api-docs.deepseek.com/.

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

Daya Guo, Qihao Zhu, Dejian Yang, and Junxiao Song. 2024. Group relative policy optimization for aligning large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. In press.

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.

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*.

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*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

10

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

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*.

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

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*.

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.

11

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 canny 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
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.

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.

Metrics	10vs1000	30vs300
Average Accuracy	78	61.6
Highest Accuracy	90	68
Majority Accuracy	84	66
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.

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).