Which LLMs Get the Joke? Probing Non-STEM Reasoning Abilities with HumorBench

Anonymous ACL submission

Abstract

We present HumorBench, a benchmark designed to evaluate large language models' (LLMs) ability to reason about and explain sophisticated humor in cartoon captions. As reasoning models increasingly saturate existing benchmarks in mathematics and science, novel and challenging evaluations of model intelligence beyond STEM domains are essential. Reasoning is fundamentally involved in text-based humor comprehension, requiring the identification of connections between concepts in cartoons/captions and external cultural references, wordplays, and other mechanisms. HumorBench includes approximately 300 unique cartoon-caption pairs from the New Yorker Caption Contest and Cartoonstock.com, with expert-annotated evaluation rubrics identifying essential joke elements. LLMs are evaluated based on their explanations towards the humor and abilities in identifying the joke elements. To perform well on this task, models must form and test hypotheses about associations between concepts, potentially backtracking from initial interpretations to arrive at the most plausible explanation. Our extensive benchmarking of current SOTA models reveals three key insights: (1) LLM progress on STEM reasoning transfers effectively to humor comprehension; (2) models trained exclusively on STEM reasoning data still perform well on HumorBench, demonstrating strong transferability of reasoning abilities; and (3) test-time scaling by increasing thinking token budgets yields mixed results across different models in humor reasoning.

1 Introduction

Recent advances in large language models and reasoning techniques have led to the saturation of many existing benchmarks, particularly in STEM domains such as mathematics and programming, where frontier models now approach or exceed human-level performance (Abdin et al., 2025a; Sun et al., 2025; Quan et al., 2025). This progression highlights the need for novel and challenging evaluations that can meaningfully differentiate model capabilities and provide insights into their reasoning processes. Non-STEM reasoning tasks, particularly those involving cultural understanding and implicit knowledge, represent underexplored territories for model evaluation.

Humor comprehension represents a particularly challenging frontier for artificial intelligence (Hessel et al., 2023; Zhang et al., 2024; Zhou et al., 2025; Kazemi et al., 2025; Liang et al., 2025). Although large language models (LLMs) excel across many domains, understanding humor still requires sophisticated reasoning that integrates context, cultural knowledge, and implicit connections. These challenges make humor an ideal testbed for evaluating advanced reasoning in AI systems.

We present **HumorBench**, a benchmark that evaluates LLMs' ability to explain sophisticated cartoon-caption humor by identifying the mental leaps connecting visuals, captions, and external knowledge (Figure 1). For each pair, we annotate the objective elements essential for comprehension, creating a ground truth focused on factual connections rather than subjective appreciation.

We benchmark both a standard set and a harder subset. On HumorBench-hard, which features more complex examples requiring multiple reasoning steps or obscure cultural knowledge, no current LLM exceeds 60% accuracy. Our benchmarking of current state-of-the-art models reveals two key findings:

1. We observe a high correlation between performance on HumorBench and existing STEM benchmarks, suggesting a significant transfer of general reasoning abilities to humor comprehension tasks.
2. Even models trained exclusively on STEM reasoning tasks (e.g., mathematical problem-solving) perform well on HumorBench, indicating that abstract reasoning skills acquired in one

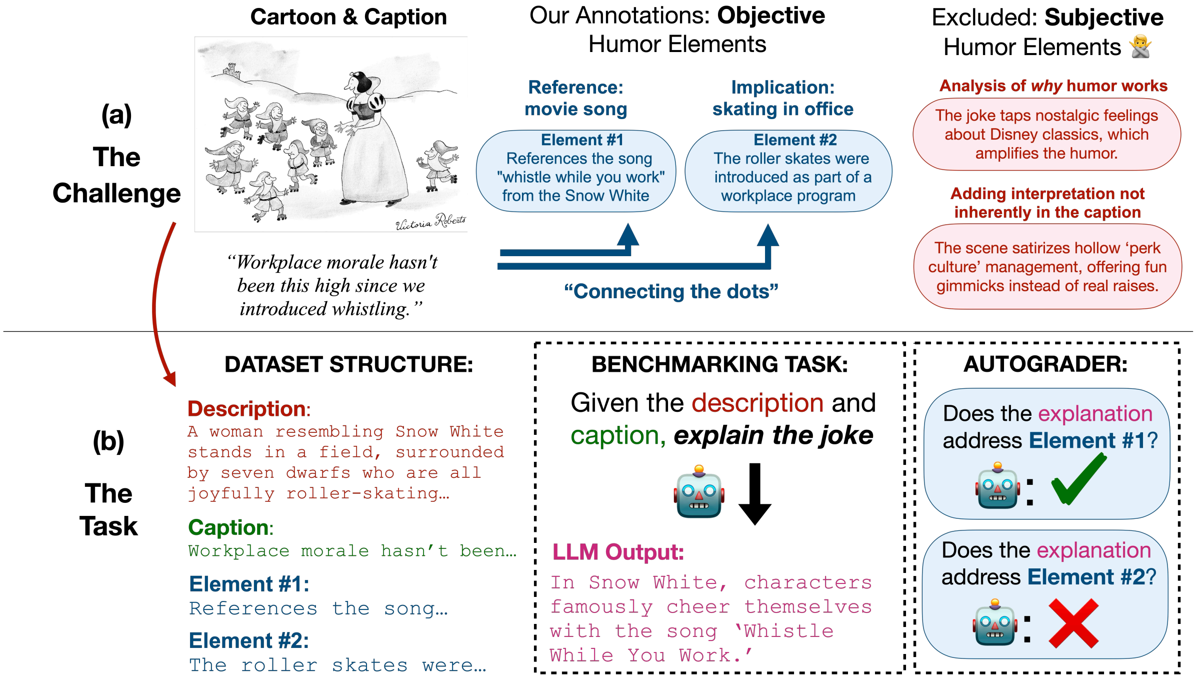

(a)
The Challenge
Cartoon & Caption
Victoria Roberts
"Workplace morale hasn't been this high since we introduced whistling."
Our Annotations: Objective Humor Elements
Reference: movie song
Element #1
References the song "whistle while you work" from the Snow White
Implication: skating in office
Element #2
The roller skates were introduced as part of a workplace program
"Connecting the dots"
Excluded: Subjective Humor Elements
Analysis of why humor works
The joke taps nostalgic feelings about Disney classics, which amplifies the humor.
Adding interpretation not inherently in the caption
The scene satirizes hollow 'perk culture' management, offering fun gimmicks instead of real raises.
(b)
The Task
DATASET STRUCTURE:
Description:
A woman resembling Snow White stands in a field, surrounded by seven dwarfs who are all joyfully roller-skating…
Caption:
Workplace morale hasn't been…
Element #1:
References the song…
Element #2:
The roller skates were…
BENCHMARKING TASK:
Given the description and caption, explain the joke
LLM Output:
In Snow White, characters famously cheer themselves with the song 'Whistle While You Work.'
AUTOGRADER:
Does the explanation address Element #1?
Does the explanation address Element #2?

Figure 1: Overview of our humor analysis approach. (a) We distinguish between **objective** and **subjective** components of a joke. To convert the open-ended task of humor explanation into a fair benchmark, we focus exclusively on **objective** elements. (b) Overview of the dataset, benchmark task, and grading scheme in HumorBench. Each cartoon-caption pair contains one or more "element" annotation. For the benchmark, an LLM is tasked with explaining the joke in the caption. An autograder evaluates if the explanation contains each element.

domain can transfer effectively to humor comprehension.

3. Test-time scaling measures for humor reasoning yield mixed results, indicating that simply increasing computational resources at inference time does not consistently improve performance on this challenging domain.

1.1 Why Another LLM Humor Benchmark?

Several benchmarks have already focused on measuring LLMs' capabilities around humor. Specifically, Hessel et al. (2023) and Zhang et al. (2024) both build upon the New Yorker Caption Contest (NYCC) dataset—a weekly feature by the New Yorker magazine where readers submit funny captions for cartoons (see Figure 1 for an example). Hessel et al. (2023) created three benchmarks using this dataset: ranking the funniness of caption pairs, matching cartoons to valid captions, and explaining the humor behind captions. However, these benchmarks simultaneously measure two distinct capabilities: (1) understanding the intended jokes (objective elements) and (2) aligning with individual and subgroup humor preferences (subjective factors). As Zhou et al. (2025) points out, performance on these previous benchmarks is heavily influenced by an LLM's ability to align with specific audience preferences rather than directly measuring its reasoning about the jokes themselves.

Our benchmark, HumorBench, addresses this limitation by focusing solely on the objective elements of humor comprehension, specifically measuring the *humor reasoning abilities* required to understand cartoons and their captions. As validated by our experimental findings, LLMs' performance on HumorBench correlates well with their performance on other reasoning benchmarks.

2 Related Work

Reasoning-focused language models. Recent advances in large language models (LLMs) have seen the emergence of specialized reasoning models that excel at logical deduction, mathematical problem-solving, and multi-step reasoning while maintaining strong general language capabilities. These reasoning-enhanced models employ various approaches: training-focused methods like those used by *Minerva* (Lewkowycz et al., 2022), *WizardMath* (Luo et al., 2023), and *Phi-4-Reasoning* (Abdin et al., 2025b) leverage carefully curated STEM-heavy corpora; inference-time techniques boost reasoning without changing model weights, including *Self-consistency* (Wang et al., 2023), *Tree-of-Thought* methods in *DeepSeek-Math* (Shao et al., 2024), *Least-to-most* prompting (Zhou et al., 2023), and *Process supervision* (Lightman et al., 2023); while hybrid approaches like *MAmmoTH* (Yue et al., 2023) combine diverse training data with structured inference protocols, *ToRA* (Gou et al., 2023) integrates formal verification systems, and *MathGLM* (Yang et al., 2024) combines symbolic computation with natural language reasoning. Models like *Gemini Ultra* (Google, 2024) and *Claude 3.5* (Anthropic, 2024) achieve strong reasoning through both architectural innovations and sophisticated training, suggesting that advances in machine reasoning now follow multiple complementary paths rather than relying solely on parameter count (Wei et al., 2024).

Humour benchmarks. Beyond simple joke generation, several resources now probe LLM humour competence. Hessel et al. (2023) introduces three New Yorker cartoon-caption subtasks that test multimodal humour understanding and explanation. For word-play, the *ExPUNations* corpus augments classic pun datasets with human-written explanations and funniness ratings (Sun et al., 2022), while Xu et al. (2024) systematically benchmarks pun recognition, explanation and creation. Complementing these datasets, (Ermakova et al., 2025) Lab provides reusable test collections for humour-aware information retrieval.

Open-ended evaluation frameworks. Automatic grading of creative, unconstrained outputs increasingly relies on the *LLM-as-Judge* paradigm. G-EVAL couples chain-of-thought GPT-4 judging with a form-filling rubric, achieving human-level reliability on summarisation and dialogue (Liu et al., 2023). MT-BENCH and its crowdsourced *Chatbot Arena* show that GPT-4 judges agree with human preferences on multi-turn instruction following in ∼80% of cases (Zheng et al., 2023). Going further, PAPERBENCH grades agents on reproducing ICML-level research papers with hierarchical GPT-4 rubrics and expert audits (Starace et al., 2025). We adopt a similar rubric-guided judging scheme but focus specifically on humour reasoning, enabling systematic comparison of explanation quality across models.

3 HumorBench

3.1 Main Benchmark Task

HumorBench frames humor understanding as an open-ended task: given a textual description of a cartoon and its caption, a model must articulate in its own words the underlying joke. We deliberately avoid the multiple-choice or ranking formats common in existing humor benchmarks because, for creative tasks, fixed answer sets can (i) inadvertently hint at the punchline and (ii) fail to accommodate the diverse range of valid explanations a competent reader might produce.

To make this free-form setting automatically gradable, we distill each cartoon into a concise rubric of 1–3 objective “elements." An element represents a single, easily verifiable fact that any correct explanation must include (e.g., in NYCC Contest #665, the observation that “the shark interprets the swimmer as groceries"), as shown in Figure 8. This approach allows for creative expression while maintaining consistent evaluation standards (see Appendix A for the complete prompt).

3.2 Dataset: Cartoon and Caption Sources

Our dataset comprises cartoons and captions from two primary sources: the New Yorker Caption Contest (NYCC) and Cartoonstock.com. We sourced NYCC captions from publicly available datasets (Hessel et al., 2023; Jain et al., 2020; Zhang et al., 2024), selecting only those ranked among the top 3 finalists to ensure each cartoon features a coherent, high-quality joke. For Cartoonstock cartoons, we utilized their original accompanying captions. Both sources specialize in dry, witty humor that demands sophisticated reasoning—often requiring multiple mental leaps to fully comprehend, as illustrated in Figure 1.

While cartoons inherently include visual elements, our benchmark focuses on testing humor comprehension rather than visual interpretation capabilities. Therefore, we created detailed textual descriptions of each cartoon, carefully capturing all information necessary to understand the caption while maintaining neutrality. These descriptions include essential details about the setting, characters, visible emotions, and speaker identification, while deliberately omitting artistic style unless directly relevant to the joke. For researchers interested in extending this to a multimodal benchmark, we provide source links to the original images—NYCC images are available through (Hessel et al., 2023; Jain et al., 2020), while Cartoonstock images require licensing.

3.3 Dataset: Element Annotation

The core labels in our dataset are the *element* annotations assigned to each cartoon–caption pair.

For every pair, we hand-annotated one to three elements—concise, direct statements that capture the objective components essential to understanding the joke. As discussed in Section 1.1, comprehending cartoon humor requires two distinct capabilities: understanding the objective content of the joke and recognizing the subjective aspects that influence audience reception. Figure 1 illustrates this distinction—subjective explanations focus on audience reactions (which vary between individuals), while objective elements center on content comprehension. Our benchmark specifically targets these objective elements, which require identifying the mental leaps necessary to "get" the joke through recognizing references, wordplay, implications, or similar mechanisms. The autograder then evaluates LLM explanations against these elements, verifying that each explanation adequately covers the fundamental objective components of the humor, ensuring a fair and consistent assessment across different models.

In summary, to make this task easily gradable, annotations follow a set of deliberate guidelines: (1) Elements must be short, direct, and easily verifiable from the description and caption; (2) An element addresses exactly one concept. Bundling multiple ideas may add noise by forcing the autograder to guess about partial correctness; (3) Elements deliberately avoid adding bias from subjective opinion about humor.

3.4 Dataset Refinement

For an LLM evaluation to provide trustworthy results, the underlying dataset must be both accurate and internally consistent. We initially collected 655 unique element annotations, but despite careful guidelines, some entries proved vague or imprecise. To systematically improve quality, we implemented an iterative refinement process.

First, we generated sample explanations for each cartoon–caption pair, alternating randomly between GPT-4o and Claude 3.7 Sonnet. Each explanation was evaluated ten times by our autograder, with elements showing verdict disagreement exceeding 30% flagged for review. These problematic cases were either refined or removed entirely. We repeated this quality control cycle until fewer than 5% of annotations triggered inconsistency flags, ultimately resulting in 499 high-quality unique element annotations forming the foundation of HumorBench.

As an additional validation step, we invited a former chief cartoon editor of the New Yorker to review a random subset of 30 annotations. The editor confirmed that all elements were fair and accurately captured the essential components of each joke. Together, these atomic, objectively verifiable criteria create a robust rubric that enables our autograder to provide consistent and reliable evaluation at scale.

3.5 Autograder and Evaluation

During evaluation, an LLM judge assesses each model's explanation against individual elements to determine whether they adequately cover the essential components of the joke. This approach allows us to efficiently evaluate open-ended text generation at scale.

However, ensuring autograder consistency presents challenges, particularly for tasks that are inherently difficult for LLMs to understand (Min et al., 2020; Starace et al., 2025). To address this, we created a separate benchmark of 300 human expert judgments on explanations from three distinct LLMs: GPT-4o, Gemini 2.5 Pro, and Claude 3.7 Sonnet. Using GPT-4o as the autograder, we achieved 92% accuracy overall:

Explainer Model	Acc. (%)	FPR (%)	FNR (%)
Overall (n=300)	92.00	14.79	6.51
Gemini 2.5 Pro	93.00	10.00	6.25
GPT-4o	92.00	14.81	5.48
Claude 3.7 Sonnet	91.00	19.57	7.80

Table 1: Autograder performance (GPT-4o judge) on 300 human-labeled explanations.

This validation provided two key insights. First, across all models, the autograder's false positive rate (FPR) substantially exceeded its false negative rate (FNR), indicating a leniency bias. This suggests that *HumorBench scores should be interpreted as an upper bound on model performance*. Second, despite using GPT-4o as the autograder, we observed no significant advantage for GPT-4o-generated explanations compared to those from other models. Together, these findings confirm that our autograder provides a valid, albeit slightly optimistic, mechanism for large-scale evaluation.

Length Control. While models were instructed to keep responses under 200 words, some models exceeded this limit, particularly when reasoning traces were included in the final output. To ensure fair comparison, we truncated all model outputs to

NYCC Contest 167
Description: In a forest clearing with tall tree trunks, a man in a suit with an American-flag lapel pin speaks at a podium while aides stand behind him. Woodland animals—a deer, snake, frog, and birds—peek from the trees and grass, watching.

Caption: “As a weasel, I need your vote.”

ELEMENT 1: References the cliché insult of calling politicians *weasels*.

ELEMENT 2: Plays on the dual meaning of “weasel” (literal animal & political pejorative), creating a pun.

Figure 2: Example HumorBench annotation. The cartoon (*left*) is paired with its description, caption, and two hand-labeled joke elements (*right*).

the last 1000 tokens.

4 Experiments

Along with creating the HumorBench evaluation, we extensively benchmarked current frontier models. For consistency, all models are given the same prompt and scaffolding describing the task (see Appendix A). We arrived at this prompt after validating across several different LLMs (Claude 3.7 Sonnet (Anthropic, 2025), GPT-4o (OpenAI, 2024a), Gemini 2.5 Pro (DeepMind, 2025)). While many LLMs had different API endpoints, we tried to maintain consistent parameters where possible. For example, all models had temperature set to 1 and external tool calling deactivated. Note, for all evaluations, autograders, and benchmarks, “GPT-4o" refers to the gpt-4o-2024-08-06 release.

4.1 Main Results

In general, the results from the main benchmarking effort were unsurprising. As shown in Figure 3, OpenAI o3 (OpenAI, 2025a) leads the pack at 87.5% accuracy, dramatically ahead of other SOTA models (Gemini 2.5 Pro, Claude 3.7 Sonnet, and Deepseek R1 (DeepSeek, 2025)), all achieving approximately 80%. In general, smaller models (like Llama 4 Maverick (Meta, 2025), Qwen 2.5 (Alibaba, 2025), and o3-mini (OpenAI, 2025b)) performed worse. We also found that newer versions of models generally dominate older versions of the same model, with o3 outperforming o1 (OpenAI, 2024b) and Gemini 2.5 pro outperforming Gemini 1.5 pro. In general, "reasoning" versions of models seemed to outperform the base versions of the same model. For example, DeepSeek R1 (79.8%) strongly outperformed Deepseek V3 (DeepSeek, 2024) (72.2%), despite being based on the same 671B parameter architecture. Similarly, Claude 3.7 Sonnet with a thinking budget of 1024 tokens (83.6%) clearly outperformed the base Claude 3.7 Sonnet (80.4%). When compared with total cost of running the benchmark, we see that more expensive models tend to outperform less expensive models, either due to a larger underlying model or using more reasoning tokens in the output.

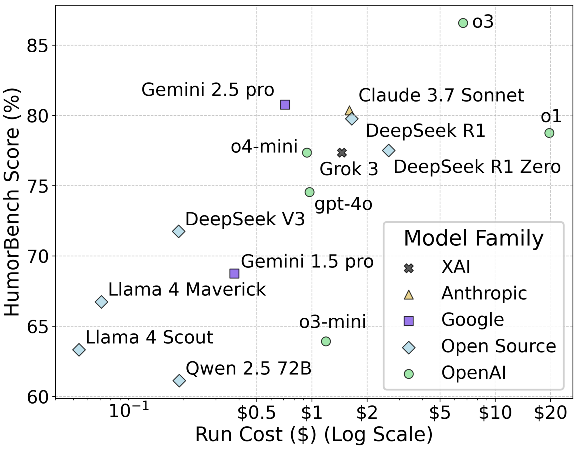


Figure 3: Benchmarking results on several frontier models

4.2 Transferability of Reasoning Skills

To gauge how well other model skills transfer to humor comprehension, we correlate HumorBench accuracy with three widely used LLM benchmarks: GPQA-Diamond (Rein et al., 2023), ARC-AGI (Chollet et al., 2025), and LM Arena ELO (Chiang

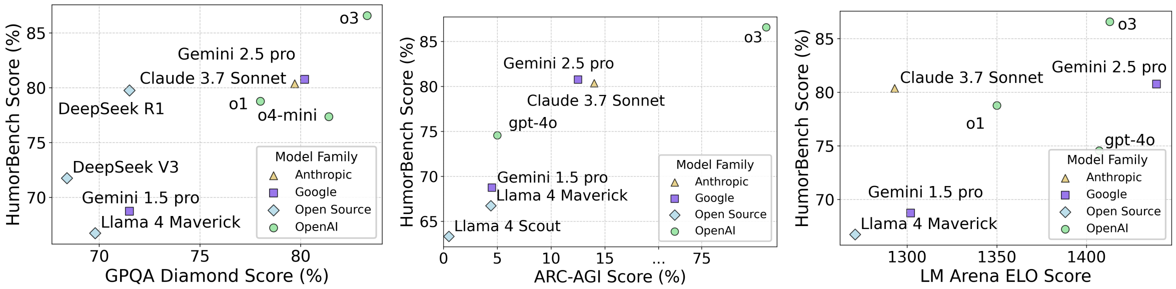


Figure 4: HumorBench performance compared to several common benchmarks. We see positive correlation with GPQA, ARC-AGI, and LMArena. In particular, ranking compared to ARC-AGI is nearly identical to that of HumorBench, indicating a strong reasoning component to the HumorBench task.

Benchmark	Corr.	p-value
GPQA Diamond	0.736*	0.024
ARC-AGI (with o-series)	0.650	0.058
ARC-AGI (w/o o-series)	0.943**	0.005
LM Arena ELO	0.714	0.071

Table 2: Spearman's rank correlations between HumorBench and other benchmarks. Asterisks indicate significance: $*p < 0.05$, $**p < 0.01$. ARC-AGI correlation shown separately for results with and without o-series models, which were fine-tuned for ARC-AGI

et al., 2024). HumorBench scores are positively associated with all three. In particular, after removing o-series models (whose reported scores come from ARC-tuned variants) the correlation with ARC-AGI rises to $\rho = 0.943$ ($p = 0.005$), underscoring the shared reasoning demands of the two tasks. The LM Arena correlation ($\rho = 0.714$) is solid but notably lower. Overall, this also suggests that LLM progress on STEM domains translates to Non-STEM reasoning as well.

STEM-only reasoning improves HumorBench Our comparison between reasoning models trained on STEM tasks via Reinforcement Learning and their base counterparts yielded particularly revealing results. As illustrated in Figure 6, R1-Zero, which developed reasoning capabilities exclusively through self-play on STEM problems, demonstrated significant improvements over its base V3 model. Remarkably, it performed nearly on par with DeepSeek R1, despite the latter being trained on non-STEM data such as reading comprehension. Similarly, Phi-4 Reasoning Plus exhibited superior performance compared to its base model (Figure 6), although its training was limited to math and coding data (Abdin et al., 2025a). These findings suggest that abstract reasoning capabilities are transferable to humor comprehension, indicating that the reasoning skills required for STEM domains may be fundamentally similar to those needed for understanding humor.

We also note that both R1-Zero and Phi-4 Reasoning Plus include their reasoning traces in their final outputs. Therefore, we evaluated their performances using the length control measure described above to ensure fair comparison across models.

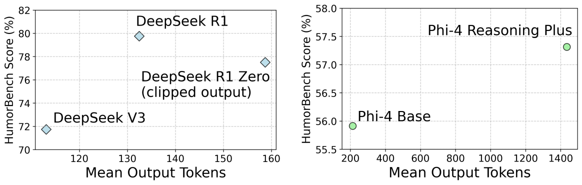
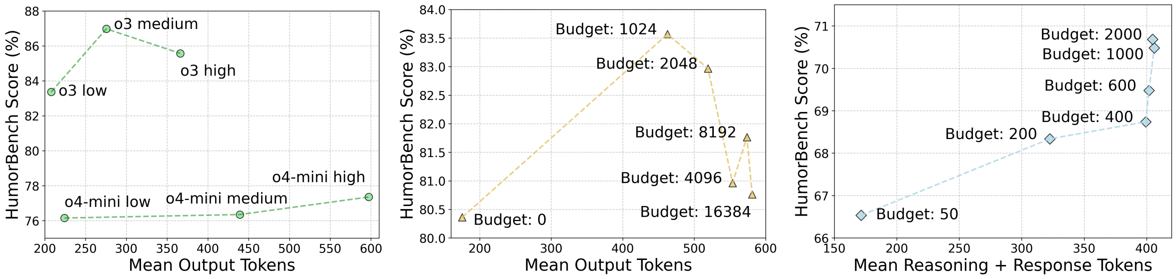


Figure 6: Deepseek R1 Zero and Phi-4 Reasoning Plus, both exclusively reasoning-trained on STEM tasks, outperform their base versions on HumorBench

4.3 Test-Time Scaling

As seen in 5, while including *some* reasoning clearly helped model performance, the effect of continuing to increase test-time compute varied significantly between models. For Qwen plus and the o- series models, increasing the reasoning parameter (reasoning budget and "effort", respectively) generally improved performance. However, for Claude 3.7 Sonnet, increasing the thinking budget beyond the minimum 1024 tokens clearly *hurt* performance.

A closer look at the reasoning trace lengths highlights that for most captions, the models did not fully exhaust their reasoning budget (see **??**). Qwen Plus, for example, rarely used more than 400 tokens, even when budgeted 2000, a trend we saw for all test-time experiments. This suggests the



(a) OpenAI o- series models, varying "reasoning effort"

(b) Claude 3.7 Sonnet performance, varying "thinking budget"

(c) Qwen Plus performance, varying "thinking budget"

Figure 5: HumorBench test-time compute experiments. Note, "mean output tokens" includes both reasoning and final response tokens

LLMs are providing final answers based on completed thinking traces, which makes the inverse test-time scaling effect more puzzling. While a few studies have looked related problems (Su et al., 2025; Shi et al., 2023; Yang et al., 2025; McKenzie et al., 2023), we defer thorough investigation of our observation to future work.

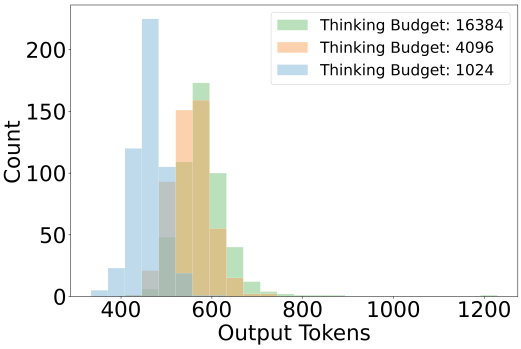


Figure 7: Token usage with different ‘thinking budget‘ parameters on Claude 3.7 sonnet. Most completions (in the hundreds of tokens) were far below their budgets.

4.4 Analysis of HumorBench Hard Subset

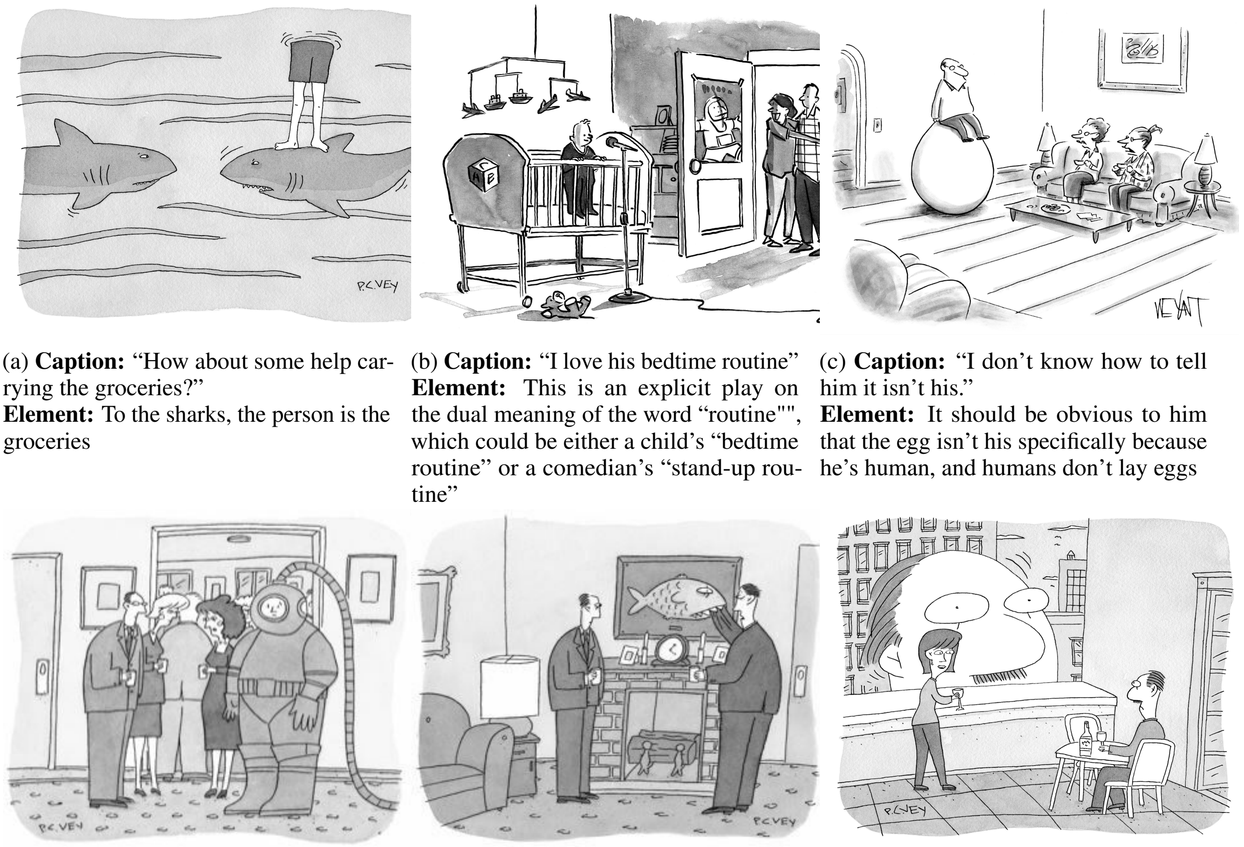
While frontier models like o3 demonstrate impressive performance on HumorBench, certain elements are persistently challenging for all models. To better understand the specific types of humor that remain challenging, we conducted a targeted analysis on the 100 unique elements that were most often missed during the benchmarking, which we call HumorBench Hard. Cartoons in this subset range from pass rates of 60% (6 in 10 models get correct) to 0% (No model gets correct). See 3 for examples of the hard subset.

To get a more granular view of the elements that constitute the hard subset, we analyzed three predefined humor categories: *wordplay*, *cultural references*, and *toxic or shocking* humor elements. These categories were annotated by an LLM categorization pipeline built on o3, which individually categorized each element as in or out of each category. Most elements did not fit into these categories, while some fit into multiple. Our analysis examined the relative representation of each category within the hard subset compared to their representation in the overall HumorBench dataset.

Humor Category	Entire Set	Hard Subset	Diff (%)
Wordplay	24.4%	19.0%	-5.4
Cultural Reference	19.0%	17.7%	-1.3
Toxic or Shocking	25.7%	26.6%	+0.9

Table 3: Representation of humor categories in the full dataset compared to the hard subset.

We summarize the main findings of this analysis in Table 3. Overall, these relatively minor deviations indicate that humor category alone is not the primary determinant of difficulty for models. Challenging examples likely hinge on subtler factors, such as the implicit conceptual leaps required or the obscurity or references. We observed that wordplay was slightly *under-represented* among the hard subset (-5.4%), suggesting current LLMs handle puns or jokes that rely on linguistic manipulation somewhat better than other elements. *Cultural references* and *toxic or shocking* humor, meanwhile, were essentially evenly represented, indicating that these styles do not disproportionately increase difficulty. These nuanced insights encourage further qualitative investigation into the underlying reasons why particular humor instances remain difficult, even for state-of-the-art LLMs trained explicitly with reasoning capabilities.



(a) **Caption:** "How about some help carrying the groceries?"
Element: To the sharks, the person is the groceries

(b) **Caption:** "I love his bedtime routine"
Element: This is an explicit play on the dual meaning of the word "routine"", which could be either a child's "bedtime routine" or a comedian's "stand-up routine"

(c) **Caption:** "I don't know how to tell him it isn't his."
Element: It should be obvious to him that the egg isn't his specifically because he's human, and humans don't lay eggs

(d) **Caption:** "This suit looked way better in the store."
Element: This is a play on words on "suit," as in a formal suit or a diving suit.

(e) **Caption:** "It's painful, but I couldn't stand another chorus of 'Take Me to the River."'
Element: This references Big Mouth Billy Bass, a classic novelty prop of a singing fish, singing "Take Me to the River"

(f) **Caption:** "How much did you spend at Macy's this year?"
Element: This implies the enormous person is a balloon for the Macy's day parade

Figure 8: Examples of elements in the HUMORBENCH hard subset. See Table 4 for full descriptions.

5 Conclusion and Future Work

In this work, we introduced **HumorBench**, the first large-scale evaluation that isolates *humor comprehension*, as opposed to subjective funniness, by grading model explanations against concise, expert-annotated *objective elements*. Our experiments with more than a dozen frontier and open-source LLMs revealed that (i) progress on STEM reasoning benchmarks translates strongly to non-STEM humor reasoning, (ii) specialized "reasoning" variants consistently outperform base models even when they were trained only on STEM corpora, and (iii) test-time compute helps, but only up to the point where the relevant background knowledge is actually present in the model. Together, these findings position HumorBench as a sensitive probe of higher-level reasoning that remains comfortably unsolved: the best model still misses over 40% of elements in our hard subset.

Looking forward, there are a number of promising avenues for extending this work. For example, a natural next step is to develop a *multimodal* version of HumorBench by reintroducing the original cartoon images. This would allow evaluation of both visual recognition and reasoning, better reflecting the complete humor comprehension task. Alternatively, the use of HumorBench's element rubrics as intermediate supervision signals offers an exciting opportunity to explore fine-tuning or reinforcement learning. Finally, improving the reliability of LLM-as-judge evaluation remains an important open challenge, particularly for creative tasks with high output variability.

We hope HumorBench spurs progress on reasoning that bridges the gap between logical deduction and nuanced human culture, and serves as a springboard for genuinely funny, culturally aware AI systems.

Limitations

HumorBench focuses on textual explanations derived from detailed image descriptions. While this controls for pure vision problems, it inevitably removes part of the challenge: recognizing visual cues. The autograder, despite a 92% agreement with humans, is mildly lenient, so reported scores should be seen as upper bounds. Additionally, while extensive validation and refinement were applied, some element annotations may still contain errors or reflect subjective interpretations that slipped through, introducing noise into model evaluation. Finally, the dataset size (499 unique elements) limits the statistical resolution of fine-grained analyses.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio C. T. Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025a. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025b. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Alibaba. 2025. Tongyi qianwen (qwen) 3 series and qwen 2.5 models. https://qwenlm.github.io/blog/qwen2.5-max/. Accessed 19 May 2025.

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

Anthropic. 2025. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed 19 May 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2025. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*.

Google DeepMind. 2025. Gemini 2.5 pro: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed 19 May 2025.

DeepSeek. 2024. Introducing deepseek v3: 671b-parameter mixture-of-experts model. https://api-docs.deepseek.com/news/news1226. Accessed 19 May 2025.

DeepSeek. 2025. Deepseek r1 release. https://api-docs.deepseek.com/news/news250120. Accessed 19 May 2025.

Liana Ermakova, Tristan Miller, Fabio Regattin, Antoine Bosselut, Saurabh Chaudhuri, Benoît Jeanjean, Sébastien Lefait, Stan Matwin, Véronique Moriceau, Patrick Saint-Dizier, and Lucas Vial. 2025. Overview of the clef 2025 joker lab: Automatic humour processing–recognition, retrieval, and generation. In *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum*.

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

Zhibin Gou, Zhihong He, Jian Wang, Xingyao Gao, Tong Wu, Jingze Bai, Juanzi Chen, Oriol Vinyals, Luke Perkins, Bing Sun, , and 1 others. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*.

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.

Lalit Jain, Kevin Jamieson, Robert Mankoff, Robert Nowak, and Scott Sievert. 2020. The new yorker cartoon caption contest dataset.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, and 1 others. 2025. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*.

Tuo Liang, Zhe Hu, Jing Li, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui Peng, Jing Ma, and Yu Yin. 2025. When 'yes' meets 'but': Can large models comprehend contradictory humor through comparative reasoning? *Preprint*, arXiv:2503.23137.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Tom Brown, Ben Mann, Sheer Agnihotri, Casey Möller, Nick Kalaitzis, and 1 others. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Yang Liu, Dan Iter, Yichong Xu, and Shuohang Wang Yelong Peng. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11069–11081. Association for Computational Linguistics.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, and 1 others. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed 19 May 2025.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*. Published as a long paper at EMNLP 2020.

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. 2025a. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed 19 May 2025.

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

Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, Zekun Wang, Jian Yang, Zeyu Cui, Yang Fan, Yichang Zhang, Binyuan Hui, and Junyang Lin. 2025. Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings. *arXiv preprint arXiv:2501.01257*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Zhihong Shao, Ganqu Lu, Jishuai Yu, Hang Wang, Cong Zheng, Yu Wang, Tao Zhao, Zherui Yang, Xipeng Chen, Xuetao Mao, Dan Su, and Zhilin Xu. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Matteo Starace, Adam Fisch, Nora Kassner, Jason Wei, Johannes Welbl, Marjan Ghazvininejad, Sebastian Riedel, and Jack Merullo. 2025. Paperbench: Benchmarking large language models for scientific paper understanding and knowledge production. *arXiv preprint arXiv:2501.12077*.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. 2025. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. *arXiv preprint arXiv:2503.21380*.

Jiao Sun, Ajay Nagesh, and Zachary C. Lipton. 2022. Expunations: Augmenting puns with keywords and explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2866–2879. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, William Fedus, Percy Liang, and Denny Zhou. 2024. Emergent reasoning in large language models. *arXiv preprint arXiv:2401.02642*.

Zixuan Xu, He He, Emmanuele Chersoni, Yi-Lin Tuan, and Qin Lu. 2024. A good pun is its own reword: Evaluating pun recognition, explanation, and generation in llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4213–4228. Association for Computational Linguistics.

Liangyi Yang, Yihuai Wang, Xinyan Du, Zhekai Zhang, Haotian Shi, Zhengying Zhu, and 1 others. 2024. Mathglm: Towards generalizable mathematical reasoning via language modeling. *arXiv preprint arXiv:2401.14273*.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*.

Xiang Yue, Boshi Ma, Yifan Wang, Junhong Chen, Xuanhe Gu, and Diyi Liang. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. In *Advances in Neural Information Processing Systems*.

Jifan Zhang, Lalit Jain, Yang Guo, Jiayi Chen, Kuan Lok Zhou, Siddharth Suresh, Andrew Wagenmaker, Scott Sievert, Timothy Rogers, Kevin Jamieson, and 1 others. 2024. Humor in ai: Massive scale crowd-sourced preferences and benchmarks for cartoon captioning. *arXiv preprint arXiv:2406.10522*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *International Conference on Learning Representations*.

Kuan Lok Zhou, Jiayi Chen, Siddharth Suresh, Reuben Narad, Timothy T Rogers, Lalit K Jain, Robert D Nowak, Bob Mankoff, and Jifan Zhang. 2025. Bridging the creativity understanding gap: Small-scale human alignment enables expert-level humor ranking in llms. *arXiv preprint arXiv:2502.20356*.

A Main Benchmark Prompt

```
You are a humor expert extraordinaire, judging the New Yorker Cartoon Caption Contest. Your
current task is to help us understand the humor in various submitted captions. Given a cartoon
description and a caption submission, explain (in less than 200 words) *what* the joke is,
focusing on the material substance of the joke. STRICTLY use the format: <explanation>explanation
goes here</explanation>
Cartoon description: description Caption: caption
```

B Autograder Prompt

```
You will receive: 1. A short cartoon description 2. A winning funny caption 3. A student’s
answer 4. A brief “anticipated answer point” that captures the crucial comedic device or element
Your job is to determine whether the student’s answer **explicitly covers** that “anticipated
answer point.”
- If the student’s answer captures or discusses the key comedic element (even if the wording is
different), **PASS**. - If the student’s answer **omits** or **contradicts** that key comedic
element, **FAIL**. - Do not penalize extra details or expansions. Synonyms or paraphrasing are
acceptable if they convey the same comedic logic. - Be mindful: if the anticipated answer point
emphasizes something specific (e.g. a pun, wordplay, or ironic twist), check that the student’s
answer includes it.
At the end of your evaluation, provide exactly two XML tags: 1. <reasoning>Short explanation of
your thought process</reasoning> 2. <judgement>PASS or FAIL</judgement>
Do not include additional commentary or deviation from this format.
Cartoon description: {description} Caption: {caption} Student’s answer: {explanation}
Anticipated answer point:{anticipated point}
```

C Hard Example Details

D Model performance on HumorBench Hard

Cartoon ID	Description	Caption	Element(s)	Pass Rate (%)
CC123065	Inside a workshop like room, three elves in pointy hats sit at a long table with open laptop computers. The middle elf appears distressed and is speaking, while the two elves on either side look toward him.	"It's from Santa, and it goes way, way beyond jolly."	Frames Santa as a boss making an inappropriate advance on an employee.	0
NYCC #40	The cartoon shows a woman in her underwear sitting up in bed, looking forward with a disgruntled expression. A large snow globe with a snowman inside is positioned next to her on the bed. The woman is speaking.	"I think the Manhattan skyline is getting suspicious."	This implies that she is cheating on her partner, a snow globe of the Manhattan skyline, with the snowman.	17
NYCC #15	In a restaurant, a man and a woman are sitting down to eat dressed in nice clothing. The man is leaning over the table with his hand on a glass looking at the woman with a soft smile. However, the man is bald and has a cartoonishly large forehead, with the outline of the woman visible on his forehead. The woman, sitting upright, is speaking.	"Well, it's a lovely gesture, but I still think we should start seeing other people."	Implies that the image on his forehead is a tattoo, as getting a tattoo of your significant other is a common practice.	20
NYCC #669	A baby leans over the side of a crib toward a microphone on a stand, as if ready to perform. The crib has letter blocks, and a mobile with boats and airplanes hangs above. A teddy bear lies on the floor. In the background, a couple stands in the doorway, looking at the baby with surprise. The woman is speaking with a smile.	"I love his bedtime routine."	Play on the dual meaning of the word "routine": a child's bedtime routine vs. a comedian's stand up routine.	33
NYCC #665	Two sharks are facing each other in the ocean. A person, visible only from the waist down, is standing on the back of one of the sharks. The sharks look bewildered; the carrying shark is speaking.	"How about some help carrying the groceries?"	To the sharks, the person is the groceries.	40
NYCC #687	A woman holding a wine glass stands on a rooftop in a city, delighted, looking back at a man sitting at a table with a bottle and glass. Behind them, an enormous face peers over the building, resembling the man. The woman is speaking with a smile.	"How much did you spend at Macy's this year?"	Implies the enormous person is a parade balloon for the Macy's Thanksgiving Day Parade.	40
NYCC #686	In a living room, a bald man is sitting on a giant egg, looking content. Two older women sipping tea, seated on a couch, are staring at him. The room has a coffee table, lamps, and a framed picture on the wall. One woman is speaking.	"I don't know how to tell him it's not his."	It should be obvious to him that the egg isn't his because humans don't lay eggs.	27
NYCC #61	A doctor wearing a head mirror stands behind a desk in a typical office. A giant hand is reaching through the doorway, palm up. The doctor is leaning over to check the enormous hand's pulse.	"I don't know why you're so jolly—your cholesterol is through the roof."	Wordplay: "through the roof" both as extremely elevated levels and literally breaking through the roof.	20

Table 4: Representative examples from the hard subset where a majority of evaluated LLMs failed to identify all required humor elements.

Model	**Accuracy (%)**
o3	59.85
Claude 3.7 Sonnet	54.27
gemini-2.5-pro-preview-03-25	52.44
deepseek/deepseek-r1-zero	51.22
deepseek-ai/DeepSeek-R1	51.22
o1	50.00
o4-mini	46.34
Grok 3	45.12
gpt-4o	42.68
deepseek-ai/DeepSeek-V3	39.63
meta-llama/Llama-4-Maverick-17B-128E	35.98
gemini-1.5-pro	35.37
o3-mini	32.93
meta-llama/Llama-4-Scout-17B-16E	29.88
Qwen/Qwen2.5-72B-Instruct-Turbo	26.83

Table 5: Accuracy on the **HumorBench-Hard** subset (100 items).