

000 WHICH LLMs GET THE JOKE? PROBING NON-STEM 001 REASONING ABILITIES WITH HUMORBENCH 002

003 **Anonymous authors**
004

005 Paper under double-blind review
006

007 ABSTRACT 008

009 We present HumorBench, a benchmark designed to evaluate large language models'
010 (LLMs) ability to reason about and explain sophisticated humor in cartoon captions
011 drawn from the New Yorker Caption Contest and Cartoonstock.com. As reasoning
012 models increasingly saturate existing benchmarks in math and science, novel and
013 challenging evaluations of model intelligence beyond STEM domains are essen-
014 tial. Reasoning is fundamentally involved in text-based humor comprehension,
015 requiring the identification of connections between concepts in cartoons/captions
016 and external cultural references, wordplays, and other mechanisms. HumorBench
017 includes 300 unique cartoon-caption pairs from the New Yorker Caption Contest
018 and Cartoonstock.com, with expert-annotated evaluation rubrics identifying essen-
019 tial joke elements. LLMs are evaluated based on their explanations towards the
020 humor and abilities in identifying the joke elements. To perform well on this task,
021 models must identify associations between concepts, potentially backtracking from
022 initial interpretations to arrive at the most plausible explanation. Our extensive
023 benchmarking of current models reveals three key insights: (1) LLM progress
024 on STEM reasoning transfers effectively to humor comprehension; (2) models
025 trained exclusively on STEM reasoning data still perform well on HumorBench,
026 demonstrating strong transferability of reasoning abilities; and (3) scaling thinking
027 token budgets yields mixed results across models in humor reasoning.
028

029 1 INTRODUCTION 030

031 Recent advances in large language models and reasoning techniques have led to the saturation of
032 many existing benchmarks, particularly in STEM domains such as mathematics and programming,
033 where frontier models now approach or exceed human-level performance (Abdin et al., 2025a; Sun
034 et al., 2025; Quan et al., 2025). This progression highlights the need for novel and challenging
035 evaluations that can meaningfully differentiate model capabilities and provide insights into their
036 reasoning processes. Non-STEM reasoning tasks, particularly those involving cultural understanding
037 and implicit knowledge, represent underexplored territories for model evaluation.
038

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

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

048 We benchmark both a standard set and a harder subset. On HumorBench-hard, which features more
049 complex examples requiring multiple reasoning steps or obscure cultural knowledge, no current LLM
050 exceeds 60% accuracy. Our benchmarking of current state-of-the-art models reveals two key findings:
051

- 052 1. We observe a high correlation between performance on HumorBench and existing STEM
053 benchmarks, suggesting a significant transfer of general reasoning abilities to humor com-
prehension tasks.

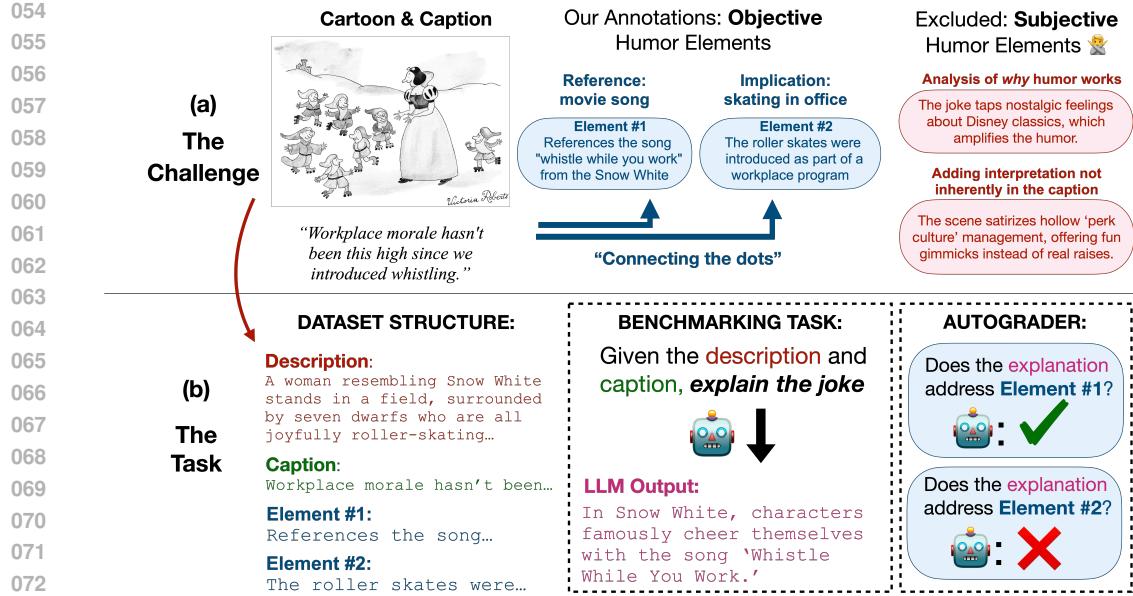


Figure 1: Overview of our humor analysis approach. (a) We distinguish between **objective** and **subjective** components of a joke. To convert the open-ended task of humor explanation into a fair benchmark, we focus exclusively on **objective** elements. (b) Overview of the dataset, benchmark task, 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.

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 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.

We acknowledge HumorBench covers a somewhat narrow domain of humor: English-language single-panel cartoons from the New Yorker Caption Contest and Cartoonstock.com, which reflect a particular Western, New Yorker-style sense of humor. However, we believe this format offers a clean, well-controlled testbed for probing non-STEM reasoning in LLMs.

108

2 RELATED WORK

110 **Reasoning-focused language models.** Recent advances in large language models (LLMs) have
 111 seen the emergence of specialized reasoning models that excel at logical deduction, mathematical
 112 problem-solving, and multi-step reasoning while maintaining strong general language capabilities.
 113 These reasoning-enhanced models employ various approaches: training-focused methods like those
 114 used by *Minerva* (Lewkowycz et al., 2022), *WizardMath* (Luo et al., 2023), and *Phi-4-Reasoning*
 115 (Abdin et al., 2025b) leverage carefully curated STEM-heavy corpora; inference-time techniques
 116 boost reasoning without changing model weights, including *Self-consistency* (Wang et al., 2023),
 117 *Tree-of-Thought* methods in *DeepSeek-Math* (Shao et al., 2024), *Least-to-most* prompting (Zhou et al.,
 118 2023), and *Process supervision* (Lightman et al., 2023); while hybrid approaches like *MAmmOTH*
 119 (Yue et al., 2023) combine diverse training data with structured inference protocols, *ToRA* (Gou
 120 et al., 2023) integrates formal verification systems, and *MathGLM* (Yang et al., 2024) combines
 121 symbolic computation with natural language reasoning. Models like *Gemini Ultra* (Google, 2024)
 122 and *Claude 3.5* (Anthropic, 2024) achieve strong reasoning through both architectural innovations
 123 and sophisticated training, suggesting that advances in machine reasoning now follow multiple
 124 complementary paths rather than relying solely on parameter count (Wei et al., 2024).

125 **Humour benchmarks.** Beyond simple joke generation, several resources now probe LLM hu-
 126 mour competence. Hessel et al. (2023) introduces three New Yorker cartoon-caption subtasks that
 127 test multimodal humour understanding and explanation. For word-play, the *ExPUNations* corpus
 128 augments classic pun datasets with human-written explanations and funniness ratings (Sun et al.,
 129 2022), while Xu et al. (2024) systematically benchmarks pun recognition, explanation and creation.
 130 Complementing these datasets, (Ermakova et al., 2025) Lab provides reusable test collections for
 131 humour-aware information retrieval.

132 **Open-ended evaluation frameworks.** Automatic grading of creative, unconstrained outputs increas-
 133 ingly relies on the *LLM-as-Judge* paradigm. G-EVAL couples chain-of-thought GPT-4 judging with
 134 a form-filling rubric, achieving human-level reliability on summarisation and dialogue (Liu et al.,
 135 2023). MT-BENCH and its crowdsourced *Chatbot Arena* show that GPT-4 judges agree with human
 136 preferences on multi-turn instruction following in ~80% of cases (Zheng et al., 2023). Going further,
 137 PAPERBENCH grades agents on reproducing ICML-level research papers with hierarchical GPT-4
 138 rubrics and expert audits (Starace et al., 2025). We adopt a similar rubric-guided judging scheme
 139 but focus specifically on humour reasoning, enabling systematic comparison of explanation quality
 140 across models.

141

3 HUMORBENCH

142

3.1 MAIN BENCHMARK TASK

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

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

153

3.2 DATASET: CARTOON AND CAPTION SOURCES

154 Our dataset comprises cartoons and captions from two primary sources: the New Yorker Cap-
 155 tion Contest (NYCC) and Cartoonstock.com. We sourced NYCC captions from publicly available
 156 datasets (Hessel et al., 2023; Jain et al., 2020; Zhang et al., 2024), selecting only those ranked among
 157 the top 3 finalists to ensure each cartoon features a coherent, high-quality joke. For Cartoonstock car-
 158 toons, we utilized their original accompanying captions. Both sources specialize in dry, witty humor

162 that demands sophisticated reasoning—often requiring multiple mental leaps to fully comprehend, as
 163 illustrated in Figure 1.

164 While cartoons inherently include visual elements, our benchmark focuses on testing humor com-
 165 prehension rather than visual interpretation capabilities. Therefore, we created detailed textual
 166 descriptions of each cartoon, carefully capturing all information necessary to understand the caption
 167 while maintaining neutrality. These descriptions include essential details about the setting, characters,
 168 visible emotions, and speaker identification, while deliberately omitting artistic style unless directly
 169 relevant to the joke. The choice to not directly include the images makes HumorBench a cleaner
 170 source of signal for humor reasoning, and not vision capabilities. For researchers interested in
 171 extending this to a multimodal benchmark, we provide source links to the original images: NYCC
 172 images are available through (Hessel et al., 2023; Jain et al., 2020), while Cartoonstock images
 173 require licensing.

175 3.3 DATASET: ELEMENT ANNOTATION

176 The core labels in our dataset are the element annotations assigned to each cartoon–caption pair. For
 177 every pair, we hand-annotated one to three elements—concise, direct statements that capture the
 178 objective components essential to understanding the joke. As discussed in Section 1.1, comprehending
 179 cartoon humor requires two distinct capabilities: understanding the objective content of the joke
 180 and recognizing the subjective aspects that influence audience reception. Figure 1 illustrates this
 181 distinction—subjective explanations focus on audience reactions (which vary between individuals),
 182 while objective elements center on content comprehension. Operationally, we treat an element as
 183 objective if it states a concrete, checkable fact or relation needed to reconstruct the intended reading
 184 of the joke from the description and caption, and treat as subjective anything that appeals to taste,
 185 funniness, or a particular theory of why the joke works. Our benchmark specifically targets these
 186 objective elements, which require identifying the mental leaps necessary to “get” the joke through
 187 recognizing references, wordplay, implications, or similar mechanisms. The autograder then evaluates
 188 LLM explanations against these elements, verifying that each explanation adequately covers the
 189 fundamental objective components of the humor, ensuring a fair and consistent assessment.

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



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.

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

211 3.4 DATASET REFINEMENT

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

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

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

228 3.5 AUTOGRADER AND EVALUATION

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

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

Explainer Model	Acc. (%)	FPR (%)	FNR (%)
<i>Overall</i> (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

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

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

253 **Length Control.** While models were instructed to keep responses under 200 words, some models
 254 exceeded this limit, particularly when reasoning traces were included in the final output. To ensure
 255 fair comparison, we truncated all model outputs to the last 1000 tokens.

257 4 EXPERIMENTS

259 Along with creating the HumorBench evaluation, we extensively benchmarked current frontier
 260 models. For consistency, all models are given the same prompt and scaffolding describing the task
 261 (see Appendix A). We arrived at this prompt after validating across several different LLMs (Claude
 262 3.7 Sonnet Anthropic (2025), GPT-4o OpenAI (2024a), Gemini 2.5 Pro DeepMind (2025)). While
 263 many LLMs had different API endpoints, we tried to maintain consistent parameters where possible.
 264 For example, all models had temperature set to 1 and external tool calling deactivated. Note, for all
 265 evaluations, autograders, and benchmarks, “GPT-4o” refers to the gpt-4o-2024-08-06 release.

267 4.1 MAIN RESULTS

269 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

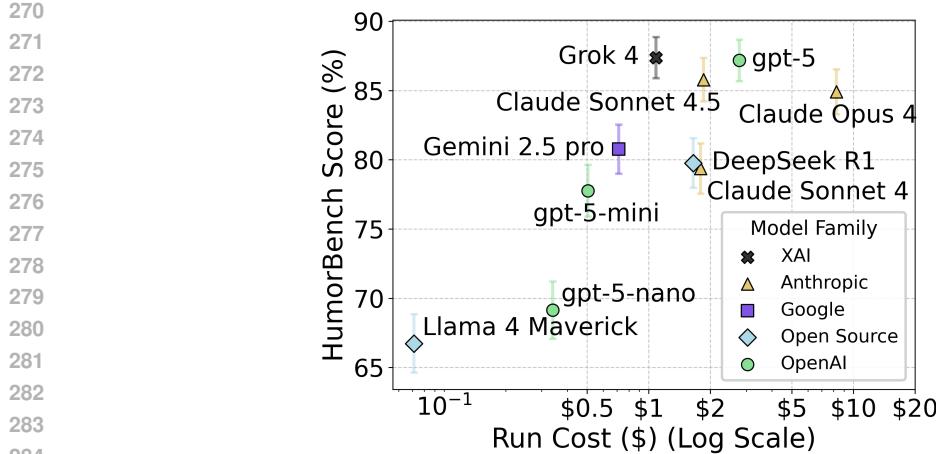


Figure 3: Benchmarking results on frontier models on HumorBench. Points show mean element-level accuracy; error bars denote ± 1 standard error across elements.

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.

4.2 TRANSFERABILITY OF REASONING SKILLS

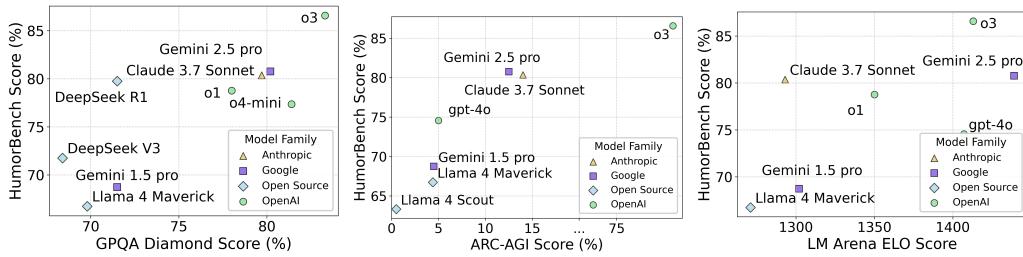


Figure 4: HumorBench performance compared to several benchmarks. We see positive correlation with GPQA, ARC-AGI, and LMArena. In particular, ranking compared to ARC-AGI is nearly identical to that of HumorBench, indicating a strong reasoning component to the HumorBench task.

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

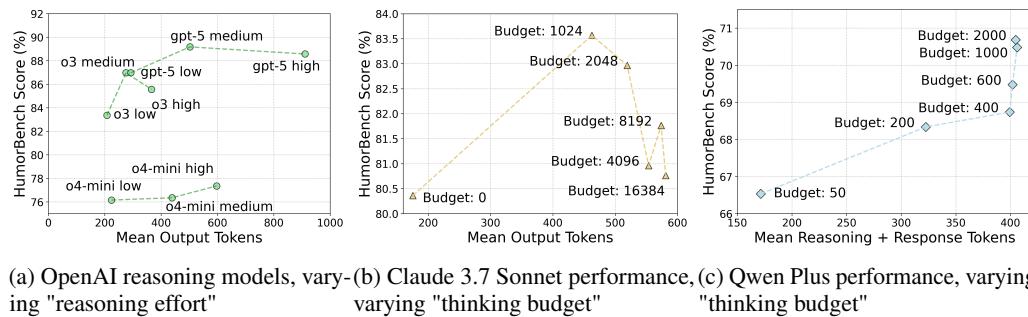
Table 2: Spearman's rank correlations between HumorBench and other benchmarks. Asterisks indicate significance: * $p < 0.05$, ** $p < 0.01$. ARC-AGI correlation shown separately for results with and without o-series models, which were fine-tuned for ARC-AGI

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 et al. (2024). HumorBench scores are positively associated with all three (see table 2). In particular, after removing o-series models (whose scores come from ARC-tuned variants) the correlation with ARC-AGI rises to $\rho = 0.943$ ($p = 0.005$), underscoring the shared skills of the two tasks. The LM Arena correlation ($\rho = 0.714$) is notably lower. Overall, this suggests that LLM progress on STEM domains transfers to Non-STEM reasoning.

331 STEM-ONLY REASONING IMPROVES HUMORBENCH

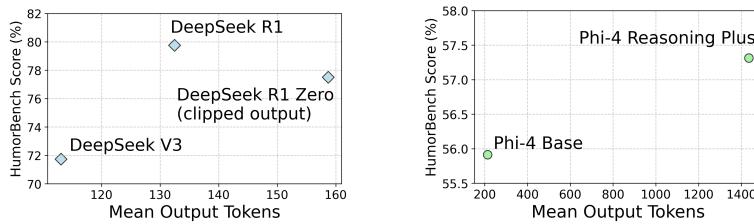
333 Our comparison between reasoning models trained on STEM tasks via Reinforcement Learning
 334 and their base counterparts yielded particularly revealing results. As illustrated in Figure 6, R1-
 335 Zero, which developed reasoning capabilities exclusively through self-play on STEM problems,
 336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

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



354 (a) OpenAI reasoning models, vary-(b) Claude 3.7 Sonnet performance, (c) Qwen Plus performance, varying
 355 "reasoning effort" varying "thinking budget" "thinking budget"
 356

357 Figure 5: HumorBench test-time compute experiments. Note, "mean output tokens" includes both
 358 reasoning and final response tokens
 359



369 Figure 6: Deepseek R1 Zero and Phi-4 Reasoning Plus, both exclusively reasoning-trained on STEM
 370 tasks, outperform their base versions on HumorBench
 371

372 4.3 TEST-TIME SCALING

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

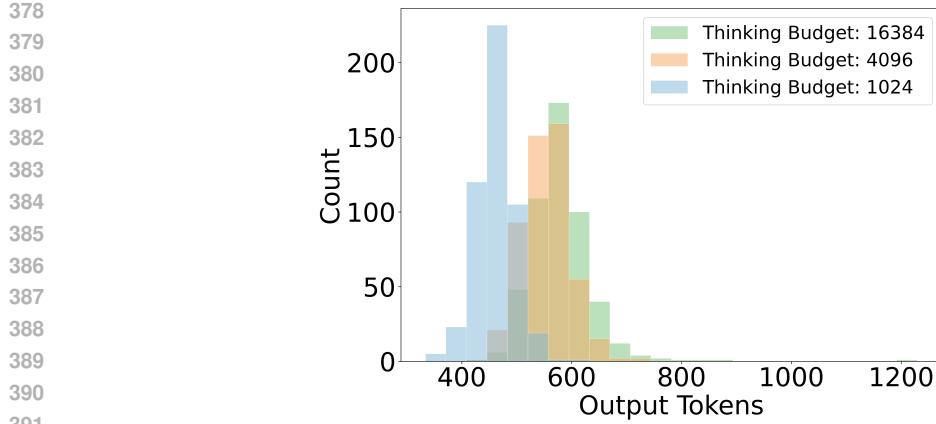


Figure 7: Token usage with different ‘thinking budget’ parameters on Claude 3.7 sonnet. Most completions (in the hundreds of tokens) were far below their budgets.

A closer look at the reasoning trace lengths highlights that for most captions, the models did not fully exhaust their reasoning budget (see 7). 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 LLMs are providing final answers based on completed thinking traces, which makes the inverse test-time scaling effect more puzzling. A small qualitative trace study for Claude 3.7 Sonnet (Appendix E) points to explicit definition of loaded terms and retrieval of obscure cultural references as one way extra budget helps. While a few studies have looked related problems (Su et al., 2025; Shi et al., 2023; Yang et al., 2025), we defer thorough investigation of our observation to future work.

4.4 ANALYSIS OF HUMORBENCH HARD SUBSET

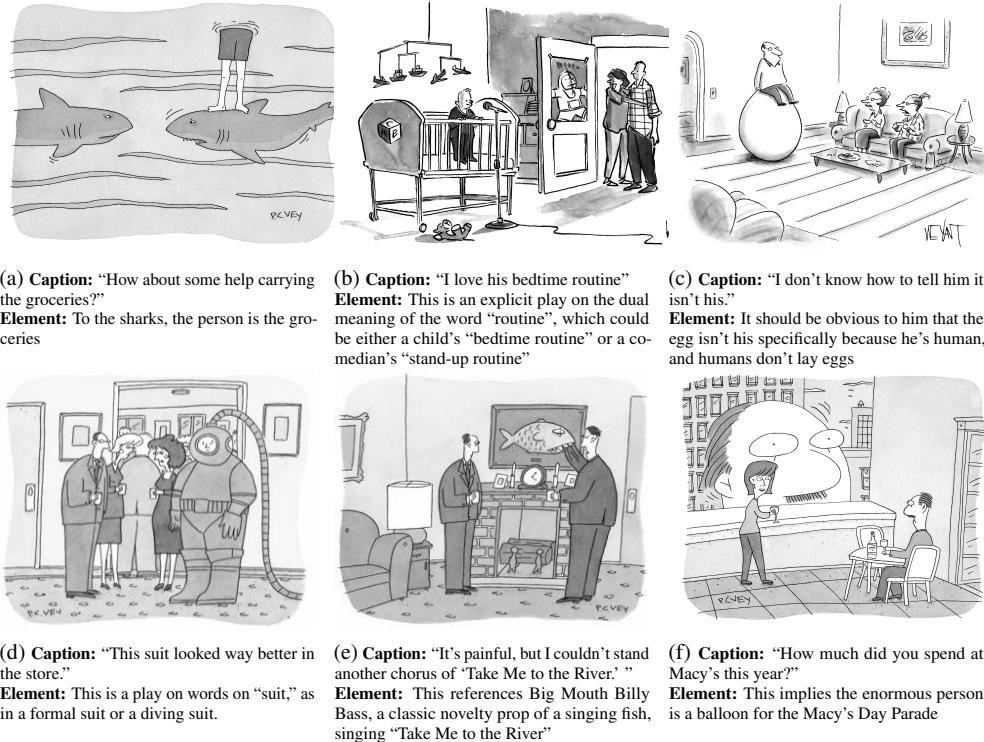


Figure 8: Examples of elements in the HUMORBENCH hard subset. See Table 5 for full descriptions.

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.

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.

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.

Last, in Figure 8, we highlight a few examples that are particularly challenging in the HumorBench hard subset. Detailed descriptions are also provided in Table 5. To our surprise, although several of them are quite intuitive and easy to recognize for humans, LLMs usually struggle on these examples. This represents a fundamental difference in reasoning abilities between humans and LLMs.

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. A natural next step is to develop a *multimodal* version of HumorBench by reintroducing the original cartoon images, allowing evaluation of both visual recognition and reasoning and better reflecting the complete humor comprehension task. Alternatively, the use of HumorBench's element rubrics as supervision signals offers an opportunity to explore reinforcement learning. Finally, improving the reliability of LLM-as-judge evaluation remains an important open challenge, particularly for creative tasks.

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

486 REFERENCES
487

488 Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao
489 Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann,
490 Yash Lara, Caio C. T. Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli
491 Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and
492 Guoqing Zheng. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025a. URL
493 <https://arxiv.org/abs/2504.21318>.

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

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

501 Anthropic. Claude 3.5 sonnet. [https://www.anthropic.com/news/
502 clause-3-5-sonnet](https://www.anthropic.com/news/clause-3-5-sonnet), 2024.

503 Anthropic. Claude 3.7 sonnet and claude code. [https://www.anthropic.com/news/
505 clause-3-7-sonnet](https://www.anthropic.com/news/
504 clause-3-7-sonnet), 2025. Accessed 19 May 2025.

506 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolaos Angelopoulos, Tianle Li, Dacheng
507 Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena:
508 An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*,
509 2024.

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

512 Google DeepMind. Gemini 2.5 pro: Our most intelligent ai model. [https://blog.google/
513 technology/google-deepmind/gemini-model-thinking-updates-march-2025/](https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/), 2025. Ac-
514 cessed 19 May 2025.

515 DeepSeek. Introducing deepseek v3: 671b-parameter mixture-of-experts model. [https://api-docs.
517 deepseek.com/news/news1226](https://api-docs.
516 deepseek.com/news/news1226), 2024. Accessed 19 May 2025.

518 DeepSeek. Deepseek r1 release. [https://api-docs.
519 deepseek.com/news/news250120](https://api-docs.deepseek.com/news/news250120), 2025.
520 Accessed 19 May 2025.

521 Liana Ermakova, Tristan Miller, Fabio Regattin, Antoine Bosselut, Saurabh Chaudhuri, Benoît
522 Jeanjean, Sébastien Lefait, Stan Matwin, Véronique Moriceau, Patrick Saint-Dizier, and Lucas
523 Vial. Overview of the clef 2025 joker lab: Automatic humour processing–recognition, retrieval,
524 and generation. In *Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum*,
525 2025.

526 Google. Gemini 2.0 flash thinkingg. [https://deepmind.google/technologies/gemini/
528 flash-thinking](https://deepmind.google/technologies/gemini/
527 flash-thinking), 2024.

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

532 Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff,
533 and Yejin Choi. Do androids laugh at electric sheep? humor “understanding” benchmarks from
534 the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.),
535 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:
536 Long Papers)*, pp. 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics.
537 doi: 10.18653/v1/2023.acl-long.41. URL <https://aclanthology.org/2023.acl-long.41/>.

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

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

543

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

548 Tuo Liang, Zhe Hu, Jing Li, Hao Zhang, Yiren Lu, Yunlai Zhou, Yiran Qiao, Disheng Liu, Jeirui
 549 Peng, Jing Ma, and Yu Yin. When 'yes' meets 'but': Can large models comprehend contradictory
 550 humor through comparative reasoning?, 2025. URL <https://arxiv.org/abs/2503.23137>.

551

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

555 Yang Liu, Dan Iter, Yichong Xu, and Shuohang Wang Yelong Peng. G-eval: NLg evaluation using
 556 gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods*
 557 in *Natural Language Processing*, pp. 11069–11081. Association for Computational Linguistics,
 558 2023.

559

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

564

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

566

567 Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering
 568 ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020. doi: 10.48550/arXiv.
 569 2004.10645. Published as a long paper at EMNLP 2020.

570

571 OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024a. Accessed 19 May
 2025.

572

573 OpenAI. Introducing openai o1-preview. <https://openai.com/index/introducing-openai-o1-preview/>, 2024b.

574

575 OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, 2025a. Accessed 19 May 2025.

576

577 OpenAI. Openai o3-mini. <https://openai.com/index/openai-o3-mini/>, 2025b.

578

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

584

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

588

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

592

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

594 Matteo Starace, Adam Fisch, Nora Kassner, Jason Wei, Johannes Welbl, Marjan Ghazvininejad,
 595 Sebastian Riedel, and Jack Merullo. Paperbench: Benchmarking large language models for
 596 scientific paper understanding and knowledge production. *arXiv preprint arXiv:2501.12077*, 2025.
 597

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

601

602 Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang,
 603 Lei Fang, and Ji-Rong Wen. Challenging the boundaries of reasoning: An olympiad-level math
 604 benchmark for large language models. *arXiv preprint arXiv:2503.21380*, 2025. URL <https://arxiv.org/abs/2503.21380>.

605

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

609

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

613

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

616

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

621

622 Liangyi Yang, Yihuai Wang, Xinyan Du, Zhekai Zhang, Haotian Shi, Zhengying Zhu, et al. Math-
 623 glm: Towards generalizable mathematical reasoning via language modeling. *arXiv preprint
 624 arXiv:2401.14273*, 2024.

625

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

628

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

632

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

636

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

641

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

645

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

648 REPRODUCIBILITY STATEMENT
649650 We release all resources needed to reproduce our results in a project repository (the repository will be
651 made public in the non-anonymized version). The repository contains (i) the full benchmark cartoon
652 descriptions and annotations, (ii) the autograder harness that contains the prompts and workflow of
653 the autograder, and (iii) scripts to run the explainer model, run the autograder, and aggregate and
654 analyze results. Detailed instructions are provided in the README655
656 ETHICS STATEMENT
657658 To respect copyright and enable lawful redistribution, we release only rubric-guided textual descrip-
659 tions and derived annotations of the cartoons, no original images. Some items may reference sensitive
660 topics common in humor; the benchmark is provided for research use and downstream applications
661 should include appropriate safety review.662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701

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>

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 MODEL PERFORMANCE ON HUMORBENCH HARD

Model	Accuracy (%)
gpt-5	61
o3	59
Claude 3.7 Sonnet	54
gemini-2.5-pro-preview-03-25	52
deepseek/deepseek-r1-zero	51
deepseek-ai/DeepSeek-R1	51
o1	50
o4-mini	46
Grok 3	45
gpt-4o	42
deepseek-ai/DeepSeek-V3	39
meta-llama/Llama-4-Maverick-17B-128E	35
gemini-1.5-pro	35
o3-mini	32
meta-llama/Llama-4-Scout-17B-16E	29
Qwen/Qwen2.5-72B-Instruct-Turbo	26

Table 4: Accuracy on the **HumorBench-Hard** subset (100 items).

756 **D HARD EXAMPLE DETAILS**
757

759 Cartoon ID	760 Description	761 Caption	762 Element(s)	763 Pass Rate (%)
764 CC123065	765 Inside a workshop like room, three elves 766 in pointy hats sit at a long table with open 767 laptop computers. The middle elf appears 768 distressed and is speaking, while the two 769 elves on either side look toward him.	770 “It’s from 771 Santa, and it 772 goes way, way 773 beyond jolly.”	774 Frames Santa as a boss making an inappropriate advance on an employee.	775 0
776 NYCC #40	777 The cartoon shows a woman in her under- 778 wear sitting up in bed, looking forward with a disgruntled expression. A large 779 snow globe with a snowman inside is positioned next to her on the bed. The woman 780 is speaking.	781 “I think the 782 Manhattan skyline is 783 getting suspicious.”	784 This implies that she is cheating on her partner, a snow globe of the Manhattan skyline, with the snowman.	785 17
786 NYCC #15	787 In a restaurant, a man and a woman are 788 sitting down to eat dressed in nice clothing. The man is leaning over the table 789 with his hand on a glass looking at the woman with a soft smile. However, the 790 man is bald and has a cartoonishly large forehead, with the outline of the woman 791 visible on his forehead. The woman, sitting upright, is speaking.	792 “Well, it’s a 793 lovely gesture, 794 but I still think 795 we should start 796 seeing other people.”	797 Implies that the image on his forehead is a tattoo, as getting a tattoo of your significant other is a common practice.	798 20
799 NYCC #669	800 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.	801 “I love his bed- 802 time routine.”	803 Play on the dual meaning of the word “routine”: a child’s bedtime routine vs. a comedian’s stand up routine.	804 33
805 NYCC #665	806 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.	807 “How about 808 some help 809 carrying the 810 groceries?”	811 To the sharks, the person is the groceries.	812 40
813 NYCC #687	814 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.	815 “How much 816 did you spend 817 at Macy’s this 818 year?”	819 Implies the enormous person is a parade balloon for the Macy’s Thanksgiving Day Parade.	820 40
821 NYCC #686	822 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.	823 “I don’t know 824 how to tell 825 him it’s not 826 his.”	827 It should be obvious to him that the egg isn’t his because humans don’t lay eggs.	828 27
829 NYCC #61	830 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.	831 “I don’t know 832 why you’re so 833 jolly—your 834 cholesterol is 835 through the 836 roof.”	837 Wordplay: “through the roof” both as extremely elevated levels and literally breaking through the roof.	838 20

801 Table 5: Representative examples from the hard subset where a majority of evaluated LLMs failed to
802 identify all required humor elements.
803804 **E REASONING TRACE CASE STUDIES**
805806 To better understand when test-time “thinking” helps, we qualitatively inspected cases where
807 claude-3-7-sonnet-latest in thinking mode (budget 2048 tokens) succeeded on an element
808 that the same model in standard mode missed. Prompts, data, and autograder remained identical; only
809 the decoding mode changed.

810 We highlight two representative anecdotes.
 811

812 **Anecdote 1: “Faith-based initiative” (NYCC 41.0).** The cartoon shows a church interior with a
 813 fully stocked bar; the caption is *“Finally—a faith-based initiative I can embrace.”* The key element
 814 is the reference to President George W. Bush’s *Faith-Based and Community Initiatives* program.
 815

816 **Base model.** The standard run defined “faith-based initiative” only in generic terms (e.g., religiously
 817 motivated social programs) and never linked it to a specific US policy or administration, so the
 818 political reference element was marked as missing.
 819

820 **Thinking model.** In the thinking trace, the model first writes a short definition of the term as a policy
 821 label and then explicitly ties it to the Bush administration before drafting the final explanation:
 822

823 “...the phrase ‘faith-based initiative’ is also a political term, *especially associated with the*
 824 *George W. Bush administration ...*”
 825

826 This additional retrieval step causes the final explanation to mention the Bush-era program and passes
 827 the element.
 828

829 **Anecdote 2: Big Mouth Billy Bass (NYCC 57.0).** The cartoon shows a mounted fish on the wall
 830 with a man’s hand in its mouth; the caption is *“It’s painful, but I couldn’t stand another chorus of*
 831 *‘Take Me to the River.’”* The key element is that this is a reference to the novelty product *Big Mouth*
 832 *Billy Bass*, a singing fish decoration that plays “Take Me to the River.”
 833

834 **Base model.** The standard run assumes the man is singing the song and interprets the joke as
 835 preferring physical pain to bad singing, never mentioning the novelty fish prop, so it fails the element.
 836

837 **Thinking model.** The thinking trace instead immediately connects the song title to the product:
 838

839 “... ‘Take Me to the River’ is a song that was *famously used for novelty singing fish decora-*
 840 *tions (like Big Mouth Billy Bass) ...”*
 841

842 The final explanation then correctly frames the joke around escaping the repeated jingle from the
 843 novelty plaque, and the element passes.
 844

845 These case studies suggest that, at least in some instances, the additional “thinking” budget is used
 846 to (i) explicitly define loaded terms and retrieve associated political context, and (ii) connect cues
 847 (a song title) to specific cultural artifacts (a novelty product), which can unlock the correct humor
 848 interpretation.
 849

850 F WEB SEARCH ABLATION

851 To assess how access to external knowledge affects humor understanding, we ran an ablation on the
 852 GPT-5 model family. For each of three sizes (Nano, Mini, and full) we compared a *base* configuration
 853 (no tools) to a *search* configuration in which the model could call a web search tool before producing
 854 its explanation. Prompts, HumorBench items, and the GPT-4o autograder were identical across
 855 conditions, and evaluated on the full benchmark as well as the hard subset.
 856

857 Model	858 Overall, Base (%)	859 Overall, Search (%)	860 Hard, Base (%)	861 Hard, Search (%)
858 GPT-5 Nano	69.1	67.9	34.0	31.0
859 GPT-5 Mini	77.8	80.2	41.0	47.0
860 GPT-5	87.2	89.4	56.0	63.0

861 Table 6: HumorBench element-level accuracy (%) for GPT-5 models with and without web search,
 862 on the full benchmark (Overall) and the 100 hardest items (Hard-100).
 863

864 Overall, web search slightly hurts GPT-5 Nano, but consistently improves GPT-5 Mini and GPT-5.
 865 This pattern suggests that external tools are most useful when the base model is already strong enough
 866 to recognize when and how to use retrieved information.
 867