HumorBench: Evaluating Reasoning in LLMs Through Humor Comprehension

Reuben Binns et al.

May 3, 2025

Abstract

Humor comprehension represents a valuable test case for evaluating reasoning capabilities in Large Language Models (LLMs). We introduce HumorBench, a benchmark designed to assess an LLM's ability to identify and explain the core comedic elements in humorous content using cartoon examples from various sources. Our approach breaks down humor comprehension into discrete mental leaps that connect references, implications, and contexts. Our evaluation framework uses an LLM-based autograder validated against human judgments, achieving 87% accuracy. Results show high correlation between performance on HumorBench and STEM-focused benchmarks, suggesting that general reasoning ability transfers to humor comprehension. We also identify a subset of particularly challenging examples (HumorBench-Hard) where even frontier models achieve limited success. HumorBench provides a novel non-STEM reasoning benchmark that complements existing metrics while highlighting LLMs' ability to make conceptual connections between disparate elements—a fundamental aspect of reasoning that extends beyond humor.

1 Introduction

Humor comprehension provides a unique window into the reasoning capabilities of Large Language Models (LLMs). Understanding humor requires sophisticated mental leaps and the ability to connect disparate concepts—skills that are fundamental to reasoning in many domains. While recent advances have shown impressive capabilities across various tasks, humor remains a valuable testing ground that reveals a model's ability to understand nuanced contexts, cultural references, and the subtle mechanisms that create comedic effects.

HumorBench uses cartoons and captions from the New Yorker Caption Contest and CartoonStock.com, sources known for sophisticated humor requiring cultural knowledge, understanding of tropes, and nuanced storytelling. For each cartoon-caption pair, we provide the model with:

- A detailed text description of the cartoon's visual elements
- The corresponding caption
- A request to explain the humor in the pairing

The key innovation of HumorBench is our approach to decomposing humor into identifiable elements that require specific mental leaps to connect. For example, understanding a cartoon showing Snow White with dwarfs on roller skates captioned "Workplace morale hasn't been this high since we introduced whistling" requires:

- Recognizing the reference to the song "Whistle While You Work" from Snow White
- Understanding that roller skates were introduced in a workplace context
- Connecting these elements through mental leaps that bridge the cartoon visuals with cultural knowledge

Our evaluation framework uses an LLM-based autograder that examines whether these specific humor elements are present in model-generated explanations. The autograder was validated against human judgments with 87% accuracy, showing a higher False Positive Rate than False Negative Rate.

Our findings reveal several key insights:

- HumorBench provides a valuable non-STEM reasoning dataset and benchmark
- Performance on HumorBench shows high correlation with STEM benchmarks, suggesting that general reasoning capabilities transfer across domains
- The mental leaps required for humor comprehension appear to rely on similar cognitive mechanisms as those used in STEM problem solving
- Analysis of challenging examples reveals insights into the boundaries of current LLMs' reasoning capabilities

HumorBench serves as a novel evaluation framework that reveals insights into LLM reasoning abilities, particularly in making the conceptual leaps required to connect disparate elements—skills essential not just for humor but for advanced natural language understanding and problem-solving more broadly.

2 Related Work

Previous work on humor in NLP has focused on three main directions: humor recognition (detecting whether content is intended to be humorous), humor generation (creating new humorous content), and humor ranking (predicting which content humans will find funnier). Notable examples include:

- **Hessel et al. (2022)** developed an LLM-based system to evaluate New Yorker cartoon caption contest submissions, focusing on predicting which captions humans would rate as funnier.
- **Yang et al. (2023)** worked on humor explanation tasks focused on why content might be considered humorous.
- **Hasan et al. (2021)** explored computational humor through generation tasks.

Our work differs from these approaches by:

- Focusing on humor as a proxy for evaluating reasoning capabilities
- Decomposing humor into specific elements that require mental leaps to connect
- Providing detailed expert annotations of humor elements for evaluation
- Using a validation-tested LLM autograder system to assess explanations
- Exploring correlations between humor comprehension and performance on STEM reasoning tasks

3 HumorBench

3.1 The Mental Leaps Challenge

The New Yorker cartoon style represents a particularly sophisticated form of humor that relies on mental leaps connecting:

- Cultural knowledge and references
- Understanding of social norms and their subversion
- Wordplay and dual meanings
- Juxtaposition of the mundane with the absurd
- Implicit connections that readers must infer

Humor comprehension requires identifying and connecting these elements through mental leaps—cognitive connections that bridge different concepts, references, and implications. For example, in a cartoon showing a woman with dwarfs on roller skates, the caption "Workplace morale hasn't been this high since we introduced whistling" requires connecting:

- The visual reference to Snow White and the dwarfs

- The song "Whistle While You Work" from the movie
- The implication that roller skating has been introduced in a workplace
- The juxtaposition of the fairy tale setting with modern office culture

These mental leaps represent reasoning challenges; they require recognizing references, understanding implications, and connecting disparate domains in ways that create humor through incongruity, subversion, or unexpected associations. The patterns of reasoning required mirror those used in other domains, including STEM problem-solving.

3.2 Our Task

The core task in HumorBench requires a model to:

1. Receive a text description of a New Yorker cartoon image
2. Receive the corresponding winning caption
3. Generate a concise explanation (under 200 words) identifying the key comedic elements of the joke

The explainer prompt instructs the model:

> "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."

3.3 Autograder Construction

To evaluate model explanations, we constructed an LLM-based autograder using GPT-4o. The autograder:

- Receives the cartoon description, caption, model-generated explanation, and the "anticipated point" (ground truth humor element)
- Determines whether the explanation covers the anticipated point
- Outputs a PASS/FAIL judgment with reasoning

We validated the autograder against 400 human judgments (PASS/FAIL labels for explanations from four different models across 100 humor elements), achieving 87% accuracy. Importantly, the validation found that the autograder has a higher False Positive Rate than False Negative Rate, meaning it's more lenient than human judges. This suggests our reported performance metrics represent an upper bound on model capabilities.

The autograder prompt instructs the model to evaluate whether the explanation captures the key humor element:

> "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.'"

4 Dataset Curation

4.1 Humor Elements as Reasoning Challenges

The primary innovation in our dataset is the identification and annotation of humor elements that serve as reasoning challenges. For each cartoon-caption pair, we identify specific elements that contribute to the humor, such as:

- References to cultural works (e.g., "the song 'whistle while you work' from Snow White")
- Context-specific implications (e.g., "roller skates were introduced as part of a workplace")
- Wordplay or puns
- Juxtaposition of unexpected elements

- Subversion of expectations
- Absurdity or exaggeration

These elements serve as the ground truth for evaluation. An explanation successfully addresses the reasoning challenge when it identifies these specific elements and the mental leaps connecting them.

4.2 Data Sources and Selection

The dataset contains:

- Unique identifiers for each data point
- Cartoon descriptions detailing visual elements
- Winning captions associated with each cartoon
- Expert-annotated humor elements identifying the specific comedic devices at work

We sourced the original cartoons and captions from:

- Nextml Caption Contest Data
- jmhessel/newyorker_caption_contest Hugging Face dataset
- CartoonStock

5 Experiments

5.1 Experimental Setup

We evaluated multiple leading LLMs on HumorBench, including:

- GPT-4o
- Claude 3.7 Sonnet
- Gemini 2.5 Pro
- Llama 4 Maverick
- Other models (details to be added)

For each model, we:

- Generated explanations using the explainer prompt
- Evaluated explanations using the autograder
- Calculated PASS rates (percentage of explanations deemed to correctly identify the humor element)
- Tracked token usage and calculated cost metrics

Implementation details:

- We used asynchronous processing to handle multiple examples efficiently
- Explainer and autograder clients supported various model APIs (OpenAI, Claude, Gemini, Together, XAI)
- Response parsing extracted explanations from within specified XML tags

5.2 Main Results

Key findings include:

- Frontier models (e.g., GPT-4o, Claude 3.7) achieved the highest PASS rates on the standard benchmark
- We identified a challenging subset (HumorBench-Hard) where even the best models show limited success
- Model performance correlated with cost, with more expensive models generally performing better
- Certain humor elements (particularly those requiring cultural knowledge or multi-step reasoning) proved challenging across all models

5.3 Correlation with STEM Benchmarks

A central finding of our work is the high correlation between performance on HumorBench and performance on traditional STEM-focused benchmarks. This suggests that general reasoning ability transfers effectively across domains, with the mental leaps required for humor comprehension drawing on similar cognitive mechanisms as those used in mathematics, science, and engineering problems.

Models that perform well on benchmarks like GPQA, ARC-AGI, and LM Arena ELO also tend to perform well on HumorBench. This correlation indicates that:

- Humor comprehension relies on general reasoning capabilities rather than domain-specific knowledge
- Improvements in STEM reasoning likely transfer to improved performance in humor understanding
- HumorBench provides a complementary evaluation framework that tests similar underlying abilities in a different context

This finding supports the idea that humor comprehension can serve as a proxy for evaluating reasoning capabilities in LLMs, providing insights that complement other benchmarks while engaging different forms of knowledge and conceptual connections.

5.4 Analysis of HumorBench-Hard

HumorBench-Hard (HBH) consists of a subset of examples that present particular reasoning challenges. Common patterns in these examples include:

- Multiple cultural references that must be simultaneously recognized
- Complex wordplay or puns requiring linguistic reasoning
- Non-obvious connections between visual elements and caption content
- Multi-step reasoning chains connecting different elements
- Nuanced understanding of social norms or expectations being subverted

These examples often require more sophisticated reasoning, with multiple mental leaps needed to connect different elements.

5.5 Test-Time Scaling and Reasoning Ability

Our experiments also revealed interesting patterns in how performance scales with additional reasoning resources:

- Models with larger context windows or more reasoning capabilities showed better performance on complex examples
- Test-time scaling techniques (e.g., Claude's thinking budget, OpenAI's reasoning effort) improved performance on the benchmark
- The improvements from these techniques further support the idea that humor comprehension relies heavily on reasoning ability

6 Conclusion and Future Work

HumorBench provides a novel framework for evaluating an essential aspect of language understanding—the ability to make mental leaps between concepts. Our results highlight the strong correlation between humor comprehension and performance on STEM reasoning tasks, suggesting that similar cognitive mechanisms underlie both domains.

The benchmark serves as a valuable tool for measuring progress in natural language understanding and reasoning, particularly in areas requiring:

- Integration of cultural and contextual knowledge
- Recognition of implicit connections
- Understanding of social norms and their subversion
- Multi-step reasoning about intentions and meanings

The correlation between performance on our benchmark and STEM-focused benchmarks suggests that reasoning ability transfers across domains, and that improving models' general reasoning capabilities enhances their performance across a variety of tasks, including humor comprehension.

Future work could extend HumorBench by:

- Including more diverse humor sources beyond New Yorker cartoons
- Developing metrics for specific dimensions of reasoning in humor understanding
- Comparing autograder performance across different LLMs
- Exploring few-shot prompting or fine-tuning to improve explainer performance
- Investigating the use of reinforcement learning to enhance model ability to draw connections between concepts

Limitations

The current version of HumorBench has several limitations:

- Focuses primarily on New Yorker-style humor, which may not generalize to other cultural contexts
- Relies on text descriptions rather than actual cartoon images
- Uses an LLM-based autograder, which has inherent limitations despite validation
- Exhibits a higher False Positive Rate than False Negative Rate in autograder judgments
- The correlation with STEM reasoning may not capture all aspects of humor comprehension