HumorBench: A Benchmark for Evaluating Humor Comprehension in Large Language Models

Reuben Binns et al.

May 3, 2025

Abstract

Humor comprehension represents a particularly challenging task for Large Language Models (LLMs), requiring sophisticated understanding of context, culture, and implicit connections. We introduce HumorBench, a benchmark specifically designed to evaluate an LLM's ability to identify and explain the core comedic elements in humorous content using New Yorker Cartoon Caption Contest examples. Unlike existing humor benchmarks that conflate subjective appreciation with objective comprehension, HumorBench isolates the latter, focusing exclusively on whether models can identify the specific elements that make a cartoon-caption pair humorous. Our evaluation framework uses an LLM-based autograder validated against human judgments, achieving 87% accuracy. Results from testing leading models reveal significant gaps in humor comprehension abilities, with even frontier models struggling on more complex examples. HumorBench serves as a novel evaluation tool that highlights specific limitations in LLMs' ability to make conceptual connections and recognize implicit meaning—skills essential not just for humor but for advanced natural language understanding.

1 Introduction

Humor comprehension represents a particularly challenging task for Large Language Models (LLMs), requiring sophisticated understanding of context, culture, and implicit connections. While recent advances have shown impressive capabilities across various reasoning tasks, humor remains a frontier challenge that tests a model's ability to understand nuanced social contexts, cultural references, and the subtle mechanisms that create comedic effects.

Our work focuses exclusively on objective humor comprehension (understanding why something is intended to be humorous) rather than subjective appreciation (finding something funny). This distinction is crucial, as existing benchmarks often conflate these separate challenges.

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 evaluation methodology. Rather than measuring a model's ability to generate humorous content or judge what humans might find funny, we assess whether the model can identify the specific, objective elements that create the humor. For each cartoon-caption pair, we have expert-annotated "humor elements" that capture the key aspects of the joke. An LLM-based autograder (validated against human judgments with

87% accuracy) then evaluates model-generated explanations based on whether they successfully identify these core elements.

Our benchmark reveals that while frontier models perform reasonably well on standard examples, they still struggle significantly with more complex humor. We also release HumorBench-Hard (HBH), a subset featuring particularly challenging examples, where no current LLM exceeds 20% accuracy.

HumorBench serves as a novel evaluation framework that reveals specific gaps in LLM comprehension abilities, particularly in making the conceptual leaps required to connect disparate elements and recognize implicit meaning—skills essential not just for humor but for advanced natural language understanding.

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 but focused on explaining why humans might find content funny rather than identifying objective humor elements.
- **Hasan et al. (2021)** explored computational humor through generation tasks, emphasizing the subjective aspects of humor production.

Our work differs from these approaches by:

- Focusing exclusively on objective comprehension rather than subjective appreciation
- Providing detailed expert annotations of humor elements for evaluation
- Using a validation-tested LLM autograder system to assess explanations
- Testing frontier models on particularly challenging examples of sophisticated humor

3 HumorBench

3.1 Why Another Humor Benchmark?

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

- 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

Previous work on New Yorker cartoons has primarily focused on:

- Ranking captions by predicted funniness
- Explanation tasks that ask models to predict why humans might find content amusing
- Generation of new captions for existing cartoons

These approaches all conflate objective and subjective factors in humor assessment. By contrast, HumorBench isolates the objective comprehension element, creating a more focused evaluation of an LLM's understanding capabilities.

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 "material substance" or key comedic element of the joke

Unlike previous explanation tasks, we don't judge the model's ability to predict what humans would find funny. Instead, we evaluate whether the explanation covers the specific, objective elements that create the humor, as identified in our expert annotations.

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 explicitly 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 Cartoon Caption 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

4.2 Comprehension Element Annotation

The "element" annotations identify the specific humor mechanism at work in each cartoon-caption pair. These annotations were created by humor experts and serve as the ground truth for evaluation.

Each element identifies the specific comedic device or concept that an explanation should capture, such as:

- Wordplay or puns
- Juxtaposition of unexpected elements
- Cultural references

- Subversion of expectations
- Absurdity or exaggeration

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
- All models showed significant performance drops on HumorBench-Hard examples
- 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 Comparison and Correlations to Other Benchmarks

5.4 Analysis of HumorBench-Hard

HumorBench-Hard (HBH) consists of particularly challenging examples where even frontier models struggle. Common failure patterns include:

- Missing cultural references that are essential to understanding the joke
- Failing to recognize wordplay or puns
- Missing connections between visual elements and caption content
- Inability to recognize subverted tropes or expectations
- Struggling with humor that requires multi-step reasoning

5.5 Human-AI Collaborative Explanation

6 Conclusion and Future Work

HumorBench provides a novel framework for evaluating an essential aspect of language understanding—the ability to comprehend humor. Our results

highlight significant gaps in current LLMs' ability to identify and explain even relatively straightforward humor elements, with performance dropping dramatically on more complex examples.

The benchmark serves as a valuable tool for measuring progress in natural language understanding, 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

Future work could extend HumorBench by:

- Including more diverse humor sources beyond New Yorker cartoons
- Developing metrics for specific dimensions of 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
- May not fully separate objective understanding from subjective factors in all cases