JEST: A Benchmark for Rating the Funniness of Short Texts

Joe Toplyn
Semantic AI and Creativity Lab
East Texas A&M University
Commerce, Texas 75428 USA
joetoplyn@twentylanemedia.com

Ori Amir[1]
Psychology Department, Fulbright University Vietnam
Semantic AI and Creativity Lab, East Texas A&M University
Commerce, Texas 75428 USA
oriacadem@gmail.com

Abstract

A conversational AI system is perceived as more humanlike when it exhibits a sense of humor. To achieve this, the system must not only recognize humorous conversational responses, but also assess how funny a typical human would find them, enabling appropriate reactions. While prior work has advanced binary humor detection in short texts, no benchmark exists for evaluating funniness on a numerical scale in conversational contexts. In this paper we introduce JEST, a benchmark derived from 80 short conversational texts generated using AI and annotated with funniness ratings by seven expert joke writers. To ensure consistency and a balanced distribution across funniness categories, we computed final ratings using the three raters whose individual ratings were most evenly distributed. JEST achieves high inter-rater reliability (ICC(2,k) = .87), with texts divided into four roughly equal categories: 0 (Not Funny) to 3 (Very Funny). As an illustrative application, JEST is used to provide evidence that an off-the-shelf LLM can rate the perceived funniness of a dataset of short, conversational texts in substantial agreement with a panel of human experts. JEST facilitates the evaluation of automated funniness rating systems, thereby supporting the development of effective automated joke generation systems.

Introduction

The ability to mimic a sense of humor is an important trait in effective conversational AI systems (Ceha et al. 2021; Clark et al. 2019; Kang et al. 2017; Mirowski, Mathewson and Branch 2025; Nijholt et al. 2017; Sun et al. 2024). For a conversational AI system to do that, it must have the ability to both automatically generate jokes and to automatically recognize them (Inoue, Lala and Kawahara 2022; Inoue et al. 2025; Quan, Ramakrishnan and Chin 2025; Winters 2021). And if the AI system recognizes humorous text, it needs to be able to rate it for funniness. Funniness ratings are essential because they determine what type and degree of humor response to provide and have been shown to correlate strongly with laughter intensity. (McKeown and Curran 2015; Mittal et al. 2021).

To evaluate the performance of automated funniness raters intended for use in a conversational AI system, a benchmark against which the raters can be tested is necessary. But to the best of our knowledge, no benchmark exists that incorporates short texts similar to conversational turns along with relatively robust measures of how funny they are. To address that research gap, we introduce in this paper a benchmark incorporating short, two-part texts in English with funniness ratings assigned by expert human joke writers. We name it the Joke Evaluation by Skilled Testers (JEST) benchmark.

Related Work

There are many humor datasets consisting of short texts (Kalloniatis and Adamidis 2024; Lemmens and De Marez 2025). But few of the texts in those datasets are similar in structure to a conversational turn in English. For example, Bogireddy, Suresh and Rai (2023) present a dataset of humorous textual posts related to COVID-19 from the subreddits r/Jokes, r/CoronavirusMemes, r/Onion headlines, and r/NottheOnion headlines. Bower and Steyvers (2021) use 60 jokes scraped from the subreddit r/Jokes. Faruqi and Shrivastava (2018) use a dataset of 400,000 sentences extracted from social media and humor-dedicated websites,

[1] Corresponding author

in addition to negative samples such as news headlines and proverbs. Kiddon and Brun (2011) study a dataset of "That's what she said" jokes. Mihalcea and Strapparava (2005) use a dataset of short, humorous sentences–"one-liners"–and non-humorous data such as news headlines and proverbs. Murakami et al. (2025) base a benchmark on a dataset derived from the Japanese creative response game Oogiri and the funniness votes of game participants. Potash, Romanov and Rumshisky (2017) introduce a dataset composed of humorous tweets that could have a given hashtag. Yang, Hooshmand and Hirschberg (2021) present a dataset of Facebook posts related to COVID-19, each post with a humor score derived from its reaction emojis. And Zangari et al. (2025) use annotated sets of puns and non-puns.

In addition, few humor datasets of short texts are annotated with ratings of funniness, because they are usually intended for use in binary classification tasks. And of those datasets that are annotated with funniness ratings, few have those ratings assigned by comedy experts. Potash, Romanov and Rumshisky (2017) do rely on the expert judgments of a television show's staff to measure the funniness of tweets. But Avetisyan, Safikhani and Broneske (2023) use only two human evaluators of indeterminate expertise to assign funniness ratings to short, humorous texts, while Chiruzzo, Castro and Rosá (2020) use crowdsourced annotators who visited a particular web page to rate humorous Spanish tweets for funniness. Other examples include Hossain, Krumm and Gamon (2019), who employ non-expert researchers and Amazon Mechanical Turk workers as judges to rate edited news headlines for funniness, and Weller and Seppi (2020), who use the upvotes of presumably non-expert Reddit users to measure the funniness of humorous texts from the r/Jokes subreddit.

Having identified the absence of a benchmark of short texts similar to conversational turns along with relatively robust funniness measures, in particular funniness ratings by experts, we built it as described in the next section.

Building the Benchmark

Constructing JEST took place in stages: designing the benchmark, creating a dataset of short texts composed of setups and responses, and obtaining expert ratings of the funniness of each short text.

Designing the Benchmark

In designing the benchmark, we made the simplifying decision that it should be useful for evaluating automated funniness raters in conversational AI systems intended for typical American adults. We also made the important observation that a standalone, "canned" joke consisting of an unfunny setup sentence followed by a punch line is equivalent, from the standpoint of joke mechanics, to a conversational turn consisting of an utterance that is not funny followed by a funny response (Toplyn 2014). Those factors led us to conclude that the benchmark would be based on a dataset of short texts in English, each of which would consist of (1) an unfunny sentence that could plausibly function as either an utterance in a conversation or the setup of a joke, followed by (2) a sentence that could function as either a funny response to the utterance or a punch line of the joke. The short texts would vary in funniness and each would be accompanied by a numerical rating reflecting how funny typical American adults would be likely to find it. The performance of a funniness-rating system against the benchmark would be evaluated by measuring the degree of agreement between the system's funniness ratings for the 80 texts and the benchmark ratings.

Creating the Short Texts

Because of the limitations of existing datasets, described above, we decided to create a completely new dataset of 80 short texts for this paper. We believed that this number was large enough to allow statistically significant conclusions to be drawn, but not so large that a human who is rating all the texts in one session would experience fatigue, resulting in lower quality ratings.

The following process was used to create short texts that may be funny to some degree and consist of a setup sentence followed by a response sentence.

Generating the Setups

To serve as unfunny setup sentences, we considered extracting headlines from a news dataset (Faruqi and Shrivastava 2018; Mihalcea and Strapparava 2005). But news headlines might result in jokes that become dated and less funny over time, and therefore seemed suboptimal for a benchmark that we intended to have a relatively long useful life. We also considered selecting utterances from an existing conversational dataset (Toplyn 2023). But the datasets we found did not include enough utterances that we believed might plausibly appear in an everyday conversation and also might plausibly elicit a funny response from a companion. In addition, existing conversational datasets are limited in size, and we wanted to identify a source of setup sentences that is potentially unlimited so as to facilitate future research.

So, we generated setup sentences with OpenAI's ChatGPT, powered by the default GPT-5 model that it was using on September 15, 2025 (Singh et al. 2025). This entailed an iterative prompt-engineering dialogue during which we guided the output based on repeated evaluations of its style, tone, topic, factuality, plausibility, and naturalness. This prompt reflects the cumulative refinement of the instructions that produced the setups selected for the benchmark:

> Generate short, conversational setups that sound like something a friend might casually mention after skimming the news or noticing a pop-culture or everyday-life trend. Keep them factually true if they involve facts and fully plausible if styled like light news items. Make them evergreen, interesting enough to invite a punch line, and natural enough for casual conversation. Focus on everyday-life observations, slice-of-life community quirks, social-media and

> lifestyle trends, and trendy but credible pop-culture stories, while omitting quirky true facts. Keep most setups to a single short sentence, with a few slightly longer but uncluttered variations for variety, and avoid multi-clause structures using "and."

This method of generating setups is noteworthy because each resulting setup contains a pair of discrete and semantically distant concepts. Research has shown that this property tends to produce original jokes—ones that are qualitatively distinct from those obtained by simply prompting an LLM to generate humor (Amir 2025; Amir et al. 2026; Veale 2024).

From the 280 potential setups generated using this method, Authors JT and OA used their expert judgment as experienced comedy practitioners to select 80 setups that they agreed encompassed a wide variety of subject matter and also resembled plausible conversational utterances that might induce a witty companion to respond with a joke.

Generating the Responses

Once the setups were selected, each was paired with a response crafted to complete a short text that could be rated as either 0 (Not Funny), 1 (Somewhat Funny), 2 (Funny), or 3 (Very Funny). This four-point scale, similar to that used by Hossain, Krumm and Gamon (2019), was chosen because it seemed to span a sufficiently broad range of perceived funniness without demanding unrealistically precise predictions of perceived funniness. In implementations, if a conversational AI system equipped with a funniness rating module rated a human's response to its utterance as 0, the system would not exhibit any amusement behavior. But as the AI system rated the human's responses as increasingly funny, it would react appropriately with, say, a smile, a chuckle, or a laugh (Inoue, Lala and Kawahara 2022).

Two different methods were employed to generate the responses to the setups: one that was likely to yield funniness ratings of 1, 2, or 3; and a different one that would probably yield a rating of 0. Here are the two methods:

1: Somewhat Funny; 2: Funny; and 3: Very Funny–All of the responses likely to yield these ratings were produced by JT using Witscript, a neural-symbolic hybrid AI system that generates short jokes (Toplyn 2023). Witscript uses a large language model (LLM) to execute joke-writing algorithms created by a human expert. The version of Witscript employed here was the one publicly available as an app on October 28, 2025 [https://witscript.com/]. Using Witscript to generate these responses was orders of magnitude faster and less expensive than paying professional comedy experts to write them would have been.

0: Not Funny–All of the responses likely to yield this rating were generated on November 14, 2025, by OpenAI's ChatGPT, powered by its default model, GPT-5.1. This prompt was used to obtain the "Not Funny" responses:

> You are a typical American adult. Treat the given text as the first sentence of a news story or interesting anecdote. Write a plausible second sentence that continues the story in a conversational, realistic way. The sentence should be approximately 11–15 words long. You can describe the situation, add context, comment on people's reactions, or include your own reaction or opinion.

This prompt elicited responses comparable in length to those generated by Witscript, thereby eliminating the chance that a funniness rating system could identify Not Funny texts in the benchmark by their length. This method of generating unfunny responses is also in the spirit of Winters and Delobelle (2020), who demonstrate the advantages of using an automated method to generate non-jokes that are close in structure and vocabulary to jokes; this approach is preferable to using non-jokes from a completely different domain, such as a dataset of real news headlines, when evaluating humor recognition systems (Annamoradnejad and Zoghi 2024; Inácio, Wick-Pedro and Oliveira 2023).

Each setup and response produced by the above methods was concatenated to form a short text. Using Witscript and ChatGPT to build the short texts in this way minimizes the chances that any of the completed texts had previously appeared online and that an AI-powered funniness-rating system had been trained on them. This allows the benchmark to be used for relatively unbiased testing.

Applying those two generation methods to the 80 setups, JT generated 91 short texts possessing what he perceived to be widely varying levels of funniness, which he and OA then rated independently. After sharing and discussing their ratings, JT substituted a few responses, after which the authors used their expert judgment to cherry-pick 80 texts that they agreed would result in expert human raters assigning approximately 25% of the total number to each of the four rating categories.

Rating the Short Texts

Having prepared the dataset of 80 short texts, we turned to the question of how to assign them robust funniness ratings. As noted above, the short texts, while conversational, also have the same form as standalone setup/punch line jokes. Therefore, delivering the texts orally in front of large, live audiences of typical American adults and measuring the resulting laughter, if any, would be the best way to assign the ratings to them (Mirowski, Mathewson and Branch 2025; Mittal et al. 2021; Romanowski, Valois and Fukui 2025; Toplyn and Amir 2025). But this method was ruled out for practical and financial reasons: obtaining a statistically significant measurement of laughter, and therefore funniness, for each of the 80 texts would mean having them all delivered by various performers in different orders to many audiences consisting of hundreds of people.

Measuring the laughter on preexisting recordings of stand-up comedy performances would also fall short because of selection bias: comics only deliver jokes to live audiences that they expect to get laughs. Therefore, those

recorded jokes do not represent the full range of funniness–from very funny to not funny at all–that is required in the desired benchmark.

An alternative to measuring laughter would be to have human evaluators assign funniness ratings to the texts. Humor datasets exist that have been annotated by non-experts, usually crowdsourced, and by experts (Kalloniatis and Adamidis 2024). We concluded that ratings assigned by non-experts would be unreliable because research indicates that non-expert humans tend to be poor judges of creative works (Lamb, Brown and Clarke 2015), open-ended generated text (Karpinska, Akoury and Iyyer 2021), and transcripts of standup comedy (Romanowski, Valois and Fukui 2025). This conclusion is consistent with neuroimaging evidence suggesting that years of professional comedy experience are associated with measurable anatomical differences in brain structure and with distinct patterns of neural activity when processing humor (Amir and Biederman 2016; Brawer and Amir 2021). In addition, the evaluation of humor is greatly affected by the world knowledge, ideological biases, and other characteristics of the evaluators, who are often recruited using crowdsourcing (Loakman, Maladry and Lin 2023). For those reasons, we hypothesized that non-experts may not be able to accurately predict how typical American adults would react to the short texts. That is a particularly important consideration, because what is relevant to developing a benchmark that represents the expected funniness perceived by typical American adults is not what the human evaluators themselves perceive as funny, but what they believe typical American adults would perceive as funny.

Therefore, instead of non-experts, we had the texts rated for funniness by seven expert joke writers. The writers collectively have a total of over 120 years of professional experience writing jokes for television shows in the U.S. This experience qualifies them to provide benchmark funniness ratings for the texts for these reasons: (1) each of those experts has been paid to write, edit, and informally rank for funniness thousands of texts of the type in the benchmark, i.e., short, self-contained texts intended to elicit laughter from the millions of typical American adults viewing the shows at home; and (2) each of those experts has absorbed ground-truth information about the funniness of those texts by listening to hundreds of live studio audiences, each consisting of hundreds of typical American adults, reacting to a live host delivering the texts, typically during a comedy monologue.

Text	Consensus Rating
Parents are pushing for healthier food in school cafeterias. Makes sense—kids spend so much time at school, they might as well eat better.	0
The city is considering turning part of the old train station into an art gallery. Finally, commuters can miss their train while staring at an abstract departure board.	1
A clothing brand just launched a line of pajamas designed to look like office wear. So now you can get fired and go straight back to bed without changing.	2
Couples who laugh together are more likely to stay together longer. Unless the laughter is because one of them just fell off the roof.	3

Table 1: Examples of texts in the JEST benchmark that have ratings on which all three human expert raters agreed.

To prevent bias, the human experts were not told where the texts came from; they were only told that their ratings would be used to compare how jokes are rated by experts and non-experts. The texts were presented to the experts in a spreadsheet, with the same random ordering for each expert, along with these instructions:

> Your task is to rate how funny a given piece of text would be to an audience of typical American adults. Use this scale:
> 0 – Not Funny: No humor present. Likely to get no reaction.
> 1 – Somewhat Funny: Mildly amusing or clever. Almost a joke. Might get a smile.
> 2 – Funny: Clearly a joke. Likely to get a chuckle.
> 3 – Very Funny: A great joke. Likely to get a laugh.

The experts, who were paid $50 each, rated the texts independently and never had access to each other's ratings. See Table 1 for an example of a text in each funniness rating category.

Results and Discussion

To derive the final benchmark ratings, we selected a subset of three raters, out of the seven experts, whose individual rating distributions were the most balanced across the four funniness categories (0–3). This approach of increasing the influence of the most consistent or balanced raters (or equivalently, selecting a subset of them) has been recommended in methodological work on annotation quality (Dawid and Skene 1979; Krippendorff 2018; Paun et al. 2018; Snow et al. 2008). The selection served two purposes: (1) it enhanced the overall consistency of the gold-standard ratings while ensuring a more even spread of ratings across the full range of funniness levels; and (2) because most conversational agents aim to be agreeable and engaging, a somewhat more generous response to user humor is preferable; selecting the higher-rating cluster of experts supports this goal without sacrificing reliability. Indeed, even when aggregating all seven raters, the relative funniness ranking of the texts remains highly similar, differing primarily in an overall downward shift. Full details of the selection of raters for the final benchmark, individual rating distributions, and resulting improvements in balance and reliability follow.

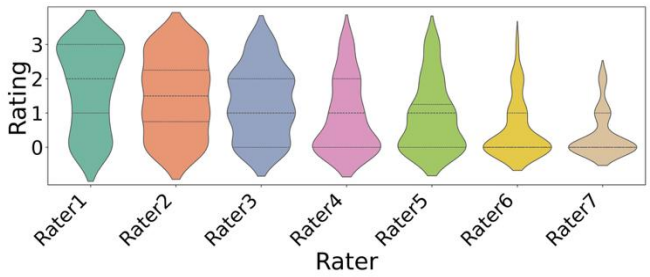

Figure 1: Violin plots displaying the distribution of funniness ratings (0–3 scale) for each of the seven expert raters across 80 texts. The plots highlight variability in rating tendencies: Raters 1-3 tend toward higher ratings and approximating an even spread across categories, while Raters 4-7 are notably conservative. Inner quartiles indicate the interquartile range.

Figure 1 shows the variability in the rating tendencies of the seven expert raters. Two raters were markedly stricter (means of 0.51 and 0.35) than the other raters (means of 0.90–1.74). Such rater clustering is common in subjective annotation tasks (Dethlefs et al. 2014) and, in particular, in humor (Ruch 1992). Notably, comedians are widely recognized within their own community as being particularly harsh critics and demanding audiences (Wilson 2014), a bias that developers of conversational agents may wish to avoid replicating.

This clustering of raters is also evident in Figure 2, which details pairwise Spearman correlations among the ratings of the seven expert raters. A core subset of three raters (anonymized as Rater1, Rater2, and Rater3) exhibits the highest correlations.

Table 2 shows the key inter-rater reliability and agreement metrics for all seven experts and for the core subset of three experts (Raters 1-3) whose individual rating distributions were the most balanced and consistent across the four funniness categories. Exact Agreement means

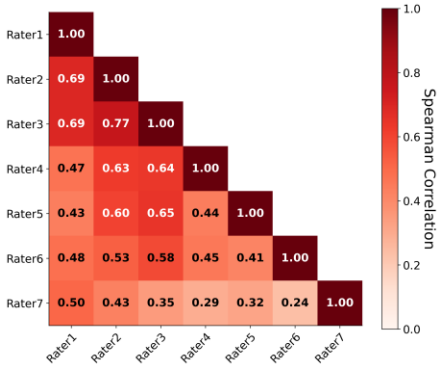

Figure 2: Pairwise Spearman correlations among the seven expert raters' text ratings, with correlations ranging from 0.24 to 0.77. The matrix reveals moderate overall agreement, with the core trio (Raters 1-3) exhibiting the highest correlations (>0.70), while Raters 6 and 7 show weaker alignment with the group, seemingly due to their conservative bias and a floor effect.

Metric	7 raters	3 raters
ICC(2,k)	0.82	0.87
Krippendorff's α (ordinal)	0.40	0.68
Exact Agreement (%)	20.0	41.3
Adjacent Agreement (%)	35.0	77.5
Mean Per-Item SD	0.74	0.45
Avg. Pairwise Spearman ρ	0.50	0.72

Table 2: Inter-rater reliability and agreement comparison (80 items, ordinal 0–3 scale).

identical ratings from the relevant raters, while Adjacent Agreement means ratings that differ by at most one point.

ICC(2,k) values indicate high reliability of the aggregated ratings in both configurations (0.82–0.87). Krippendorff's α was substantially higher for the three-rater subset (0.68 vs. 0.40), reflecting greater ordinal consistency. This improvement stems from the systematic differences in rating severity among the full set of seven experts. We consider ICC(2,k) to be the primary metric because it captures agreement on the actual ratings rather than merely their order, which is important here because each rating corresponds to a specific level of humor reaction from an AI agent. Additionally, ICC(2,k) is well-suited to our crossed design, in which all raters evaluated all items.

Selecting the subset of three raters with the most balanced individual distributions yielded sharper consensus: mean per-item standard deviation decreased from 0.74 to 0.45, exact agreement increased from 20% to 41%, and adjacent agreement rose from 35% to 78%.

Average pairwise Spearman correlation also improved substantially from 0.50 (all seven raters) to 0.72 (subset of three raters), indicating stronger rank-order consistency within that subset. Figure 3 provides a graphical illustration of the high degree of agreement in the ratings given by the subset of three raters.

In addition to possessing high reliability, we also wanted the benchmark to incorporate ratings distributed as evenly

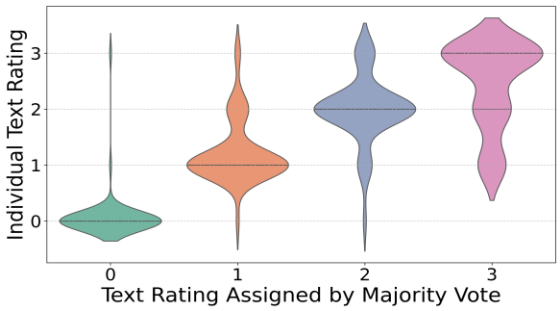

Figure 3: Violin plot showing the number of individual ratings in each category that were assigned to texts by the three selected raters, conditioned on the ratings that were assigned to those texts by the majority of the three selected raters. The plot illustrates how individual ratings cluster around consensus labels, with tighter distributions in smaller subsets indicating higher agreement.

as possible across the full 0–3 funniness range, to support fair evaluation of automated funniness raters at all levels of perceived funniness. With this additional objective in mind, we examined the majority-vote distributions (with ties resolved toward the higher rating) for all seven raters and for the three-rater subset:

- All seven raters: 41% Not Funny, 20% Somewhat Funny, 21% Funny, 18% Very Funny
- Three-rater subset: 25% Not Funny, 20% Somewhat Funny, 20% Funny, 35% Very Funny

The distributions showed that the three-rater subset provided the closest approximation to equal quarters while avoiding the pronounced skew toward Not Funny observed with all seven raters included. Although we selected only three out of the seven raters for the official JEST benchmark, mean ratings between the two configurations (average ratings of the seven vs. three raters) correlated extremely highly across texts (Spearman $\rho = .95$, $p < .001$), confirming that relative funniness rankings remained virtually unchanged despite subset selection.

For the above reasons, the official JEST benchmark defines its reference ratings as the median rating given by the subset of three selected experts (Rater1, Rater2, Rater3) for each of the 80 texts. We used the median instead of the mean because the 0–3 scale is coarse and the median is more robust to occasional outlier ratings and better reflects consensus. This three-rater configuration yields ICC(2,k) = .87, a substantial agreement for such a highly subjective task.

Illustrative Application

To demonstrate the utility of the JEST benchmark, we propose this hypothetical scenario: The XYZ Company is developing an artificial companion for typical American adults and wants to give its conversational AI system the ability to deliver an appropriately scaled humor reaction when a user responds to one of its utterances with a joke.

So, the company sets out to identify the best off-the-shelf LLM to power a funniness-rating module that can be deployed locally in its artificial companion so as to deliver low-latency, humanlike humor reactions to the user at reasonable cost. For business reasons, the company wants to use an LLM in OpenAI's GPT-5 family of models.

Model selection

To assist the hypothetical company, we evaluated models that span the range of capability of the GPT-5 family, from lightweight variants (GPT-5-nano and GPT-5-mini) to the core generation (GPT-5 and its incremental updates GPT-5.1 and GPT-5.3) and the high-capability GPT-5-pro model. By "capability" we mean a broad measure of a model's overall performance across a variety of tasks. We also included a coding-optimized model, GPT-5-codex, to see whether a model optimized for structured reasoning tasks would perform better than general-purpose models.

The GPT-5 family models evaluated in this study were all released following the official GPT-5 launch period beginning in August 2025.

Evaluation procedure

When evaluated as funniness raters, the LLMs were used off-the-shelf and were not prompted with any examples or any instructions related to humor theories. Instead, all of the above LLMs were given the following prompt.

> System message:
> You are evaluating short texts for their potential to be funny to a typical American adult audience.
>
> Use the following 0–3 scale:
> 0 = not funny (likely to get no reaction)
> 1 = almost funny (might get a smile)
> 2 = funny (likely to get a chuckle)
> 3 = very funny (likely to get a laugh)
>
> User message:
> Text: {{text}}
> Output only the numeric rating.

For evaluation, all LLMs were accessed via API in early March 2026. The performance of each LLM was evaluated by measuring the degree of agreement between the LLM's funniness ratings and the expert ratings across the 80 texts in the JEST benchmark.

Unlike the inter-rater reliability analysis above, where ICC(2,k) and Krippendorff's α were appropriate for multiple comparable raters evaluating the same items, the two-way ICC framework is less suitable for the analysis here because it assumes symmetric, exchangeable raters drawn from the same population. That assumption does not hold when one "rater" is a carefully constructed human expert consensus benchmark and the other is an LLM operating via a fundamentally different mechanism.

We therefore treat JEST as a fixed reference standard and measure each LLM's agreement against it using quadratic weighted Cohen's κ as the primary metric, which accounts for the ordinal structure of the 0–3 scale, penalizes larger disagreements more heavily than smaller ones, and is standard practice in NLP for evaluating system outputs against a gold-standard reference. We additionally report Spearman's ρ to capture rank-order consistency. Unweighted κ was not used because it treats all disagreements as equivalent regardless of magnitude.

Performance assessment

Table 3 shows how the seven LLMs compare on key characteristics and agreement against the JEST benchmark. According to the widely used scale established by Landis and Koch (1977), κ values above .60 indicate substantial agreement with a reference standard. Using that guideline, most of the evaluated GPT-5 family models demonstrated substantial agreement with the benchmark. But the least-capable models, GPT-5-mini (κ = .58) and GPT-5-nano (κ = .29), did not reach that threshold. This result is consistent

Model	Capability	Speed	Cost Eff.	Cohen's κ	Spearman's ρ
gpt-5-pro	5	2	2	.66*** [.53, .76]	.66*** [.49, .79]
gpt-5.3	5	3	3	.73*** [.63, .80]	.77*** [.66, .85]
gpt-5.1	4	3	3	.74*** [.64, .81]	.78*** [.68, .85]
gpt-5	4	3	3	.73*** [.63, .80]	.76*** [.64, .84]
gpt-5-codex	4	3	3	.69*** [.56, .79]	.69*** [.51, .82]
gpt-5-mini	3	4	4	.58*** [.46, .66]	.73*** [.60, .82]
gpt-5-nano	2	5	5	.29*** [.17, .41]	.52*** [.37, .65]

Table 3. Comparison of the LLMs, with the agreement between their funniness ratings and the JEST benchmark as measured by quadratic weighted Cohen's κ and Spearman's ρ. 95% confidence intervals are in square brackets, with bootstrap p-values all ***: $p < .001$. Relative qualitative ratings of Capability, Speed, and Cost Efficiency were provided by OpenAI in early March 2026; 5 = highest in model group, 1 = lowest.

with Loakman, Thorne and Lin (2025), who found that smaller LLMs (e.g., Gemini Flash and GPT-4-mini) underperformed their full-size counterparts on a joke-explanation task. They attributed this at least partly to the smaller models' failure to retrieve the cultural references that jokes rely on, possibly because such references were underrepresented in their training data. It is also possible that smaller models do not have the ability to perform the deep semantic analysis required for detecting humor structures in text.

The remaining five models all exceeded the .60 threshold, suggesting that larger models generally have the contextual understanding needed to approximate expert judgments of funniness. Among them, GPT-5.1 achieved the highest agreement ($\kappa = .74$), followed closely by GPT-5 and GPT-5.3 (both $\kappa = .73$). The specialized GPT-5-codex ($\kappa = .69$) trailed those top performers slightly, which may reflect its optimization for structured coding tasks instead of conversational funniness evaluation. More surprisingly, GPT-5-pro ($\kappa = .66$) also performed somewhat below the three leaders. Narad et al. (2025) found that post-training models such as DeepSeek R1 on general STEM reasoning tasks improved their ability to rate and explain humorous New Yorker cartoon captions. However, the models they tested predate the GPT-5 series, which already incorporates extensive reasoning-focused post-training. It is possible that the heavier reasoning emphasis of GPT-5-pro may encourage it to "overthink the joke," rather than mimic the fast, associative behavior that humans exhibit when deciding whether some text is funny.

Table 3 also shows that Spearman correlations were similarly strong across all the models except for the least-capable, GPT-5-nano, with substantially overlapping confidence intervals. This suggests consistent ordinal agreement on joke funniness regardless of model capability.

Taken together, these results suggest that GPT-5.1, with the highest κ, offers the best overall balance of agreement with expert funniness ratings while still being speedy and cost effective enough for deployment. Therefore, the XYZ Company should consider GPT-5.1 to be its top candidate for powering the funniness rating module in its artificial companion. However, if later testing reveals that GPT-5.1 does not meet the required speed or cost constraints, GPT-5-mini may be a viable alternative: it's faster and more economical than GPT-5.1 while still performing close enough to the substantial-agreement threshold to plausibly "get" a joke the way a typical American adult would.

Contributions

This paper makes the following contributions:

1. It introduces the first benchmark for evaluating funniness rating systems, comprising a dataset of 80 short texts annotated with non-binary funniness ratings by human experts, along with a primary evaluation metric: the quadratic weighted Cohen's κ. The benchmark data will be made available by the corresponding author, upon request, to legitimate researchers who agree not to redistribute it publicly, in order to prevent its use as training data and preserve its intended use as an evaluation resource. The benchmark data consists of two datasets. Dataset A presents the official JEST benchmark data: the 80 short texts, including setups and responses, ordered by descending median funniness rating (computed from the ratings of the three selected raters whose ratings were most evenly distributed across the four funniness categories), together with the median rating, exact and adjacent agreement percentages, and mean absolute deviation for each text. Dataset B provides, for transparency, the complete data from all seven expert raters, including medians, agreement metrics, descriptive statistics, and individual ratings for each text.
2. It presents a speedy, economical method of using an LLM and an automated joke generator to synthesize an unlimited number of short texts with varying degrees of funniness. Those texts can greatly increase the amount of data available for humor research.
3. It provides evidence that an off-the-shelf LLM can rate the perceived funniness of a dataset of short, conversational texts in substantial agreement with a panel of human experts.

Conclusion

The JEST benchmark is a resource that can be used to evaluate and advance the development of automated funniness-rating systems that perceive the funniness of short texts as typical American adults would. Such systems could be used to rapidly and economically evaluate the output of automated joke generation systems, thereby facilitating the development of the latter.

An effective funniness-rating system would also enable a conversational AI system to judge how funny a user's utterance is and respond with an appropriately scaled humor reaction. Paired with an automated joke generator, an automated funniness rater could enable a conversational AI, like a chatbot, to strike typical humans as having a humanlike sense of humor, thus making the chatbot more likeable and enhancing its interactions with users.

In the future, we will use the JEST benchmark to evaluate a wide range of additional methods for rating perceived funniness, including numerical ratings by crowdsourced, non-expert evaluators; by both naive and humor-theory-specific LLMs from multiple providers; by an ensemble of LLMs-as-judges; and by a novel, neural-symbolic, hybrid AI system. Our goal is to identify the most effective system for automatically rating the perceived funniness of a two-part text similar to a turn in a conversation.

Limitations

In this section we list limitations of this work and of the JEST benchmark it presents.

1. The utility of the benchmark is limited to funniness rating systems designed for short, conversational texts and typical American adults. It may be of little use in evaluating rating systems designed for other languages, cultures, demographics, socio-economic contexts, personality types, or forms of humor. Acknowledging that humor is highly subjective and context-dependent, we make no claim that JEST is a benchmark of funniness as perceived by everyone everywhere at all times.
2. This work treats the aggregated expert judgments in the JEST benchmark as a practical reference standard for funniness as perceived by a substantial segment of the U.S. adult population, but such judgments do not constitute an objective ground truth.
3. The benchmark is based on a relatively small number of texts (80), reflecting a tradeoff between the size of the dataset and the potential fatigue of the human expert raters.
4. The benchmark incorporates the ratings of only three experts. As we have shown in this paper, given the subjective nature of humor, a different group of experts might arrive at somewhat different consensus ratings.
5. The benchmark is based on the ratings of expert joke writers who judged the texts on their potential to amuse a broad U.S. audience; as a result, conversational jokes that rely on shared context, i.e., "inside jokes," may not be well-represented.
6. While the benchmark is based on texts designed to be relatively evergreen, evolving cultural references or shifts in humor tastes over time could affect its long-term utility.
7. Finally, the benchmark is intended solely as a resource for evaluating funniness rating systems rather than as a training corpus, which may limit its usability for some applications.

Acknowledgments

We would like to thank the expert joke writers who rated the texts in the JEST benchmark.

References

Amir, O. 2025. Are AI-generated jokes truly original? Charting the "Joke Space." In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*. Campinas, Brazil. Association for Computational Creativity.

Amir, O., and Biederman, I. 2016. The neural correlates of humor creativity. *Frontiers in Human Neuroscience*, 10:597.

Amir, O.; Toplyn, J.; Ngo, H.D.; and Hickerson, K.P. 2026 (in press). Navigating the Joke Space: Towards automated originality assessment of AI-generated humor. In *Proceedings of the 2nd Workshop on Computational Humor (CHum)*, Online. Association for Computational Linguistics.

Annamoradnejad, I., and Zoghi, G. 2024. ColBERT: Using BERT sentence embedding in parallel neural networks for computational humor. *Expert Systems with Applications*, 249, PB.

Avetisyan, H.; Safikhani, P.; and Broneske, D. 2023. Laughing out loud – Exploring AI-generated and human-generated humor. In *Proceedings of the 4th International Conference on NLP Artificial Intelligence Techniques (NLAI 2023)*.

Bogireddy, N.R.; Suresh, S.; and Rai, S. 2023. I'm out of breath from laughing! I think? A dataset of Covid-19 humor and its toxic variants. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion)*, 1004–1013, New York, NY, USA. Association for Computing Machinery.

Bower, A.H., and Steyvers, M. 2021. Perceptions of AI engaging in human expression. *Scientific Reports*, 11:21181.

Brawer, J., and Amir, O. 2021. Mapping the 'funny bone': Neuroanatomical correlates of humor creativity in professional comedians. *Social Cognitive and Affective Neuroscience*, 16(9): 915-925.

Ceha, J.; Lee, K.; Nilsen, E.S.; Goh, J.; and Law, E. 2021. Can a humorous conversational agent enhance learning experience and outcomes? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–14, New York, NY, USA. Association for Computing Machinery.

Chiruzzo, L.; Castro, S.; and Rosá, A. 2020. HAHA 2019 dataset: A corpus for humor analysis in Spanish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5106–5112. Marseille, France. European Language Resources Association.

Clark, L.; Pantidi, N.; Cooney, O.; Doyle, P.; Garaialde, D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C.; Wade, V.; and Cowan, B.R. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–12. New York, New York, USA. Association for Computing Machinery.

Dawid, A.P., and Skene, A.M. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Dethlefs, N.; Cuayáhuitl, H.; Hastie, H.; Rieser, V.; and Lemon, O. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *14th Conference of the European Chapter of the Association for Computational Linguistics 2014*, 702–711. Gothenburg, Sweden. Association for Computational Linguistics.

Faruqi, F., and Shrivastava, M. 2018. "Is this a joke?": A large humor classification dataset. In *Proceedings of the 15th International Conference on Natural Language Processing*, 104–109, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Hossain, N.; Krumm, J.; and Gamon, M. 2019. "President vows to cut hair": Dataset and analysis of creative text editing for humorous headlines. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 133–142, Minneapolis, Minnesota. Association for Computational Linguistics.

Inácio, M.L.; Wick-Pedro, G; and Oliveira, H.G. 2023. What do humor classifiers learn? An attempt to explain humor recognition models. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Inoue, K.; Elmers, M.; Lala, D.; and Kawahara, T. 2025. Why do we laugh? Annotation and taxonomy generation for laughable contexts in spontaneous text conversation. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, 318–323, Bilbao, Spain. Association for Computational Linguistics.

Inoue, K.; Lala, D.; and Kawahara, T. 2022. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and AI*, 9:933261.

Kalloniatis, A., and Adamidis, P. 2024. Computational humor recognition: A systematic literature review. *Artificial Intelligence Review*, 58:43. Published online 2025.

Kang, S.; Krum, D.M.; Khooshabeh, P.; Phan, T.; Chang, C.; Amir, O.; and Lin, R. 2017. Social influence of humor in virtual human counselor's self-disclosure. *Computer Animation and Virtual Worlds*, 28.

Karpinska, M.; Akoury, N.; and Iyyer, M. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kiddon, C., and Brun, Y. 2011. That's what she said: Double entendre identification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 89–94, Portland, Oregon, USA. Association for Computational Linguistics.

Krippendorff, K. 2018. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Lamb, C.; Brown, D.G.; and Clarke, C.L.A. 2015. Human competence in creativity evaluation. In *Proceedings of the Sixth International Conference on Computational Creativity*, 102-109. Park City, Utah, USA. Association for Computational Creativity.

Landis, J. R., and Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lemmens, J., and De Marez, V. 2026. Computational humor modeling: A survey on the state of the art. *ACM Computing Surveys*, 58 (7), 1-37.

Loakman, T.; Maladry, A.; and Lin, C. 2023. The iron(ic) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6676–6689, Singapore. Association for Computational Linguistics.

Loakman, T.; Thorne, W.; and Lin, C. 2025. Comparing apples to oranges: A dataset & analysis of LLM humour understanding from traditional puns to topical jokes. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 9502–9518, Suzhou, China. Association for Computational Linguistics.

McKeown, G., and Curran, W. 2015. The relationship between laughter intensity and perceived humour. In *Proceedings of the 4th Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalisations in Speech,* 27–29, Enschede, Netherlands.

Mihalcea, R., and Strapparava, C. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Mirowski, P.; Mathewson, K.; and Branch, B. 2025. The theater stage as laboratory: Review of real-time comedy LLM systems for live performance. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, 88–95, Online. Association for Computational Linguistics.

Mittal, A.; Jeevan, P.; Gandhi, P.; Kanojia, D.; and Bhattacharyya, P. 2021. "So you think you're funny?": Rating the humour quotient in standup comedy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10073–10079, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Murakami, S.; Kamigaito, H.; Takamura, H.; and Okumura, M. 2025. Oogiri-master: Benchmarking humor understanding via Oogiri. *arXiv preprint arXiv:2512.21494*.

Narad, R.; Suresh, S.; Chen, J.; Dysart-Bricken, P.S.L.; Mankoff, B.; Nowak, R.; Zhang, J.; and Jain, L. 2025. Which LLMs get the joke? Probing non-STEM reasoning abilities with HumorBench. *arXiv preprint arXiv:2507.21476.*

Nijholt, A.; Niculescu, A.I.; Valitutti, A.; and Banchs, R.E. 2017. Humor in human-computer interaction: A short survey. In Joshi, A.; Balkrishan, D. K.; Dalvi, G.; and Winckler, M., eds., *Adjunct Proceedings INTERACT 2017 Mumbai: 16th IFIP TC.13 International Conference on Human Computer Interaction*, 192–214.

Paun, S.; Carpenter, B.; Chamberlain, J.; Hovy, D.; Kruschwitz, U.; and Poesio, M. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Potash, P.; Romanov, A.; and Rumshisky, A. 2017. Semeval-2017 task 6: Hashtagwars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 49–57, Vancouver, Canada. Association for Computational Linguistics.

Quan, K.; Ramakrishnan, P.; and Chin, J. 2025. Can AI take a joke—or make one? A study of humor generation and recognition in LLMs. In *Proceedings of the 2025 Conference on Creativity and Cognition*, 431-437, New York, New York, USA. Association for Computing Machinery.

Romanowski, A.; Valois, P.H.V.; and Fukui, K. 2025. From punchlines to predictions: A metric to assess LLM performance in identifying humor in stand-up comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Ruch, W. 1992. Assessment of appreciation of humor: Studies with the 3 WD humor test. *Advances in Personality Assessment*, 9, 27–75.

Singh, A.K. et al. 2025. OpenAI GPT-5 system card.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A.Y. 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Sun, X.; Teljeur, I.; Li, Z.; and Bosch, J.A. 2024. Can a funny chatbot make a difference? Infusing humor into conversational agent for behavioral intervention. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24)*, 1–19, New York, New York, USA. Association for Computing Machinery.

Toplyn, J. 2014. *Comedy Writing for Late-Night TV*. Rye, New York, USA: Twenty Lane Media, LLC.

Toplyn, J. 2023. Witscript 3: A hybrid AI system for improvising jokes in a conversation. *arXiv preprint arXiv:2301.02695*.

Toplyn, J., and Amir, O. 2025. Can AI make us laugh? Comparing jokes generated by Witscript and a human expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, 71–78, Online. Association for Computational Linguistics.

Veale, T. 2024. From symbolic caterpillars to stochastic butterflies: Case studies in re-implementing creative systems with LLMs. In *Proceedings of the 15th International Conference on Computational Creativity (ICCC)*. Jönköping, Sweden. Association for Computational Creativity.

Weller, O., and Seppi, K. 2020. The rJokes dataset: A large scale humor collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6136–6141, Marseille, France. European Language Resources Association.

Wilson, G. 2014. *The Complete Guide to Stand-Up: Everything You Need to Know, from Open-Mics to Going Pro!* Gregory D Wilson / The Comedy Institute.

Winters, T. 2021. Computers learning humor is no joke. *Harvard Data Science Review*, 3(2).

Winters, T., and Delobelle, P. 2020. Dutch humor detection by generating negative examples. In *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC 2020) and the 29th Belgian-Dutch Conference on Machine Learning (Benelearn 2020)*, 313–323.

Yang, Z.; Hooshmand, S.; and Hirschberg, J. 2021. Choral: Collecting humor reaction labels from millions of social media users. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4429–4435, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zangari, A.; Marcuzzo, M.; Albarelli, A.; Pilehvar, M.T.; and Camacho-Collados, J. 2025. Pun unintended: LLMs and the illusion of humor understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 27924–27959, Suzhou, China. Association for Computational Linguistics.