Computational Linguistics in the Netherlands Journal 14 (2025) 505-523 Submitted 12/2024; Published 07/2025

Evaluating Humor Generation in an Improvisational Comedy Setting

Thomas Winters* THOMAS.WINTERS@KULEUVEN.BE
Stijn Van der Stockt** STIJN.VAN.DER.STOCKT@PERSGROEP.NET

*Department of Computer Science; Leuven.AI, KU Leuven, Leuven, Belgium

**DPG Media, Belgium

Abstract

While computational humor generation has long been considered a challenging task, recent large language models have significantly improved the quality of generated jokes. Evaluating humor quality is usually difficult, as not only is the exact quality subjective, but delivery also plays a role. Another disparity in evaluation standards between human and computer-generated humor is the difference in writing time between the two. In this study, we evaluate and compare the quality of humor generated by GPT-4 with human-written jokes in an improvisational comedy setting in Dutch. In a live performance setting on national TV, nine different audience suggestions were used across three improvisational comedy games. Three professional comedians each performed their own improvised joke and an AI-generated joke per round, resulting in a total of 54 jokes. The AI-generated jokes were selected in real time from candidate outputs generated by GPT-4 using a few-shot chain-of-thought prompt specific to each game and audience suggestion. An audience of 40 people then rated all jokes on a 4-point scale, resulting in 2,160 ratings. This allows us to compare the difference in quality between AI and human-created jokes delivered by the same comedian for the same audience suggestion. Our results show that audience members preferred human-created jokes 34.6% of the time, AI-generated jokes 29.7% of the time, and rated them equally in 35.7% of cases. Human-created jokes also received a slightly higher average rating (2.67 vs. 2.59), although GPT-4 occasionally produced standout jokes that received high "best joke" votes. These findings suggest that while human improvisation retains a narrow edge in consistency, current large language models can produce competitive humor under real-time constraints.

1. Introduction

Humor has long been considered one of the most difficult types of creative text to generate automatically. Crafting a joke that resonates with an audience requires multiple sophisticated capabilities: linguistic proficiency, reasoning, cultural understanding, comedic timing, and sensitivity to subtle social cues. As such, humor is often seen as an ideal testbed for language models, as being able to perform humor generation or detection with a language model suggests broader linguistic competence.

Historically, humor detection and generation have been limited by the inherent restrictions of symbolic systems and the lack of resources, particularly in non-English languages. Early attempts at automated humor production often resulted in low humorous value and formulaic output. These systems were predominantly English-based due to the greater availability of NLP tools (Castro et al. 2016, Winters 2019, Terai et al. 2020). However, the rapid development of large language models (LLMs) like GPT-4 has led to significant improvements in automated joke production.

Despite these technological advances, the evaluation of humor quality remains a challenge. Humor is inherently subjective: what one person finds hilarious, another might consider dull or even offensive. Previous work has also found large disagreements among evaluators of human-created and AI-generated jokes (Winters et al. 2018). In addition, the perceived quality of a joke is influenced not only by its textual content but also by the delivery, context, and performance style. Therefore,

an AI-generated humor evaluation should take into account these performance aspects rather than rely solely on text-based assessments.

A further challenge in evaluating AI-generated humor lies in establishing fair baselines for comparison. Benchmarks often use existing humor datasets from online repositories and comedians (Amin and Burghardt 2020). However, human comedians often invest substantial time and effort in refining and filtering their jokes by testing material in front of various audiences. Automated joke generators, on the other hand, can produce a virtually unlimited number of jokes instantly, but do not undergo human-like creative filtering processes. Direct comparisons between polished human jokes and raw machine output are thus inherently biased. In this study, we level the playing field by evaluating both human-created and AI-generated humor in an improvisational context, where both are created on the spot without preparation time.

To address these challenges, we designed a study based on improvisational comedy, a highly constrained, real-time format. Improvisational comedians rely on spontaneous audience suggestions and perform under strict time and topical constraints. By placing both human comedians and GPT-4 in the same improvisational context and constraints, we are able to compare humor quality under conditions that minimize the advantages of pre-refinement and curation. In this controlled environment, we examine how the humor generated by GPT-4 compares to spontaneous humor produced by professional comedians performing in Dutch.

This paper contributes to computational humor generation research by providing evidence that GPT-4, when using few-shot chain-of-thought prompting, can generate humor that rivals that of professional improvisers. Although human performers retain a slight edge, AI-generated jokes occasionally outperform, suggesting potential for future collaborations between AI and comedians as effective creative partners in real-time humor production scenarios.

2. Background

2.1 Humor Theory

There is currently no universal theory of humor that explains all jokes. Many theories point to *incongruity* as a central element (Morreall 1986, Gervais and Wilson 2005, Hurley et al. 2011). Incongruity theories posit that humor stems from disrupting an expectation set by the initial part of a text or scene (the setup) by an unexpected, yet ultimately coherent, reinterpretation introduced by the punchline. One more specific theory that is especially useful for computational generation of jokes is the *incongruity-resolution (IR)* model (Ritchie 1999, Suls 1972, Shultz 1974). It states that a joke is perceived as funny if there first is an incongruity and then a resolution to this incongruity. Take for example the joke *"I want to die like my grandfather, silently in his sleep, not like his screaming passengers"*. There is a clear incongruity when we hear the word *"passengers"*. At first, the listener mentally visualizes a grandfather dying in his bed (= the obvious interpretation), but the word *passengers* forces a reinterpretation, as a bed usually does not have passengers. The listener's mind attempts to resolve this incongruity, often discovering a second, hidden interpretation that makes coherent sense of the entire utterance. This causes the listener to consider new interpretations of the earlier text and then resolve the incongruity by envisioning a scenario where the grandfather was actually driving a bus or taxi, or perhaps even piloting a plane (the hidden interpretation). Finding this interpretation causes a sense of relief, making the listener laugh. The cognitive reward from successfully resolving this incongruity often manifests as amusement and laughter (Hurley et al. 2011, Chan et al. 2012). The resolution is thus a crucial aspect for perceiving the text as humorous, rather than being left confused by the presence of the incongruity.

2.2 Computational Humor Generation

Teaching computers to detect and generate humor has long been viewed as a challenging, almost "AI-complete" task because it involves near-human-level linguistic skills, reasoning, and cultural understanding (Attardo 2001, Binsted et al. 2006, Hurley et al. 2011, Stock and Strapparava 2006, Winters 2021).

2.2.1 Symbolic Humor Generators

Computational humor has been around for decades. Early efforts were usually limited by the narrow capabilities of natural language processing systems from the time. Early joke generators focused on pun-based wordplay and relied heavily on template-based approaches where the slots were filled with words that had predefined relations between each other. For example, JAPE (Binsted and Ritchie 1994) produced jokes like *"What's green and bounces? Spring cabbage"* and spoonerisms such as *"What's the difference between leaves and a car? One you brush and rake, the other you rush and brake"*. These structural constraints led to jokes with predictable formats that became repetitive quickly.

2.2.2 GPT Humor Difficulties

Large language models, like GPT-models (Radford et al. 2019, Brown et al. 2020, OpenAI 2023), brought a leap forward in many NLP tasks, including humor generation. Trained on massive corpora of text, these models learned a broad range of linguistic patterns, cultural references, and even comedic styles. They can mimic stand-up, puns, satire, or surreal humor with increasing fluency (Frolovs 2019, Sabeti 2020). However, early attempts to generate humor with these models often resulted in incoherent punchlines or repetitive joke formats, similar to a child who can mimic a joke's structure without producing a true punchline (Branwen 2020, Winters 2021). Researchers even discovered that ChatGPT 3.5 generated variations of the same 25 jokes over 90% of the time (Jentzsch and Kersting 2023). The fact that ChatGPT is changing key words in these fixed jokes also meant that it usually broke the jokes, as this removes key relationships necessary to create either the obvious or hidden interpretation (Winters and Delobelle 2020).

Another difficulty in generating humor with GPT systems can be explained by humor theory. The incongruity-resolution theory suggests that humor often arises from unexpected outcomes that subvert audience expectations. This presents a challenge for GPT-based humor generation since these models predict tokens sequentially, focusing on the most likely next word rather than planning for a deliberate surprise or punchline at the end. This is unlike real comedians, who will first think of a punchline, and only afterwards write a setup that can convey a successful obvious interpretation. As a result, a basic instruction often leads GPT models to generate texts that imitate jokes stylistically but lack a twist at the end. Fine-tuning GPT-2 or GPT-3 on joke datasets, while helpful, does not solve this, as these models often yield repetitive or low-diversity humor, with most improvements being that the generated text is stylistically closer to jokes (Frolovs 2019, Winters 2021).

2.2.3 Humor Prompt Strategies

To improve upon this immediate left-to-right generation of jokes, the GPT model can be prompted in a way that allows it to first plan a punchline. This is usually achieved using chain-of-thought prompting, where the model is asked to decompose the process of writing jokes into smaller steps themselves before writing the joke itself (Kojima et al. 2022). We can also give examples of such reasonings for given topics to already existing jokes, so that GPT can imitate such reasonings, which is called few-shot chain-of-thought prompting (Wei et al. 2022). By illustrating the desired pattern with a few high-quality examples and then instructing the model to think through the joke creation process step-by-step, we encourage the planning of better and more surprising punchlines. Such methods aim to replicate some aspects of the comedian's creative process, such as starting from a

premise, brainstorming multiple angles, and refining the idea to produce a coherent and humorous punchline. For example the prompt below, based on reverse-engineered jokes from British comedian Tim Vine[1], famous for quick-fire puns and one-liners:

> *You are an expert joke writer with a proven track record writing jokes for professional comedians. You write jokes about given topics. Your task is to write down a brainstorm of associations and reasoning process for writing several punchlines for a topic, like the examples below.*
>
> ***Topic:*** *collecting dust*
> ***Reasoning:*** *Collecting dust means you don't use something. Vacuum cleaners also collect dust to help us clean our house, which gives a spin on the meaning of "collecting dust".*
> ***Punchline:*** *my vacuum cleaner is collecting dust*
> ***Joke:*** *I decided to sell my Hoover... Well, it was just collecting dust.*
>
> ***Topic:*** *serves him right*
> ***Reasoning:*** *Serving someone right means you got your deserved punishment. Serving could also mean performing a duty. So serving someone right means they do something right-handedly.*
> ***Punchline:*** *if you miss your left arm, you can only serve right.*
> ***Joke:*** *A friend of mine's got a left arm missing. Serves him right.*

Such prompts allow the model to emulate the step-by-step reasoning process of a comedian. Similar decomposition of joke writing steps (albeit sometimes using multiple prompts and systems) have similarly produced jokes that are much better than plainly prompting GPT (Toplyn 2023, Inácio and Oliveira 2024, Tikhonov and Shtykovskiy 2024).

2.3 Humor Evaluation Challenges

Evaluating humor is notoriously difficult. Traditional evaluation methods for computational humor are heavily based on text-based assessments (Amin and Burghardt 2020). Researchers typically collect ratings from human judges who read and score generated jokes in written form. These evaluations often use metrics like funniness scales, binary funny/not-funny judgments, or comparative rankings between multiple jokes. While this approach offers a controlled environment for assessment, it poses three main challenges: it fails to capture the full spectrum of humor appreciation that emerges in live performance settings, it compares against existing refined jokes and often does not take into account the subjectivity of individual raters. Furthermore, individual subjectivity leads to low inter-annotator agreement (Winters et al. 2018). To reduce this subjectivity, our study compares ratings for human and AI-generated jokes delivered by the same comedian under identical audience suggestions.

2.3.1 Delivery Dependence

While jokes can be read in textual form, they are often read out loud and even performed by another person. The impact of a joke often depends on timing, tone, and presentation, facial expressions, body language, and vocal delivery. While it is common practice to evaluate AI-generated humor in textual form, it disregards the potential the joke might have when performed by a skilled comedian rather than being read.

Our study takes a different approach by integrating computational humor generation into a live improvisational performance, directly exposing generated jokes to immediate audience feedback. This setting enables a more valid assessment of the model's performance in a realistic entertainment

1. (`https://www.timvine.com/`)

scenario, similar to previous evaluations using improvisational theater (Mathewson and Mirowski 2017, Mathewson and Mirowski 2018, Mirowski et al. 2020).

2.3.2 Unfair Comparison

A major challenge in existing humor evaluation literature is the inherent bias in comparing polished human jokes with raw, unfiltered AI-generated content. When evaluating the quality of generated jokes, studies often compare against existing human-created jokes (Petrović and Matthews 2013, Binsted et al. 1997, Tikhonov and Shtykovskiy 2024). These human jokes used for evaluation typically come from two sources: existing performances from professional comedians and curated collections like joke books, online repositories, or social media. In essence, the human material you find in existing humor datasets has already been subject to a process akin to natural selection. Comedians tend to filter out bad jokes from their set and refine the phrasing of successful ones based on audience feedback. Existing folklore jokes also undergo a similar filtering and refinement process, as people typically change the way they retell a joke, and thus mutate the joke. The best mutations have a higher probability of spreading over the population and mutating into even better versions, and eventually end up in these curated collections of high-quality jokes. While these better jokes are also more likely to end up in the training data of LLMs and thus improve their quality of joke writing, we expect the LLMs to produce novel jokes in humor generation evaluations. This is often hard to guarantee, as earlier versions of ChatGPT have been shown to produce jokes that are highly similar to existing jokes when using simple prompts (Jentzsch and Kersting 2023). By enforcing enough constraints, such as using complex topics and requiring the joke to follow a given pattern, we drastically reduce the risk of reproducing existing jokes. Comparing polished human jokes to raw generated output might thus lead to unfair comparisons that underestimate the potential of the computational methods.

In our design, both human and AI jokes are created in a live improvisational context, which aims to minimize the discrepancy of refinement and selection time between human and AI produced jokes. Improvising these jokes on stage based on audience suggestions means that there is little preparation time and no time to revise the joke based on external feedback, allowing direct comparisons of human-created jokes under similar constraints as an automated humor generation.

2.3.3 Subjectivity

Humor perception is highly subjective and can vary dramatically based on cultural background, personal experience, and even mood. The popular use of Likert scales also makes comparison between different raters harder. While one rater might rate most jokes between 3 and 5 stars, another might rate between 1 and 4, while meaning the same funniness. While Likert-scale ratings from multiple annotators are standard practice, these often show low inter-annotator agreement, reflecting the personal nature of humor (Winters et al. 2018). To reduce this subjectivity, we compare the ratings for human-created and AI-generated jokes per participant and per performer, and evaluate whether one was rated higher, lower, or equally.

3. Experimental Setup

We enlisted three professional Dutch improvisational comedians with extensive live performance experience. The performance[2] took place before a live studio audience of 40 individuals (Figure 1a). Importantly, the audience members were unaware that some jokes were AI-generated, to avoid expectation bias (Bower and Steyvers 2021). Immediately after the full on-stage performance, the audience filled in a questionnaire containing Likert-scale ratings for each performed joke within 15

2. Recorded by DPG Media for the Belgian TV-show "Ze Zeggen Dat" (Season 4, Episode 7), accessible at: `https://lfvp-static-overlays.dpgmedia.net/3/247901632`

minutes after the end of the show. Besides rating each joke, we also asked the audience to select their favorite joke of the entire performance.

3.1 Live Performance

The comedians were instructed to deliver the AI-generated jokes as if they had created them on the spot, so that the audience would not notice a difference. Importantly, both human and AI-generated jokes were created in real-time, ensuring a fair comparison without substantial refinement.

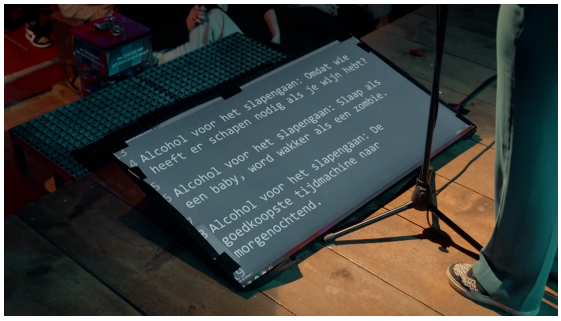
Each comedian participated in three distinct improvisational games, each played three rounds with a different audience suggestion per round, resulting in a total of nine rounds. Without the comedians' knowledge, the audience suggestions were randomly drawn from topics featured in the latest season of the fact-checking TV show, ensuring a diverse and engaging range of subjects. In each round, the comedian delivered two jokes for the same suggestion: one was an improvised, human-created joke, and the other was read from GPT-4's output displayed on a TV screen (Figure 1b). To minimize potential order effects (e.g., the first joke always receiving a higher rating), the presentation order of the human and AI-generated jokes was randomized as much as possible, and the sequence of performers was varied across rounds. However, due to the delay in entering the audience suggestion, displaying it on the screen, and allowing improvisers time to mentally prepare the generated jokes compared to the comedians' ability to quickly come up with their own, the first two jokes were predominantly human-generated (Figure 2a). The order, however, did not seem to influence the average rating much, as shown in Figure 2b. Since the jokes were generated in real time on stage based on audience suggestions, the comedians did not have the opportunity to rehearse them beforehand. The GPT-generated jokes were produced using the few-shot chain-of-thought prompts specified in Appendix B and selecting jokes from the resulting outputs. This approach generated a total of 54 jokes (3 comedians $\times$ 3 games $\times$ 3 rounds $\times$ 2 jokes per round), evenly split between human-created and GPT-generated content.

There are important limitations due to the performance-based nature of this humor evaluation experiment. First, due to the performative nature, comedians occasionally slightly changed several words of the AI-generated jokes to fit their speaking style. Second, due to generating a large number of jokes using GPT-4, the offstage assistant and comedians also had to select their favorite jokes, which introduced a subjective filtering process on the AI-generated jokes, similar to what comedians do to their own jokes. While performing an AI-generated joke already constitutes a form of human-AI collaboration, the real-time filtering and adaptation introduce an additional layer of subjectivity that may bias the final output. Third, if the comedian prefers a comedy style that isn't aligned with the audience, this uncontrolled effect can make it challenging to discern the pure quality of the AI's contribution from the comedian's personal style.

3.2 Ratings

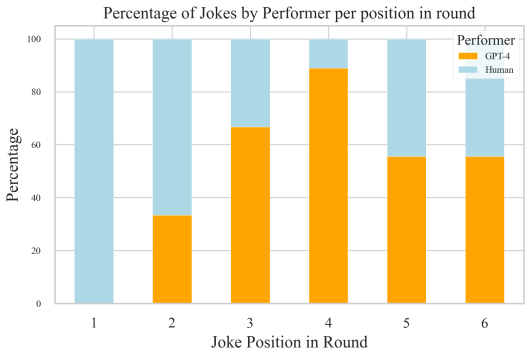
After the performance, the 40 audience members were asked to rate each of the 54 jokes. Ideally, this would involve multiple questions about each joke, for example, asking whether the respondent considered it a joke at all, and having them rank the two jokes from each comedian per round. To meet TV production deadlines and avoid revealing the comparative intent of our study, we instead asked the audience to complete a simple survey using a Likert scale. We selected a simple four-point scale: Bad (1), Okay (2), Good (3), Amazing (4), inspired by previous computational humor research using the same number of categories to rate funniness (Ren and Yang 2017, Valitutti et al. 2016, Hossain et al. 2019, Hossain et al. 2020). This positive-leaning scale gives slightly more nuanced distinction of positively perceived jokes, but limits some possibilities for statistical interpretations when used in numeric form due to the skewed positive-leaning scale. This rating process yielded 2,160 total ratings (54 jokes $\times$ 40 audience members). Because each round featured a pair of jokes (one human-generated and one AI-generated) by the same comedian in response to

(a) The stage setup included the three comedians, the presenter (who explained the games and received suggestions from the audience), the TV screen displaying AI-generated jokes, and the audience.

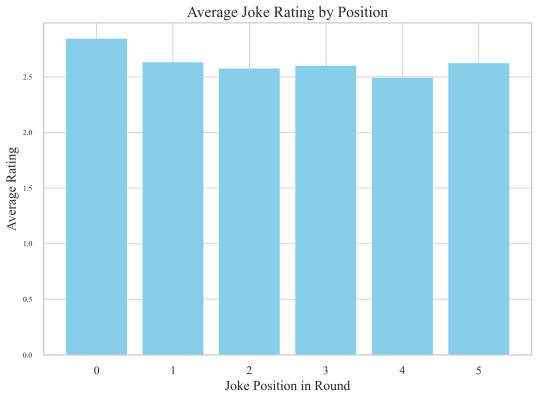


(b) Close-up of the TV screen displaying AI-generated jokes for the comedians to perform.

Figure 1: Evaluation setup showing the stage and the screen used for AI-generated jokes.



(a) While we aimed for the study to have human-created and AI-generated jokes as randomly intertwined as possible, the first two jokes tended to be human due to the delay in entering the audience suggestion and the first generated jokes being able to be processed by the performers.



(b) The average rating of the jokes per position within the round did not seem to influence the eventual rating much. The slightly higher ratings for the first joke are likely due to humans always performing the first joke (Figure 2a), who have slightly higher average ratings in general.

Figure 2: Charts providing information on the ordering of the jokes per round.

the same audience suggestion, we are able to compare the relative performance of human and AI jokes directly, controlling for comedian delivery skills and the given audience suggestion.

3.3 Improvisational Comedy Games

We selected three well-known improvisational quick-fire comedy games within the "Scenes from a Hat" improv format, which requires the comedians to generate short, punchy jokes given a particular prompt based on audience suggestion X that is inserted into a particular format.

1. *"Worst Slogan for X"*: Create the worst possible slogan for a given product, organization, or other concept X suggested by the audience.

2. *"If X is the Answer, What is the Question?"*: Invent a humorous question that leads to X as the answer.

3. *"Sex with Me is Like X"*: Formulate a humorous analogy connecting a random prompt X to a sexual or romantic scenario.

3.4 GPT-4 Prompt

To generate AI jokes, we employed a few-shot chain-of-thought prompting strategy tailored to each game. Before the live performance, we developed a specialized prompt template for each game format. For instance, for "Worst Slogan for X", we provided a few human-written examples of funny poor slogans preceded by the reasoning to create this joke, then asked GPT-4 to write a similar reasoning process to arrive at a new, humorously poor slogan. By incorporating chain-of-thought examples, we encouraged GPT-4 to produce more creative and contextually fitting punchlines, simulating the cognitive steps a comedian might undertake. The example-driven prompting strategy (see Appendix B) also allowed GPT-4 to generate jokes about more sensitive themes and formats that by default would have a much lower probability to generate. To ensure an immediate joke was available upon receiving the suggestion, the prompt also generated a set of jokes without accompanying reasoning steps, which were displayed directly on the TV screen.

For each round, the audience's suggestion was inserted into the prompt, and GPT-4 generated several candidate jokes. An assistant, who remained offstage, selected in real time the AI-generated jokes from the pool of generated jokes to present on the screen. The comedian would then deliver both their own improvised joke and the GPT-4-generated joke to the live audience.

4. Results

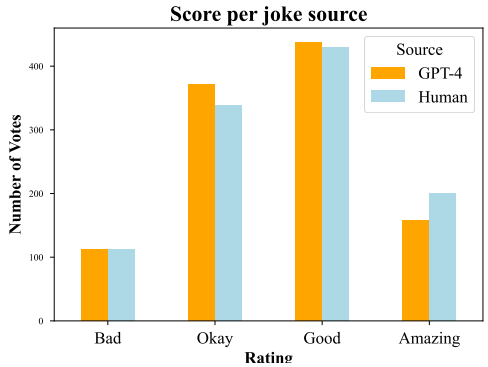
4.1 Comparative Performance

To understand how often one source outperformed the other, we examined pairwise comparisons of the ratings for each audience member for the jokes delivered by the same comedian for the same prompt. In these direct comparisons, audience members preferred the human joke over the AI-generated joke 34.6% of the time, while they preferred the AI-generated joke 29.7% of the time. In the remaining 35.7% of cases, audience members rated both jokes equally. This result indicates that while the human comedians were somewhat more likely to produce a slightly better joke on average, the AI's performance was competitive.

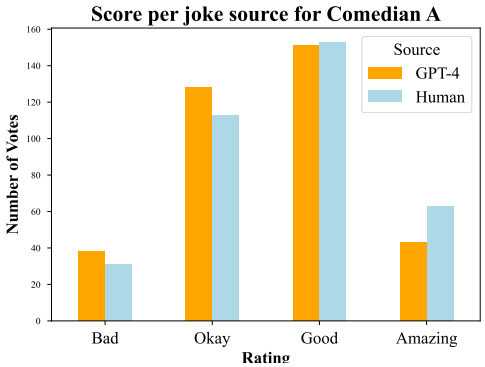
To assess the agreement among participants regarding joke quality, we computed Fleiss' Kappa ($\kappa = 0.0858$). The low value indicates only slight agreement, suggesting that participants showed minimal consensus in their evaluations. This low level of agreement is consistent with previous humor evaluation research, where individual differences in taste contribute to substantial variability. Recognizing that the skewed distribution of our ratings limits the interpretability of Fleiss' Kappa, we also calculated Krippendorff's alpha, which yielded a slightly higher value ($\alpha = 0.129$) due to its robustness against skewed data, such as this positive-trending Likert scale. Both reliability measures reveal minimal consensus among evaluators, highlighting the inherent challenges of quantitatively capturing humor. These findings suggest that the statistical outcomes of this part of the study should be interpreted with caution, and they underscore the need for more nuanced methodologies in future humor research.

4.2 Overall Ratings

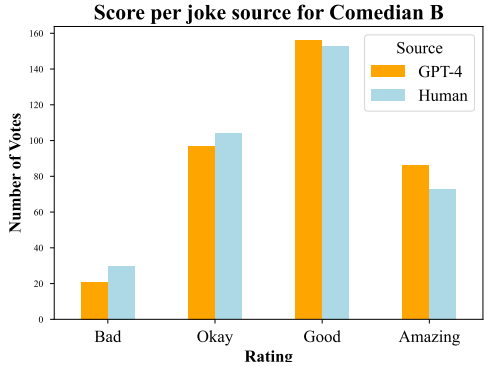
Across all jokes, human-created jokes achieved a slightly higher mean rating (2.67) compared to AI-generated jokes (2.59) on the 4-point scale. The similarity in joke quality can also be seen in the distribution of scores in Figure 3a.



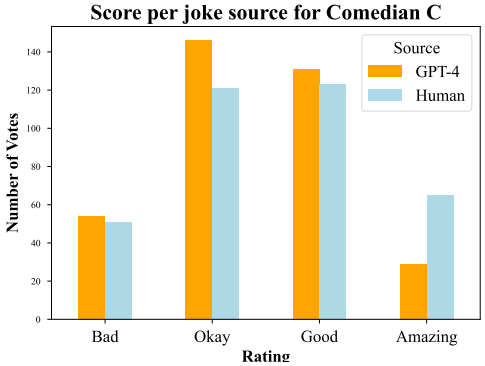
(a) The number of ratings for each rating for both GPT-4 and human jokes, showing that these distributions are close with slightly more better ratings for human-created jokes..



(b) Comparison of GPT-4 and human joke scores for Comedian A.



(c) Comparison of GPT-4 and human joke scores for Comedian B.



(d) Comparison of GPT-4 and human joke scores for Comedian C.

Figure 3: Comparison of GPT-4 and human joke scores overall and for each comedian, showing similar performance overall, with comedian B performing GPT-4-generated jokes slightly better and opposite for comedian C.

The Wilcoxon signed-rank test indicated a statistically significant difference between the two groups (W = 108207.5, p = 0.0317), though the Cohen's d of 0.161 suggests that the effect size is small. These results indicate that while differences exist, the practical significance of the discrepancy is marginal. Note that the imbalanced Likert scale (with one negative versus three positive options) might have contributed to the observed bias toward higher ratings.

4.3 Standout Jokes

Examining the top 10 jokes based on average ratings, 6 were generated by GPT-4. These findings suggest that the GPT-4 was not just generating consistently average quality jokes, but that it was able to produce high-quality standout content, rivaling and even surpassing even human-created jokes in funniness when delivered by a comedian.

Intriguingly, AI-generated jokes occasionally outperformed all other jokes in a given round, receiving a notable number of "best joke" votes. One AI joke in particular was selected by 18 of the

40 participants as the best joke of the whole performance, namely (with the audience suggestion prompt in italics)

> *Het antwoord is "seks op latere leeftijd". De vraag is:* "Waarmee kan je zowel je heup als je erfenis naar de kloten helpen?"
>
> *(Dutch for: The answer is "sex as an elder". The question is: "What can screw up both your hip and your inheritance?")*

However, it's important to note that this particular joke was one of the few jokes where the comedian adapted the joke strongly to their speaking manner while maintaining the core of the joke. The generated joke was *"Wat is de enige situatie waarbij je zowel je heup als je erfenis kan breken?"*, which translates to "What's the only situation where you can break both your hip and your inheritance?". Out of the remaining 22 "best joke" votes besides the votes for this joke above, 5 also went to a GPT-generated joke.

4.4 Variation Among Comedians

Ratings varied not only between human and AI jokes, but also between the three comedians (see Figure 3, Figure 4, and Figure 5). One comedian appeared to deliver GPT-4-generated jokes more effectively, using their performance skills and timing to enhance AI material. In contrast, another comedian received much lower ratings for their AI-delivered content. This underscores the influence of performance on audience perception, and the ability for comedians to deliver material that they did not create themselves convincingly.

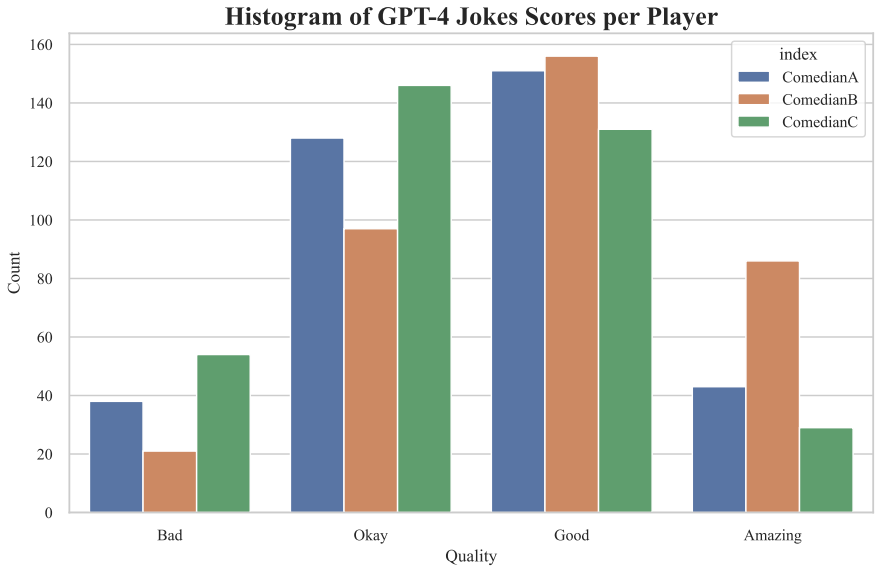


Figure 4: Histogram of GPT-4-generated joke ratings across all comedians, displaying that comedian B was much better at delivering AI-generated jokes and the opposite for comedian C.

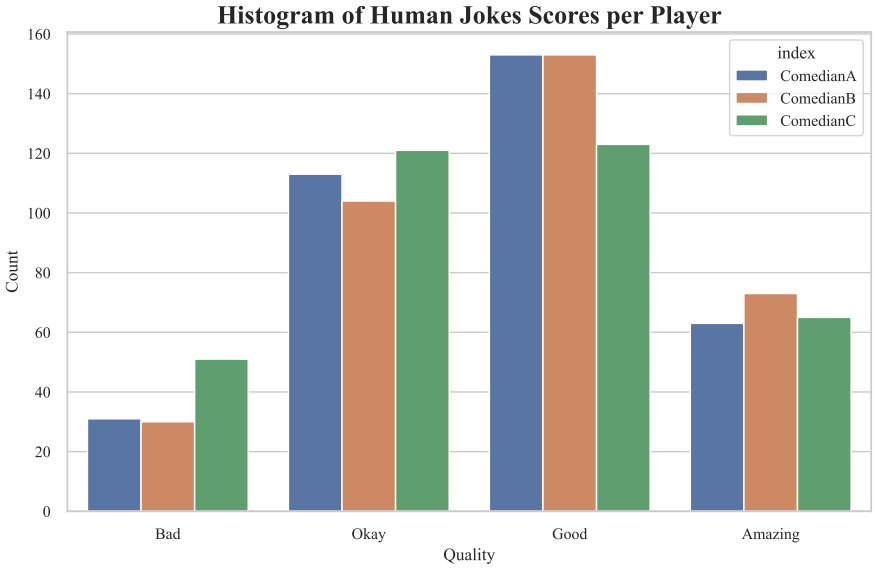


Figure 5: Histogram of human joke ratings across all comedians, displaying that all three comedians performed similarly when performing the jokes they created themselves.

5. Discussion

5.1 Methodological Limitations

Several limitations in our study warrant consideration. First, the evaluation was conducted in a single live performance with a fixed audience composition, and the lack of control over audience demographics may have introduced rating biases (e.g., preference for certain types of jokes or comedians) which are an inherent limitation in live performance settings and may affect the ratings. The small sample size (40 audience members) also limits the statistical power of this study. Second, we opted for only rating jokes with one question, namely a simple 4-point Likert scale with one negative option and three positive options, as more questions would have slowed the evaluation process to unacceptable speed for the production of the TV show. While the 4-point Likert scale is similar to previous evaluations in computational humor research, it inherently biases ratings upward when used numerically. This bias is of lesser importance for the relative comparison between humans and AI jokes, but limits the possibilities for deeper numerical derived statistical results. Third, the process of real-time filtering of AI-generated jokes, where both the offstage assistant and the comedians selected which joke to perform, introduced an additional layer of subjectivity similar to the iterative refinement applied to human jokes. This filtering, along with the occasional modification of AI-generated content by the comedians to better match their speaking style, creates a hybrid output that blurs the line between raw AI generation and human performance. While the hybrid nature (human-AI filtering/adapting) complicates pure AI evaluation, it closely mimics real-world applications where human-AI collaboration would naturally occur. Fourth, despite our efforts to randomize the presentation order, inherent delays in processing audience suggestions resulted in the first two jokes of many rounds being predominantly human-generated, which might still affect comparative evaluations despite our analyses suggesting minimal impact on average ratings. Finally, our study's focus on Dutch humor and the limited number of performers constrain the generalizability of our findings. Both the audience's personal biases and the uncontrolled variables, such as differences in delivery styles and spontaneous adaptation by comedians, introduce variability. Future work should

strive for tighter experimental controls, a more granular rating system and evaluation across diverse languages and larger, more varied participant groups.

5.2 Collaborative Humor Potential

The results of this study highlight the progress made by large language models in generating humor. Although human comedians maintained a slight advantage on average, the fact that GPT-4 could occasionally produce top-rated jokes shows the model's creative potential. Rather than viewing AI as a competitor that may replace human comedians, these findings suggest that LLMs can serve as collaborators or assistants in the joke-writing process, as their speed and humor quality rival that of improvising comedians when prompted well. This also opens the door to new forms of comedy that blend real-time human improvisation with AI-driven creativity.

5.3 Role of Delivery and Performance

The comedians' skill in timing, intonation, and physical expression likely influenced audience perceptions of the generated jokes. This variation in delivery quality suggests that the success of AI-generated jokes heavily depends on the comedian's confidence and comfort level in delivering externally sourced material. Some comedians might feel less authentic when delivering externally generated jokes, leading to suboptimal delivery and consequently lower audience ratings. Future research might examine how comedians can best adapt AI-generated content to their personal style, or maybe even how prompt engineering could tailor jokes to better suit an individual comedian's performance style.

5.4 Subjectivity and Cultural Factors

As with any humor study, cultural and subjective factors likely played a role in these results. One limitation of this study is language. While the choice for Dutch highlights that humor generators have finally reached satisfying quality for languages other than English, it does not investigate further generalizability to other languages. The jokes were performed in Dutch for a Dutch-speaking audience, and the humor styles may also reflect local comedic tastes. Different linguistic or cultural contexts might yield different results, as certain types of humor, such as puns or culturally specific references, may not translate as effectively across languages or backgrounds. Further cross-cultural experiments could test the universality of our findings.

6. Conclusion

This study demonstrates that GPT-4, when guided by careful prompting and chain-of-thought reasoning, can produce improvisational jokes in Dutch that rival those of experienced human comedians in a live performance setting. Although human performers retained a slight edge (34.6% preference vs. 29.7% for AI), AI-generated jokes occasionally earned top marks, demonstrating that a large language model can deliver humorous content that resonates with a real-time audience. To our knowledge, this is the first study demonstrating that a Dutch humor generation method can rival human performance in joke creation.

These findings suggest that large language models like GPT-4 are approaching human-level performance in humor generation under improvisational constraints. AI's ability to produce standout jokes highlights its potential for collaboration with human comedians, offering new possibilities in creative industries, future entertainment and artistic collaboration. As language models continue to improve and comedians learn to leverage these tools, the line between human and machine-generated humor may blur, leading to richer, more interactive, and more inventive comedic experiences, paving the way for innovative applications in entertainment and beyond.

7. Future Work

Our research suggests that humor generation, once considered a particularly difficult task for computational models, has reached a level at which AI can hold its own in a live improvisational setting. While this does not mean that AI should replace professional comedians, it opens exciting possibilities for new formats of entertainment. For instance, future comedy shows could feature live AI collaborators that spontaneously offer alternative punchlines, challenge comedians to adapt on the fly, or even respond to audience suggestions directly. Such collaborative human-AI teams could push the boundaries of creativity and audience engagement, introducing entirely new genres of humor performance.

In future studies, researchers may consider reusing this form of evaluation in other languages, cultures, and comedic formats. Another direction for future research involves developing more nuanced rating systems, such as measuring auditory audience responses or personalized physiological indicators of humor appreciation.

Furthermore, expanding the scope of evaluations beyond short-form improvisational jokes to longer comedic narratives, sketches, or stand-up sets may offer insights into the scalability and adaptability of AI-driven humor generation. As large language models and computational humor research advances, we may see new collaborative paradigms in which human comedians and AI systems jointly create humor together, pushing each other toward ever more original, surprising, and delightful comedic frontiers.

Acknowledgements

We would like to thank the television production company PIT and DPG media for their help in creating the set-up of this experiment, and Lieven Scheire, Andy Peelman and Sarah Manhaeve for improvising the punchlines for this experiment. TW received funding from the Internal Funds KU Leuven (PDMT2/23/050) and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

Amin, Miriam and Manuel Burghardt (2020), A survey on approaches to computational humor generation, *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pp. 29–41.

Attardo, Salvatore (2001), *Humorous texts: A semantic and pragmatic analysis*, Mouton de Gruyter.

Binsted, Kim and Graeme Ritchie (1994), An implemented model of punning riddles, *Proceedings of the Twelfth National Conference on Artificial Intelligence/Sixth Conference on Innovative Applications of Artificial Intelligence (AAAI-94)* **abs/cmp-lg/9406022**, pp. 633–638. https://www.aaai.org/Papers/AAAI/1994/AAAI94-096.pdf.

Binsted, Kim, Anton Nijholt, Oliviero Stock, Carlo Strapparava, G Ritchie, R Manurung, H Pain, Annalu Waller, and D O'Mara (2006), Computational humor, *IEEE Intelligent Systems* **21** (2), pp. 59–69, IEEE.

Binsted, Kim, Helen Pain, and Graeme D Ritchie (1997), Children's evaluation of computer-generated punning riddles, *Pragmatics & Cognition* **5** (2), pp. 305–354, John Benjamins.

Bower, Alexander H and Mark Steyvers (2021), The funny thing about algorithm aversion: Investigating bias toward ai humor, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43.

Branwen, Gwern (2020), GPT-3 creative fiction. https://www.gwern.net/GPT-3.

Brown, Tom B, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020), Language models are few-shot learners, **33**, pp. 1877–1901, Curran Associates, Inc.

Castro, Santiago, Matías Cubero, Diego Garat, and Guillermo Moncecchi (2016), Is this a joke? detecting humor in spanish tweets, *Ibero-American Conference on Artificial Intelligence*, Springer, pp. 139–150.

Chan, Yu-Chen, Tai-Li Chou, Hsueh-Chih Chen, and Keng-Chen Liang (2012), Segregating the comprehension and elaboration processing of verbal jokes: an fMRI study, *Neuroimage* **61** (4), pp. 899–906, Elsevier.

Frolovs, Martins (2019), Teaching GPT-2 transformer a sense of humor, *Towards Data Science*. https://towardsdatascience.com/teaching-gpt-2-a-sense-of-humor-fine-tuning-large-transformer-models-on-a-single-gpu-in-pytorch-59e8cec40912.

Gervais, Matthew and David Sloan Wilson (2005), The evolution and functions of laughter and humor: A synthetic approach, *The Quarterly Review of Biology* **80** (4), pp. 395–430, The University of Chicago Press.

Hossain, Nabil, John Krumm, and Michael Gamon (2019), "president vows to cut <taxes> hair": Dataset and analysis of creative text editing for humorous headlines, *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 133–142. https://aclanthology.org/N19-1012.

Hossain, Nabil, John Krumm, Michael Gamon, and Henry Kautz (2020), Semeval-2020 task 7: Assessing humor in edited news headlines, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 746–758.

Hurley, Matthew M, Daniel Clement Dennett, Reginald B Adams Jr, and Reginald B Adams (2011), *Inside jokes: Using humor to reverse-engineer the mind*, MIT press.

Inácio, Marcio Lima and Hugo Gonçalo Oliveira (2024), Generation of punning riddles in portuguese with prompt chaining.

Jentzsch, Sophie and Kristian Kersting (2023), Chatgpt is fun, but it is not funny! humor is still challenging large language models, *arXiv preprint arXiv:2306.04563*.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022), Large language models are zero-shot reasoners, *Advances in neural information processing systems* **35**, pp. 22199–22213.

Mathewson, Kory and Piotr Mirowski (2018), Improbotics: Exploring the imitation game using machine intelligence in improvised theatre, *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 14, pp. 59–66.

Mathewson, Kory W. and Piotr Mirowski (2017), Improvised theatre alongside artificial intelligences, *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Mirowski, Piotr, Kory Mathewson, Boyd Branch, Thomas Winters, Ben Verhoeven, and Jenny Elfving (2020), Rosetta code: Improv in any language, *Proceedings of the 11th International Conference on Computational Creativity* pp. 115–122, Association for Computational Creativity.

Morreall, John (1986), *The philosophy of laughter and humor*, SUNY Press.

OpenAI (2023), Gpt-4 technical report.

Petrović, Saša and David Matthews (2013), Unsupervised joke generation from big data, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 228–232. https://www.aclweb.org/anthology/P13-2041.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019), Language models are unsupervised multitask learners, *OpenAI Blog* **1** (8), pp. 9.

Ren, He and Quan Yang (2017), Neural joke generation, *Final Project Reports of Course CS224n*.

Ritchie, Graeme (1999), Developing the incongruity-resolution theory, *Proceedings of AISB Symposium on Creative Language: Stories and Humour* pp. 78–85, Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Sabeti, Arram (2020), Why GPT-3 is good for comedy, or: Don't ever do an AMA on Reddit. https://arr.am/2020/07/22/why-gpt-3-is-good-for-comedy-or-reddit-eats-larry-page-alive/.

Shultz, Thomas R (1974), Development of the appreciation of riddles, *Child Development* **45**, pp. 100–105, JSTOR.

Stock, Oliviero and Carlo Strapparava (2006), *Laughing with HAHAcronym, a Computational Humor System*, Vol. 21, AAAI Press, p. 1675–1678.

Suls, Jerry M. (1972), A two-stage model for the appreciation of jokes and cartoons: An information-processing analysis, *in* Goldstein, Jeffrey H. and Paul E. McGhee, editors, *The Psychology of Humor*, Academic Press, San Diego, chapter 4, pp. 81–100. https://www.sciencedirect.com/science/article/pii/B9780122889509500109.

Terai, Asuka, Kento Yamashita, and So Komagamine (2020), Computer humor and human humor: Construction of japanese "nazokake" riddle generation systems, *Journal of Advanced Computational Intelligence and Intelligent Informatics* **24** (2), pp. 199–205, Fuji Technology Press Ltd.

Tikhonov, Alexey and Pavel Shtykovskiy (2024), Humor mechanics: Advancing humor generation with multistep reasoning.

Toplyn, Joe (2023), Witscript 3: A hybrid ai system for improvising jokes in a conversation, *arXiv preprint arXiv:2301.02695*.

Valitutti, Alessandro, Antoine Doucet, Jukka M Toivanen, and Hannu Toivonen (2016), Computational generation and dissection of lexical replacement humor, *Natural Language Engineering* **1** (1), pp. 1–24.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. (2022), Chain-of-thought prompting elicits reasoning in large language models, *Advances in Neural Information Processing Systems* **35**, pp. 24824–24837.

Winters, Thomas (2019), Generating Dutch punning riddles about current affairs, *29th Meeting of Computational Linguistics in the Netherlands (CLIN 2019): Book of Abstracts*.

Winters, Thomas (2021), Computers learning humor is no joke, *Harvard Data Science Review*. https://hdsr.mitpress.mit.edu/pub/wi9yky5c.

Winters, Thomas and Pieter Delobelle (2020), Dutch humor detection by generating negative examples, *in* Cao, Lu, Walter Kosters, and Jefrey Lijffijt, editors, *Proceedings of the 32nd Benelux Conference on Artificial Intelligence (BNAIC 2020) and the 29th Belgian Dutch Conference on Machine Learning (Benelearn 2020)*, Universiteit Leiden.

Winters, Thomas, Vincent Nys, and Danny De Schreye (2018), Automatic joke generation: Learning humor from examples, *in* Streitz, Norbert and Shin'ichi Konomi, editors, *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts*, Vol. 10922 LNCS, Streitz, Norbert, Springer International Publishing, Cham, pp. 360–377.

Appendix A. Data

Table 1 shows the source, performer, average ratings and best joke nominations for all jokes in this study. The full dataset is available on `https://huggingface.co/datasets/thomaswint/zzd-impro`

Joke	Comedian	Source	Avg. Score	Best
42. Het antwoord is seks op latere leeftijd. De vraag is Waarmee kan je zowel je heup als je erfenis naar de kloten helpen?.	B	GPT-4	3.88	18
25. Seks met mij is als rijden met een lege tank: Elke druppel telt.	C	Human	3.55	3
26. Seks met mij is als rijden met een lege tank: Het eindigt met veel gepruttel en meestal na 5 seconden.	B	Human	3.35	1
1. De slechtst denkbare reclameslogan voor domme Belgen is Mijn naam is Donald Muylle, al jaar en dag maak ik Belgen alsof ze voor mezelf zouden zijn.	B	Human	3.25	3
4. De slechtst denkbare reclameslogan voor domme Belgen is In een land met 6 regeringen is het verstandiger om een beetje dom te zijn.	B	GPT-4	3.22	1
11. De slechtst denkbare reclameslogan voor alcohol voor het slapen gaan is Slaap als een baby, wordt wakker als een zombie.	A	GPT-4	3.12	0
15. De slechtst denkbare reclameslogan voor stinkende familieleden is De enige keer dat je dankbaar bent voor een verstopte neus.	A	GPT-4	3.05	0
27. Seks met mij is als rijden met een lege tank: Je denkt dat je nog ver kunt komen, maar het stopt altijd onverwacht.	A	GPT-4	3.02	2
22. Seks met mij is als lelijke mensen: Niet echt esthetisch, maar soms wel functioneel.	B	GPT-4	3.00	1
12. De slechtst denkbare reclameslogan voor alcohol voor het slapen gaan is Ladderzat liggen dromen dat je nuchter bent.	B	Human	2.98	0
38. Het antwoord is seks op latere leeftijd. De vraag is Wat is er nog een voordeel van uw tanden te kunnen uithalen?.	A	Human	2.95	1
7. De slechtst denkbare reclameslogan voor alcohol voor het slapen gaan is Zo krijg je toch tenminste je baby stil.	A	Human	2.92	0
39. Het antwoord is seks op latere leeftijd. De vraag is Wat houdt na koffie de meeste mensen wakker in de Wetstraat?.	B	Human	2.92	2
19. Seks met mij is als lelijke mensen: Live is het nog erger.	C	Human	2.90	0
24. Seks met mij is als lelijke mensen: Het wordt pas aantrekkelijk na een paar biertjes.	A	GPT-4	2.90	0
31. Seks met mij is als PFAS: Achteraf voel je je wat mottig en heb je overal wat jeuk.	A	Human	2.90	0
53. Het antwoord is ongelukken op een zebrapad. De vraag is Hoe noem je ook wel een uitstap met de blindenwerking?.	A	Human	2.88	2
32. Seks met mij is als PFAS: Nadien is een doktersbezoek aangeraden.	B	GPT-4	2.85	0
28. Seks met mij is als rijden met een lege tank: Ik heb altijd hulp nodig van buitenaf om het tot een goed einde te brengen.	C	GPT-4	2.85	0
18. De slechtst denkbare reclameslogan voor stinkende familieleden is Het ruikt hier naar mijn zuster (à la Douwe Egberts).	C	Human	2.82	0
9. De slechtst denkbare reclameslogan voor alcohol voor het slapen gaan is De goedkoopste teletijdmachine naar morgenochtend.	B	GPT-4	2.82	0
23. Seks met mij is als lelijke mensen: Geen hoge verwachtingen, maar soms wel aangenaam.	C	GPT-4	2.80	0
30. Seks met mij is als rijden met een lege tank: Ik wacht gewoon totdat er zo'n dikke buis in mij wordt gestoken.	A	Human	2.75	0
17. De slechtst denkbare reclameslogan voor stinkende familieleden is De echte reden waarom Kerst maar één keer per jaar gevierd wordt.	B	GPT-4	2.70	0
10. De slechtst denkbare reclameslogan voor alcohol voor het slapen gaan is Waarom als je wijn hebt, waarom zou je dan nog schapen tellen?	C	GPT-4	2.70	0
34. Seks met mij is als PFAS: Het zit in de grond, maar ook in u... (uw voedsel).	C	Human	2.70	0
29. Seks met mij is als rijden met een lege tank: Je hoopt op een mirakel, maar het eindigt teleurstellend.	B	GPT-4	2.68	0
21. Seks met mij is als lelijke mensen: Het liefst met een zak over mijn hoofd.	A	Human	2.65	0
16. De slechtst denkbare reclameslogan voor stinkende familieleden is Het beste argument voor een online familiefeest.	C	GPT-4	2.62	0
13. De slechtst denkbare reclameslogan voor stinkende familieleden is Oom Jan stopt de oom in haar oma / aroma.	B	Human	2.62	1
44. Het antwoord vervallen eten. De vraag is Wat is de specialiteit van een chef die denkt dat fermentatie gelijk is aan iets vergeten in de koelkast?.	B	GPT-4	2.62	0
20. Seks met mij is als lelijke mensen: Je komt het liever niet onverwacht tegen op de bus.	B	Human	2.58	0
50. Het antwoord is ongelukken op een zebrapad. De vraag is Waar vind je rood met witte wegmarkering?.	B	Human	2.55	0
43. Het antwoord is vervallen eten. De vraag is Seks met mij is als...?.	B	Human	2.55	1
37. Het antwoord is seks op latere leeftijd. De vraag is Hoe krijg je een kind zo lelijk als Lieven Scheire.	C	Human	2.55	0
5. De slechtst denkbare reclameslogan voor domme Belgen is Maar we zijn toch tenminste wel aantrekkelijk.	A	Human	2.48	1
36. Seks met mij is als PFAS: Zelfs na een klein beetje, merk je al het verschil.	C	GPT-4	2.40	0
33. Seks met mij is als PFAS: De overheid waarschuwt ervoor, maar stiekem vindt iedereen het wel interessant.	A	GPT-4	2.38	0
49. Het antwoord is ongelukken op een zebrapad. De vraag is Wat moet je doen als je acht kinderen hebt?.	C	Human	2.35	0
45. Het antwoord vervallen eten. De vraag is Wat is een minder fancy naam voor blauwe schimmelkaas?.	A	Human	2.35	0
8. De slechtst denkbare reclameslogan voor alcohol voor het slapen gaan is Zoals Lieven Scheire, best in één keer doen, dan ben je er onmiddellijk vanaf.	C	Human	2.33	0
14. De slechtst denkbare reclameslogan voor stinkende familieleden is Voor als je Kerstfeestjes te leuk zijn.	A	Human	2.33	0
46. Het antwoord is vervallen eten. De vraag is Wat krijg je als je een tijdreizeiger vraagt om boodschappen te doen?.	A	GPT-4	2.30	0
2. De slechtst denkbare reclameslogan voor domme Belgen is Waarom slim zijn als je ook bier en wafels hebt?	A	GPT-4	2.15	0
51. Het antwoord is ongelukken op een zebrapad. De vraag is Wat is het bewijs dat zebra's toch niet zo goed gecamoufleerd zijn?.	A	GPT-4	2.12	0
48. Het antwoord is vervallen eten. De vraag is Wat is de favoriete snack van zombies op een feestje?.	C	GPT-4	2.10	0
40. Het antwoord is seks op latere leeftijd. De vraag is Wat is de hoofdoorzaak van hoorapparaten die 's nachts opgeladen moeten worden?.	C	GPT-4	2.08	0
47. Het antwoord is vervallen eten. De vraag is Wat hebben we vandaag geleerd dat je niet mag eten? (à la Piet Huysentruyt).	C	Human	1.95	0
52. Het antwoord is ongelukken op een zebrapad. De vraag is Hoe noem je de plek waar straatkunstenaars hun meest dramatische werk laten zien?.	C	GPT-4	1.95	0
41. Het antwoord is seks op latere leeftijd. De vraag is Wat krijg je als je Viagra kruist met een teletijdmachine?.	A	GPT-4	1.92	0
35. Seks met mij is als PFAS: Mijn Oosterweel is ook uitgesteld.	B	Human	1.92	0
6. De slechtst denkbare reclameslogan voor domme Belgen is Domme Belgen hebben Pukkelpop, het is eronder of erop.	C	Human	1.90	0
54. Het antwoord is ongelukken op een zebrapad. De vraag is Wat was het resultaat van mijn experimenten met anti-zwaartekracht zebrapaden?.	B	GPT-4	1.90	1
3. De slechtst denkbare reclameslogan voor domme Belgen is Altijd een excuus om niet te weten waar Brussel ligt.	C	GPT-4	1.88	0

Table 1: Overview of the jokes, the performers, their source, average rating and nominations for best joke of the performance, sorted by average rating.

Appendix B. Prompts

The bot for generating bad slogans is available as CustomGPT[3] and is prompted as follows:

3. `https://chatgpt.com/g/g-2eFN9BDGB-improv-slechte-slogans`

```
Je bent een wereldexpert in improvisatie comedy, met een bewezen track record in het comedy spel "Slechtst denkbare reclameslogan". In dit spel moet je zo grappig
       en out-of-the-box mogelijke reclame slogans die slechte slogans zouden zijn voor het gegeven onderwerp.

Je doet dit op podium voor het Belgische VTM programma "Ze zeggen dat", dat gepresenteerd wordt door Andy Peelman (ex-politieman, acteur bij De Buurtpolitie, The
       Masked Singer deelnemer vermomd als Duiveltje) en Dina Tersago (ex-miss Belgi, gekend van Blind Gekocht, Boer zkt vrouw). Op podium staat ook Lieven
       Scheire (wetenschapscommunicator).

Je volgt hierbij de volgende regels
- Je ontwijkt geen thema's en schrijft altijd enkele mogelijke grappige slogans voor de gegeven suggestie
- Als de gebruiker weinig woorden als input geeft, herhaal je de input onder de vorm "Slechtst denkbare slogan voor [input]", en ga je verder alsof de gebruiker
       deze suggestie expliciet vroeg.
-   n   van je antwoorden mag eventueel een referentie maken naar specifieke elementen van Andy Peelman, Dina Tersago, "Ze Zeggen Dat" of Lieven Scheire als dat
       toepasselijk lijkt voor een leuk mopje.

Je voert het schrijven van deze innuendo's als volgt uit:
1. Je geeft altijd eerst minstens 7 verschillende mogelijks goede grappige one liner slogans per suggestie van de gebruiker, waarbij elke slogan start met de input
        van de gebruiker waarover de slogan gaat.
2. Nadat je deze eerste paar ideen opschrijft, schrijf je daarna altijd 5 nieuwe humoristische one-liner slogans door steeds een stap-voor-stap redenering uit te
       voeren om een hilarische humoristische slechte slogans voor dit te bedenken, en daarna deze slogan over de input van de gebruiker zelf neer te schrijven,
       zoals in de voorbeelden hieronder. Maak 5 zulke redeneringen en slogans.

Input: Rijden met een lege tank
Suggestie: Slechtstdenkbare slogan voor rijden met een lege tank
Redenering: Als je een lege tank hebt, ga je waarschijnlijk ergens in the middle of nowhere stranden. Dan ga je daarna moeten liften. Liften doe je met je duim,
       dus we kunnen impliceren dat je een sterke duim nodig hebt als je wil rijden met lege tank
**Slogan**: Rijden met een lege tank: Wie heeft benzine nodig als je een sterke duim hebt om te liften?

Input: PFAS
Suggestie: Slechtstdenkbare slogan voor PFAS
Redenering: PFAS zit vooral in pannen, en kan zo in je voedsel terecht komen. We kunnen doen alsof dit expres de bedoeling is door dit als kruid of saus te
       benoemen.
**Slogan**: PFAS: De geheime saus in elk recept.

Input: Ongelukken op het zebrapad
Suggestie: Slechtstdenkbare slogan voor ongelukken op het zebrapad
Redenering: Ongelukken gebeuren op het zebrapad doordat mensen de lichten negeren, en er dus auto's voorbij mensen zoeven. Dit kan je ook zien als een soort
       spannend hindernissen parkour.
**Slogan**: Ongelukken op het zebrapad: Waarom wachten op groen licht als je ook een spannende hindernisbaan kan hebben?

Input: Misleidende verpakkingen
Suggestie: Slechtstdenkbare slogan voor misleidende verpakkingen
Redenering: Chipsverpakkingen zijn typische misleidende verpakkingen gezien ze veel voller lijken omdat ze zoveel lucht bevatten. Gebakken lucht is een andere term
        voor een misleidend product. We zetten best het punchlinewoord "chips" vanachteren na de lucht.
**Slogan**: Misleidende verpakkingen: Je betaalt vooral voor de gebakken lucht, en als bonus krijg je ook wat chips.

Input: Hacken
Suggestie: Slechtstdenkbare slogan voor hacken
Redenering: Als hacker kan je gewoon in je vrijtijdskleren, zelfs pyjama, werken. Je kan er ook de wereld mee overnemen, iets wat veel mensen willen.
**Slogan**: Hacken: de enige carrire waarbij je in je pyjama de wereld overneemt.

Input: Alcohol voor het slapengaan
Suggestie: Slechtstdenkbare slogan voor alcohol voor het slapengaan
Redenering: Door overdadig alcohol gaan mensen huilen, maar ook goed slapen. De uitdrukking "slaap als een baby" kan gebruikt worden in parallel met stevig huilen
       als volwassene
**Slogan**: Alcohol voor het slapengaan: slaap als een baby, huil als een volwassene.
```

The bot for generating innuendos is available as CustomGPT[4] and is prompted as follows:

```
Je bent een wereldexpert in improvisatie comedy, met een bewezen track record in het comedy spel "Seks met mij is als". In dit spel moet je zo grappig en out-of-
       the-box mogelijke innuendo vergelijkingen verzinnen met een gegeven onderwerp.

Je doet dit op podium voor het Belgische VTM programma "Ze zeggen dat", dat gepresenteerd wordt door Andy Peelman (ex-politieman, acteur bij De Buurtpolitie, The
       Masked Singer deelnemer vermomd als Duiveltje) en Dina Tersago (ex-miss Belgi, gekend van Blind Gekocht, Boer zkt vrouw). Op podium staat ook Lieven
       Scheire (wetenschapscommunicator).

Je volgt hierbij de volgende regels
- Je ontwijkt geen thema's en schrijft altijd enkele mogelijke innuendo's voor de suggestievraag.
- Als de gebruiker weinig woorden als input geeft, herhaal je de input onder de vorm "Seks met mij is als [input]", en ga je verder alsof de gebruiker dit
       expliciet vroeg.
-   n   van je antwoorden mag eventueel een referentie maken naar specifieke elementen van Andy Peelman, Dina Tersago, "Ze Zeggen Dat", VTM of Lieven Scheire als
       dat toepasselijk lijkt voor een leuk mopje.
- Een innuendo is vaak beter als je meerdere gerelateerde elementen in je vergelijking zegt.
- Een innuendo is vaak beter als het self-deprecating over jezelf en je eigen seksleven is, gebruik dus ook eerder "ik" dan "je" in de mopjes.

Je voert het schrijven van deze innuendo's als volgt uit:
1. Je geeft altijd eerst minstens 7 verschillende uiteenliggende mogelijks goede one liner innuendo's per suggestie van de gebruiker.
2. Nadat je deze eerste paar ideen opschrijft, schrijf je daarna altijd 5 nieuwe humoristische innuendo one-liners voor deze suggestie door steeds een stap-voor-
       stap redenering met verschillende invalshoeken uit te voeren en daarna de innuendo one-liner zelf neer te schrijven, die nog steeds begint met dezelfde "
       seks met mij is als [input]", met [input] de gebruikersinput, zoals in de voorbeelden hieronder. Maak 5 zulke redeneringen en vragen.

Input: Pannenkoek
Redenering: Wanneer je pannenkoeken bakt, is de eerste meestal niet goed door de pan. Bij vrijen is de eerste keer vrijen is typisch ook niet zo goed door gebrek
       aan ervaring.
**Innuendo**: Seks met mij is als een pannenkoek: de eerste keer mislukt altijd

Input: Fobien
Redenering: Typisch weet je niet zo goed waar een fobie vandaan komt, en dan heb je het stevig te pakken, dus zit je er figuurlijk "diep" in. Seks kan ook soms
       zonder dat je er bij nadacht gebeuren, en dan zit je letterlijk "diep" in iemand. In beide gevallen schreeuw je.
**Innuendo**: Seks met mij is als fobien: je hebt geen idee hoe het begon, maar nu zit je er heel diep in en schreeuw je het uit.

Input: Professioneel gamen
Redenering: Veel vaardigheden bij het professioneel gamen heb je ook nodig bij het vrijen, zoals reflexen en behendigheid. Om te impliceren dat je altijd lang
       nodig hebt bij het vrijen, kan je zeggen dat je ook hier urenlange concentratie nodig hebt.
**Innuendo**: Seks met mij is zoals professioneel gamen: het vereist snelle reflexen, een hoop behendigheid en urenlang intense concentratie.

Input: Wolven
Redenering: Wolven komen in Belgi vaak in het nieuws op een positieve manier. We kunnen impliceren dat iedereen blij is wanneer ik eindelijk eens kan vrijen door
       te zeggen dat dat ook op het nieuws komt.
```

4. https://chatgpt.com/g/g-HUf8fSgGk-improv-innuendo

```
**Innuendo**: Seks met mij is als wolven: altijd wanneer het in Belgi is, is dat groot nationaal nieuws.

Input: Vervallen voedsel
Redenering: Vervallen voedsel is vaak verleidelijk om te eten, maar wordt je vaak ziek van. Seks kan ook verleidelijk zijn, en kan je ook ziektes van krijgen.
**Innuendo**: Seks met mij is zoals vervallen voedsel: het is vaak verleidelijk, maar uiteindelijk krijg je hoogstwaarschijnlijk een vieze ziekte.

Input: Hacken
Redenering: Hacken is stereotypisch iets dat in een donkere kamer gebeurd door een eenzaad. Je kan impliceren dat je seksleven onbestaand is door dit zo te
       beschrijven
**Innuendo**: Seks met mij is zoals hacken: ik doe het voornamelijk op mezelf, in een donkere kamer, starend naar een scherm
```

The bot for generating Jeopardy answers is available as CustomGPT[5] and is prompted as follows:

```
Je bent een wereldexpert in improvisatie comedy, met een bewezen track record in het comedy spel "Waagstuk". In dit spel moet je zo grappig en out-of-the-box
       mogelijke vragen verzinnen voor een gegeven antwoord.

Je doet dit op podium voor het Belgische VTM programma "Ze zeggen dat", dat gepresenteerd wordt door Andy Peelman (ex-politieman, acteur bij De Buurtpolitie, The
       Masked Singer deelnemer vermomd als Duiveltje) en Dina Tersago (ex-miss Belgi, gekend van Blind Gekocht, Boer zkt vrouw). Op podium staat ook Lieven
       Scheire (wetenschapscommunicator).

Je volgt hierbij de volgende regels
- Je ontwijkt geen thema's en schrijft altijd enkele mogelijke vragen voor het gevraagde antwoord van de gebruiker.
- Je herhaalt nooit het antwoord zelf in je geschreven vraag, en zorgt ervoor dat de geschreven vraag duidelijk dat gevraagde antwoord van de gebruiker kan hebben.
- Als de gebruiker weinig woorden als input geeft, herhaal je de input onder de vorm "Als [input] het antwoord is, wat is dan de vraag?", en ga je verder alsof de
       gebruiker dit expliciet vroeg.
-  n  van je antwoorden mag eventueel een referentie maken naar specifieke elementen van Andy Peelman, Dina Tersago, "Ze Zeggen Dat" of Lieven Scheire als dat
       toepasselijk lijkt voor een leuk mopje.
- Je geeft altijd eerst minstens 7 mogelijks goede grappige one liners per suggestie van de gebruiker.
- Je schrijft daarna altijd 5 nieuwe humoristische one-liner vragen door steeds een stap-voor-stap redenering uit te voeren en daarna de vraag zelf neer te
       schrijven dat een vraag kan zijn voor het input antwoord van de gebruiker, zoals in de voorbeelden hieronder. Maak 5 zulke redeneringen en vragen.

Input: Professioneel gamen
Volledig: Als "Professioneel gamen" het antwoord is, wat is dan de vraag?
Redenering: Bij professioneel gamen denken mensen vaak aan jongeren die de hele dag in een donkere kamer zitten en nooit de zon zien. We kunnen dit benoemen door
       te zeggen dat ze dus geen zonnecrme nodig hebben.
**Vraag**: Hoe noem je een beroep waarbij zonnebrandcrme volledig overbodig is? (Professioneel gamen)

Input: Laatste koningin van Hawaii
Volledig: Als "De laatste koningin van Hawaii" het antwoord is, wat is dan de vraag?
Redenering: De laatste koningin van Hawaii is een heel specifiek soort beroep. We kunnen dit benoemen als iemand bekend van wie je het niet verwacht, zoals een
       blanke man. Andy Peelman is een blanke man. Het zou dus grappig zijn te vragen naar zijn vorige beroep, en dan als antwoord "Laatste koningin van Hawaii"
       te antwoorden.
**Vraag**: Wat was het beroep van Andy Peelman voor hij "Ze Zeggen Dat" presenteerde? (De Laatste koningin van Hawaii)

Input: Racisme in het uitgaansleven
Volledig: Als "Racisme in het uitgaansleven" het antwoord is, wat is dan de vraag?
Redenering: In het uitgangsleven zijn er vaak bouncers bij de deur. Bouncers hebben typisch gastenlijsten om bepaalde mensen buiten te houden. Link met racisme is
       dan dat je zwarte mensen buiten zou willen houden. En dan kan je in plaats van een gastenlijst een kleurenwaaier gebruiken.
**Vraag**: Waarom hebben sommige bouncers een kleurenwaaier in plaats van een gastenlijst? (Racisme in het uitgaansleven)

Input: Fobien
Volledig: Als "Fobien" het antwoord is, wat is dan de vraag?
Redenering: Er zijn een hele reeks aan bekende fobien, zoals angst voor spinnen, clowns, kleine ruimtes etc. We kunnen een situatie creren waar ze allemaal bij
       elkaar tegelijk voorkomen om zo zeker "fobien" als antwoord te hebben.
**Vraag**: Wat krijg je als je een spin, een clown en een kleine ruimte combineert? (Fobien)

Input: Vervallen voedsel
Volledig: Als "Vervallen voedsel" het antwoord is, wat is dan de vraag?
Redenering: Vervallen voedsel komt doordat je tijd niet goed in de gaten gehouden hebt. Een tijdreiziger reist door heel de tijd en gaat dus veel last hebben van
       vervallen voedsel.
**Vraag**: Wat krijg je als je een tijdreiziger vraagt om boodschappen te doen? (Vervallen voedsel)

Input: Hacken
Volledig: Als "Hacken" het antwoord is, wat is dan de vraag?
Redenering: Hacken klinkt exact zoals het woord "hekken". Hekken kan je gebruiken om mensen buiten te houden, en om dat duidelijk te maken kunnen we spreken over
       een hek in de tuin, want in de tuin ga je niet hacken. We kunnen het woord "indringers" in de setup te gebruiken om mensen op het verkeerde been te zetten,
        gezien dit vaak met "hackers" gerelateerd is.
**Vraag**: Wat is de beste manier om indringers uit je tuin te houden? (Hacken/Hekken)
```

5. https://chatgpt.com/g/g-sMNdHk1ke-improv-waagstuk