# Can AI Make Us Laugh? Comparing Jokes Generated by Witscript and a Human Expert

**Joe Toplyn**[1]**, Ori Amir**[2]
[1]Twenty Lane Media, LLC; [2]HaHator LLC & Pomona College
joetoplyn@twentylanemedia.com, oriacadem@gmail.com

## Abstract

Evaluating the effectiveness of a joke-generating AI system ultimately comes down to one question: are its jokes as funny as those crafted by humans? Prior studies have typically relied on numerical ratings assigned by human evaluators—a method with inherent limitations—and few have directly compared the quality of AI-generated jokes to that of jokes created by professional human joke writers. In this study, we measured audience laughter—a direct and fundamental response to jokes—to assess the funniness of jokes produced by a specialized AI joke-writing system. We also compared those jokes to those written by a professional human joke writer to determine which elicited more laughter. Our findings reveal that the AI-generated jokes got as much laughter as the human-crafted ones. This suggests that the best AI joke generators are now capable of composing original, conversational jokes on par with those of a professional human comedy writer.

## 1 Introduction

Generating humor is often regarded as an AI-complete problem, one that requires full human intelligence to solve (Hurley, 2011; Winters, 2021). Generating original humor is even a challenge for humans (Amir, 2022; Tikhonov, 2024), and the brains of professional comedians are distinct functionally and structurally (Amir, 2016; Brawer, 2021). Witscript is one of the few AI systems that can generate contextually integrated jokes, like the jokes a human might improvise in a conversation (Toplyn, 2021b). The last time Witscript was systematically evaluated (Toplyn, 2023), human evaluators judged its responses to input sentences to be jokes 44% of the time. Since then, the Witscript system has been improved. This paper puts the current system to a more challenging test, comparing the funniness of its jokes to the funniness of jokes written by a professional human joke writer.

To determine whether Witscript is as funny as a human expert, a reliable method for evaluating the funniness of jokes is necessary. Goes et al. (2022) use an AI model, but papers on the computational generation of humor almost always evaluate the generated text using non-expert humans recruited on crowdsourcing platforms like Amazon Mechanical Turk (Loakman, 2023). This evaluation method is probably common because it is relatively low-cost and easy to carry out.

Nevertheless, using non-expert humans to rate jokes on a numerical scale has significant limitations (Amin, 2020; Hossain, 2020; Inácio, 2024; Valitutti, 2013). Evidence indicates that non-expert humans cannot appropriately evaluate the quality of creative text (Lamb, 2015). In the case of jokes, this apparent inability to judge quality may arise because jokes are designed to elicit laughter, not high numerical ratings. Indeed, a common definition of "joke" is "something said or done to provoke laughter" (`www.merriam-webster.com/dictionary/joke`). And "voiced laughter is correlated with highly amusing multimedia content" (Petridis, 2009). So, we believe that measuring the laughter elicited by a joke is a better way to measure its funniness.

Laughter is strongly influenced by social context. We laugh the most when we interact with someone in person, instead of via voice or text (Scott, 2014). Therefore, to ensure a stronger laughter signal that can be more accurately measured, a joke should be delivered to a group of people by someone in their presence. Delivering the joke to a group would also help compensate for the fact that evaluating humor is subjective: if a



*Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 71–78
January 19, 2025. ©2025 Association for Computational Linguistics

joke elicits a big laugh from a group, that means many people thought it was funny and, therefore, that the joke can objectively be assigned a high funniness rating. We decided, then, that the most reliable way to measure the funniness of jokes like those generated by Witscript is to measure how much laughter they elicit when they are delivered by professional standup comics in front of live audiences.

## 2 Related Work

Other authors have tasked human evaluators with comparing the funniness of jokes written by humans to that of jokes generated by AI systems. But those authors used numerical scales, not measurements of laughter, to rate the funniness of the output (Gorenz, 2024; He, 2019; Mittal, 2022; Petrović, 2013; Tikhonov, 2024; Zhang, 2020). To the best of our knowledge, this paper represents the first time that jokes generated by an AI system have been formally evaluated in the context of standup comedy performances.

## 3 Description of the Witscript System

Witscript is a neural-symbolic hybrid AI system designed to work in American English (Toplyn, 2023). It's symbolic because it incorporates joke-writing algorithms created by a human expert (Toplyn, 2014). And it's neural because it executes those algorithms, and other joke-production methods, by calling on a large language model in the GPT family from OpenAI (Brown et al., 2020). The Witscript jokes used in this research were generated by the version of the Witscript app that was publicly available on October 9, 2024, from `www.witscript.com`. The algorithms are based on formulas described in Toplyn (2014) and several patents (Toplyn 2020a, 2020b, 2021a).

## 4 System Evaluation

### 4.1 Input Selection

Author OA selected 16 current news headlines for use in evaluating Witscript. Author JT, a professional comedy writer, eliminated any of those headlines that were strongly associated with events occurring after the knowledge cutoff date of the GPT model used by Witscript. That way, Witscript's performance wouldn't be adversely affected by the system's dependence on non-current training data.

From the remaining news headlines, JT selected eight that, in his expert opinion, had two characteristics that made them particularly well-suited for joke writing: (1) they were likely to capture most people's interest, as good joke topics do (Toplyn, 2014); and (2) they were relatively "evergreen"—likely to seem fresh indefinitely—so jokes based on them wouldn't get stale and unfunny before testing was completed.

Then JT edited each of the eight selected news headlines into a form that he believed, in his expert opinion, would make it a useful joke topic. Each resulting topic had the following characteristics: (1) it was one sentence; (2) it was likely to be easily understood by its intended audience of adult Americans; and (3) it was relatively simple, with only one or two attention-getting elements, which Toplyn (2014) calls "topic handles."

### 4.2 Joke Production

The human expert—a longtime joke writer for a well-known, U.S.-based, late-night comedy/talk show—and Witscript, operated by JT, independently generated jokes based on the eight edited topics. They were given three days to complete the task to the best of their ability, so that the speed of their joke production would not be a factor.

The human expert and JT each selected from all of their own output the one joke for each topic that they believed would elicit the most laughter from an audience of typical American adults. They submitted their eight chosen jokes to a third-party data manager without sharing them with each other. All of Witscript's selected jokes were submitted exactly as they were output by Witscript.

### 4.3 Laughter Measurement

Experienced standup comics performed two comedy sets in front of live audiences in comedy venues in the U.S. The comics did not reveal the sources of the jokes and did not know which jokes had been written by AI. In each set, jokes based on all of the eight topics were performed, with half of the punchlines written by the human expert and half by Witscript. Both the order of the topics and which punchline was selected for each topic were determined randomly and counterbalanced between sets. As a cover story, the comics explained that they would be performing some jokes written by a friend.

To measure the quantity of laughter elicited by each joke, the recording of each set was labeled to mark the segments in which laughter occurred. The original audio was then converted to a graph of decibels over time using Formula 1.

$$dB = 20 * log10(|s| + 1e^{-6}) \qquad (1)$$

In the formula, s is the original sound wave, and $1e^{-6}$ is the lowest sound level perceivable by humans. The area under the curve, representing the "quantity of laughter," was then computed using Simpson's numerical integration method implemented in Python (Matthews, 2004). We refer to the measure as Total Laughter; its units are decibel-seconds (see Figure 1). We believe this method best captures the quantity of laughter compared to other potential methods such as the average, median, or max, as those other methods would be poor at capturing situations in which different individuals in the audience "get the joke" at different times, resulting in the same amount of laughter spread over a longer period of time.

For the present analysis, we used the audio of two high-quality sets performed at the same North Hollywood venue by the same comedian, Mike Perkins, with audience sizes of 35 and 15. The sets were performed a month apart at the same time of day (at the end of the comedian's 10-minute set opening the 8 p.m. show). Two other sets were excluded from the analysis either because of poor venue quality or small audience size (N<10).

The audio tracks were annotated to select the segments of laughter associated with each joke. In a typical set, sounds unrelated to the laughter, such as heckling, would mix with the laughter. However, these interferences were not an issue in the sets we analyzed. Additionally, comedians might speak over the laughter to make a comment or start the next joke. But in the sets we analyzed, the comedian made an effort to let the audience laugh uninterrupted, though he often did start the next joke when he felt the laughter was dying down. We always ended the laugh segment before the comedian resumed talking, so the audio segment contained laughter only. Importantly, the comedian was not aware which jokes had been written by AI, so any such interference affected all jokes equally.

We compared the performance of Human vs. AI jokes within sets and between sets. The between-sets comparison required some form of normalization of the laughs to remove any bias resulting from the size of the audience or other characteristics affecting the overall loudness of its laughter. That normalization was achieved by:

1. Prior to conducting a paired t-test, we compared two joke versions across sets. The Loudness measure of all the jokes within a set was normalized by the median Loudness across all jokes in the set.
2. For the GLM the Set was included as a regressor of no interest.

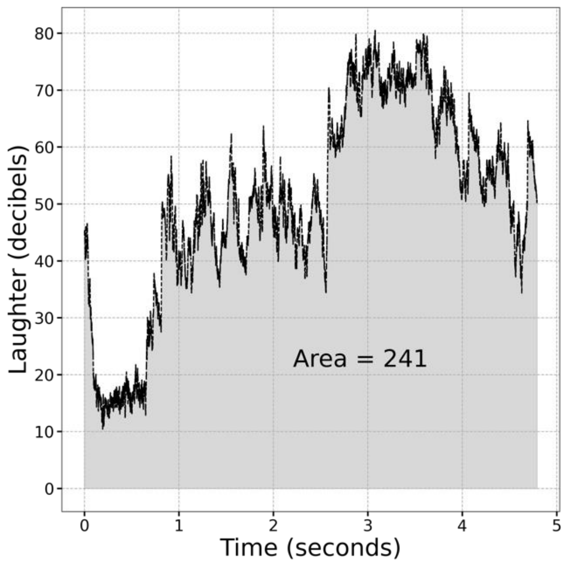


Figure 1: A demonstration of how the "Total Laughter" of a single joke is measured. The sound wave of the laughter segment following the joke is converted to dB over time. The area under the curve (here 241) is the Total Laughter in decibel-seconds.

## 4.4 The Hypothesis

Historically, the standard for demonstrating that AI had reached a certain milestone against human performance involved only a few data points. For example, Kasparov played only six games with Deep Blue (AI) in 1997 (scoring 2.5-3.5). In 2011, Watson (AI) competed only once against two human champions on *Jeopardy!*, and won. While such events would not meet the nominal standards of statistical significance required to determine that AI was "consistently" better than the human champions, they are nevertheless considered meaningful milestones, since before those events it was considered inconceivable that AI would perform at the level of those human champions even once.

If generating jokes for a comedy/talk show-style monologue, where the quality is judged by

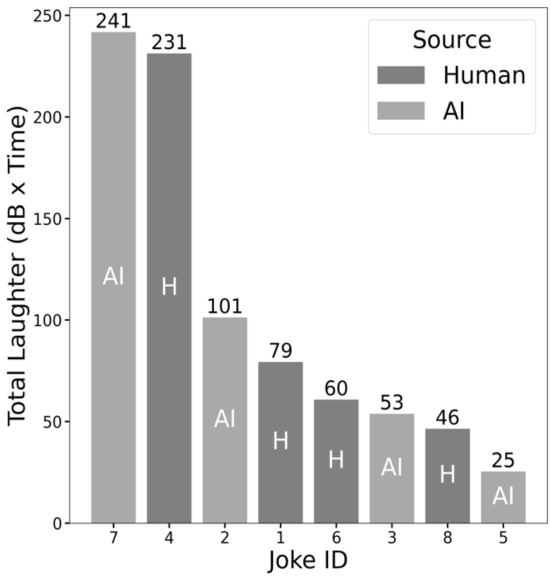


Figure 2: The jokes written by the human expert (H) and Witscript (AI) in order of the Total Laughter they elicited in Set 1. Joke ID corresponds to the actual order in which the jokes were told. The jokes are listed in the Appendix.

audience laughter, was an AI-complete problem, we would expect that:

$H_0$: *None* of the AI-generated jokes would perform better than *any* of the professional human writer's.

We could reject this hypothesis if:

$H_1$: Some of the AI-generated jokes performed better than some of the Human's.

## 5 Results and Discussion

### 5.1 Analysis Within a Set

Figure 2 displays the eight jokes performed in Set 1[1] ranked by the Total Laughter they elicited. Three of the four jokes written by AI elicited more laughter than at least one joke written by the human expert. Additionally, the joke that elicited the most laughter was AI-written.

This result is in line with $H_1$, in that some of the AI-written jokes did better than some of the Human's. The same pattern held true for Set 2; see the Appendix for the data. If we deem this result to be reliable, we can conclude that writing the type of humor analyzed here is not AI-complete. How can we determine this reliability?

How reliable is the measure itself? The measure captures the total laughter of an audience of N=35 and 15 in Sets 1 and 2, respectively. In a classical experiment, jokes are rated by a handful of raters. While audience members' responses are not entirely independent (e.g., laughter is contagious) whatever effect audience members had on each other was present for all jokes and presumably had the effect of signal amplification rather than of cancellation of individual differences. Additionally, unlike with raters, it is not possible to tease apart the contributions of individual raters (here, audience members). Despite these drawbacks, the number of raters/audience members is much larger than in a typical study in the field, suggesting higher reliability than the standard. The validity of the measure is arguably higher since the measure is of a natural response to jokes in a natural environment. However, there may be other forms of humor for which a traditional approach using numerical ratings would be better suited than our measurement method.

How did the Human and AI jokes compare? The funniest joke (area under the curve = 241) was written by AI. On average, AI did slightly better ($M$ = 106, $SD$ = 96) than the Human ($M$ = 104, $SD$ = 86) in Set 1, with the reverse true in Set 2 (AI: $M$ = 66, $SD$ = 21; Human $M$ = 99, $SD$ = 93). However, these differences were not significant (both sets: Mann-Whitney U(4,4) = 8.0, ns). The lack of statistical difference between the groups is not meaningful with the present sample size. Instead, as explained above (see the hypotheses), we rely on a standard similar to Deep Blue's and Watson's, that of a limited live demonstration of equivalence to human performance, which we have met.

### 5.2 Comparison Between the Sets

As described above, the two sets had the same eight topics, for which half of the punchlines were written by AI and half by the Human. The jokes were counterbalanced so that if a particular topic had a punchline written by the Human in Set 1 it would have a punchline written by AI in Set 2, and vice versa.

The audience size for Set 1 was bigger than for Set 2 (35 vs. 15), resulting in longer laugh times (M = 2.16 sec. vs. 1.71 sec.) and greater values on our Total Laughter metric (M = 105 vs. 83). But

[1] Set 1 had the bigger audience (N=35). It would be inappropriate to display jokes from both sets in this figure because of the difference in the Loudness baseline.

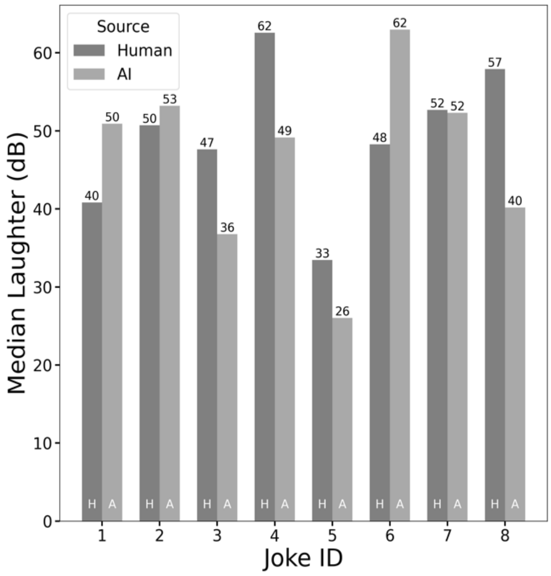


Figure 3: The Median Laughter Loudness (over the duration of the laugh) elicited by the Human (H)- vs. AI (A)-written jokes for each topic across the two sets. The lack of pattern suggests equivalent performance by the Human and AI sources.

Median Laughter Loudness showed no difference (for both sets, M = 48). Controlling for that baseline, no significant differences were observed between the AI and human-written versions of the joke for each topic. This was true for our Total Laughter metric as well as for other measures, including Mean Loudness, Median Loudness, and Length of Laugh (all paired t values < 1, ns; a GLM statistically controlling for Set effects returned the same result). Since there was no difference in the Median Laughter Loudness, that metric lends itself to a bar graph comparing the two sets which has no distortions resulting from normalization; see Figure 3.

Overall, comparing the AI and Human jokes on the same topic between sets mirrors the result of comparing the AI and Human jokes within the sets—there is no difference in the effectiveness of the jokes.

## 6 Contributions

This paper makes the following contributions:

1. It introduces a novel method of evaluating the funniness of jokes—measuring the laughter they elicit.
2. It demonstrates a way to compare the joke-writing ability of an AI system to that of a human expert in the real-world setting of standup comedy.
3. It provides further evidence that computational joke generation is best accomplished by taking a hybrid neural-symbolic approach.
4. It provides further evidence that at least one type of humor, generating monologue-style jokes for an American audience, is not AI-complete.

## 7 Conclusion

AI-written jokes, performed in front of a live audience, elicited laughter within the same range as jokes written by a professional human comedy writer.

Some AI-written jokes ranked higher than some of the human-written jokes, and the funniest joke, as measured by quantity of laughter, was written by AI.

The study provides naturalistic, real-world evidence that when it comes to generating comedy/talk show monologue-style humor, an AI system can perform at the level of a professional human comedy writer.

## 8 Limitations

1. Several aspects of the performances may have contributed to a joke's funniness beyond the quality of its writing. These include the comic's vocal delivery and any gestures and facial expressions he chose to make. We assume these factors influenced AI and human-written jokes equally, since the comic did not know which jokes had been written by AI. This kind of noise is the price of conducting an arguably more valid naturalistic study. It is not likely to reflect systematic bias.
2. Our measure captures the funniness ratings of the ~50 audience members for the two sets. However, the audience members cannot be considered fully independent (e.g., laughter is contagious). That acknowledged, whatever influence audience members had on each other, it was likely a constant factor of amplification affecting all jokes similarly.
3. The Witscript jokes submitted for evaluation were cherry-picked by a human expert from all of the jokes generated by Witscript on the assigned topics. However, we don't consider that to be a major limitation because the human jokes submitted for evaluation were similarly cherry-picked from multiple joke candidates crafted by the human writer.

## Acknowledgments

We would like to thank the following comedians for their insights and help: Mike Perkins, Kevin Hickerson, Ajitesh Srivastava, and the comedy/talk show writer who wrote the Human jokes for the experiment.

## References

Miriam Amin and Manuel Burghardt. 2020. A Survey on Approaches to Computational Humor Generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Ori Amir et al. 2022. The elephant in the room: attention to salient scene features increases with comedic expertise. *Cognitive Processing,* 23(2), 203-215.

Ori Amir and Irving Biederman. 2016. The Neural Correlates of Humor Creativity. *Frontiers in Human Neuroscience, 10*(597).

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

Tom B. Brown et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Fabricio Goes, Zisen Zhou, Piotr Sawicki, Marek Grzes, and Daniel G. Brown. 2022. Crowd score: A method for the evaluation of jokes using large language model AI voters as judges. arXiv preprint arXiv:2212.11214.

Drew Gorenz and Norbert Schwarz. 2024. How funny is ChatGPT? A comparison of human- and AI-produced jokes. PLoS ONE 19(7): e0305364. https://doi.org/10.1371/journal.pone.0305364.

He He, Nanyun Peng and Percy Liang. 2019. Pun Generation with Surprise. North American Chapter of the Association for Computational Linguistics.

Nabil Hossain, John Krumm, Michael Gamon, and Henry Kautz. 2020. SemEval-2020 Task 7: Assessing Humor in Edited News Headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Matthew M. Hurley, Daniel C. Dennett, and Reginald B. Adams. 2011. Inside Jokes: Using Humor to Reverse-Engineer the Mind. MIT Press.

Marcio L. Inácio and Hugo G. Oliveira. 2024. Generation of Punning Riddles in Portuguese with Prompt Chaining. 15th International Conference on Computational Creativity (ICCC'24).

Carolyn Lamb, Daniel G. Brown, and Charles L.A. Clarke. 2015. Human Competence in Creativity Evaluation. Sixth International Conference on Computational Creativity.

Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. The Iron(ic) Melting Pot: Reviewing Human Evaluation in Humour, Irony and Sarcasm Generation. Conference on Empirical Methods in Natural Language Processing.

John H. Matthews. 2004. Simpson's 3/8 Rule for Numerical Integration, Numerical Analysis-Numerical Methods Project.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating Humorous Puns with Ambiguous Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

Stavros Petridis and Maja Pantic. Is this joke really funny? Judging the mirth by audiovisual laughter analysis. 2009. In *2009 IEEE International Conference on Multimedia and Expo*, New York, NY, USA, pp. 1444-1447, doi: 10.1109/ICME.2009.5202774.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Sofia, Bulgaria. Association for Computational Linguistics.

Sophie Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. 2014. The social life of laughter. *Trends in Cognitive Sciences*, 18(12), 618–620. https://doi.org/10.1016/j.tics.2014.09.002.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor Mechanics: Advancing Humor Generation with Multistep Reasoning. arXiv preprint arXiv:2405.07280.

Joe Toplyn. 2014. *Comedy Writing for Late-Night TV: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-Form Comedy*. Twenty Lane Media, LLC, Rye, New York.

Joe Toplyn. 2020a. Systems and Methods for Generating Jokes. U.S. Patent No. 10,642,939. Washington, DC: U.S. Patent and Trademark Office.

Joe Toplyn. 2020b. Systems and Methods for Generating Comedy. U.S. Patent No. 10,878,817. Washington, DC: U.S. Patent and Trademark Office.

Joe Toplyn. 2021a. Systems and Methods for Generating and Recognizing Jokes. U.S. Patent No. 11,080,485. Washington, DC: U.S. Patent and Trademark Office.

Joe Toplyn. 2021b. Witscript: A System for Generating Improvised Jokes in a Conversation. In *Proceedings of the 12th International Conference on Computational Creativity*, 22–31. Online: Association for Computational Creativity.

Joe Toplyn. 2023. Witscript 3: A Hybrid AI System for Improvising Jokes in a Conversation. arXiv, abs/2301.02695.

Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M. Toivanen. 2013. "Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 243–248, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Winters. 2021. Computers Learning Humor Is No Joke. *Harvard Data Science Review*, *3*(2). doi.org/10.1162/99608f92.f13a2337.

Hang Zhang, Dayiheng Liu, Jiancheng Lv, and Cheng Luo. 2020. Let's be Humorous: Knowledge Enhanced Humor Generation. *Annual Meeting of the Association for Computational Linguistics*.

## Appendix. The Jokes

Below is a full list of the jokes and their Joke ID, which indicates the order in which they were told in the sets. Each joke has a topic that serves as a prompt/setup for both the AI- and human-written punchlines. Each set randomly includes half of the punchlines written by AI. Next to each joke, we also provide these metrics for the laughter it elicited: Total Laughter, in decibel-seconds (TL); total laugh Time, in seconds (T); and Median Laughter Loudness over the duration of the laugh, in decibels (ML).

**Joke 1**
**Topic**:
A new report says that NASA officials are worried about a leak on the International Space Station.
**Human**: (TL: 79, T: 2.00, ML: 40)
Will they fix it? Naw, even in space, landlords don't fix leaks.
"But Houston, we have a potty problem."
That's on them, we subcontracted to Boeing.
**AI**: (TL: 51, T: 1.04, ML: 50)
They're especially concerned since the leak is coming from one of their astronauts' space diapers.

**Joke 2**
**Topic**:
Why do TV stations air false political ads?
**Human**: (TL: 51, T: 1.04, ML: 50)
That's so after the election, we welcome the sound of "Attention, Hemorrhoid Sufferers!"
**AI**: (TL: 101, T: 1.96, ML: 53)
Because they want to make sure the viewers are just as confused as the candidates!

**Joke 3**
**Topic**:
A company just introduced a virtual dog leash that uses wireless technology.
**Human**: (TL: 73, T: 1.58, ML: 47)
Wifi can control my dog's movements? So where's his virtual pooper scooper?
**AI**: (TL: 53, T: 1.54, ML: 36)
But I'm pretty sure that's just a fancy way of saying 'I don't want to walk my dog.'

**Joke 4**
**Topic**:
Bob Yerkes, a stuntman who appeared in "Star Wars," died at the age of 92.
**Human**: (TL: 231, T: 3.92, ML: 62)
In his long career, he broke so many bones, his grave says Rest in Pieces. But true Star Wars fan to the end, he asked to be buried in his parent's basement.
**AI**: (TL: 48, T: 1.00, ML: 49)
He passed away surrounded by his loved ones and a strategically placed pile of mattresses.

**Joke 5**
**Topic**:
BuzzFeed put out a list of 31 things to buy when you finally decide to update your kitchen.
**Human**: (TL: 36, T: 1.08, ML: 33)
If you ask me, appliances are too smart already. The clock on my coffee maker flashes 12 12 12... What'll it do smarter--snicker? "Tsk tsk tsk. So much for caffeine increasing brain function."
**AI**: (TL: 25, T: 0.92, ML: 26)
Number 32 on the list: a new Buzzfeed article on 31 ways to use all the unnecessary gadgets you bought from the first list.

**<u>Joke 6</u>**
**Topic**:
Scientists have discovered a sixth ocean more than 400 miles below the surface of the Earth.
**Human**: (TL: 60, T: 1.29, ML: 48)
Great, I was just looking for a gnarly new place to surf. (mime surfing around dangers) "Stalactite! Stalagmite! Bats! Gollum!!"
**AI**: (TL: 94, T: 1.54, ML: 62)
Looks like Aquaman's commute just got a whole lot longer.

**<u>Joke 7</u>**
**Topic**:
Scientists are studying whether astronauts in the future could transform rocks into food.
**Human**: (TL: 236, T: 4.62, ML: 52)
Hey, don't give Fruity Pebbles any ideas. Rocky Road with real rocks? You could chip a tooth on Stone Ground Mustard!
**AI**: (TL: 241, T: 4.79, ML: 52)
Which is great news for anyone who's ever had a craving for a pebble pie.

**<u>Joke 8</u>**
**Topic**:
A new study says that young children in the UK get almost half their calories from ultra-processed food.
**Human**: (TL: 46, T: 0.88, ML: 57)
If you think that's bad, the other half is British cooking.
**AI**: (TL: 70, T: 1.79, ML: 40)
The most popular kids' meals in the UK are now the Happy Meal, the Crispy Chicken Sandwich, and Uncle Nigel's Deep-Fried Crumpets.