The Roast of GPT4o: Experiments in Generating, Detecting and Evaluating Celebrity Roast Comedy

Jens Lemmens, Jérémy Genette, Walter Daelemans
University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
firstname.lastname@uantwerpen.be

Tony Veale
University College Dublin
Belfield, Dublin 4, Ireland
tony.veale@ucd.ie

Abstract

We present exploratory experiments in the comedic roasting capabilities of GPT4o. Specifically, @ComedyCentral roasts were scraped to design a survey in which participants blindly evaluated snippets of human and AI roasts, and had to predict the author (AI/human) in a second round of reviewing. The results show that there is no significant difference in how the barbs in human- and AI-generated roasts are rated. Further, a qualitative analysis showed that although the model utilizes specific recurrent phrases to imitate the style of human comedians, both generative LLM detectors and humans performed suboptimally in predicting the true author of the roasts.

1 Introduction and related work

While the most recent language models have shown complex logical reasoning capabilities, mixed results have been reported on the problem of humour generation, which requires world knowledge, semantic reasoning, and linguistic insight. On the one hand, these mixed results may reflect the variety of humour genres that are explored and the prompting strategies utilized in earlier work. Specifically, previous studies have chiefly focused on stand-alone jokes, e.g., Toplyn and Amir (2025); Horvitz et al. (2024); He et al. (2024); Bogireddy et al. (2023); Zeng et al. (2024); Zhang et al. (2024); Mittal et al. (2022). Such out-of-context jokes, however, account just for an estimated 17% of our daily experience of humour (Martin and Ford, 2018). In contrast, performance humour is relatively understudied in Natural Language Generation (NLG) research and has mainly focused on stand-up comedy or sitcom humour, e.g., Li et al. (2023); Mirowski et al. (2024). In this paper, we thus aim to provide new insights into the humour generation problem by presenting initial experiments with GPT4o (OpenAI, 2024) on the unexplored task of comedy roast generation. A *roast* is a form of insult comedy in which guests hurl affectionate barbs at a guest of honour. Specifically, the unit of analysis in this paper is a specific fragment (or "burn") that is uttered by a certain speaker at such a roast event.

Apart from a strong focus on specific genres of humour, the inconclusive findings regarding the humor capabilities of LLMs could also be attributed to a lack of a standardized evaluation paradigm (Lemmens and De Marez, 2026). Previous work has used human evaluators, automatic methods, or a combination of both. In the case of human evaluators, both A/B testing, e.g., Chen et al. (2024), and scoring on Likert scales has been used, e.g., Gorenz and Schwarz (2024), although these scales have not been used consistently in terms of values (5-point vs. 7-point scale) and in what they measure (overall score, humorousness, originality, style, etc.). An advantage of human evaluators is that they can consider a variety of fine-grained aspects of humour when evaluating content, and can offer qualitative insights into their evaluation, in contrast to automatic methods. Disadvantages, on the other hand, include greater expense and slower work-rates, as well as the subjectivity of personal preferences (Romanowski et al., 2025). In this paper, we alleviate this problem by including a large number of human evaluators and by analyzing these personal preferences using a mixed effects statistical analysis.

We begin by scraping Comedy Central roast transcripts from YouTube, and use AI to generate new roast materials. A survey is then used to allow participants to blindly rate the human- and AI-written roasts to determine the comedic abilities of GPT4o. In addition, we investigate the extent to which AIs can predict if a roast was written by another AI, which may produce insights into how to automatically evaluate the quality of AI-generated humour.

We explore the following research questions and hypotheses: (1) Can GPT4o generate roasts that are as funny as human roasts? (2) How accurately can the author (AI or human) of a roast be recognized?

Proceedings of the 2nd Workshop on Computational Humor (CHum 2026), pages 65–71

We hypothesize that (1) due to model alignment, and the creative nature of the task, GPT4o is unable to generate roasts of the same funniness as humans. Thus, ratings for human roasts will be significantly higher than those for AI-generated ones. In addition, (2) humans and LLM detectors can both predict if a roast was created by AI or not with relatively high accuracy due to the quality difference between the human and AI generated roasts.

Contributions[1]: Firstly, we provide the materials to scrape all ComedyCentral roasts transcripts on Youtube. Secondly, the results of this study serve as a proof-of-concept that GPT4o can generate roast content that is as funny as that of human comedians, in opposition to our initial hypothesis, while adhering to a specific writing style, as explained in a qualitative analysis. Thirdly, we show that LLM detectors exhibit sub-optimal results on comedy data, although LLM detectors still perform better at this task than humans, who do not exceed random chance.

2 Methodology

A survey was designed to let 64 participants (see Appendix A) blindly rate fragments of human- and AI-generated roasts. To collect data for the survey, we used the Google API to scrape all videos of the @ComedyCentral channel on YouTube containing 'roast' in the title. Then, twenty fragments from this collection of transcriptions were manually selected, following 2 criteria: First, for purposes of comparison, the fragment must be self-contained, and understandable without additional video/audio data. Secondly, 10 distinct roast targets were represented in the subset, i.e. 2 fragments per target.

A random selection of 10 of these 20 fragments were included in the survey, while the remaining 10 were used to generate new roast fragments with GPT4o (OpenAI, 2024). The prompt (Appendix C) first explains what a roast is, using a short version of the Wikipedia definition, and names the specific celebrity to be targeted by the roast. A short Wikipedia biography of the target is then given in the prompt as background information, and a 1-sentence paraphrase of an original roast is provided as a topic. An example roast with a different target is also provided so that the model acquires a sense of the general style and structure of a roast.

Since it was the intention to ask participants to predict whether a roast was written by AI or not, post-processing was required to avoid unintentionally identifying the author with highly revealing text features. Therefore, profane words were removed from the human roasts (if grammatically possible), or replaced with a non-profane synonym, since LLMs rarely generate profanity, even when role-playing, due to model alignment (Huang et al., 2025). In addition, em-dashes were either removed or replaced by commas in the AI-generated roast, as these are a common feature of LLM outputs (Sukhareva, 2025). At this point, our mini-dataset pairs 10 human roasts with 10 AI roasts.

A survey was then created to let participants evaluate these roast fragments[2] in a within-subjects design. This survey consists of two parts: First, participants are asked to indicate whether they are familiar with the target of the roast (variable name: 'familiarity') and to then rate the roasts on a Likert scale (variable name: 'rating') from 1 (not funny) to 5 (extremely funny). To avoid biasing them in their ratings, participants are told that the purpose of the survey is to compare the humour preferences of humans to those of LLMs. This masks the true goal of the study, so as to not reveal that half of the fragments are AI-generated.

In part 2 of the survey, participants were told that half of the roasts were generated by AI, but not told which ones. They were then presented with the same content in the same order as in part 1, and were asked to indicate whether they had seen the roast before on Comedy Central (variable: 'prior encounter') and to predict whether an AI or human wrote it (variable: 'author prediction').

To evaluate the author predictions of participants (AI or human), those predictions were compared with two state-of-the-art generative AI classifiers: Binoculars (Hans, Abhimanyu and Schwarzschild, Avi and Cherepanova, Valeriia and Kazemi, Hamid and Saha, Aniruddha and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom, 2024) and GPTZero (OpenAI, 2023). GPTZero is a multi-step commercial LLM detector that specializes in detecting content generated by ChatGPT, GPT4, Gemini, Claude, and LLaMa models. Binoculars, in contrast, is an open-source zero-shot LLM detection method for text data that leverages normalized cross-perplexity scores between two LLMs to predict whether a text was written by AI or not. In this experiment, we used Falcon-7B and Falcon-

[1]The code and survey data are available on Github: https://github.com/LemmensJens/roast_of_gpt4o.

[2]See Appendix B for an example of the data.

7B-instruct (Almazrouei et al., 2023), and a perplexity threshold of 0.90, as proposed in Hans, Abhimanyu and Schwarzschild, Avi and Cherepanova, Valeriia and Kazemi, Hamid and Saha, Aniruddha and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom (2024).

3 Results

3.1 Statistical analysis

The analysis of the survey data is summarized in Figure 1, where it can be observed that participants predicted that the funnier roasts were human-generated. Additionally, roasts were rated as funnier when participants were more confident that they had seen the roast before. These observations are confirmed by the results of the following statistical analysis (conducted using R (R Development Core Team, 2022) and the *ordinal* package (Christensen and Christensen, 2015)).

Cumulative Link Mixed-Effects Models were used to examine whether 'author', 'familiarity', 'author prediction', and 'prior encounter' have an effect on 'funniness'. Models of increasing complexity were built step-by-step by incrementally including fixed and random effects, with the funniness rating used as the dependent variable. The inclusion of a predictor was assessed by a likelihood ratio test (Baayen, 2008).

The final model included 'author prediction' and 'prior encounter' as fixed effects, and a random intercept for each participant and each roast. The inclusion of 'author' and 'familiarity' as predictors did not significantly improve model fit and were therefore excluded. This suggests that these variables did not have a significant effect on the funniness ratings. That is, there was no statistically significant difference between the funniness ratings of the human-generated roasts and the AI-generated roasts, indicating that, in general, the participants found the AI roasts and the human roasts equally funny. Similarly, whether participants were familiar with the subject of the roast had no influence on how funny they found the roast.

However, 'author prediction' and 'prior encounter' had a significant effect on funniness ratings, as shown in Table 1, which presents the results of the fixed effects from the modelling procedure. From these results, it can be observed that there was a significant positive main effect of 'author prediction' ($\beta = 0.648$, $SE = 0.118$, $z = 5.488$, $p < 0.001$). This indicates that participants tended to find a roast more funny if they thought the roast was human-generated rather than AI-generated.

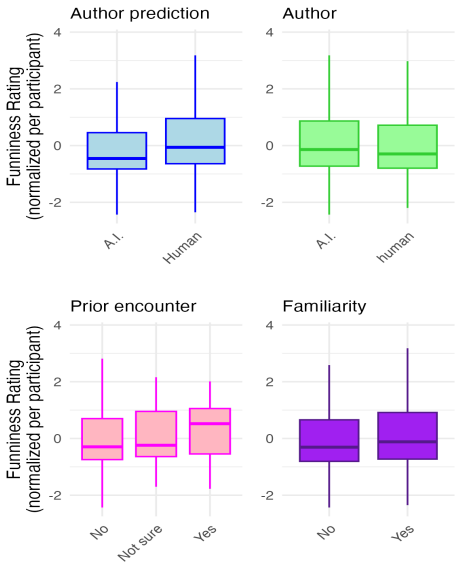

Figure 1: Overview of the funniness ratings: Likert scores normalized per participant (mean and standard deviation). Non-normalized results are included in Appendix E.

The participants were first asked to provide ratings without knowing that some roasts were AI-generated. After it was revealed that not all roasts were human-generated, participants were then asked to predict the author of each roast. As such, this suggests that a participant's perception of a roast's funniness influenced their prediction of its author. Given that there was no significant effect of the actual author on the funniness rating, however, this indicates that there was a strong mismatch between the actual authors and the predicted authors of the roasts. The low performance of the participants on the author prediction task, as further described in Table 2, supports this observation.

In addition, 'prior encounter' (whether they think they have seen the roast before) had a significant linear effect on the funniness rating ($\beta = 0.601$, $SE = 0.192$, $z = 3.136$, $p = 0.002$). This indicates that the more confident a participant's belief is in having encountered a roast before, the funnier they tend to rate it. So prior exposure not only plays a significant role in humour evaluation but also complicates automatic humour assessment, as it increases the subjectivity of such evaluations.

Variable	Estimate	Std. Error	z value	Pr(>|z|)
Author prediction	0.648	0.118	5.488	< 0.001***
Prior encounter [Linear]	0.601	0.192	3.136	0.002**
Prior encounter [Quadratic]	0.032	0.194	0.163	0.870

Signif. codes: 0'***' 0.001'**' 0.01'*'

Table 1: Fixed effects of the final model.

Class	Participants			LLM Detection Systems	
	Precision	Recall	F1-score	GPTZero	Binoculars
Human	48.5 (12.4)	57.3 (19.0)	51.9 (14.2)	58.8 - 100.0 - 74.1	66.7 - 60.0 - 63.2
AI	48.2 (17.3)	40.6 (17.6)	43.0 (16.2)	100.0 - 30.0 - 46.2	63.6 - 70.0 - 66.7
Macro	48.4 (13.8)	49.0 (11.8)	47.5 (12.6)	79.4 - 65.0 - 60.1	65.2 - 65.0 - 64.9

Table 2: Author prediction task. Left: human participants (mean (std.)). Right: LLM detectors (pre, rec, F1).

The results in Table 2 show that the averaged performance of the participants on the author prediction task was not higher than random chance, reflecting both the difficulty of the task and the success of GPT4o. As shown in Table 2, the LLM detectors scored substantially higher than random chance, but by no means showed high accuracy.

3.2 Qualitative analysis

Now, we shall discuss the best and worst rated AI roasts in detail, and analyze why participants predicted either the right or wrong author for these roasts (recall that participants were allowed to put their reasons into words in the questionnaire). In addition, we attempt to identify general trends in AI roasts, in terms of both style and content.

In the best rated AI roast, 3 jokes were made about the height of Kevin Hart. Although this roast was rated the funniest, there are a number of considerations to be made: first, the joke "your big break was playing a Lego in Toy Story" is not factual, since Kevin Hart did not feature in that movie. Further, "You're so tiny, you could hang-glide with a Dorito", seems original, but after a Google search, it became clear that this is an existing "yo mama" joke. The final joke, in contrast, appears to be original and based on relevant facts: "I heard they tried to cast you in Jumanji as a rock, but even Dwayne couldn't find you." This joke builds upon the knowledge that Kevin Hart starred in Jumanji, while incorporating a size-joke (a rock is small, Kevin Hart is small; Kevin Hart is a rock). Simultaneously, there is also a reference to Dwayne "The Rock" Johnson, who also stars in Jumanji.

Conversely, participants who believed this roast was written by a human highlighted that the sentences were rather short, as in spoken language, that the first person view was prominent, and that the imagery was too striking to be AI-generated.

The lowest-rated AI roast was that of Pete Davidson: "Alright, folks, let's talk about Pete Davidson. This guy's career has skyrocketed, but he still lives in his dad's shadow. I mean, Pete, you lost your father on 9/11, but look on the bright side, you're still the second biggest disaster to come out of that day." This roast riffs on a single joke, but does not develop it logically: There is a reference to 9/11, an event in which Davidson's father died, but GPT4o falsely implies that Davidson was also born on the same day. This illogical joke development, however, was not the main reasons why participants incorrectly predicted the author of this roast. Instead, they reported that the main reasons for their decision are that the roast contains "spoken" language ("alright folks") and remarkable brutality.

Considering all roasts used for the survey, we found that GPT4o relies on a number of frequent tics and phrases to imitate the style of comedians: ("..., (am I) right?", "Let's talk about [target]", "But hey, (at least) ..."; "I mean, ..."; "But let's be honest, ..."). Regarding the content, on the other hand, two general trends were observed. First, GPT4o often mentions specific statistics (e.g., "You [Charlie Sheen] were the highest-paid actor on TV at $1.8 million an episode."; "$2 million in settlements? That's one expensive 'extracurricular activity"' [about James Franco]"). Secondly, the model often ends a roast on a positive (if sometimes sarcastic) note. For example: "[David Hasselhoff] You're proof that you don't need to act well when you look good in red trunks, bravo."

4 Conclusion

A small-scale survey indicated that, contrary to our initial hypothesis, GPT4o can generate roast content that is perceived to be as funny as human content, since no statistically significant difference was found in the funniness ratings by the participants of the survey. Interestingly, we found that there were significant effects of "prior encounter" and "author prediction", highlighting important evaluation considerations when using human raters.

Further, a qualitative analysis also showed that GPT4o used specific stylistic tactics to construct roasts, including mentioning specific statistics or facts retrieved via the prompt, ending the roast on a positive note, and using phrases commonly used by comedians to link sentences or jokes. This indicates that there are certain features that can be detected in the writing style of GPT4o, although the

participants did not pick up on this, as evidenced by their no-better-than-random performance on this task. Generative AI predictors, on the other hand, showed higher accuracy than the participants, although performance was still sub-optimal, contradicting our initial hypothesis. However, this result is presumably the effect of the specificity of the genre and the relatively short text lengths.

5 Limitations

The number of roasts included in the survey was limited to 20 fragments to avoid annotator fatigue: participants reported that completing the survey took between 1 and 2 hours. In future work, however, a similar study may be conducted with a larger sample size, using a between-subjects design, and/or using complete roast contributions of specific comedians.

In addition to the limited size of the survey in terms of roasts, the background of participants was relatively homogeneous, given that they were all students from the same Master's program.

Finally, note that the human generated roasts were written for a live performance, i.e. for a specific audience, format, and for verbal use. The GPT generated roasts on the other hand, has no information about the audience, or does not write with the purpose of live, verbal performance. As mentioned, however, the human data was selected to be interpretable without additional context.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *Preprint*, arXiv:2311.16867.

R. H. Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Neha Reddy Bogireddy, Smriti Suresh, and Sunny Rai. 2023. I'm out of breath from laughing! I think? A dataset of COVID-19 Humor and its toxic variants. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 1004–1013, New York, NY, USA. Association for Computing Machinery.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a Joke Master? Pun Generation via Multi-Stage Curriculum Learning towards a Humor LLM. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 878–890, Bangkok, Thailand. Association for Computational Linguistics.

Rune Haubo Bojesen Christensen and Maintainer Rune Haubo Bojesen Christensen. 2015. Package 'ordinal'. *Stand*, 19(2016).

Drew Gorenz and Norbert Schwarz. 2024. How funny is ChatGPT? A comparison of human- and A.I.-produced jokes. *PLOS ONE*.

Hans, Abhimanyu and Schwarzschild, Avi and Cherepanova, Valeriia and Kazemi, Hamid and Saha, Aniruddha and Goldblum, Micah and Geiping, Jonas and Goldstein, Tom. 2024. Spotting llms with binoculars: zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, and Naihao Deng. 2024. Chumor 1.0: A Truly Funny and Challenging Chinese Humor Understanding Dataset from Ruo Zhi Ba. *Preprint*, arXiv:2406.12754.

Zachary Horvitz, Jingru Chen, Rahul Aditya, Harshvardhan Srivastava, Robert West, Zhou Yu, and Kathleen McKeown. 2024. Getting Serious about Humor: Crafting Humor Datasets with Unfunny Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 855–869, Bangkok, Thailand. Association for Computational Linguistics.

Tiansheng Huang, Sihao Hu, Fatih Ilhan, Selim Furkan Tekin, Zachary Yahn, Yichang Xu, and Ling Liu. 2025. Safety tax: Safety alignment makes your large reasoning models less reasonable. *Preprint*, arXiv:2503.00555.

Jens Lemmens and Victor De Marez. 2026. Computational Humor Modeling: A Survey on the State of the Art. *ACM Comput. Surv.*, 58(7).

Jianquan Li, XiangBo Wu, Xiaokang Liu, Qianqian Xie, Prayag Tiwari, and Benyou Wang. 2023. Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7581–7596, Toronto, Canada. Association for Computational Linguistics.

Rod A. Martin and Thomas E. Ford. 2018. *The Psychology of Humor*. Elsevier Inc.

Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A Robot Walks into a Bar: Can Language Models Serve as Creativity Support-Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *Proceedings of the*

2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, page 1622–1636, New York, NY, USA. Association for Computing Machinery.

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.

OpenAI. 2023. GPTZero.

OpenAI. 2024. GPT-4o.

R Development Core Team. 2022. *R: A Language and Environment for Statistical Computing*.

Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. 2025. From Punchlines to Predictions: A Metric to Assess LLM Performance in Identifying Humor in Stand-Up Comedy. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Maria Sukhareva. 2025. Why em-dashes are common in llm outputs. https://folkertstijnman.com/blog/why-em-dashes-are-common-in-llm-outputs/. Accessed: 2026-04-08.

Joe Toplyn and Ori Amir. 2025. Can AI Make Us Laugh? Comparing Jokes Generated by Witscript and a Human Expert. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 71–78, Online. Association for Computational Linguistics.

JingJie Zeng, Liang Yang, Jiahao Kang, Yufeng Diao, Zhihao Yang, and Hongfei Lin. 2024. "Barking up the Right Tree", a GAN-Based Pun Generation Model through Semantic Pruning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2119–2131, Torino, Italia. ELRA and ICCL.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr. 2024. Creating a Lens of Chinese Culture: A Multimodal Dataset for Chinese Pun Rebus Art Understanding. *Preprint*, arXiv:2406.10318.

A Survey participants

In total, 64 responses from unique participants were collected. All participants were Master's students in Professional Communication and Management at the University of Antwerp, of which virtually all students were native speakers of Dutch, but were proficient in English due to their multilingual university education.

B Roast example

Bruce Willis. What a career, right? "The Fifth Element," "The Sixth Sense," "The Whole Nine Yards," "Twelve Monkeys," zero Oscars. And it's not just action movies that made Bruce a star. He's actually a great dramatic actor, too. I loved "The Sixth Sense." It's a great movie and it's a really impressive performance. I don't know how you pretended not to be embarrassed while a 10 year old kid acted circles around you, but you did it. And the ending, I did not see that twist coming. I mean, I shouldn't spoil it, but, it's been like 20 years. It's so good. Okay, so at the end of "The Sixth Sense," Bruce goes back to making bad movies.

C Prompt

You are a stand-up comedian, and it is your task to generate a comedy roast about roastee. A roast is a form of comedy in which a specific individual is subjected to jokes at their expense, intended to amuse the event's wider audience. Roasts are intended to honor a specific individual in a unique way. In addition to jokes and insult comedy, such events may also involve genuine praise and tributes. The assumption is that the roastee can take the jokes in good humour and not as serious criticism or insult. A short example of such a roast can be found below, in which {example_roastee} is the subject:

{example_roast}

To generate a comedy roast, you will draw inspiration from the Wikipedia page of {roastee}, which is included below:

{short_wiki_bio}

Your output must only contain the roast, which must be between 50 and 100 words long. The topic of the roast must be about {summary of the human roast snippet}.

Figure 2: Prompt used for the roast generation.

D Temperature

Initial prompt fine-tuning experiments indicated that a temperature setting of 0.7 was optimal, since the model started to misinterpret the task or generate gibberish at higher temperature settings.

E Non-normalized Likert scores

Author	Mean	Std.
Human	2.31	0.36
AI	2.40	0.25
Combined	2.35	0.30

Table 3: Funniness ratings obtained in the survey (mean and standard deviations).