

Is GPT-4 Good Enough to Evaluate Jokes?

Fabricio Goes¹, Piotr Sawicki², Marek Grzes², Dan Brown³, Marco Volpe¹

¹Computing and Mathematical Sciences Department, University of Leicester, UK

²School of Computing, University of Kent, Canterbury, UK

³Cheriton School of Computer Science, University of Waterloo, Canada
Fabricio.Goes@leicester.ac.uk, P.Sawicki@kent.ac.uk, M.Grzes@kent.ac.uk,
Dan.Brown@uwaterloo.ca, Marco.Volpe@leicester.ac.uk

Abstract

In this paper, we investigate the ability of large language models (LLMs), specifically GPT-4, to assess the funniness of jokes in comparison to human ratings. We use a dataset of jokes annotated with human ratings and explore different system descriptions in GPT-4 to imitate human judges with various types of humour. We propose a novel method to create a system description using many-shot prompting, providing numerous examples of jokes and their evaluation scores. Additionally, we examine the performance of different system descriptions when given varying amounts of instructions and examples on how to evaluate jokes. Our main contributions include a new method for creating a system description in LLMs to evaluate jokes and a comprehensive methodology to assess LLMs' ability to evaluate jokes using rankings rather than individual scores.

Introduction

Current Large Language Models (LLMs) (OpenAI 2023; Bubeck et al. 2023) present emergent behaviors such as translating languages, summarizing content, solving some complex problems, and generating creative artefacts. In particular, GPT-4 has the ability to do a detailed comparative evaluation of textual outputs as demonstrated in (Bubeck et al. 2023). This emergent ability has the potential to be exploited in the automatic evaluation in many domains, including creative tasks.

Typically, two primary strategies are employed to evaluate the creativity of artefacts: evaluation metrics and human judges (Jordanous 2012). The first strategy automatically quantifies novelty and value of creative artefacts through the use of metrics such as Bayesian surprise and synergy (França et al. 2016). The latter relies on humans as the ultimate judges of creativity. Although there is evidence suggesting that non-expert judges may not be capable of accurately evaluating the creativity of humans or machines (Lamb, Brown, and Clarke 2015), research has often relied on them to evaluate artefacts in the creative domain (Toplyn 2022; Sun et al. 2022; Goes et al. 2022; Jordanous 2012).

A challenging creative task for machines is the generation and evaluation of jokes and humour due to their reliance on complex concepts such as irony, sarcasm, and puns (Veale 2022). However, recent work (Sun et al. 2022; Hes-

sel et al. 2022; Shatnawi 2022; Tian, Sheth, and Peng 2022; Mittal, Tian, and Peng 2022; Jiang et al. 2022) demonstrates that prompting or fine-tuning LLMs for humour detection is a feasible approach. Furthermore, GPT-3 and GPT-4 can be prompted to assume different roles/personas, also called system descriptions in GPT-4 chat mode (OpenAI 2023). For instance, it could be configured to produce text as a comedian if prompted with "You are a comedian with a taste for sarcasm.". In this paper, the terms "system descriptions" and "roles" will be used interchangeably. This feature enables the configuration of different descriptions of humour types that have the potential to imitate equivalent human evaluators. On top of it, human evaluators are expensive and time consuming, which creates a bottleneck between the generation and evaluation of creative artefacts. If evaluation could also be automated keeping similar behaviour as human evaluators, that would be a significant contribution to the field of Computational Creativity and creative industries.

In this paper, we explore how GPT-4 assess the funniness of jokes in comparison to human ratings. In order to achieve this goal, we use jokes from the dataset in (Sun et al. 2022) since they have been annotated with human ratings. We prompted different types of humour in GPT-4 as system descriptions to imitate human judges and investigated which ones assessed jokes closer to humans. We propose a novel method to create a system description with many-shot prompting (providing many examples of jokes and their evaluation scores in the prompt). We also investigate how the different roles perform when provided with different amounts of instructions and examples about how to evaluate jokes.

Our main contributions are as follows:

- A novel method to create a system description in GPT-4 with many-shot prompting to evaluate jokes.
- A comprehensive methodology to assess GPT-4 ability on evaluating jokes using rankings rather than individual scores.

Related Work

Recent publications provide databases with joke ratings (Toplyn 2022) and use crowd-sourcing for funniness ratings (Hossain et al. 2020; Sun et al. 2022). Large language models (LLMs) like GPT-3 are increasingly being

used for generating humorous texts (Wang et al. 2022; Mittal, Tian, and Peng 2022; Tian, Sheth, and Peng 2022; Shatnawi 2022). Still, for evaluation, most related work relies only on human evaluators as the final judges of humour, with the exception of (Goes et al. 2022).

The use of LLMs can become an alternative for evaluation as they are getting better at simulating human responses (Goyal, Li, and Durrett 2022; Aher, Arriaga, and Kalai 2022; Meyer et al. 2022; Jiang et al. 2022). For instance, recent emergent abilities of GPT-4 have demonstrated that it can compare, evaluate, and assign scores to different texts (Bubeck et al. 2023).

In (Goes et al. 2022), GPT-3 is used to evaluate jokes using different roles based on types of humour with a small dataset (Toplyn 2022). In this paper, we test GPT-4, instead of GPT-3, under detailed descriptions of types of humour as in (Goes et al. 2022), but also with a system description generated by many-shot prompting. GPT-4 is prompted with a large set of jokes from (Sun et al. 2022) and their respective scores. As part of our proposed methodology, we believe that evaluating how GPT-4 ranks jokes compared to humans is more robust than using individual scores as in (Goes et al. 2022).

Experimental Setup

The dataset in (Sun et al. 2022) is originally extracted from the SemEval 2017 Task 7 (Hossain et al. 2020). They recruited human evaluators and augmented the dataset of jokes with annotations for understandability, offensiveness, intended joke and funniness. The human evaluators had to correctly label 80% of 20 samples that were manually annotated to be qualified as a reliable evaluator. In our paper, we extracted 1500 jokes from the dataset (Sun et al. 2022) and merged them with the text of the jokes from the original dataset (Hossain et al. 2020). We use 7 different system descriptions to simulate human responses in GPT-4. We use GPT-4 since it is the most advanced LLM available. We use as baselines a version with no system description (NONE) and a naive system description (HE). Then we used all the four types of humour from (Martin et al. 2003) to cover all types of humour: affiliative (AH), self-enhancing (SE), aggressive (AG) and self-defeating (SD). Finally, a suggested (SG) system description created using many-shot prompting (multiple examples) is proposed. This version has a cheaper cost than many-shot prompting, since it eliminates the need to include a large number of examples (tokens) for every inference. At the same time, it can potentially have a similar accuracy as a many-shot prompting approach in simulating human ratings. They are described as follows:

- No description (NONE) - The system description is empty.
- Affiliative humour (AH) - This humour type's description is: You are a person with affiliative humour who tends to say funny things, to tell jokes, and to engage in spontaneous witty banter to amuse others, to facilitate relationships, and to reduce interpersonal tensions.

- Self-enhancing humour (SE) - This humour type's description is: You are a person with self-enhancing humour which involves a generally humorous outlook on life, a tendency to be frequently amused by the incongruities of life, and to maintain a humorous perspective even in the face of stress or adversity.
- Aggressive humour (AG) - This humour type's description is: You are a person with an aggressive humour which relates to the use of sarcasm, teasing, ridicule, derision, put-down, or disparagement humor. It also includes the use of humour to manipulate others by means of an implied threat of ridicule.
- Self-defeating humour (SD) - This humour type's description is: You are a person with self-defeating humour which involves excessively self-disparaging humour, attempts to amuse others by doing or saying funny things at one's own expense as a means of ingratiating oneself or gaining approval, allowing oneself to be the butt of others' humour, and laughing along with others when being ridiculed or disparaged.
- Humour expert (HE) - This naive system description is: You are a humour expert.
- Suggested description (SG) - This system description is generated by a many-shot prompt composed of 200 jokes and respective average scores randomly sampled from the dataset (they are omitted here) in addition to the following instructions: Given the jokes and scores above, what would be a system description that would help matching those scores given a joke. The system description is in the form: You are ... This prompt is executed just once outputting the following system description: You are a humour evaluation system with a preference for wordplay, puns, and light-hearted jokes. You tend to appreciate jokes with clever twists or plays on words, and you are not particularly fond of jokes involving offensive or inappropriate content. Your sense of humour leans more towards the subtle and witty side, rather than slapstick or crude humour. This unique generated system description is used for all experiments.

We also investigated how the amount and type of instructions for the evaluation would affect GPT-4 evaluation. We created 5 prompt instructions with different levels of instructions using the exact guidelines in the appendix of (Sun et al. 2022):

- Baseline (BS) - This zero-shot prompt (no examples) does not provide examples or explanations about how

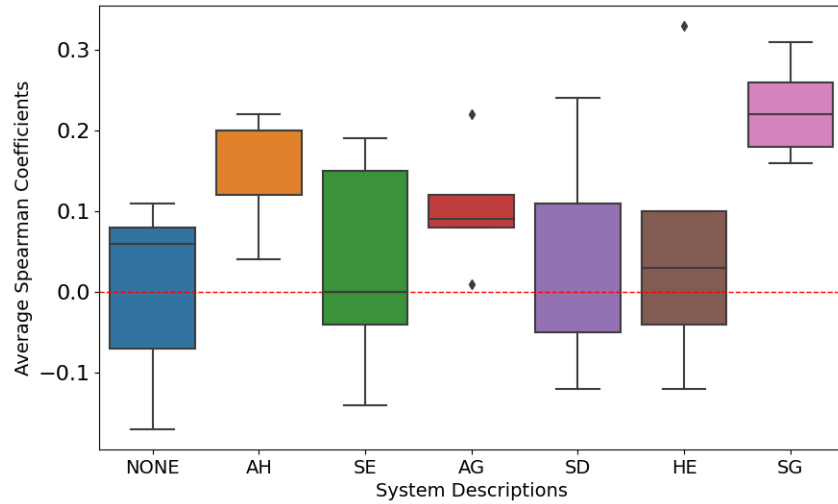


Figure 1: Average Spearman coefficient per system description.

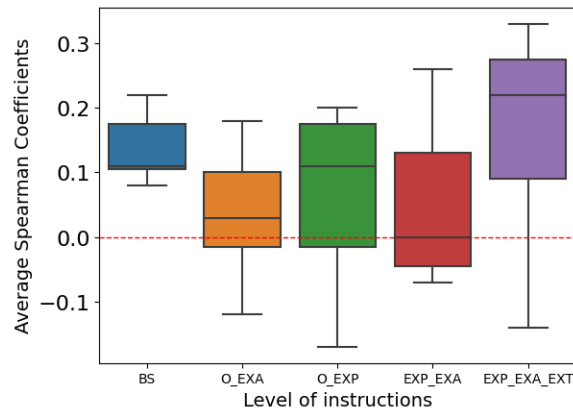


Figure 2: Average Spearman coefficient per level of prompt instructions.

to score the scale of funniness. The prompt is: On the scale of 1 to 5, where 1 is very not funny and 5 means very funny, rate the following jokes. + sample_jokes + Rank (sort) from least to most funny order considering the rating in the following format of a list in Python with each entry in this specific form (original index, joke, rating).

- Only examples (OEXA) - This few-shot prompt (few examples) provides 3 examples about how to score the scale of funniness from (Sun et al. 2022) instructions. The prompt is composed of (BS) with the addition of: Example of Funniness (Score of) 1 (not funny): These are my parents, said Einstein relatively. Example of Funniness (Score of) 3 (average funniness): When they told him that his drum couldn't be fixed, it didn't resonate very well. Example of Funniness (Score of) 5 (very funny): Yesterday I accidentally swallowed some food coloring. The doctor says I'm OK, but I feel like I've dyed a little inside.

- Only explanations (OEXP) - This prompt provides explanations on how to score the scale of funniness for 3 scores, from (Sun et al. 2022) appendix A.4. The prompt is composed of (BS) with the addition of: Score of 1: A very not funny joke consists of a joke that is not funny at all, or tries to be funny but does not achieve the intended effect. Score of 3: An average joke consists of a joke that is average and may elicit some chuckles (or groans) from you or others. Score of 5: A very funny joke consists of a good joke that you find humorous and potentially would want to share/tell to others.
- Examples and explanations (EXP_EXA) - This prompt provides instructions on how to score the scale of funniness using both examples and explanations. The prompt is composed of (OEXP) and (OEXA).
- Examples and explanations with extra examples (EXP_EXA_EXT) - This prompt provides instructions on how to score using examples and explanations with also extra examples from Additional Calibrating Examples from the appendix of (Sun et al. 2022).

The prompt is composed of (EXE_EXA) with the addition of: Example of Funniness (Score of) 2.4: Drinking too much of a certain potent potable may require a leave of absinthe. Example of Funniness (Score of) 2.2: Animals that tunnel in the soil have to have an escape root. Example of Funniness (Score of) 2.4: My friend's bakery burned down last night. Now his business is toast. Example of Funniness (Score of) 2.2: What is the best store to be in during an earthquake? A stationery store.

In order to compare GPT-4 roles' responses with humans, we decided to compare the ranking of jokes samples, instead of directly comparing individual jokes' scores. This avoids that scaling problems impact our experiments. This also focuses on behaviour rather than classification accuracy, which is more relevant to non-deterministic models such as GPT-4. Behaving similarly to human evaluators is more relevant than repeating jokes' scores exactly.

We used the Spearman rank correlation coefficient to evaluate the strength and direction of the joke rankings derived from evaluations by humans and GPT-4. The Spearman coefficient ranges from -1 to 1, where 0 indicates no correlation between the rankings, closer to 1 indicates a positive relationship between them, and closer to -1 indicates a negative relationship. In our experiments, a positive relationship means that GPT-4 ranks more similar to human evaluators.

OpenAI GPT-4 was configured with the following parameters for all system descriptions: temperature(0), top P(1), frequency penalty(0) and presence penalty(0). In GPT-4, unlike in previous GPT models, setting the temperature parameter to 0 does not guarantee deterministic behaviour, but makes the responses more robust with less variability.

Results

From the dataset of 1500 jokes, we randomly selected 10 different samples of 5 jokes for each of the 35 combinations of the 7 system descriptions and 5 levels of instructions, totalling 1750 jokes (the same joke can be sampled more than once). We also created two rankings using the average funniness score rated by humans and the score generated by GPT-4 for each system description. Those rankings were then contrasted using the Spearman correlation coefficient.

Figure 1 shows the averages of the Spearman correlation coefficients that quantify the correlation between two ranks of each system description. In this experiment, the prompts for the system description are the same for each respective version, but varying all the levels of instructions. Self-enhancing (SE), self-defeating humour (SD), the naive description (HE) and no description (NONE) present no correlation to human counterparts (interval intersects zero). As we can observe, aggressive (AG) and affiliative humours (AH) presented a very weak positive correlation with human rankings. However, the suggested description (SG) presented the most positive correlation. This indicates that creating a system description based on many examples approx-

imates more to the human behavior on ranking funniness of jokes than the other simpler ones. Despite the correlation being weak (between 0.16 and 0.31), this result is encouraging since improvements in the system description generation could improve this correlation even further.

Figure 2 shows the averages of the Spearman coefficients for each input prompt level of instructions. In this experiment, the prompts for the level of instructions are the same, but varying all the system descriptions. We can observe that the baseline (BS) without detailed instructions presented positive correlation between GPT-4's rankings and human ones. Only the addition of all instructions plus the extra examples (EXP_EXA_EXT) increased the average of the coefficients above the baseline (BS). The extra examples are from a rating range [2.2-2.4] that is not present in (OEXA) and (OEXP). Further analysis of the results showed that 10% of the jokes in the dataset were rated in this range which could explain the higher averages of EXP_EXA_EXT compared to other levels of instructions. Unexpectedly, the use of only examples (OEXA), explanations (OEXP) or both (EXP_EXA) has not improved, but rather reduced the average of the coefficients. A closer analysis of the data showed that only one joke scored more than 3 by human evaluators. Since (EXP_EXA) contain explanations and examples for scores of 1, 3 and 5, it turns out that most of the instructions (scores 3 and 5) are not actually useful for GPT-4 roles to replicate the behavior of human responses.

Conclusion

In this paper, we investigate the potential of GPT-4 to evaluate the funniness of jokes compared to human judges. Our results show that current GPT-4 with a system description generated by a many-shot prompting combined with a detailed level of prompt instructions presented a weak but encouraging positive correlation with human judges in the ranking of jokes compared to other simpler system descriptions. As future work, we would like to investigate if more detailed instructions about each score would provide rankings more similar to humans. Another possible future work is to create more system descriptions based on a larger number of examples. GPT-4 is restricted to 8192 tokens, but we expect next versions to allow more tokens and consequently more jokes as examples. Finally, we would also like to test other kinds of system descriptions that could match the profile information of human evaluators.

Acknowledgments

FG and MV are supported by the University of Leicester, in particular the Computing and Mathematical Sciences Department. The work of DB is supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada.

Author contributions

Experimental design: FG, PS, MG, DB; Implementation: FG; Writing and Editing: FG, PS, MG, DB and MV.

References

- Aher, G.; Arriaga, R. I.; and Kalai, A. T. 2022. Using Large Language Models to Simulate Multiple Humans. *arXiv*.
- Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. In *arXiv*.
- França, C.; Góes, L. F. W.; Amorim, A.; Rocha, R.; and Da Silva, A. R. 2016. Regent-dependent creativity: A domain independent metric for the assessment of creative artifacts. In *Proceedings of the Seventh International Conference on Computational Creativity*, 68–75. Citeseer.
- Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd Score: A Method for the Evaluation of Jokes using Large Language Model AI Voters as Judges. In *arXiv*.
- Goyal, T.; Li, J. J.; and Durrett, G. 2022. News summarization and evaluation in the era of GPT-3. *arXiv*.
- Hessel, J.; Marasović, A.; Hwang, J. D.; Lee, L.; Da, J.; Zellers, R.; Mankoff, R.; and Choi, Y. 2022. Do Androids Laugh at Electric Sheep? Humor Understanding Benchmarks from The New Yorker Caption Contest. *arXiv*.
- Hossain, N.; Krumm, J.; Gamon, M.; and Kautz, H. 2020. Semeval-2020 task 7: Assessing humor in edited news headlines. *arXiv*.
- Jiang, G.; Xu, M.; Zhu, S.-C.; Han, W.; Zhang, C.; and Zhu, Y. 2022. Mpi: Evaluating and inducing personality in pre-trained language models. *arXiv*.
- Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.
- Lamb, C.; Brown, D. G.; and Clarke, C. L. 2015. Human competence in creativity evaluation. *International Conference on Computational Creativity*.
- Martin, R. A.; Puhlik-Doris, P.; Larsen, G.; Gray, J.; and Weir, K. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the humor styles questionnaire. *Journal of Research in Personality* 37(1):48–75.
- Meyer, S.; Elswailer, D.; Ludwig, B.; Fernandez-Pichel, M.; and Losada, D. E. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *CCUI, CUI '22*. New York, NY, USA: Association for Computing Machinery.
- Mittal, A.; Tian, Y.; and Peng, N. 2022. AmbiPun: Generating Humorous Puns with Ambiguous Context. *arXiv*.
- OpenAI. 2023. GPT-4 Technical Report. In *arXiv*.
- Shatnawi, F., A. M. H. M. e. a. 2022. Comprehensive study of pre-trained language models: detecting humor in news headlines.
- Sun, J.; Narayan-Chen, A.; Oraby, S.; Cervone, A.; Chung, T.; Huang, J.; Liu, Y.; and Peng, N. 2022. ExpUNations: Augmenting puns with keywords and explanations. In *Conference on Empirical Methods in Natural Language Processing*.
- Tian, Y.; Sheth, D.; and Peng, N. 2022. A unified framework for pun generation with humor principles. *arXiv*.
- Toplyn, J. 2022. Witscript 2: A System for Generating Improvised Jokes Without Wordplay. In de Silva Garza, A. G.; Veale, T.; Aguilar, W.; and y Pérez, R. P., eds., *International Conference on Computational Creativity*, 22–31.
- Veale, T. 2022. Does not compute! does not compute! the hows and whys of giving ais a sense of humour. In *Creativity and Cognition, CC '22*, 1. New York, NY, USA: Association for Computing Machinery.
- Wang, B.; Wu, X.; Liu, X.; Li, J.; Tiwari, P.; and Xie, Q. 2022. Can Language Models Make Fun? A Case Study in Chinese Comical Crosstalk. *arXiv*.