The 7th International Conference on Interdisciplinary Research
on Computer Science, Psychology, and Education (2023)

A Study on a Large Language Model's Ability to Solve Riddles

Sugyeong Eo[1], Sungmin Ahn[2], Jeongbae Park[3]*
[1]Department of Computer Science and Engineering, Korea University, [2]O2O Inc. [3]Human-inspired AI Research
{djtnrud,insmile}@korea.ac.kr, {smahn}@o2o.kr

Abstract

Recently, Large Language Models (LLMs) have been attracting global attention by displaying strong performance and high user satisfaction. In this paper, we explore the wit of GPT-4 by utilizing riddles. We investigate how the GPT-4 model responds to riddle questions. Through error type analysis, we conclude that GPT-4 possesses creativity while also having an issue of hallucination.

Keywords: Natural Language Processing, Language Model, Large Language Model, Riddle, Creativity

1. Introduction

Recently, the emergence of large language models (LLMs) has garnered global attention due to their impressive performance improvements and user satisfaction. LLMs have showcased superior performance in both quantitative and qualitative assessments, as well as in human-like evaluations, compared to pre-trained language models. Notably, various LLMs such as T5 [1], GPT-3 [2], OPT [3], FLAN-T5 [4], LLaMA [5, 6], InstrucGPT [7], ChatGPT [8], and GPT-4 [9] have sequentially emerged, consistently breaking records in tasks such as question-answering, conversational systems, and summarization. In this paper, in terms of creativity, we explore the wit of GPT-4 [9]. Wit is an ability grounded in abstract thinking and creativity, and this paper aims to evaluate whether GPT-4 possesses such capabilities. We choose riddles as the evaluation task, as wit is connected to intelligent humor and wordplay. Riddles are composed of witty and humorous question-and-answer pairs, and the model's inference results for the questions can be used to judge GPT-4's creativity and abstract thinking abilities [10, 11, 12, 13, 14].

The experiments in this paper are conducted by presenting riddle questions to GPT-4 in the form of prompts and obtaining results through model inference. Through this, we aim to analyze the types of errors generated by GPT-4.

We confirm that while GPT-4 has considerable creativity, it still cannot escape issues like hallucination. Thus, we conclude that there is a need to perform tasks using wit appropriately without inducing hallucination.

2. Methods

The riddle data is collected through web crawling. For the model, we leverage GPT-4, a state-of-the-art model as far as we know. We input the query of the riddle with a definition of the "riddle" and its data samples. For the number of data samples, we select 16 shots after the empirical experiments.

3. Experiments

We analyzed the types of errors made by GPT-4, and the results are shown in Table 1.

In the first row in the Table, the GPT-4 outputs an existing entity as the inference result; however, this is completely unrelated to the query. Conversely, in the second row, it outputs the term "유돈날" which appears to be related to the query but is a non-existent entity, thus inducing hallucination.

There are also results where an entity that was neither relevant nor existent was generated; for example, when asked about the hottest fruit in the world, the model responds with "파이낸초(파이어+망고)" which is completely irrelevant. Through this, we confirm that the model generates multiple hallucinations when it cannot solve problems using its intrinsic knowledge. Hallucination is a consistent issue

*Corresponding author

Question	Answer	LLM Inference Results	Relevance to Query	Entity Existence	-
직접 만든 총은?	손수건	셀프카메라	X	O	
설날에 용돈을 하나도 못받으면?	설거지	유돈날	O	X	Hallucination
세상에서 가장 뜨거운 과일은?	천도복숭아	파이낸초(파이어+망고)	X	X	Hallucination
아마존의 창업자는?	아마존	제프 베조스	O	O	Common-sence knowledge-based answer
걸어가면서 길 위에 도장 찍는 것은?	지팡이	발자국	O	O	Answer with wit

Table 1. Error analysis on the GPT-4 model's response

with LLMs and can occur when thinking variously, such as creativity or wit, leads to excessive divergence. Therefore, we conclude that responses should be generated that possess appropriate wit without threatening consistency, fluency, and relevance.

The fourth row is an incorrect answer considered from a perspective that does not base the task understanding on riddles; in other words, the LLM was considered incorrect for generating a common-sense answer rather than a witty one. Finally, although it was considered incorrect through EM and F1 scores in the fifth row, qualitative analysis find that the model has a considerably excellent ability for wit and inference. Therefore, we conclude that, overall, the LLM can perform tasks that go beyond simple tasks and can generate answers that appropriately utilize inference and wit.

4. Conclusion

This paper conducts an exploration of the wit of GPT-4 based on their ability to solve riddles. The analysis found that while some of the responses from GPT-4 were considered incorrect, they were notably creative and witty. These results have been validated through various examples, and it is assumed that the GPT-4 generated them based on its capabilities in creativity, abstract thinking, wit, and reasoning. Although the model shows excellent capabilities in divergent thinking, excessive divergence leads to hallucinations and generates non-existent entities or expressions. Therefore, we conclude that there is a need to perform tasks requiring creativity while appropriately utilizing wit without causing hallucinations.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425) and this work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2022R1A5A7026673). Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

Reference

[1] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, Vol. 21, No. 1, pp. 5485–5551, 2020.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[3] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[6] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-

tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[7] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, Vol. 35, pp. 27 730–27 744, 2022.

[8] OpenAI-Blog, "Chatgpt: Optimizing language models for dialogue," 2022. [Online]. Available: https://openai.com/blog/chatgpt/

[9] OpenAI, "Gpt-4 technical report," 2023.

[10] R. A. Georges and A. Dundes, "Toward a structural definition of the riddle," *The Journal of American Folklore*, Vol. 76, No. 300, pp. 111–118, 1963.

[11] Y. Zhang and X. Wan, "Birdqa: A bilingual dataset for question answering on tricky riddles," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 10, pp. 11 748–11 756, 2022.

[12] 신명선, "수수께끼의 메타언어적 성격에 대한 국어 교육적 고찰," *국어교육*, No. 110, pp. 4–90, 2003.

[13] 김열규, "수수께끼라는 언어 전략이 텍스트 상관성에 던지는 문제 몇 가지," *배달말*, Vol. 14, pp. 315–336, 1989.

[14] 안혜리, 황민아, and 최경순, "학령기 경계선지능아동의 수수께끼 유머 이해 능력," *특수교육논총*, Vol. 37, No. 4, pp. 27–41, 2021.