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

Assessing the Retrieval-based Generation Capabilities of Large Language Models: A Call for a New Benchmark

Jungseob Lee[1◦], Junyoung Son[1◦], Taemin Lee[2◦], Chanjun Park[3], Myunghoon Kang[1], Yuna Hur[2*]
[1]Department of Computer Science and Engineering, Korea University, [2]Human-inspired AI Research, [3]Upstage
{omanma1928, s0ny, taeminlee}@korea.ac.kr,
{bcj1210}@naver.com, {chaos8527, yj72722}@korea.ac.kr

Abstract

This paper proposes the necessity for a new benchmark to evaluate the retrieval-based generation capabilities of large language models. By assessing the ability of existing large language models to generate answers based on untrained fictitious information, we aim to highlight the issue that these models may not accurately reflect the capabilities of real-world retrieval-based answer systems, as they generate responses based on facts seen during pre-training. Hence, an accurate evaluation method is crucially needed.

Keywords: Large Language Models, Hallucination, Retrieval-based Generation

1. Introduction

The emergence of large language models (LLMs) like ChatGPT, capable of understanding and generating human-like language, has seen an increasing trend in their application across various services [1, 2].

Despite the advent of these LLMs, there are limitations to their use, such as in the case of chatbots for specific companies. This is because LLMs generate responses based on a common range of knowledge, which may not be sufficient for specialized applications [3, 4]. Additionally, LLMs suffer from the problem of hallucination, wherein the model generates information or facts that don't actually exist [3].

To overcome this issue, retrieval-based LLMs that incorporate a restricted set of documents into the prompt and generate responses from these documents are extensively utilized in the industry[1] [1, 5].

However, there is no existing benchmark for selecting such retrieval-integrated LLMs. Even if such a benchmark were developed, it would be challenging to accurately evaluate whether the LLM inferred the answer from its inherent learned information or skillfully extracted the relevant information from the given document in the prompt. This problem persists even when the benchmark is based on factual, up-to-date documents, as future LLMs may infer results from the data embedded in the benchmark, making it difficult to accurately evaluate the model's retrieval-based generation performance.

This paper discusses the design of benchmarks considering various real-service cases to provide insights for appropriate service selection based on specific objectives.

2. Suggestion of Retrieval-based Benchmarks

This paper advocates for the necessity of a novel benchmark to assess the retrieval-based answer generation capabilities of large-scale language models. It proposes the creation of a dataset based on fictitious facts, as opposed to actual ones. This is intended to compensate for the inadequacy of models that generate answers reflecting facts encountered during pre-training, which may not accurately evaluate the actual proficiency of retrieval-based answer systems.

To examine the retrieval generation performance of large-scale language models across multiple cases, the necessity for five cases is argued: long document, two gold documents containing answers, one gold document and similar distractor document (with keywords), one gold document and similar distractor document (without keywords), and cross-reference multi-hop reasoning.

2.1 Long Document

The long document case targets situations where a single long (over 800 characters) gold document is given. This case

*Corresponding author
[1]https://tools.zmo.ai/webChatGPT

is intended to evaluate the language model's ability to effectively extract and comprehend necessary information from lengthy documents.

2.2 Two Gold Documents Containing Answers

The two gold documents containing answers case targets situations where two gold documents, each containing an answer, are given. This case checks whether the language model can properly reference one or more of the two gold documents when multiple pieces of relevant information are retrieved for answer generation.

2.3 Gold Document and Similar Distractor Document with Keywords

The case of one gold document and similar distractor document with keywords targets situations where a single gold document and a similar distractor document that would likely be retrieved in a retrieval system are given. This case assumes that the retrieval model has retrieved the correct document and a corresponding candidate document, and both documents contain keywords.

2.4 Gold Document and Similar Distractor Document without Keywords

The case of one gold document and similar distractor document without keywords also targets situations where a single gold document and a similar distractor document that would likely be retrieved in a retrieval system are given, similar to the previous case. However, in this case, the documents do not contain keywords.

3. Conclusion

In this paper, we advocate for the need for a new benchmark to evaluate the retrieval-based answer generation capability of large-scale language models. This benchmark aims to assess the pure retrieval-based answer generation ability of the model based on fictitious facts, not real ones, on which the model has not been pre-trained. The benchmark should be structured to comprehensively judge five cases. Through this, it is possible to evaluate whether large-scale language models, when combined with a retrieval system, can effectively retrieve relevant information from the given document information without using the model's inherent information. We anticipate that this benchmark will allow for a more accurate evaluation of the retrieval-based answer generation ability of large-scale language models, facilitate their use according to the intended purpose, and suggest directions for the improvement of the latest large-scale language models.

However, there are clear limitations to the development of such a benchmark. Firstly, a benchmark based on fiction can reinforce the illusion knowledge that large-scale language models encounter. Secondly, the performance of the benchmark may vary depending on the various prompts provided by the large-scale language models. Thirdly, the benchmark is designed as a multiple-choice format for easy evaluation, which does not fully assess the model's accurate generation ability.

4. Acknowledgement

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation). 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).

Reference

[1] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.

[2] N. M. S. Surameery and M. Y. Shakor, "Use chat gpt to solve programming bugs," *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, Vol. 3, No. 01, pp. 17–22, 2023.

[3] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, Vol. 13, No. 4, p. 410, 2023.

[4] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding chatgpt: A taxonomy of existing research, current challenges, and possible future directions," *Journal of King Saud University-Computer and Information Sciences*, p. 101675, 2023.

[5] S. Liu, J. Wang, Y. Yang, C. Wang, L. Liu, H. Guo, and C. Xiao, "Chatgpt-powered conversational drug edit-

ing using retrieval and domain feedback," *arXiv preprint arXiv:2305.18090*, 2023.