Qeios

Peer Review

Review of: "Creating a Biomedical Knowledge Base by Addressing GPT's Inaccurate Responses and Benchmarking Context"

Jie Zheng[1]

1. School of Information Science and Technology, ShanghaiTech University, Shanghai, China

This paper presents GNQA, a biomedical knowledge base that leverages a retrieval augmented generation (RAG) system driven by GPT to address the challenge of efficiently discovering and summarizing scientific findings in the fields of aging, dementia, Alzheimer's, and diabetes. A key innovation is the implementation of a context provenance tracking mechanism that allows researchers to validate responses against original material and obtain references to the original papers, thereby reducing inaccuracies and 'hallucinations' in GPT-generated responses. The study also introduces RAGAS, a combined human expert and AI-driven evaluation system, to measure the effectiveness of RAG systems, achieving high scores in answer relevance and faithfulness. The GNQA knowledge base is integrated into the GeneNetwork web service, providing a valuable resource for researchers and enabling continuous performance assessment through benchmarking.

Specific comments:

1. In the initial part of the introduction, it was mentioned that efficiently exploring and summarizing the constantly evolving scientific discoveries poses a significant challenge. However, this framework utilizes **outdated** research literature as the database for RAG, and the answers to questions are constrained to the knowledge provided by these literatures, lacking the capacity to access newly discovered knowledge. Under such circumstances, how can we address the aforementioned challenge? It is contended that only by enabling the RAG database to be updated in real time and continuously supplemented with new literature can we truly achieve the intended objectives.

doi.org/10.32388/F1Y39Z

2. When conducting performance evaluation using GPT, it is inevitable that the generation of large language models is unstable, and the evaluation results may vary each time. As such, repeating the experiment for each score and taking the average would make the results more persuasive. The paper does not show whether repeated experiments were carried out. If not, it is recommended to conduct supplementary experiments.
3. This study uses RAG to build a dialogue knowledge base in the biomedical field and comprehensively evaluates the performance of the database in answering questions. It combines GPT evaluation and expert evaluation and conducts detailed analysis and discussion of the evaluation results. However, the overall study lacks innovation. The main RAG and RAGAS parts in the framework are from existing research. Moreover, the study does not compare the results with using only GPT to answer questions, failing to justify the necessity of the research. If GPT itself can answer well enough, is there a need to use RAG? Although RAG can indeed address the hallucination issue of large language models, this paper has not proven through experiments that using RAG is truly better, thus lacking persuasiveness.

Declarations

Potential competing interests: No potential competing interests to declare.