Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240498

PubMed Retrieval with RAG Techniques

Alex THOMO[a,1]
[a] *University of Victoria, USA*
ORCiD ID: Alex Thomo https://orcid.org/0000-0002-3020-2258

Abstract. This study explores the application of Retriever-Augmented Generation (RAG) in enhancing medical information retrieval from the PubMed database. By integrating RAG with Large Language Models (LLMs), we aim to improve the accuracy and relevance of medical information provided to healthcare professionals. Our evaluation on a labeled dataset of 1,000 queries demonstrates promising results in answer relevance, while highlighting areas for improvement in groundedness and context relevance.

Keywords. Retriever-Augmented Generation (RAG), LLM, PubMed

1. Introduction

The rapid advancement of medical knowledge challenges healthcare professionals to stay current with research findings and guidelines. Traditional information retrieval methods often struggle to effectively extract and formulate knowledge from document collections, leading to gaps in medical insight application. Retriever-Augmented Generation (RAG) with Large Language Models (LLMs) presents a promising solution to this issue. RAG combines retrieval systems with generative models to enhance the quality of the information retrieval [3]. Our study leverages RAG to improve medical information retrieval from a PubMed database [2], evaluating its effectiveness through metrics such as groundedness, context relevance, and answer relevance on a dataset of 1,000 PubMed queries. The results show sensible answer relevance, highlighting RAG's potential to bridge the gap between medical research and knowledge acquisition for medical personnel.

2. Methods

Our implementation[2] focused on developing RAG systems to effectively update doctors' knowledge. Our RAG model embeds chunks of our data collection into vectors using GPT-3.5, stores them in a vector database, and retrieves the most relevant chunks based on a query prompt. These chunks are then provided as context for answering the query. We employed a labeled dataset of 1,000 queries from PubMed to evaluate the performance of our model. The evaluation metrics included groundedness, context relevance, and answer relevance, which were computed using TruEra [1].

Groundedness measures the authenticity and accuracy of retrieved information, emphasizing the importance of credible sources for medical topics. The answers should

[1] Corresponding Author: Alex Thomo; E-mail: thomo@uvic.ca.

[2] https://github.com/thomouvic/PubmedRAG.

reflect current research findings. Contextual relevance assesses the system's ability to understand query nuances and retrieve precise information, focusing on information specific to the query's context. Answer relevance evaluates whether the answer is factually correct and addresses the core concern of the query, involving the generator's skill in incorporating relevant information into a coherent and responsive answer.

3. Results

The evaluation results for the standard RAG model showed an answer relevance score of 0.87, indicating a high degree of accuracy in addressing the core concerns of the queries. However, the groundedness score of 0.52 suggests room for improvement in the authenticity and accuracy of the retrieved information. Context relevance scored 0.26, highlighting a need for better understanding of query nuances. We further investigated advanced RAG techniques such as Sentence Window Retrieval [4] and Auto-Merging Retrieval [3]. Surprisingly, these methods did not outperform the baseline system, with Sentence Window Retrieval achieving a groundedness score of 0.11 and a context relevance score of 0.12, while Auto-Merging Retrieval scored 0.26 in groundedness and 0.23 in context relevance. This could be due to the nature of the PubMed dataset, where article excerpts are typically short and lack interconnectivity. The results indicate that while RAG shows promise in improving answer relevance in medical information retrieval, there is a need for further refinement in groundedness and context relevance. The lack of improvement with advanced RAG techniques suggests that further research is needed to optimize these methods for medical applications.

4. Conclusions

This study showcases the potential of Retriever-Augmented Generation (RAG) in enhancing medical information retrieval, crucial for healthcare professionals to keep pace with rapid advancements in medical research. Our results indicate significant answer relevance, demonstrating RAG's ability to bridge the gap between extensive medical research and knowledge acquisition for medical personnel.

However, there is room for improvement in groundedness and context relevance, essential for ensuring the accuracy and specificity of retrieved information. Future efforts should focus on refining these aspects to further optimize RAG's performance in medical information retrieval. By addressing these challenges, we can advance towards providing healthcare professionals with timely, accurate, and contextually relevant medical information, thereby improving clinical decision-making and patient care.

References

[1] Truera. A suite of tools for evaluating rag systems. 2023.
[2] Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146. 2019 Sep 13.
[3] Ke Y, Jin L, et al. Development and Testing of Retrieval Augmented Generation in Large Language Models--A Case Study Report. arXiv preprint arXiv:2402.01733. 2024 Jan 29.
[4] Olivares DG, Quijano L, Liberatore F. Enhancing Information Retrieval in Fact Extraction and Verification. InProceedings of the Sixth Fact Extraction and VERification Workshop (FEVER) 2023 May (pp. 38-48).