2024 International Conference on AI x Data and Knowledge Engineering (AIxDKE) | 979-8-3315-1704-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIXDKE63520.2024.00032

Retrieval-Augmented Generation for Pharmacopoeia: Application and Evaluation

Rizaldi Fahmi*, Soojin Cheon†, Yesol Park†, Joonho Kwon‡
* *Dept. of Information Convergence Engineering*, ‡ *Dept. of Data Science*,
Pusan National University
{fahmi.rizaldi, jhkwon}@pusan.ac.kr
† *AI Strategy Team*, Daewoong
{soojin3108, heute0201}@gmail.com

***Abstract*—Pharmacopoeia documents are used in pharmaceutical companies as references for ensure safe handling of chemical compounds. However, manual review of these documents can be time-consuming and prone to human error. To assist users in working with these documents, we designed PharmChat, a retrieval-augmented generation (RAG)-based system which enables efficient question-and-answer interactions with multiple short pharmacopoeia documents using natural language. In contrast to other conventional RAG systems, PharmChat converts original pharmacopoeia into separate documents to avoid context mixing during document chunking. PharmChat leverages embedding-based similarity searches between the documents as knowledge-base and user queries using an open-source vector database, ChromaDB. Upon receiving a user query, the most relevant chunks are retrieved and provided as context for a large language model (LLM) to be used as reference to minimize hallucination. Experimental results with real pharmacopoeia short documents demonstrate that our PharmChat system provides quick response times and achieves a BERT-F1 score of 76%.**

***Index Terms*—Document Search, Retrieval-Augmented Generation, Large Language Model**

I. Introduction

A pharmacopoeia is a comprehensive document that provides methods for identifying drugs and medicinal compounds, serving as a key reference for pharmaceutical companies in drug development and testing. Large language model (LLM)-based chatbots can help reduce inefficiencies and human error by enabling document querying, but LLMs in specialized domains like pharmacopoeias are prone to hallucinations. Retrieval-Augmented Generation (RAG) [1] mitigates these issues by retrieving relevant context, leading to more accurate responses. While RAG has shown promise in various tasks, no research has yet applied it to pharmacopoeia documents. Prior studies have used naive data preprocessing that disregards document structure, which is unsuitable for pharmacopoeias where context integrity for each substance is essential. To address this, we developed PharmChat, the first RAG-based chatbot tailored for pharmacopoeia review, introducing a document-splitting technique that preserves context. Experimental results on pharmacopoeia datasets demonstrate fast response times and sufficient retrieval precision and recall for short document contexts.

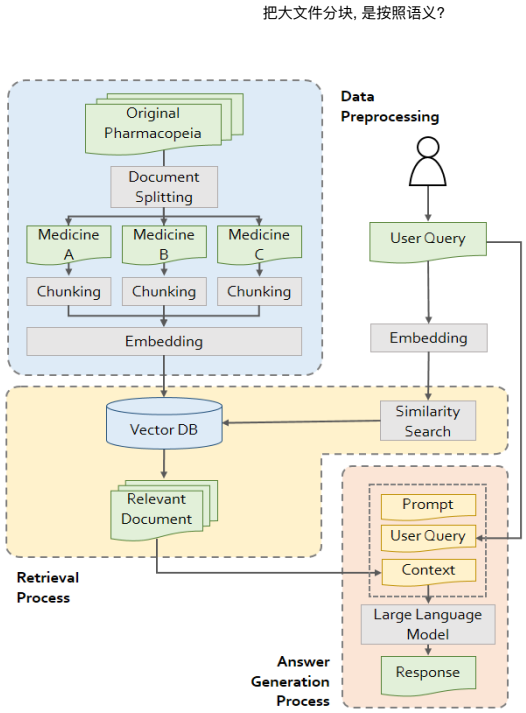

Fig. 1. Architecture of PharmChat RAG system

II. The Proposed PharmChat System

Figure 1 illustrates the architecture of the PharmChat system, which consists of three stages: Data Preprocessing, Retrieval Process, and Answer Generation. Data Preprocessing, performed once at setup and repeated only with new pharmacopoeia documents, involves converting documents into high-dimensional vectors for semantic similarity searches. To manage multiple substances on a single page, documents are split so that each substance is treated independently, avoiding context mixing. Chunking with overlap is applied to maintain context across segments.

The retrieval process embeds the user query and searches for the most relevant document chunks in the vector database using cosine similarity, returning these chunks as context for

2831-7203/24/$31.00 ©2024 IEEE
DOI 10.1109/AIxDKE63520.2024.00032

the LLM. In the final Answer Generation stage, three inputs are provided to the LLM: (1) a System Prompt defining the task, (2) the User Query, and (3) the retrieved document Context. The GPT-4 variant of OpenAI's ChatGPT model then generates the response.

III. Experimental Results

A. Dataset and Metrics

The dataset comprises internal pharmacopoeia documents that define and outline testing methods for various medical substances. These documents vary in length, with some pages covering multiple compounds, adding complexity to document processing.

We evaluate the retrieval system using Mean Average Precision (`MAP`) and Recall, where `MAP` assesses precision with a focus on document ranking, and recall measures the system's ability to retrieve all relevant information. The quality of the final response is measured using `ROUGE` [2] and `BERTScore` [3]. Specifically, `ROUGE-L` considers the longest common sequences between generated and reference text, which is essential in pharmacopoeia contexts, while `BERTScore` captures semantic similarity between embeddings.

B. Result and Discussion

TABLE I
Retrieval task evaluation

Metric	Result
MAP@4	0.937
Recall@4	0.486
MAP@10	0.858
Recall@10	0.618

1) Retrieval Task: Table I presents the retrieval results for different top@K document values. High `MAP` scores (MAP@4 = 0.937) indicate effective ranking of relevant documents, but low recall scores (Recall@4 = 0.486) suggest missing contexts for some queries. In pharmacopoeia, missing context can lead to inaccuracies, as full document context is often needed to answer queries accurately. One reason for this issue is the limitation on the number of documents returned, where longer documents split into multiple chunks prevent full retrieval. While increasing the number of returned documents can capture more context, it may decrease `MAP`. A re-ranking mechanism is needed to keep the most relevant documents at the top. Additionally, increasing the number of retrieved documents may introduce irrelevant chunks, particularly in shorter documents, reducing the effectiveness of the RAG system. Prompt engineering is essential to ensure the LLM processes only the relevant context.

2) Answer Generation Task: Table II shows the evaluation of two LLM models based on `ROUGE-L` and `BERTScore`. Both models show unsatisfactory performance in precision and recall, with the best model reaching only a precision of 0.556, meaning much of the generated text did not appear in the context, and recall was similarly low. Despite this, both models perform well on `BERTScore`, indicating semantic similarity between the generated responses and the reference text, even if exact sequences were missed. Comparing gpt-4o to gpt-4o-mini, there is no significant difference in their `ROUGE-L` or `BERTScore` performance. However, Table III shows that gpt-4o-mini generates responses 78% faster than gpt-4o and is currently cheaper for answer generation.

TABLE II
Answer generation evaluation (Precision(P), Recall(R), F1-score(F1)) of different LLM models

Model	ROUGE-L			BERTScore		
	P	R	F1	P	R	F1
gpt-4o	0.556	0.106	0.177	0.759	0.763	0.761
gpt-4o-mini	0.487	0.109	0.178	0.755	0.765	0.760

TABLE III
Response Generation Time

Model	Time (s)
gpt-4o	12.7
gpt-4o-mini	7.1

IV. Conclusion

In this paper, we propose PharmChat, a RAG system designed to assist users in reviewing and querying pharmacopoeia documents. PharmChat preprocesses documents by splitting them into individual medicine entries, storing document embeddings in a vector database for similarity-based searches, and using retrieved documents as context for LLM-generated responses. Experimental results on real pharmacopoeia datasets show that PharmChat achieves sufficient precision and recall for both document retrieval and query answering with short document contexts.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(No. 2020R1I1A3072457) and MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2023-00259967) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

[1] Orlando Ayala and Patrice Bechard. Reducing hallucination in structured outputs via retrieval-augmented generation. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[2] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[3] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.