The Application of RAG Technology in Traditional Chinese Medicine

Yufeng Liu
[1]Artificial Intelligence, Shanghai Normal University, Shanghai, 200233, China
email: 1000517572@smail.shnu.edu.cn

Abstract. The inheritance and development of traditional Chinese medicine are facing limitations and significant challenges in today's society. This article combines the Large Language Model (LLM) model with the Retrieval Augmented Generation (RAG) model to address this issue, helping medical students to quickly retrieve highly specialized knowledge and to some extent, assisting in the modernization and inheritance of traditional Chinese medicine. Taking the Compendium of Materia Medica as an example, this article divides the text into blocks and vectorizes them, splitting the initial text into different blocks, and then selecting an optimization model to embed the text blocks. Afterwards, an index is established for the text, and after retrieval, filters and reordering techniques are used to further refine the search results. Next, this article constructs a chat engine to provide chat logic for the RAG system, using query compression technology to solve problems related to subsequent support and pronoun reference, and helping the chat engine to consider the context of the conversation. Finally, this article sends the context of the unlocked results block by block to LLM to optimize the answer, while summarizing the retrieved context, adapting prompts, and generating multiple answers based on different context blocks, connecting and summarizing them. Based on the relevant experimental results, the model's text matching degree is relatively high. At the same time, the model's recall rate and text accuracy rate are acceptable, indicating that prompt word engineering is still needed to improve the accuracy of knowledge retrieval in traditional Chinese medicine.

Keywords: Large language model, retrieval augmented generation, Chinese tradition medicine.

1 Introduction

Since ancient times, Traditional Chinese Medicine has been an integral part of China's excellent traditional culture, deeply rooted in the soil of Chinese culture. Traditional Chinese Medicine is a traditional medical system that the Chinese nation has accumulated and developed over a long period of life and practice, with a unique theoretical style and rich diagnosis and treatment experience. However, with the introduction of Western medicine into China, Western medicine has also rapidly developed in China. Compared with Traditional Chinese Medicine, Western medicine's diagnosis of diseases reflects the nature of the lesion to a greater extent,

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_43

which greatly determines the treatment methods and measures. This has led to a wealth of search content, entries, and knowledge on the internet for Western medicine [1]. In contrast, searching for Traditional Chinese Medicine on the internet often produces inaccurate or false information, leading to computer vision. This phenomenon has a significant negative impact on the inheritance, development, and application of Traditional Chinese Medicine.

The wave of Large Language Models (LLMs) has already swept through various industries in recent years. However, when it comes to specific industries such as medicine, LLMs often exhibit a lack of professional knowledge. Even the field of vector search have been driven by this trend, although there have been search engines based on faiss as early as 2019 [2]. Vector database startups like Chroma, Weavaite.io, and Pinecone are built on existing open-source search indexes, mainly faiss and nmslib [3] and have recently added additional storage for input text and other tools. In the field of pipelines and applications based on large language models, the two most famous open-source libraries are LangChain and LlamaIndex, which were established in October and November 2022, respectively, with only one month apart [4]. The creation of these two libraries was inspired by the release of ChatGPT and gained widespread adoption in 2023.

In recent years, Retrieval Augmented Generation (RAG) has become an undisputed choice. RAG combines search technology with the prompt word function of large language models, which means that questions are posed to the model and the information found by the search algorithm is used as background context. These queries and retrieved contextual information are integrated into the prompts sent to the large language model [5]. RAG can be applied in many fields through query enhancement, data block extraction, recursive knowledge graph, and response enhancement. For example, in the legal field, RAG can help lawyers quickly retrieve legal codes and be more proficient in preparing cases. Additionally, entrepreneurs can use this technology to build AI intelligent customer service, enterprise intelligent knowledge base, AI search engine, and other applications at a lower cost, enabling natural language input and dialogue with various forms of knowledge organization.

This article aims to solve the problem of retrieving Traditional Chinese Medicine knowledge using Retrieval Augmented Generation technology. Taking "Compendium of Materia Medica" as an example, which is the earliest book on drug classification in Chinese medicine and laid the foundation for the theoretical system of Chinese herbal medicine. Technically, this article uses Milvus vector database as the foundation and trains the ChatGPT model. The results demonstrated the effectiveness of the employed method.

2 Method

2.1 Data Preparation

The data for this article comes from the work "Compendium of Materia Medica" by Li Shizhen in the Ming Dynasty. Li Shizhen spent 27 years humbly consulting doctors, farmers, woodcutters, hunters, fishermen, and others while practicing

medicine. He based his work on various ancient herbal books, combined with his own observations and understanding, to classify drugs and ultimately write this masterpiece. The book not only corrected several errors in the past study of herbal medicine, but also integrated a large amount of scientific data, proposed a more scientific method of drug classification, incorporated advanced biological evolution ideas, and reflected rich clinical practice.

The entire book consists of 52 volumes and approximately 2 million words, with over 1.9 million words in total. It records 1,892 types of drugs (including 374 new additions) and is divided into 60 categories. This article obtained the relevant knowledge content of " Compendium of Materia Medica " and obtained a txt file [6]. This was used as the source file for this research and training.

2.2 RAG-based Compendium of Materia Medica

The core technology of RAG, which includes the core algorithms and steps is described in this section. The following Fig. 1 shows the approximate framework of the core technology.

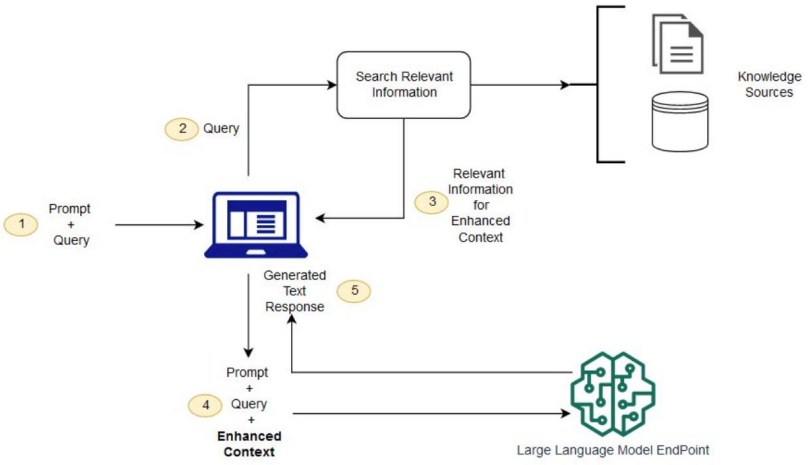

Fig. 1. Conceptual flow of RAG in conjunction with LLM [7].

Firstly, RAG technology creates vector indexes for the indexed document content, and uses the cosine distance between the search vector and the query vector as the nearest vector index. Through querying, the document with the closest semantic meaning can be found [8]. In this article, the NodeParser class in Llamaindex is used to effectively partition the article data content, so that the initial long document can be split into text blocks of a certain size without losing its meaning [9]. Furthermore, the end2end method in LlamaIndex was used to optimize the search model, making it easier to embed blocks and obtain vectorized content.

Next, this article uses Zilliz's vector database to automatically ingest data, making it easier for the RAG algorithm to efficiently index and store metadata for filtering information according to certain conditions. Then, two indexes are created for the content, one composed of summaries and the other composed of blocks segmented

from the document content data. The content is then enhanced by using smaller blocks for better search quality. Finally, this study refined the search results obtained from the above algorithms through filtering and reordering to obtain the final answer.

Thirdly, this article creates chat logic for the RAG system, taking into account the context of the conversation. This problem can be solved by using query compression technology to consider the chat context together with the user's query. In this study, this article used the functional API of OpenAI in Langchain to implement this. The records in the chat interaction can be put into the new query, and the retrieved context can be passed to LLM to generate the answer, providing a more flexible chat mode.

The last step is to generate the corresponding answer based on the retrieved text content data and the user's initial query. This study sent the retrieved context block by block to LLM to optimize the answer, summarize the retrieved context, match it with prompts, and connect the generated answers together.

After providing much high-quality information about Compendium of Materia Medica as possible, it is better to choose the final reading comprehension stage of the big model. OpenAI' s newly released gpt4 version, gpt4-turbo-1205, is significantly better compared to gpt-35-turbo-1106, both in terms of formatting requirements and overcoming illusions, and this study chose gpt4-turbo as the final reader.

2.3 Implementation Details

When users query related content, this article set '-d' as the query parameter and '-overwrite' as the overwrite parameter, with 'remove' only accepting one parameter. When splitting the text, chunk_size=600, chunk_overlap=20. This article uses gpt-3.5-turbo or gpt-4-turbo as the generator, and the iteration=10000.

The performance evaluation of the RAG system needs to include multiple frameworks, so it is required to use multiple independent indicators to evaluate the retrieval performance of the RAG solution. For example, this study used classic context precision, model recall rate [10], and answer relevance to evaluate this application project.

3 Results and Discussion

3.1 The Evaluation Results of the Application of RAG

This paper uses more traditional metrics such as Context Precision, Recall Rate, and Answer Relevance to evaluate the retrieval performance of the RAG model in the search of Compendium of Materia Medica. From Table 1, it can be seen that the Answer Relevance value of this model is relatively high, at 0.856, indicating that the retrieved answers have a high correlation with the inquiry information and there is less irrelevant content. The text accuracy rate is 0.532 and the recall rate is 0.561, both of which are acceptable, indicating that most of the retrieved answers are correct and relevant.

Table 1. The evaluation results of the application of RAG

Method	Context Precision	Recall	Answer Relevance

RAG	0.532	0.561	0.856

3.2 The Use of RAG in the Search of Compendium of Materia Medica

C：\Users\Lenovo>(rag) question：橘子皮的功效
痰隔气胀，水煎服。下焦冷气，蜜丸服。
Source:[果部]

Fig. 2. Question and Answer about the benefits of orange peel (Chinese version).

C：\Users\Lenovo>(rag) question：The benefits of orange peel:
Answer: When a patient starts coughing up phlegm and experiencing bloating, boiling orange peel in hot water and consuming the cooked peel can treat phlegm in the throat. When a patient feels cold entering their body, making sweet pills out of orange peel and taking them can treat the coldness inside the body.

Fig. 3. Question and Answer about the benefits of orange peel (English version).

C：\Users\Lenovo>(rag) question：铅白霜的功效
治鼻衄不止，铅白霜末，新汲水(调)服一字。
Source:[十全博旧方]

Fig. 4. Question and Answer about the benefits of lead white frost (Chinese version).

C：\Users\Lenovo>(rag) question：The benefits of lead white frost:
Answer: When a patient experiences continuous nosebleeds, taking a portion of lead white frost powder and brewing it with freshly boiled water can alleviate the symptoms. After drinking a cup of the solution, the patient will feel better.

Fig. 5. Question and Answer about the benefits of lead white frost (English version).

C：\Users\Lenovo>(rag) question：如何治疗头疼：
方法一：取当归 10 克，川芎 5 克，白芷 6 克，葱白 3 根。水煎 2 次，混合后分上、下午服，每日 1 剂。
方法二： 刀豆根 15 克，黄酒 50 毫升。加适量的水煎服，每日 1 剂，分 2 次服。
方法一：辛夷花 10-12 克，鸡蛋 2 个，加清水适量同煮，蛋熟后去壳再煮片刻，饮汤吃蛋。
方法二：干白菜根 1 块，红糖 60 克，生姜 3 片。水煎服，日服 3 次。
方法三：川芎 6 克，绿茶 6 克，红糖适量。清水一碗半，煎至一碗，去渣饮用。

Fig. 6. Question and Answer about How to treat a headache (Chinese version).

C: \Users\Lenovo>(rag) question: How to treat a headache :
Answer: Method 1: Take 10g of Angelica sinensis, 5g of Chuanxiong, 6g of Dahurian Angelica, and 3 scallions. Boil twice, mix and divide into morning and afternoon doses, take once daily.
Method 2: Take 15g of Knife Bean Root and 50ml of Huangjiu. Add an appropriate amount of water and boil, take once daily, divided into two doses.
Method 1: Take 10-12g of Magnolia liliflora, 2 eggs, and an appropriate amount of water. Cook together until the eggs are cooked, remove the shells, and cook for a while longer. Drink the soup and eat the eggs.
Method 2: Take 1 piece of dried Chinese cabbage root, 60g of brown sugar, and 3 slices of ginger. Boil in water and take three times a day.
Method 3: Take 6g of Chuanxiong, 6g of green tea, and an appropriate amount of brown sugar. Boil in one and a half bowls of water until reduced to one bowl, remove the residue and drink.

Fig. 7. Question and Answer about How to treat a headache (English version).

According to Fig. 2, Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, shown above RAG can provide relatively accurate and relevant text for most answers. For example, in Fig. 3 and Fig. 5, for smaller and more precise questions such as the benefits of orange peel and lead white frost, RAG can provide correct answers - specific usage methods and diseases with strong targeting. This shows that RAG can be of great help in searching for traditional Chinese medicine. However, for larger and less targeted questions, RAG still inevitably produces computer illusions, such as "How to treat a headache" in Fig. 7. According to the answers provided in the figure, RAG may provide answers for treating the same disease manifestations in different chapters, resulting in confusion in the method numbering and misunderstand in context. This is because the Compendium of Materia Medica contains many chapters, each of which contains the same disease manifestations and different disease causes and treatments. If only a few prompts are given, RAG cannot distinguish between them, and like traditional GPT, it will provide answers that lack targeting. Therefore, even with an enhanced retrieval method for RAG, people still need to build prompt engineering to improve the accuracy of search success when testing related data, as the values of Context Precision and Recall will not be very high.

4 Conclusion

In this study, this article aims to combine RAG technology with traditional Chinese medicine techniques, using Compendium of Materia Medica as an example, and further promote the development and inheritance of traditional Chinese medicine with advanced information technology. RAG technology has surpassed traditional semi-structured techniques in obtaining enhanced data in the professional medical field, and has improved and reduced the model's dependence on external knowledge sources for structured data preprocessing. At the same time, RAG technology has an adaptable retrieval process, which improves the efficiency and accuracy of answering

questions by autonomously judging and using LLM to determine whether to re-retrieve. In this experiment, RAG technology has a high answer relevance and accuracy and a moderately suitable recall rate for Compendium of Materia Medica, which has to some extent accelerated the retrieval speed of ancient Chinese pharmacy. In the future, this technology can expand the application scope of RAG technology to more multimodal fields, ensuring the correctness of RAG retrieval and expanding it to handle various data forms such as code, images, and videos, so that RAG has more important practical significance in the deployment of artificial intelligence.

References

1. Chen, J.: Chinese Medicine. People's Medical Publishing House, p2, Beijing (1998).
2. Facebook AI Research: Faiss. Available at: https://github.com/facebookresearch/faiss (2019).
3. Searchiavrius: Non-Metric Space Library (NMSLIB). Available at: https://github.com/nmslib/nmslib (2020).
4. Gao, L., Ma, X., Lin, J., Callan, J.: Precise Zero-Shot Dense Retrieval without Relevance Labels (2023).
5. AI Technology Community: In-depth article! The most comprehensive overview of large model RAG technology (2024).
6. Book Download Website: Compendium of Materia Medica. Available at: http://wap.bookdown.info/down/684.html (2018).
7. Retrieval-based Language Models and Applications. Available at: https://acl2023-retrieval-lm.github.io/ (2023).
8. Amazon Web Services: What is RAG. Available at: https://aws.amazon.com/cn/what-is/retrieval-augmented-generation/ (2024).
9. iyacontrol: Illustrated Advanced RAG Techniques. Available a t: https://zhuanlan.zhihu.com/p/674755232 (2023).
10. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. Available at: https://arxiv.org/abs/2310.11511 (2023).