2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC) | 979-8-3315-3462-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/EIECC64539.2024.10929529

TCM MLKG-RAG: Traditional Chinese Medicine Intelligent Diagnosis Based on Multi-Layer Knowledge Graph Retrieval-Augmented Generation

Qi Chen
Department of Electronic Engineering and Information Science
University of Science and Technology of China
Hefei, China
chenqi1113@mail.ustc.edu.cn

Lin Ni*
Department of Electronic Engineering and Information Science
University of Science and Technology of China
Hefei, China
*nilin@ustc.edu.cn

Abstract—**Traditional Chinese Medicine (TCM) search engines often struggle with the issue of redundant data volumes, making it difficult to meet users' demands for precise information retrieval. Large Language Models (LLMs) excel in understanding questions and summarizing key points due to their vast number of parameters. However, keeping pace with updates in TCM knowledge requires significant computational resources and time for finetuning these large models. Retrieval-augmented generation (RAG) allows LLMs to generate more accurate, specialized, and timely responses without the need to update their parameters. TCM knowledge is characterized by its dispersed nature and a blend of classical and vernacular language, which makes traditional RAG unsuitable for the field of TCM. To address this, we have developed a TCM knowledge graph RAG that integrates multi-layered knowledge bases, with the lower layer consisting of TCM-specific terminology and explanations, and the upper layer comprising clinical diagnosis and treatment cases. Furthermore, we have proposed two retrieval methods: keyword retrieval and therapy retrieval. Keyword retrieval is designed to search for information on TCM-specific terms, while therapy retrieval locates diseases based on medical and patient information and provides corresponding treatment methods. We have validated the effectiveness of our methods across various datasets.**

Keywords—traditional Chinese medicine; retrieval-augmented generation; knowledge graph

I. INTRODUCTION

As Internet and information technology advance, a variety of TCM websites have emerged. However, conventional search engines rely on keyword matching, which has limitations. First of all, the accuracy of the search results is insufficient, and it is often accompanied by a large amount of redundant data, because they do not assess the relevance of content to user queries beyond the presence of specific keywords. Additionally, the proliferation of false information on the Internet makes it challenging for users to verify the authenticity of TCM knowledge. These issues significantly hinder the quality and effectiveness of search results, causing difficulties in locating information.

With the huge amount of parameters, LLMs excel in dialogue generation and professional examinations, aiding in information organization and summarization. However, updating LLMs through fine-tuning to accommodate changes in TCM treatment schemes is resource-intensive and time-consuming. In response to this problem, RAG [1] integrates retrieval algorithms to feed relevant knowledge information into LLMs, enabling the generation of more precise, professional, and up-to-date responses without the need for further parameter updates.

However, the knowledge of TCM has the characteristics of scattered knowledge, the text is difficult to understand, and the words are multi-meaning. There is also numerous redundant information in the corpus of the doctor-patient dialogue. Therefore, the traditional RAG is not suitable for TCM. We propose a multi-layer knowledge graph-integrated RAG that leverages the TCM knowledge system and its linguistic characteristics to enhance information retrieval and response accuracy. The work content is as follows:

- Considering the decentralization of TCM knowledge and the abundance of redundant information, we transform the original unstructured data into structured data, specifically a knowledge graph, to enhance retrieval capabilities.
- The linguistic characteristics of TCM, often described as 'semi classical Chinese and semi vernacular,' can pose challenges for LLM comprehension, so we construct multi-layer knowledge base. The bottom layer consists of TCM-specific terminology and their definitions; The upper layer is a knowledge graph extracted from doctor-patient dialogue, and the proprietary noun entities in it will be mapped to the entities in the bottom layer knowledge base to help LLM understand.
- Propose two retrieval methods, one is keyword retrieval, which is used to retrieve information on TCM proprietary terms. Another method is therapy retrieval, which locates diseases and provides treatment methods based on input medical and patient information. In order to speed up retrieval efficiency, the upper level case is classified.

II. RELATED WORKS

With the advent of LLMs, generative language models have ascended to the forefront, showcasing formidable capabilities across a spectrum of linguistic tasks. Currently, numerous models leverage TCM knowledge for pre-training or fine-tuning purposes. For instance, ShenNong TCM [2] has been instrumental in advancing the deployment of large-scale language model. Initiatives like Huatuo [3] have open-sourced a suite of large language models that have been fine-tuned with TCM directives, including LLaMA, Alpaca Chinese, Bloom, and others, which has improved the auxiliary diagnostic effect of the base mode. Qibo [4] primarily focuses on establishing evaluation metrics for the TCM field based on educational materials and offers objective multiple-choice questions across various subjects to gauge foundational knowledge and competencies in TCM. However, the problem with this method is that when the TCM knowledge base is updated, pre-training or fine-tuning methods can bring a lot of time and computational costs.

RAG can find textual information related to the input problem in the knowledge base through a retriever, and then input this information together with the problem into LLM to obtain updated knowledge without updating parameters. GraphRAG [5] utilizes entities and relationships in knowledge graphs to provide more accurate and context relevant information, thus performing well in tasks such as question answering and text generation. It also uses community detection algorithms to cluster similar content in articles, generate community reports, and help searchers locate target content more accurately. Self-RAG [6] conducts retrieval and self-reflection based on questions to improve the quality of LLMs generation, including their factual accuracy, without compromising their generality. HyDE [7] generates a document using the original query, which will be used as the query for retrieval. The generated document contains richer relevant information, which helps to retrieve more accurate results. However, these methods are all used for universality issues, and when applied to the vertical field of traditional Chinese medicine, they cannot adapt to the unique language environment of traditional Chinese medicine, and there are also problems such as slow retrieval speed.

III. Method

We construct a multi-layer knowledge base (Section A), then retrieve knowledge related to the input content from the knowledge base (Section B), and finally input the retrieval results along with the input to LLM for response. The overall framework diagram is shown in *Fig. 1*.

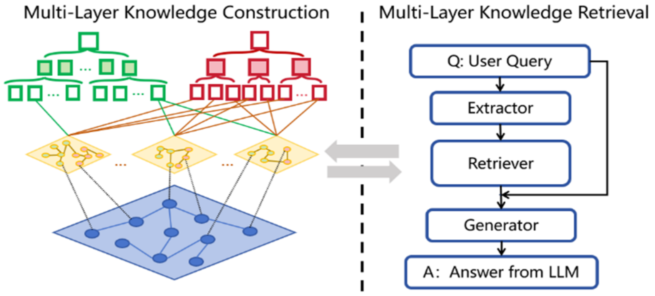
Multi-Layer Knowledge Construction
Multi-Layer Knowledge Retrieval
Q: User Query
Extractor
Retriever
Generator
A: Answer from LLM

Figure 1. Overall framework diagram

A. Multi-Layer Knowledge Construction

The framework diagram of multi-layer knowledge construction is shown in *Fig. 2*.

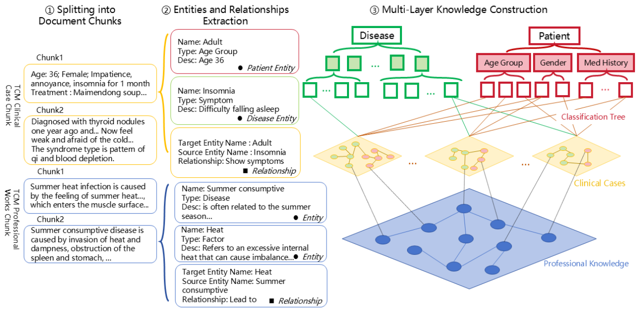
① Splitting into Document Chunks
② Entities and Relationships Extraction
③ Multi-Layer Knowledge Construction
TCM Clinical Case Chunk
Chunk1
Age: 36; Female; Impatience, annoyance, insomnia for 1 month Treatment : Maimendong soup...
Chunk2
Diagnosed with thyroid nodules one year ago and... Now feel weak and afraid of the cold... The syndrome type is pattern of qi and blood depletion.
TCM Professional Works Chunk
Chunk1
Summer heat infection is caused by the feeling of summer heat..., which enters the muscle surface...
Chunk2
Summer consumptive disease is caused by invasion of heat and dampness, obstruction of the spleen and stomach, ...
Name: Adult
Type: Age Group
Desc: Age 36
● Patient Entity
Name: Insomnia
Type: Symptom
Desc: Difficulty falling asleep
● Disease Entity
Target Entity Name : Adult
Source Entity Name : Insomnia
Relationship: Show symptoms
■ Relationship
Name: Summer consumptive
Type: Disease
Desc: is often related to the summer season...
● Entity
Name: Heat
Type: Factor
Desc: Refers to an excessive internal heat that can cause imbalance...
● Entity
Target Entity Name: Heat
Source Entity Name: Summer consumptive
Relationship: Lead to
■ Relationship
Disease
Patient
Age Group
Gender
Med History
Classification Tree
Clinical Cases
Professional Knowledge

Figure 2. Multi-Layer knowledge construction diagram

- **Splitting into Document Chunks.** For the database of TCM professional works, divide them into chunks according to paragraphs; For consultation records, divide them into chunks based on the patient's dimensions.
- **Entities and Relationships Extraction.** We use the method of prompt words to enable LLM to extract entities $E = \{e_1, e_2, \dots, e_N\}$ and relationships $R = \{(e_i, TheReferenceOf, e_j) | e_i \in E, e_j \in E\}$ from each chunk, where entities include entity names, types, and entity descriptions, and relationships include source entity names, target entity names and relationship descriptions. In addition, the entities in the clinical inquiry record chunk will be divided into patient entities or disease entities, such as gender and age belonging to patient entities, while fever and headache belong to disease entities.
- **Multi-Layer Knowledge Construction.** TCM's blend of ancient and modern Chinese, coupled with specialized terminology and metaphorical expressions, can lead to comprehension challenges for LLMs. For instance, ancient medical texts avoided direct references to constipation due to its sensitive nature, using euphemisms like "not changing clothes" (Ancient Chinese expression of constipation) instead. We construct a multi-layer knowledge base to solve this problem. The underlying knowledge comes from TCM textbooks and works, and is an explanation of TCM professional knowledge, while the upper layer knowledge comes from clinical consultation records. The two-layer knowledge base is linked through TCM professional terms, and the upper layer of consultation cases are classified. This structure clarifies ambiguities in TCM texts and facilitates swift retrieval of relevant diagnostic and treatment cases.

We establish entity connections by assessing the similarity between two knowledge base layers. First, vectorize the entities using the fine-tuned "bge-m3"[1], denoted as $emb(\cdot)$, then calculate cosine similarity and define the entity similarity threshold as δ. Entity linking occurs when the distance is below δ and is shorter than the distance to any other entity. This process can be formulated as follows:

[1] https://huggingface.co/BAAI/bge-m3

$$Distance\left(e_i^{exp}, e_j^{case}\right) = sim\left(emb\left(e_i^{exp}\right), emb\left(e_j^{case}\right)\right)$$
$$e_i^{exp} \in E^{exp}, e_j^{case} \in E^{case} \quad (1)$$

$$e_i^{exp} \leftrightarrow e_j^{case} \;\; iff \;\; e_j^{case} \;=\; \{argmin_{e_k^{case} \in E^{case}}$$
$$Distance\left(e_i^{exp}, e_k^{case}\right) | \, Distance\left(e_i^{exp}, e_j^{case}\right) < \delta\} \quad (2)$$

where E^{exp} is the bottom layer professional knowledge entity collection, and E^{case} is the upper layer TCM clinical case entity collection.

We use contrastive learning to finetune the embedding model. Firstly, based on [8], 5000 samples are constructed using the in-batch negatives method. Each sample contains a question and its correct answer (positive sample) as well as an incorrect answer (negative sample). Since that similar diseases and their explanations are grouped in the same section in [8], diseases within the same section are batched together. In the construction of positive and negative samples, the explanation of the diseases is rewritten as a question by LLM, the diseases are set as positive sample answers, and the positive samples of other samples in the same batch are set as negative samples.

In actual TCM consultations, identical symptoms can warrant different treatments based on the patient's age, gender, and medical history, as seen in the distinct treatments for "pediatric cold" versus "adult cold." Hence, we create separate indices for patient and disease characteristics. For disease classification, we follow the first and second level directories in [8] and use prompt words to guide the large language model in classifying TCM cases. Patient characteristics are categorized by age group, gender, and medical history.

B. *Multi-layer knowledge retrieval*

Two retrieval methods can be provided, one is to search for information on TCM terms (such as diseases, Chinese medicine) based on keywords, as shown in *Fig. 3*; The other method is to input patient and their disease information to query treatment or prevention plans, as shown in *Fig.4*.

- **Keyword Retrieval.** Firstly, use prompt words to have LLM extract entities E^Q from user query Q. Then use $emb(\cdot)$ to vectorize E^Q, calculate cosine similarity with entities in the underlying professional knowledge, and find the top K entities with the smallest distance. Finally, find all triplets that contain these entities, and combine these triplets with Q to form a prompt word and input it into LLM for response. The process is shown in *Fig. 3*.

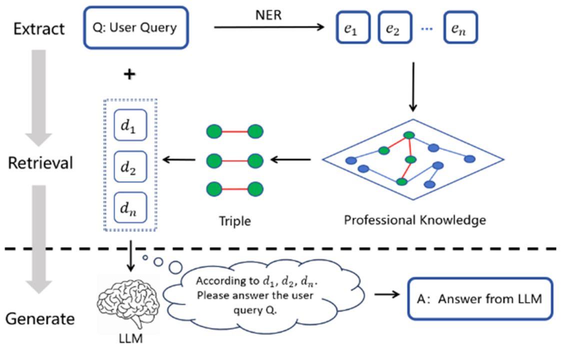

Figure 3. Keyword retrieval process diagram

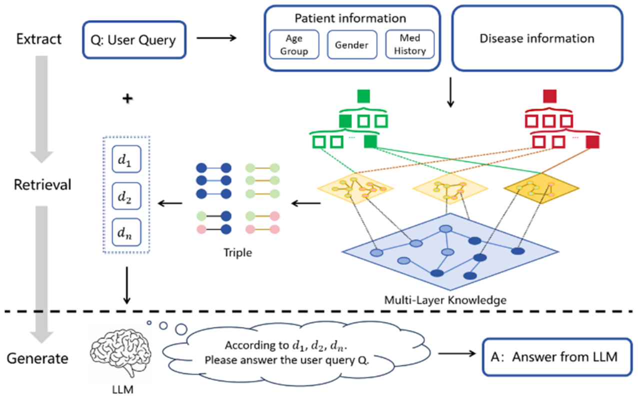

Figure 4. Therapy Retrieval Process Diagram

- **Therapy Retrieval.** This is a top-down method of querying through classification index. First, use prompt learning to extract patient information Q^P and disease information Q^D from Q. For patient information, matching cases can be located in the knowledge base with gender, age group and medical history. For disease information, first extract entity E^{QD}. Assuming that the $j-th$ class name in the $i-th$ layer is C_j^i, vectorize all entities in E^{QD} using $emb(\cdot)$, and calculate cosine similarity with C_j^i to find the class name in the $i-th$ layer with the smallest distance from E^{QD}, which serves as the root node C^{i+1} in the lower layer. The process formula is:

$$Distance\left(E^{QD}, C_j^i\right) = \sum_{e_i^{QD} \in E^{QD}} sim\left(emb\left(e_i^{QD}\right), emb\left(C_j^i\right)\right) \quad (3)$$

$$C^{i+1} = argmin_{C_j^i \in C^i} \; Distance\left(E^{QD}, C_j^i\right) \quad (4)$$

Traverse from layer C^0 until reaching the bottom C^n of the classification tree. Finally, all clinical cases under C^n will be found. If the cases contain professional terminology explanations of underlying knowledge, they will be concatenated with the cases and combined with Q to form prompt words for input to LLM for response. The process is shown in *Fig. 4*.

Let's clarify the therapy retrieval process with a case: "A 42-year-old with a small, red swelling on the left little finger, no significant pain or itching." The patient information is an adult with no medical history, and the disease information is a red swelling on the same finger without discomfort. We vectorize the disease information with $emb(\cdot)$ and find "tumor and cancer terminology" in the primary catalog is closest. Further, "tumor diseases" matches in the secondary catalog. With all information, we scour our knowledge bases for "tumor diseases" cases in adults with no history. We then extract these cases, including specialized terms, and feed them into our LLM along with the query to get the response.

IV. Experiment

A. Dataset

The TCM knowledge base is constructed using data from the "Chinese Traditional Chinese Medicine Information Query Platform"[2] and 688 authoritative texts, including classics and works like [8]. Clinical consultation records are sourced from the ShenNong TCM dataset [2], which comprises over 110,000 entries.

Our test dataset includes the TCM section of CMMLU [9], TCM-SD dataset [10], and TCM doctor-patient dialogues from "Good Doctor"[3] between 2023 and 2024. The CMMLU dataset features 11,528 questions across 67 disciplines, with a minimum of 105 questions per discipline, all in multiple-choice format with four options each. TCM-SD is the inaugural publicly available dataset for TCM diagnosis, encompassing 54,152 clinical records for 148 syndromes, detailing patient histories, symptoms, and associated diagnoses.

B. Evaluation Index

For CMMLU dataset, we employ the accuracy as a metric to assess the performance of various models in addressing single-choice questions. For TCM-SD dataset, we utilize top-1 accuracy and top-5 accuracy to evaluate model performance. Additionally, we apply BLEU and GLEU scores to assess the quality of the "Good Doctor" TCM consultation dataset.

Accuracy: This is the ratio of the number of correctly predicted samples to the total number of samples.

Top-1 Accuracy: This is equivalent to Accuracy, indicating the proportion of predictions that match the actual outcome.

Top-5 Accuracy: This metric requires the model to output the five most probable categories. If the correct answer is among these five, the prediction is considered correct.

BLEU: This stands for Bilingual Evaluation Understudy and is used to quantify the similarity between generated sentences and reference sentences by assessing n-gram overlaps to evaluate fluency at the sentence level.

GLEU: This metric automatically evaluates the fluency of generated responses, taking into account richness and smoothness in addition to fluency.

C. Experimental Results

We conducted separate tests to evaluate the performance of two distinct search methods: keyword retrieval and therapy retrieval.

For keyword retrieval, we tested the performance on CMMLU dataset by inputting both the question stem and the four options as search queries to obtain results. These results were then fed into LLM to generate answers. The final accuracy rates are presented in *Table. 1*.

TABLE I. Accuracy of LLMs Utilizing Different RAG Methods on CMMLU Dataset

	GPT-3.5 turbo	**GPT-4**	**LLaMa1 65B**	**LLaMa2 70B**
Without RAG	0.54	0.59	0.45	0.46
RAG	0.59	0.65	0.53	0.55
Graph-RAG	0.61	0.68	0.54	0.56
CRAG	0.61	0.67	0.53	0.55
KG-RAG	0.62	0.69	0.55	0.58
Self-RAG	0.61	0.7	0.53	0.55
HyKGE	0.62	0.72	0.56	0.58
TCM MLKG-RAG	**0.65**	**0.76**	**0.57**	**0.6**
TCM MLKG-RAG (w/o Multi-Layer)	0.62	0.71	0.54	0.58

We tested the performance of therapy method retrieval on both TCM-SD dataset and "Good Doctor" dataset. The results are respectively presented in *Table. 2* and *Table. 3*.

TABLE II. Top-1 and Top-5 Accuracy of LLMs Utilizing Different RAG Methods on TCM-SD Dataset

	GPT-3.5 turbo	**GPT-4**	**LLaMa1 65B**	**LLaMa2 70B**
Without RAG	0.79/0.85	0.85/0.91	0.68/0.83	0.7/0.84
RAG	0.83/0.88	0.87/0.93	0.7/0.84	0.71/0.86
Graph-RAG	0.85/0.89	0.88/0.93	0.7/0.85	0.72/0.87
CRAG	0.84/0.88	0.88/0.93	0.7/0.84	0.71/0.86
KG-RAG	0.86/0.9	0.9/0.95	0.72/0.87	0.74/0.89
Self-RAG	0.84/0.88	0.88/0.93	0.7/0.84	0.71/0.86
HyKGE	0.86/0.93	0.9/0.96	0.72/0.88	0.74/0.9
TCM MLKG-RAG	**0.9/0.96**	**0.93/0.98**	**0.79/0.92**	**0.8/0.93**
TCM MLKG-RAG (w/o Multi-Layer)	0.85/0.92	0.89/0.94	0.73/0.87	0.75/0.89

TABLE III. BLEU-4 and GLEU of LLMs Utilizing Different RAG Methods on "Good Doctor" Dataset

	GPT-3.5 turbo	**GPT-4**	**LLaMa1 65B**	**LLaMa2 70B**
Without RAG	1.86/5.01	3.86/6.45	1.04/4.15	1.26/4.68
RAG	3.56/6.24	5.78/8.72	1.24/4.47	1.66/4.84
Graph-RAG	4.05/6.74	6.43/9.31	1.82/4.98	1.97/5.21
CRAG	4.34/6.97	6.72/9.43	1.87/5.02	2.1/5.21
KG-RAG	3.85/6.55	5.89/8.93	1.71/4.85	1.78/4.85
Self-RAG	4.11/6.83	6.64/9.36	1.82/4.99	1.95/5.17
HyKGE	5.47/8.49	7.52/9.82	2.16/5.77	2.36/5.81
TCM MLKG-RAG	**7.14/9.37**	**9.13/10.26**	**2.54/5.86**	**2.76/6.03**
TCM MLKG-RAG (w/o Multi-Layer)	5.68/8.51	7.02/9.61	2.19/5.77	2.4/5.82

V. Conclusion

Leveraging the unique linguistic characteristics of TCM, we have developed an enhanced RAG that integrates multi-layer knowledge base. We have also proposed two retrieval methods tailored for different scenarios, supporting both keyword retrieval and therapy retrieval. The performance of these methods has been validated across various types of datasets. In the future, TCM diagnosis and treatment methods of "observation, hearing, questioning, and cutting" can be used to strengthen therapy retrieval.

[2] https://www.dayi.org.cn

[3] https://www.haodf.com

REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[2] W. Zhu, W. Yue, X. Wang. "ShenNong-TCM: A traditional Chinese medicine large language model" https://github.com/michael-wzhu/ShenNong-TCM-LLM (2023).

[3] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, et al. "Huatuo: Tuning llama model with chinese medical knowledge." arxiv preprint arxiv:2304.06975 (2023).

[4] H. Zhang, X. Wang, Z. Meng, Z. Chen, P. Zhuang, Y. Jia, et al. "Qibo: A Large Language Model for Traditional Chinese Medicine." arxiv preprint arxiv:2403.16056 (2024).

[5] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, et al. "From local to global: A graph rag approach to query-focused summarization." arxiv preprint arxiv:2404.16130 (2024).

[6] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi. "Self-RAG: Self-reflective retrieval augmented generation." NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following (2023).

[7] L. Gao, X. Ma, J. Lin, J. Callan. "Precise zero-shot dense retrieval without relevance labels." arxiv preprint arxiv:2212.10496 (2022).

[8] State Administration for Market Regulation. "Clinic terminology of traditional Chinese medical diagnosis and treatment—Part 1: Diseases" GB/T 16751.1-202 (2023).

[9] H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, et al. "Cmmlu: Measuring massive multitask language understanding in chinese." arxiv preprint arxiv:2306.09212 (2023).

[10] R. Mucheng, H. Heyan, Z. Yuxiang, C. Qianwen, B. Yuan, G. Yang. "TCM-SD: a benchmark for probing syndrome differentiation via Natural Language processing." Proceedings of the 21st Chinese National Conference on Computational Linguistics (2022).