# Chinese Medicine Question Answering Robot Based on RAG and Self-Built Dataset

Enpu Zuo[1,*], Chenxi Pan[2], Junyu Chen[3], Zihan Yi[4]

{2296681461@qq.com[1], 1640827701@qq.com[2], hanschen777@outlook.com[3], 2823365782@qq.com[4]}

School of Future Technology, Dalian University of Technology, Dalian 116081, China[1]
Leicester International Institute, Dalian University of Technology, Dalian 116081, China[2]
School of Economics and Management, Xidian University, Xi'an 710126, China[3]
Dalian University of Technology-Ritsumeikan University International School of Information Science & Engineering, Dalian University of Technology, Dalian 116081, China[4]
*corresponding author

**Abstract.** Traditional Chinese Medicine (TCM) is a cornerstone of China's medical heritage, renowned for its unique methods of diagnosis and treatment. Despite its long history, TCM faces challenges in the modernization process due to its reliance on doctors' expertise and lack of systematic knowledge integration. This paper introduces two major innovations: the development of the most comprehensive Chinese medicine database and the first application of search-enhanced generation Retrieval-Augmented Generation(RAG) technology. In this paper, the most comprehensive TCM database was established by crawler and OCR, and the model's understanding of TCM knowledge was enhanced through the integration of Large language models(LLMs) and RAG technology, and the ability to systematically retrieve relevant prescriptions and literature was realized to achieve more personalized and accurate treatment recommendations. We tested it on a test set and invited TCM experts to evaluate it, which validated the accuracy and reliability of our model.

**Keywords:** Traditional Chinese Medicine (TCM), Large language models (LLMs), Retrieval-Augmented Generation (RAG), Question & Answer robot(Q&A robot)

## 1 Introduction

TCM is an essential part of the Traditional Chinese medicinal heritage. After its development and accumulation through thousands of years, it has formed unique theories and distinct remedies [1]. Through overall observation and analysis, TCM focuses on maintaining self-adjusting and balance and then tailors individual treatment plans [2]. In recent years, especially during the period of the pandemic of COVID-19, TCM has demonstrated distinct advantages in curing modern conditions [3]. However, in light of the complexity of the knowledge of TCM and the diversity of diagnostic procedures, TCM diagnosis mainly relies on individual TCM practitioner’s experience. In today's rapidly developing society, this dependence poses major problems for the popularization and dissemination of TCM.

With the rapid development of LLMs, the modernization of TCM has ushered in new chances [4]. LLMs have powerful natural language processing capabilities and are able to understand and generate complex texts. They are increasingly applicable to multiple fields. LLMs have

CONF-MLA 2024, November 21, Adana, Turkey

DOI 10.4108/eai.21-11-2024.2354617

been applied to diverse tasks in the field of medicine, ranging from automated analysis of medical literature to intelligent management of medical records and facilitation of clinical decision-making [5][6].

Although LLMs has made great success in the field of natural language processing, there still exists deficiencies in processing TCM prescriptions. Due to the limitations of training data and complexity of TCM combinations, LLMs may output incomplete prescription recommendations or incorrectly combine different drug components, thereby affecting the reliability.

All these challenges lead researchers to build our own TCM databases in order to provide more systematized and comprehensive informational support. We choose the BERT model, BART model, LLaMA3-Chinese model [7], and other LLMs to fine-tune them with our self-built TCM database. By fine-tuning all these LLMs and comparing the results of analysis, the research planned to demonstrate the superior performance across various metrics.

Furthermore, in order to deal with the problem that LLMs encounter knowledge limitations and the risk of inaccurate output in the case of complex TCM prescriptions, RAG technology is introduced to enhance performance and raise the accuracy.

RAG techniques effectively utilize the knowledge in databases by fusing information retrieval and generation models [8]. Based on the symptoms entered by users, it first retrieves the relevant prescriptions from the constructed TCM databases. Then, together with LLMs, it integrates the extracted information and generates personalized prescription recommendations. This process guarantees the production of scientifically valid and reliable results.

This research paper is aimed to implement an intelligent TCM question-and-answer system based on LLMs and integrate the most optimized large language model into the QA system in order for patients to swiftly obtain reliable TCM prescriptions based on given symptoms.

In this paper, we make attempts to explore the techniques in LLMS for TCM Question Answering robot: symptom recognition and analysis, prescription matching, and generation strategies. It is expected to open new avenues toward the intelligent development of TCM through deep integration between traditional medicine and modern technology.

## 2 Related Work

### 2.1 Database Construction for TCM

The construction of datasets in TCM is fundamental to promoting its modernization and applications in LLMs. Creating systematic and structured datasets facilitates data management and analysis more effectively. This process provides support for the digitalization of TCM.

### 2.2 Development and Current Status of Datasets

Nowadays, several comprehensive databases have been established in the field of TCM. The integration of databases shares the resources in this area. For example, Chen et al. initially set up a TCM database containing over 400 types of Chinese Medicinal Herbs, including over 20,000 purified compounds extracted from these herbs [9]. Afterward, Lin et al. intended to modernize and standardize TCM by associating prescriptions, medicinal herbs, the diseases they

were targeted for, prescription ingredients, and Chinese medicinal formulation's target mechanisms [10]. Xu et al. constructed the Encyclopedia of TCM ETCM that provides basic properties and quality control standards for herbal medicines, formula compositions, ingredients, and drug similarities [11]. The chemical constituents and pharmacological actions of medicinal herbs are demonstrated by ETCM and provide important data support for the study of the TCM mechanism. In this connection, Lyu et al. integrated the previous data sets to clarify the active composition of herbal medicines and constructed the biggest database in TCM up to now, which includes more than 9000 herbs and over 60,000 herbal components [12].

### 2.3 Current Challenges

Currently, the development of TCM databases mainly focuses on detailed research at biological and chemical levels. The study of prescriptions and medicinal ingredients has been done in relation to the chemical composition and its pharmacological effects. Although this approach provides an important basis for precision medicine research and new drug development, the amount of data obtained is small and lacks diversity. Therefore, all these factors limit the optimization of LLMs. They require massive and diverse datasets to deal with complex natural language tasks. Hence, to solve these problems, extension in breadth and diversity by inputting actual clinical data is necessary to meet modern intelligent TCM research demands.

### 2.4 The Development of the Combination of RAG and LLMs

One of the major recent developments in the area of natural processing is the technology of RAG, RAG. RAG joins the strengths of information retrieval and generative models to boost the applicability of LLMs on complex tasks.

### 2.5 the Fundamental Theory of RAG Technology

The RAG technology enhances traditional generative models through an information retrieval module, allowing access to external databases and thereby providing richer knowledge support during the process of text generation. It was initially proposed by Lewis et al. and represented superior performance in knowledge-intensive tasks [8].

Retrieval Module: This module is charged with the responsibility of retrieving relevant information from external literature or knowledge bases about a query.

Generation Module: All parts like LLaMa3 and BERT are combined in generative models to generate results with higher accuracy and contextually relevant outputs.

### 2.6 Incorporating RAG into LLMs

LLMs combined with RAG technology have a number of advantages in text generation and answering systems. According to Izacard and Grave et al., using RAG technology allows models to know when to retrieve relevant information from vast collections of documents to come up with answers that are both more accurate and contextually relevant [13]. This kind of integration significantly improves the performance of the model in open-domain question-answering tasks, retaining consistency and accuracy in comparison with traditional purely generative models on complex questions that need a large amount of background knowledge.

### 2.7 The Combination of LLMs and TCM

The application of LLMs in the field of TCM is gradually increasing and offers new possibilities for the modernization and intelligent development of traditional medicine.

### 2.8 Development and Current Status

The application of large language models in the field of TCM has gradually matured. By learning a large number of TCM prescription records, case data, and corresponding literature, these models can automatically generate treatment plans for specific symptoms. One of the more salient features of these models is to handle complex natural language inputs to provide accurate diagnostic recommendations based on changing symptoms and individual differences. Hua et al. achieved the most advanced performance with the "lingdan" model, pretrained on datasets for herbal prescriptions [14]. Zhang et al. improved their accuracy considerably using the guided fine-tuning "Qibo" model, similarly based on TCM datasets [15]. Also, Yang et al. use the model "MedchatZH" that pre-trained on Chinese medical book corpora and the model could very accurately predict the outputs [16].

### 2.9 Current Challenges

Modern LLMs are already capable of providing a satisfactory answer to the diagnosis and providing solutions. However, for some special prescriptions, especially the details of all medicinal ingredients included, the results may show a bias in the results produced. In order to solve the problem, a solution is to introduce RAG [8]. RAG technology fuses modules for retrieval and generation. However, RAG has not been widely implemented in the field of TCM

## 3 Data Sources and Collection Methods

### 3.1 Data Sources

To develop a robust and precise TCM question-and-answer system, data was collected from three main sources:

1. TCM Professional Books: Classical TCM texts such as "Huangdi Neijing," "Shennong Bencao Jing," "Shanghan Zabing Lun," and "Bencao Gangmu" were used as primary knowledge bases.

2. Online TCM Databases: Data was gathered from publicly available TCM databases and websites like the ETCM database, Zhongyi Shijia, and Gu Yi Mima to obtain modern TCM research and drug information.

3. Actual Medical Consultation Records: Real consultation dialogues between doctors and patients, along with the prescribed formulas, were collected through collaboration with TCM hospitals.

Table 1 shows the number of medicines and prescriptions from each dataset source. Figure 1 shows the proportion of herbs from each source in the crawled datasets. Figure 2 shows the share of prescriptions from each source in the crawled datasets.

**Table 1.** Summary of Data Sources

| Data Source | Formulas Count | Herbs Count |
|---|---|---|
| Shennong Bencao Jing | 21 | 365 |
| Shanghan Zabing Lun | 112 | 82 |
| Bencao Gangmu | 11096 | 1095 |
| ETCM | 3959 | 402 |
| Zhongyi Shijia | 25451 | 9651 |
| Gu Yi Mima | 52116 | 18344 |
| TCM Hospital Records | 2170 | 510 |

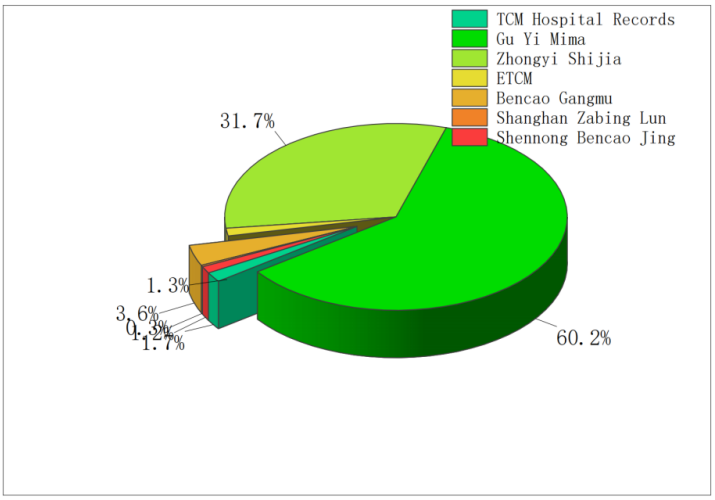



**Fig. 1.** Herbs Count

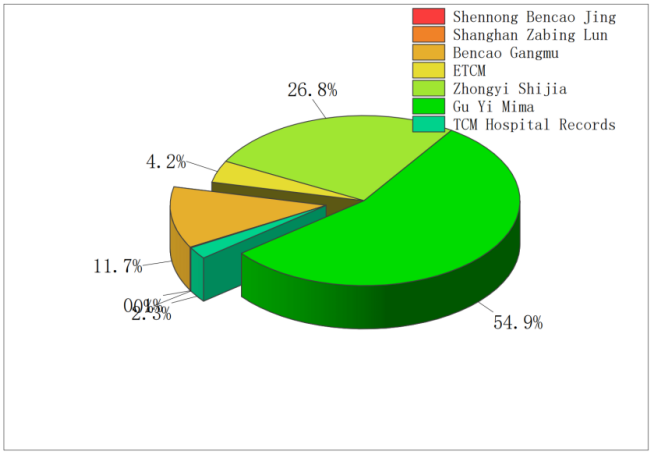



**Fig. 2.** Formulas Count

### 3.2 Data Collection Methods

1. Data Acquisition: Electronic books and websites were converted into XLSX format and transformed into Q&A format JSON files using web crawlers and OCR (Optical Character Recognition).

2. Data Preprocessing: Text data underwent cleaning to remove irrelevant characters, standardize terminology, and correct OCR errors. Additionally, natural language processing (NLP) techniques were used for tokenization, part-of-speech tagging, and entity recognition.

3. Data Annotation: TCM experts and manual annotators reviewed and labeled the generated Q&A pairs to ensure their medical accuracy and practical application. Annotations included formula names, corresponding symptoms, and formula compositions.

4. Data Augmentation: The dataset was expanded using methods such as synonym replacement, random insertion, random deletion, and back-translation.

Figure 3 shows the complete process of data set collection

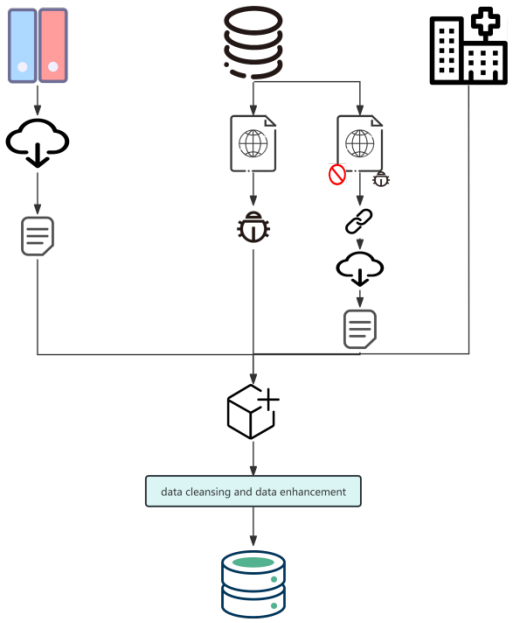



**Fig. 3.** Data processing flow

### 3.3 Dataset Construction and Validation

Training and Testing Set Division: The annotated Q&A pairs were randomly divided into training (80%) and testing (20%) sets to evaluate the system's performance.

System Testing and Feedback Loop: In the initial development phase, TCM experts and general users were invited to test the system, providing feedback for iterative optimization.

### 3.4 Ethics and Compliance

All medical consultation records were de-identified to ensure patient privacy. This study was approved by the research ethics committee of the involved hospitals and complied with relevant data protection regulations.

## 4 Methodology

In this part, we introduce our research methods. We first construct TCM-related data sets and then fine-tune the LLMs; finally combine RAG with LLMs to build a Q&A robot that can answer TCM-related questions more accurately. After training the Bert and Llama3, we found a series of problems, such as mismatched answers and questions, the model generating prescriptions by itself, model hallucinations, etc. In order to solve these problems and improve the quality and reliability of the answers, we decided to apply RAG. RAG is a technology that combines two strategies, which are retrieval and generation. In this way, RAG can improve the performance of the model in handling more difficult tasks. Furthermore, RAG can provide more accurate and comprehensive context information for the LLM, by retrieving strongly relevant information from the external knowledge base. Therefore, it is able to improve the quality of the model's answers. And for those questions that require comprehensive reasoning among different information fragments, it is able to provide more comprehensive information to help the model better understand the questions and give a more appropriate answer. After combining RAG with Bert and Llama3, we used evaluation metrics such as top-k, EM, and BLEU to evaluate the quality of the model's answers. At the same time, we also invited volunteers with backgrounds of TCM to evaluate different answers. From these two aspects, we verify that the combination of RAG and a large language model improves the answer quality of the model in TCM.

### 4.1 Model Selection and Training Procedure

In this part, we select several pre-trained LLMs to construct our QA-bot, including Bert-Base-Chinese and Llama-3-chinese-8b-instruct-v3. Eventually, we utilized Bart(combined with GPT3 and Bert)[17]. Compared to the former models. BART has higher computational efficiency and faster convergence of the error variance[18]. The TCM dataset we built earlier serves as the original data source for the entire process. In the process of data augmentation, we translated text from ancient Chinese into English and then double-translated the TCM Bert 1.0, 2.0, and Llama3-TCM.

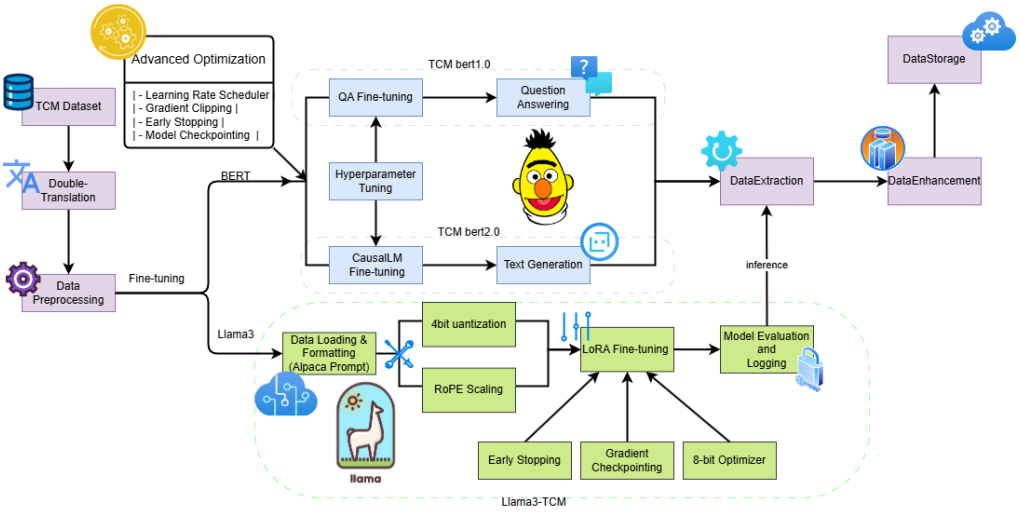



**Fig. 4.** Overall fine-tuning procedures for TCM Bert 1.0, 2.0, and Llama3-TCM.

### 4.2 TCM Bert

As shown in Figure 4, we chose a few optimization methods, including learning rate scheduling (step decay and cosine annealing), gradient clipping (clip the gradient norm to a preset threshold), early stopping (training stops if the validation error does not decrease for max iterations), and model checkpointing (save the model every 500 training iterations). These are used to improve the stability and accuracy of model training. For Bert 1.0, we utilized QA Fine-tuning, which simply relies on supervised learning. The model learns to associate input questions with the given instructions from our dataset and generate appropriate answers when asked a new question and instruction. However, we found our TCM Bert 1.0 in a bad accuracy rate. Our model should not use the question-answering interface. This interface is to do extractive question answering according to the context content of the question to answer, but the context has no background information.

Therefore, CausalLM ought to be used for inference finetune for TCM Bert2.0. CausalLM is a type of autoregressive model. "Causal models are conceptual models that depict the causal mechanisms in a system. By exploring Causal models, they are capable of answering certain questions with Existing observational data instead of the need for an extra study [19]." In the process of Autoregressive Training, the models are trained to predict the next token by sifting them the preceding tokens in the sequence. For example, if the input is "The cat is", the model will predict the next word, which might be "eating". However, despite the improved logical flow and better structure of the output, the results still demonstrated a disappointingly low level of accuracy. "It is proved that CausalLM converges to stationary points; these points may not be optimal for the learning task at hand, indicating that causalLM is not the best choice for in-context learning [19]"

As mentioned above, the performance of training BERT model in our experiments is less than satisfactory because the parameters of Bert are much lower than other transformer-based models. As a consequence, the limitation appears when dealing with complex tasks or large, diverse

datasets, just like ours. A model with more parameters would perform better to capture the nuances of the data and yield more accurate results. Therefore, we then train a model based on llama3.

### 4.3 Llama3-TCM

This section of the flowchart illustrates the training process for the LLaMA3 model. Here's a breakdown of each component. After double-translation and Formatting (Alpaca Prompt), we employed RoPE Scaling. "RoPE encodes absolute positions using a rotation matrix while explicitly incorporating relative position dependencies in the self-attention mechanism. This approach enhances the model's ability to capture dependencies between different positions in a sequence [20]"

In order to save computational resources and train larger models on limited hardware, we utilize LoRA, a technique that reduces the number of trainable parameters by introducing low-rank adaptations to the model. "LoRA reduces the memory and computational requirements. This allows for faster training and lower resource consumption compared to traditional methods that do not utilize quantization [21]." After fine-tuning with other related techniques such as Early Stopping, Gradient Checkpointing, and 8-bit Optimizer, the Llama3-TCM is able to generate the answer-based on the given context.

### 4.4 Bart with RAG

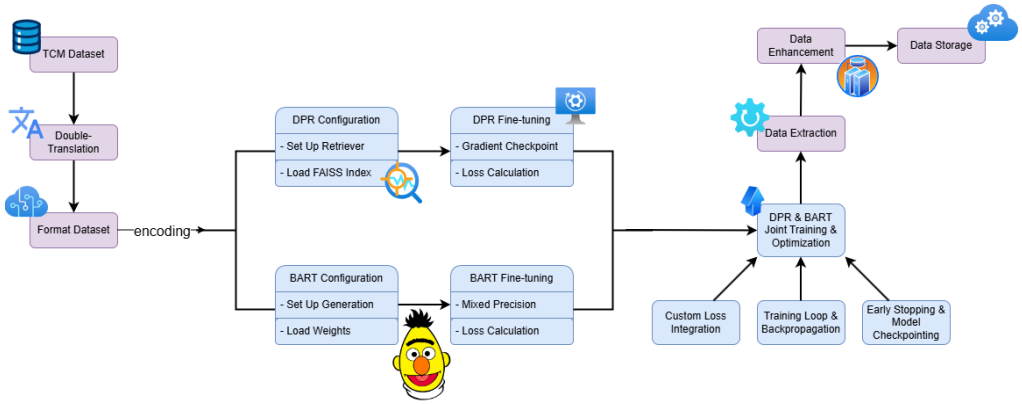



**Fig. 5.** Overall fine-tuning procedures for BART with RAG.

Although our TCM Bert and Llama3-TCM demonstrate great performance in generating appropriate answers based on context, these generation models have limited parameters, so it is particularly inefficient for handling complex or factual questions. Therefore, we decide to introduce RAG to complete the mission of Information retrieval enhancement for document classification.

"RAG models achieve state-of-the-art results on various knowledge-intensive tasks, particularly in open-domain question answering (QA). They combine the flexibility of generation with the performance benefits of retrieval-based approaches, outperforming traditional parametric models and task-specific architectures [22]." As Figure 5 shown, we utilize the RAG model,

which combines Dense Passage Retrieval (DPR) as a retriever and BART as a generator. Particularly, the key innovation lies in a loss function, which tightly integrates the outputs of the retriever (DPR) with the inputs to the generator (BART). The custom loss function plays a crucial role in this co-adaptation by guiding the retriever to focus on passages that are most relevant for the generator. This function will punish the overall model if the passages retrieved by DPR can't correctly contribute to generating accurate responses in BART. Through DPR & BART Joint Training & Optimization, it ensures that both DPR and BART are not only fine-tuned individually but can also be co-adapted. This approach creates a close relationship between the retriever and the generator. By putting the outputs of DPR directly into BART's input pipeline, we achieve more contextually relevant and accurate text generation and enhance the overall performance of the model. The relevant score of these models will be shown in the next section.

### 4.5 The Combination of the Trained Llama3-Chinese Model and RAG

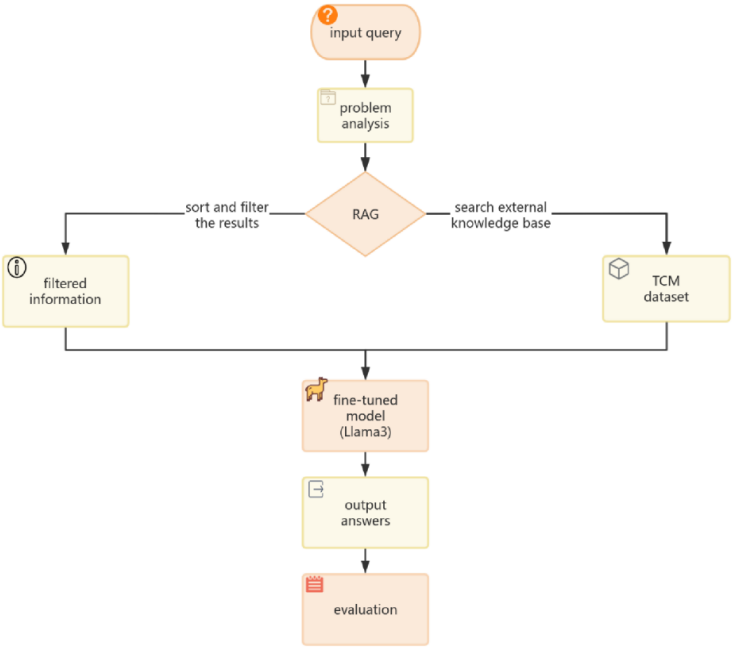



**Fig. 6.** The whole process of the combination

As shown in Figure 6, we combined RAG with the Llama 3 model. In the whole process, the user initially inputs a question. Then, RAG retrieves knowledge fragments strongly related to the user's question from our self-constructed database. After that, RAG feedback on the retrieved information to the Llama3-Chinese model. As a result, the llama model can give more accurate answers after receiving this more relevant information.

## 5 Results

### 5.1 Dataset Overview

The dataset, which includes real-world doctor-patient records and web-scraped TCM data, is partially displayed in the figure. The figure 7 shows the formula type composition of the dataset: soup: 26.53%, pill: 20.96%, powder: 20.35%, else: 19.97%, crystal: 6.49%, drugs for external use: 5.69%. The datasets show TCM names, their herbal compositions, and corresponding symptoms. Common prescriptions for liver-stomach Qi disorders include An Dong San, Chen Xiang Hua Qi Wan, Chen Xiang Qu, Ding Tong Wu Xiang San, Fo Shou Wan, and Gan Wei Qi Tong San. For instance, An Dong San comprises five herbs: Su Luo Zi, Wa Leng Zi, Chen Xiang Yuan, Chen Mu Gua, and Sheng Ge Ke. The prescribed dosage is 3 qian daily, mixed with brown sugar, with a reduced dosage for individuals with weak constitutions.

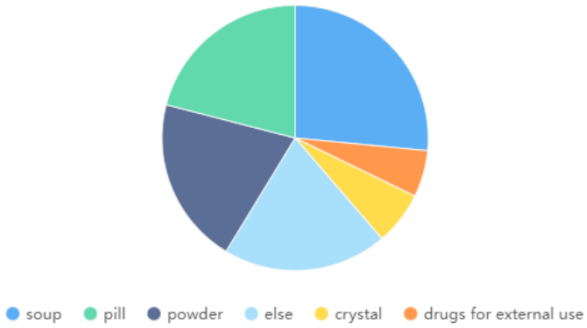



**Fig. 7.** Final Dataset Results

### 5.2 RAG

We evaluate the results given by the model from two aspects. On the one hand, we use evaluation metrics for the answer quality of LLMs, such as top-k, Exact Match, F1 Score, etc. On the other hand, we ask volunteers with backgrounds related to TCM to conduct manual evaluations of the model's answers.

The following is the formula for calculating these evaluation indicators:

$$R_1 = \frac{\sum_{n=1}^{N} \max(n_{n,m})}{\sum_{m=1}^{M} n_{m,m}} \tag{1}$$

$$F_1 = \frac{(1-\beta^2)P_1+\beta^2 R_1}{1+\beta^2(P_1-R_1)} \tag{2}$$

$$P_1 = \frac{\sum_{n=1}^{N} \max(n_{n,m})}{\sum_{n=1}^{N} n_{n,n}} \tag{3}$$

$$BARTScore = \frac{1}{N}\sum_{i=1}^{N} \cos\left(\mathrm{BERT}_{\mathrm{ref},i}, \mathrm{BERT}_{\mathrm{gen},i}\right) \tag{4}$$

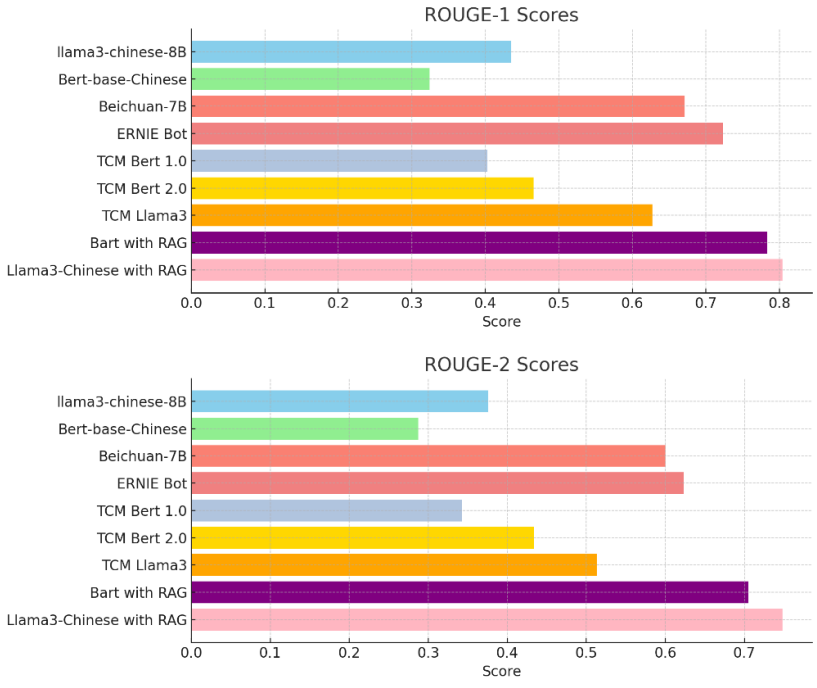
$$BLEU-4 = BP \cdot \exp\left(\sum_{n=1}^{4} \frac{P_n}{n}\right) \tag{5}$$

Here are the results of our tests:

Table 2: Evaluation results

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | BARTScore | BLEU-4 |
|---|---|---|---|---|---|
| llama3-chinese-8B | 0.4352 | 0.3757 | 0.4590 | -2.9823 | 0.3741 |
| Bert-base-Chinese | 0.3242 | 0.2874 | 0.3762 | -3.6495 | 0.2984 |
| Beichuan-7B | 0.6712 | 0.5998 | 0.6592 | -2.6473 | 0.6834 |
| ERNIE Bot | 0.7231 | 0.6237 | 0.7095 | -2.6158 | 0.7024 |
| TCM Bert 1.0 | 0.4028 | 0.3429 | 0.4158 | -3.2547 | 0.3852 |
| TCM Bert 2.0 | 0.4656 | 0.4337 | 0.4429 | -3.0479 | 0.4829 |
| TCM Llama3 | 0.6273 | 0.5132 | 0.5893 | -2.8537 | 0.5683 |
| **Bart with RAG** | 0.7836 | 0.7048 | 0.7692 | -2.4828 | 0.7427 |
| **Llama3-Chinese with RAG** | 0.8042 | 0.7483 | 0.7959 | -2.5841 | 0.7629 |

For a more intuitive comparison of results, we use bar charts:



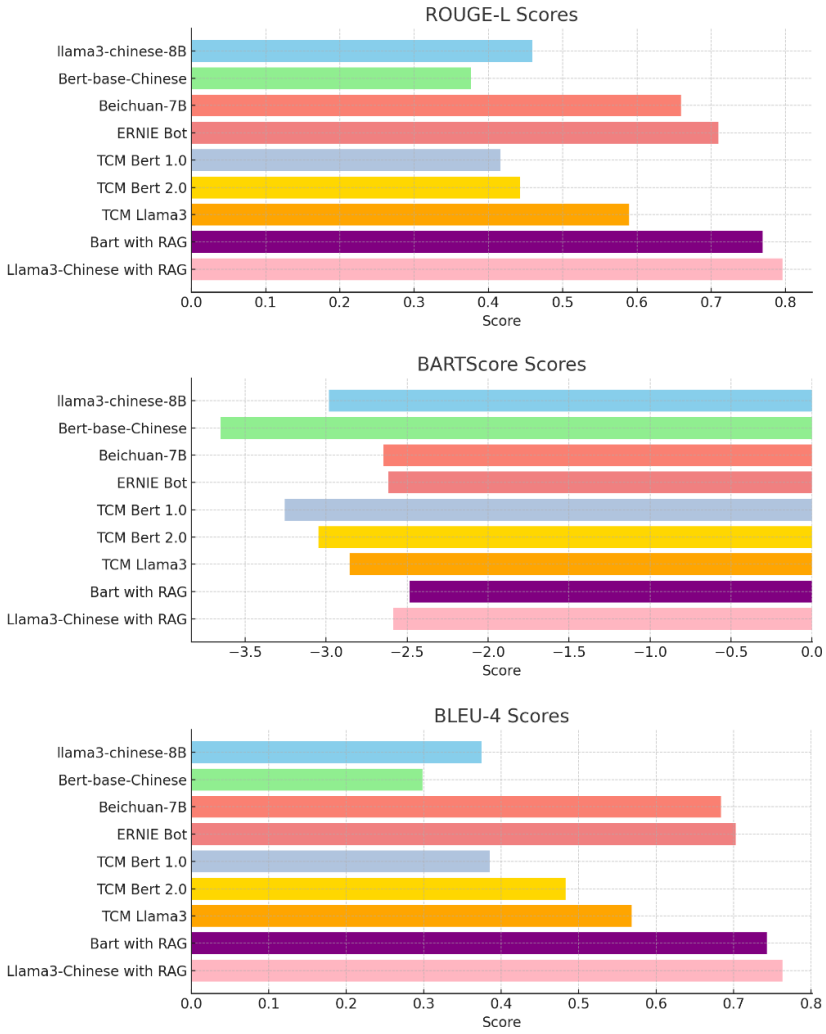



**Fig. 8.** Evaluation results

As shown in Table 2 and Figure 8, it is obvious from the bar charts that the model combined with RAG works best, and the trained models perform better than the untrained ones. At the same time, the results of our various indicators are basically consistent, which also verifies the superiority of our RAG and model combination.

## 6 Conclusion

In the study, we constructed a large dataset of TCM, which includes prescriptions, drugs, dosages and indications. This innovation not only fills the gap in the data resources of TCM but also lays a foundation for promoting the intelligent development of TCM.

With the data set constructed by ourselves, we trained many LLMs, including LLaMA and BERT. Then, we make a TCM question-answering robot. At the same time, we further optimized the accuracy and reliability of the model answers by introducing RAG technology. This is also the first time RAG technology has been applied in the field of TCM.

Our contribution lies not only in the construction of datasets and technological innovation, but also in the promotion of TCM cultural inheritance and innovation. Through intelligent means, we make the ancient wisdom of TCM full of new vitality, so that more people can easily obtain professional knowledge of TCM, and promote the popularization and application of TCM.

We know that this is just a starting point for the intelligent development of TCM. Next, we will continue to explore the integration of multi-modal technology applications, aiming to fully integrate the four concepts of TCM "look, smell, ask, and cut" into intelligent robots to achieve more comprehensive and accurate TCM diagnosis and treatment services. We believe that with the continuous progress of technology and the deepening of application, the intelligence of TCM will usher in a broader development prospect and contribute more wisdom and strength to the cause of human health.

## Acknowledgement

Enpu Zuo, Chenxi Pan, Junyu Chen, and Zihan Yi contributed equally to this work and should be considered co-first authors.

## References

[1] Xu, Judy, and Yue Yang. "Traditional Chinese medicine in the Chinese health care system." Health policy 90, no. 2-3 (2009): 133-139.

[2] Jiang, Wen-Yue. "Therapeutic wisdom in traditional Chinese medicine: a perspective from modern science." Trends in pharmacological sciences 26, no. 11 (2005): 558-563.

[3] Huang K, Zhang P, Zhang Z, Youn JY, Wang C, Zhang H, Cai H. Traditional Chinese Medicine (TCM) in the treatment of COVID-19 and other viral infections: Efficacies and mechanisms. Pharmacol Ther. 2021 Sep;225:107843. doi: 10.1016/j.pharmthera.2021.107843. Epub 2021 Mar 31. PMID: 33811957; PMCID: PMC8011334.

[4] Naveed, Humza, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. "A comprehensive overview of large language models." arXiv preprint arXiv:2307.06435 (2023).

[5] Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, Ashrafian H, Darzi A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med. 2021 Apr 7;4(1):65. doi: 10.1038/s41746-021-00438-z. PMID: 33828217; PMCID: PMC8027892.

[6] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8. Epub 2023 Jul 17. PMID: 37460753.

[7] Cui, Yiming, Ziqing Yang, and Xin Yao. "Efficient and effective text encoding for Chinese llama and alpaca." arXiv preprint arXiv:2304.08177 (2023).

[8] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[9] Chen, Calvin Yu-Chian. "TCM Database@ Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico." PloS one 6, no. 1 (2011): e15939.

[10] Xu HY, Zhang YQ, Liu ZM, Chen T, Lv CY, Tang SH, Zhang XB, Zhang W, Li ZY, Zhou RR, Yang HJ, Wang XJ, Huang LQ. ETCM: an encyclopaedia of traditional Chinese medicine. Nucleic Acids Res. 2019 Jan 8;47(D1):D976-D982. doi: 10.1093/nar/gky987. PMID: 30365030; PMCID: PMC6323948.

[11] Fang, ShuangSang, Lei Dong, Liu Liu, JinCheng Guo, LianHe Zhao, JiaYuan Zhang, DeChao Bu et al. "HERB: a high-throughput experiment-and reference-guided database of traditional Chinese medicine." Nucleic acids research 49, no. D1 (2021): D1197-D1206

[12] Lin Huang, Duoli Xie, Yiran Yu, Huanlong Liu, Yan Shi, Tieliu Shi, Chengping Wen, TCMID 2.0: a comprehensive resource for TCM, Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D1117–D1120, https://doi.org/10.1093/nar/gkx1028

[13] Izacard, Gautier, and Edouard Grave. "Leveraging passage retrieval with generative models for open domain question answering." arXiv preprint arXiv:2007.01282 (2020).

[14] Hua, Rui, Xin Dong, Yu Wei, Zixin Shu, Pengcheng Yang, Yunhui Hu, Shuiping Zhou et al. "Lingdan: enhancing encoding of traditional Chinese medicine knowledge for clinical reasoning tasks with large language models." Journal of the American Medical Informatics Association (2024): ocae087.

[15] Zhang, Heyi, Xin Wang, Zhaopeng Meng, Yongzhe Jia, and Dawei Xu. "Qibo: A Large Language Model for Traditional Chinese Medicine." arXiv preprint arXiv:2403.16056 (2024)

[16] Tan, Yang, Zhixing Zhang, Mingchen Li, Fei Pan, Hao Duan, Zijie Huang, Hua Deng et al. "MedChatZH: A tuning LLM for traditional Chinese medicine consultations." Computers in Biology and Medicine 172 (2024): 108290..

[17] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." *arXiv preprint arXiv:1910.13461* (2019).

[18] Sparapani, Rodney, Charles Spanbauer, and Robert McCulloch. "Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package." *Journal of Statistical Software* 97 (2021): 1-66.

[19] Ding, Nan, Tomer Levinboim, Jialin Wu, Sebastian Goodman, and Radu Soricut. "CausalLM is not optimal for in-context learning." arXiv preprint arXiv:2308.06912 (2023).

[20] Su, Jianlin, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. "Roformer: Enhanced transformer with rotary position embedding." *Neurocomputing* 568 (2024): 127063.

[21] Xu, Yuhui, Lingxi Xie, Xiaotao Gu, Xin Chen, Heng Chang, Hengheng Zhang, Zhensu Chen, Xiaopeng Zhang, and Qi Tian. "Qa-lora: Quantization-aware low-rank adaptation of large language models." *arXiv preprint arXiv:2309.14717* (2023).

[22] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.