ShuZhiQiHuang

Anonymous ACL submission

Abstract

In recent years, large language models (LLMs) have made significant progress in understanding and responding to user intentions. However, their application in specialized fields such as Traditional Chinese Medicine (TCM) still faces challenges, primarily due to the lack of domain-specific knowledge and the unique theoretical system of TCM. To address the specific needs of the TCM field, we propose a novel model that combines pre-training, instruction-supervised fine-tuning, and retrieval-augmented generation (RAG) techniques. This model aims to resolve the problem of hallucinations in domain-specific knowledge and enhance the understanding and application of TCM knowledge. In this study, we developed "ShuZhiQiHuang," an LLM-driven Q&A platform focused on TCM. This platform not only intelligently answers questions related to Chinese herbs, prescriptions, and syndromes but also covers tasks such as prescription retrieval, personalized recommendations, disease diagnosis reasoning, and health recipe suggestions. We extensively collected TCM-related data for pre-training and constructed a high-quality TCM knowledge graph, using ChatGPT-4 to generate a rich instruction dataset. Comprehensive evaluation shows that ShuZhiQiHuang achieves a 71% accuracy rate in TCM knowledge Q&A, significantly outperforming mainstream open-source models, and has successfully passed the TCM practitioner qualification mock exam for the first time, with an accuracy improvement of 16.39% compared to ChatGPT-4. Additionally, the retrieval-augmented generation technique further enhances its Q&A capabilities in the TCM field, effectively addressing the issue of knowledge hallucinations. These results highlight the contributions of our integrated approach in developing LLM applications for TCM.

1 Introduction

In the field of artificial intelligence, open-source large language models (LLMs) like PaLM(Chowdhery et al., 2023), LLaMA(?), Bloom(Workshop et al., 2022), the GPT series(Brown et al., 2020)(Achiam et al., 2023), and ChatGLM(Zeng et al., 2022)(Du et al., 2021) have shown significant progress in natural language processing (NLP) tasks. These advancements have prompted efforts to adapt general-purpose LLMs to specialized fields such as biomedicine, law, and finance. However, these models often underperform in these domains due to the complexity of domain-specific knowledge and computational efficiency issues.

In traditional Chinese medicine (TCM), several TCM LLMs have been proposed ((Zhang et al., 2024),(Touvron et al., 2023b),(Zhang et al., 2023),(Chen et al., 2023),(Du et al., 2023b)), primarily through continued pre-training (CPT) and supervised fine-tuning (SFT). This approach can lead to issues such as erroneous outputs and hallucinations(Rawte et al., 2023)(Umapathi et al., 2023), which hinder real-world application. Additionally, the timeliness limitation of fine-tuning data challenges the continuous updating of knowledge in these models. Due to the complexity and professional requirements of medical knowledge, developing successful LLMs in medicine demands higher accuracy and safety(Singhal et al., 2023a). Retrieval-Augmented Generation (RAG) (Gao et al., 2023) addresses these issues by retrieving authoritative knowledge bases, thereby reducing hallucinations and maintaining output timeliness and accuracy.TCM has a unique and complex theoretical framework, integrating cultural, philosophical, and historical aspects, making the acquisition and organization of TCM data challenging. Moreover, TCM often lacks large-scale clinical trials and data validation, complicating the development of LLMs in this field. Hence, developing reliable and valuable LLMs for TCM to meet modern medicine's needs is an urgent task.

We introduce the "ShuZhiQiHuang" agent to promote AI in the medical field, focusing on traditional Chinese medicine (TCM). The model, encompassing extensive TCM knowledge, is built through continuous pre-training and fine-tuning. Initially, a large corpus of classical and modern TCM literature was compiled and pre-trained using Qwen1.5-14B(Bai et al., 2023b) to understand TCM language and concepts. The model was then fine-tuned with datasets for natural language processing tasks and dialogues from the TCM knowledge graph, covering TCM diagnosis, treatment recommendations, and herbal prescriptions. Supervised fine-tuning (SFT) enabled accurate TCM information and complex interactive dialogues. Retrieval-Augmented Generation (RAG) technology further enhanced response accuracy and effectiveness.Additionally, the "ShuZhiQiHuang" website offers a multi-modal platform using AI extraction and mining algorithms to build an extensive TCM knowledge graph. This includes biomedical and TCM entities, such as syndromes, prescriptions, and ingredients, with relational data. The platform provides comprehensive medical information retrieval and multi-modal question-answering, including diagnostic decision-making, clinical prescription formulation, and herbal formula analysis. ShuZhiQiHuang aims to assist TCM professionals in research and practice, popularize TCM knowledge, and promote the modernization and internationalization of TCM, benefiting more people globally.

The main contributions of this paper are as follows:

- We developed "ShuZhiQiHuang," a multimodal knowledge retrieval and question-answering web platform tailored for traditional Chinese medicine (TCM). This Large Language Model (LLM) integrates extensive TCM knowledge and continuously improves through pre-training and fine-tuning. Using Retrieval-Augmented Generation (RAG) technology, it supports TCM Q&A and diagnostic decision-making. Based on TCM knowledge graphs and relevant literature, it provides intelligent health dialogue counseling, aiding TCM professionals in scientific medical decision-making and treatment planning.
- The ShuZhiQiHuang model has successfully passed the Traditional Chinese Medicine Practitioner Qualification Examination for the first time, surpassing existing mainstream large-scale models in the medical field and the best general large language model, GPT-4, in areas such as basic theory of traditional Chinese medicine, traditional Chinese medicine diagnostics, internal medicine, and medical ethics.
- We also develop a WeChat Mini Program to provide a more convenient user experience, allowing users to access medical information and engage in health consultations more easily.

2 Related Work

2.1 Large language model

The remarkable achievements of Large Language Models (LLMs) such as ChatGPT (Brown et al., 2020) and GPT-4 (Achiam et al., 2023) have garnered substantial attention, igniting a new wave in the field of AI. Although OpenAI has not disclosed their training strategies or weights, with the continuous development of LLMs technology, an increasing number of open-source projects have emerged. For instance, Meta's open-source LLaMA (Touvron et al., 2023a)(Touvron et al., 2023b) has surpassed GPT-3(Floridi and Chiriatti, 2020) on multiple benchmarks, and there are also open-source LLMs like ChatGLM (Zeng et al., 2022)(Du et al., 2021), Baichuan(Yang et al., 2023a), Qwen(Bai et al., 2023b), and Yi(AI et al., 2024) that support multiple languages including Chinese and English, which have quickly caught the attention of the research community.

2.2 Large language model in medical Domain

In the medical field, LLMs typically face challenges such as limited domain-specific data, adapting to new knowledge, behavior alignment, ethical, legal, and safety issues, and hallucination generation. Notable initiatives include BioBERT(Lee et al., 2020), which was pre-trained on Pubmed and PMC data. GatorTron(Yang et al., 2022) leveraged the University of Florida (UF) Health electronic health records (EHRs), containing over 50 million interactions from 2 million patients, for training. DoctorGLM(Xiong et al., 2023) and ChatDoctor(Yang et al., 2023b) were obtained by fine-tuning general-purpose LLMs ChatGLM(Zeng et al., 2022)(Du et al., 2021) and

LLaMA(Touvron et al., 2023a)(Touvron et al., 2023b) on doctor-patient dialogue data, respectively. Additionally, Med-PaLM(Singhal et al., 2023a) and Med-PaLM-2(Singhal et al., 2023b) had qualified clinicians construct instruction data to fine-tune PaLM. HuaTuo(Zhang et al., 2023) and ChatGLM-Med(Haochun Wang, 2023) built knowledge-based instruction data from knowledge graphs to inject medical knowledge into LLMs, thus improving downstream performance. Qilin-Med(Ye et al., 2023) demonstrated significant performance improvements through a training process involving Domain-Continued Pre-Training (DCPT), Supervised Fine-Tuning (SFT), and Direct Preference Optimization (DPO). Chinese medical LLMs, such as HuaTuo(Wang et al., 2023) and Zhongjing(Yang et al., 2024), were fine-tuned on LLaMA(Touvron et al., 2023a)(Touvron et al., 2023b) by incorporating the knowledge graph CMeKG(Byambasuren et al., 2019). In the field of TCM, CMLM-ZhongJing(Kang et al., 2023) is the first TCM LLM fine-tuned from Qwen1.5-1.8B-Chat(Bai et al., 2023b). Qibo(Zhang et al., 2024) is a large language model specifically constructed for the TCM field, based on the LLaMA(Touvron et al., 2023a)(Touvron et al., 2023b) model, with organization and training on a specialized corpus providing it with TCM theoretical knowledge. On the other hand, TCM-GPT(Yang et al., 2023b) constructed a TCM-specific corpus, TCM-Corpus-1B(Yang et al., 2023b), and utilized LoRA(Hu et al., 2021) technology for pre-training and fine-tuning the model to better adapt to TCM-related tasks.

3 Approach

3.1 architecture

In this research, we propose a customizable intelligent question-answering system for integrating Traditional Chinese Medicine (TCM). This system combines advanced technologies such as large language models (LLM), retrieval-augmented generation (RAG) (Gao et al., 2023), and tool-calling (Huang et al., 2024). The framework is structured into three principal layers: the Foundation Layer, the Domain Layer, and the Application Layer, as shown in Figure 1. Foundation Layer: Utilizes Qwen1.5-14B (Bai et al., 2023b) as the TCM question-answering agent, enhanced with extensive TCM knowledge. This includes a TCM knowledge graph with 150,000 nodes and 5 million relationships, over 9,000 medicinal herbs, 40,000 chemical constituents, 2,000 syndromes, 80,000 prescriptions, 110,000 diagnostic dialogues, and 1,000 classical TCM texts. Domain Layer: Uses RAG and tool-calling techniques to connect LLMs with practical TCM knowledge. It synthesizes data from the Foundation Layer into domain-specific knowledge, enabling intelligent agents to provide accurate, context-aware responses in medical question-answering tasks. Application Layer: Applies the integrated TCM knowledge to real-world medical challenges, including prescription retrieval, personalized recommendations, disease diagnosis, and health recipe suggestions. This layer utilizes the organized data and advanced computational models from the underlying layers to deliver effective medical solutions.

3.2 Continuous Pre-training

Continuing pre-training involves using a large amount of new, unstructured data to refine the foundational model. At this stage, utilizing a high-quality corpus can significantly enhance the performance of large language models (LLMs), potentially surpassing traditional scaling laws (Gunasekar et al., 2023). In the field of Traditional Chinese Medicine (TCM), continuous pre-training adjusts the model parameters to absorb domain-specific expertise, style, terminology, and principles. Given the complexity and breadth of TCM, emphasizing data diversity and quality is crucial. The TCM field encompasses rich knowledge and skills, requiring comprehensive training akin to that of professional doctors and pharmacists.To this end, we have collected various authentic and relevant text data, primarily including modern medical books, TCM textbooks, TCM prescription datasets, TCM reading comprehension test data, TCM treatment plan standards, classic TCM medical literature, TCM encyclopedias, and other corpora related to TCM theoretical characteristics. These datasets cover various departments and aspects of the medical field, providing a rich knowledge base for the model. The statistics of pre-training data are shown in Table 1. These datasets cover various aspects of TCM, with a total size of approximately 1.5GB, providing a rich TCM knowledge base for the model. To reduce the risk of catastrophic forgetting (Ren et al., 2024), where the model might lose its original knowledge capacity when learning a large amount of non-domain-specific data,

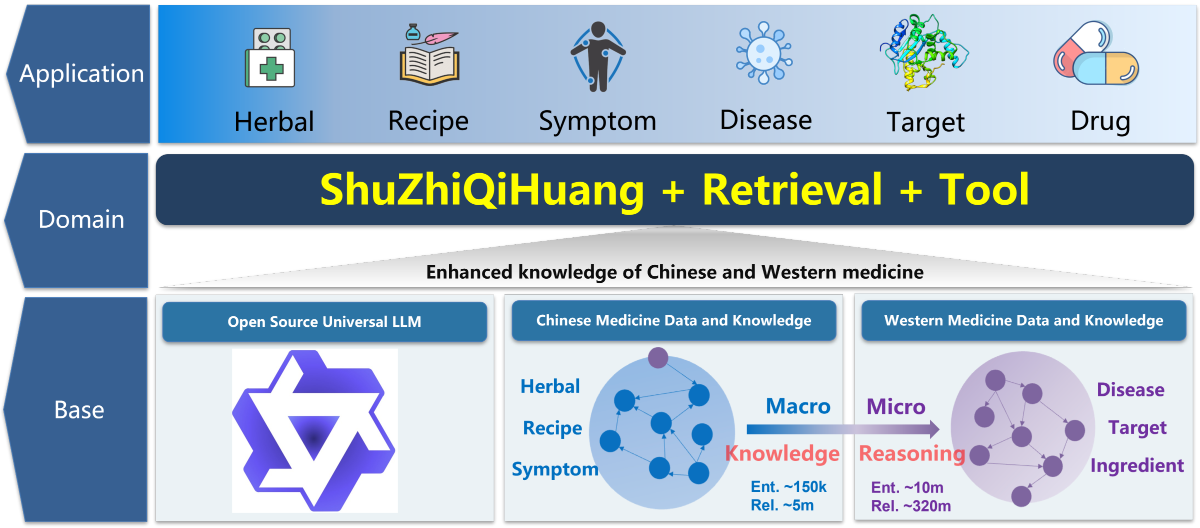

Figure 1: Framework for an Intelligent Question-Answering System Integrating Traditional Chinese Medicine and Western Medicine.

Table 1: Statistics of pre-training data for the TCM model.

Dataset	Type	Size
TCM Books	Textbook	394MB
ChatMed	Q&A	385MB
Medical Wikidoc	Wiki_Data	10MB
CMtMedQA	Q&A	151MB
Chinese Medical dialogue data	Medical report	564MB
General Data	General Knowledge	3GB

we have included some general data in the pre-training process. This general data comes from the high-quality corpus "Wanjuan-CC" (Qiu et al., 2024) released by the Shanghai Artificial Intelligence Laboratory. The ratio of domain-specific data to general data is set at 1:2, with approximately 3GB of data extracted from the corpus as general domain data.

3.3 Supervised Instruction Fine-Tuning

Supervised fine-tuning (SFT) involves adapting a pre-trained Language Model (LLM) to a specific downstream task using labeled data (Shumailov et al., 2023). Unlike unsupervised techniques, where data is not pre-validated, supervised fine-tuning employs a dataset of responses that have been meticulously validated beforehand. This distinction is crucial, as LLM training is typically unsupervised, whereas fine-tuning processes are generally supervised. During the SFT process, the pre-trained LLM undergoes further training on the labeled dataset utilizing supervised learning techniques (Dong et al., 2023). The model's weights are adjusted based on gradients derived from the task-specific loss, which quantifies the discrepancy between the model's predictions and the ground truth labels. This adjustment allows the model to internalize task-specific patterns and nuances embedded in the labeled data. By fine-tuning its parameters to align with the specific data distribution and requirements of the task, the LLM becomes proficient in executing the target task. This specialized training process enhances the model's ability to generate accurate and contextually relevant outputs, thereby improving its performance on domain-specific applications.

3.3.1 Knowledge graph-based data for Chinese medicine instruction dataset

In constructing TCM instruction data, we utilized a comprehensive knowledge graph we built, which contains over 80,000 prescriptions, 9,000 medicinal herbs, 40,000 chemical constituents, 1,000 classical TCM texts, and 2,000 syndromes. We employed an entity-centric self-instruction method

(Wang et al., 2022), focusing on the diverse entities within the knowledge graph. Utilizing ChatGPT-4, we generated a rich variety of Q&A pairs related to these entities, ensuring the comprehensiveness and contextual relevance of the instruction content. This approach enhances the model's ability to understand and apply TCM knowledge. The generated instruction data provides a solid foundation for training the model to effectively handle the complexity of the TCM field, thereby improving its utility in both research and clinical settings. Ultimately, this process resulted in 185,315 TCM-related instruction data points, with the data sources shown in Figure 2.

3.3.2 Instruction data filter

Recent work (Cao et al., 2023)(Du et al., 2023a) has shown that what determines model performance is not purely the amount of data but the quality of the data. Even a limited amount of high-quality data collected manually can improve a model's instruction-following capability. Therefore, the challenge of automatically identifying high-quality data from a vast amount of available data has become a classic problem. We will use a method (Li et al., 2023) that quantifies the difficulty of each sample for the model by calculating an "Instruction Following Difficulty" (IFD) score to autonomously select training samples from the collected data. The construction of instruction data for Traditional Chinese Medicine (TCM) involves three core stages. First, the model learns from simple experiences by training on a diverse subset of the target dataset to develop basic instruction-following capabilities. This is achieved by embedding instructions and clustering them to ensure diversity, followed by initial training. Second, the model is evaluated based on experience by calculating the Conditioned Answer Score:

$$L_\theta(A|Q) = \frac{1}{N}\sum_{i=1}^{N} \log P(w_i^A|Q, w_1^A, \ldots, w_{i-1}^A; \theta),$$

and the Direct Answer Score:

$$s_\theta(A) = \frac{1}{N}\sum_{i=1}^{N} \log P(w_i^A|w_1^A, \ldots, w_{i-1}^A; \theta),$$

quantifies the difficulty of following each instruction. Finally, in the re-training phase, the model is re-trained on samples ranked by their IFD scores, allowing the model to refine its instruction-following abilities based on self-guided experience. This structured approach ensures the model remains highly specialized and effective in handling complex TCM instructions.

3.4 Retrieval-augmented Generation

Retrieval-augmented generation (RAG) provides large language models (LLMs) with supplementary information from external knowledge sources, enabling them to generate more accurate and contextually appropriate answers while reducing hallucinations. The process involves embedding the user's query into a high-dimensional vector space using an advanced embedding model(Shahmirzadi et al., 2019), integrating the query into a structured prompt template with retrieved context information, and feeding the enhanced prompt into an LLM to generate a response. To optimize the retrieval of relevant traditional Chinese medicine (TCM) knowledge, text is segmented using Chinese recursive chunking(Chen et al., 2006) and semantic segmentation methods(Zhang et al., 2021). Chinese recursive chunking divides text into meaningful chunks by identifying nested structures, while semantic segmentation slices text based on semantic coherence. The combination of these methods allows for the extraction of contextually relevant segments, aiding the LLM in better extracting and summarizing knowledge. We use Elasticsearch(Elasticsearch, 2018), an open-source search engine, to store and retrieve data, employing a hybrid method that combines vector and keyword search.

Retrieval-augmented generation (RAG) enhances model responses' accuracy and credibility by integrating over 1,000 Traditional Chinese Medicine (TCM) documents, including ancient texts, prescriptions, and modern books. This retrieval of precise, context-relevant information from a comprehensive TCM database improves the reliability and transparency of TCM-focused LLMs in research and clinical settings. To ensure real-time, up-to-date responses, we integrate Bing as a search engine within the retrieval-generation framework. This involves retrieving search results based on user queries, converting these results into document objects, and integrating them with a language model to generate context-appropriate responses using predefined prompt templates. This approach ensures continuous knowledge updating and the integration of domain-specific information. LangChain(Topsakal and Akinci, 2023) provides

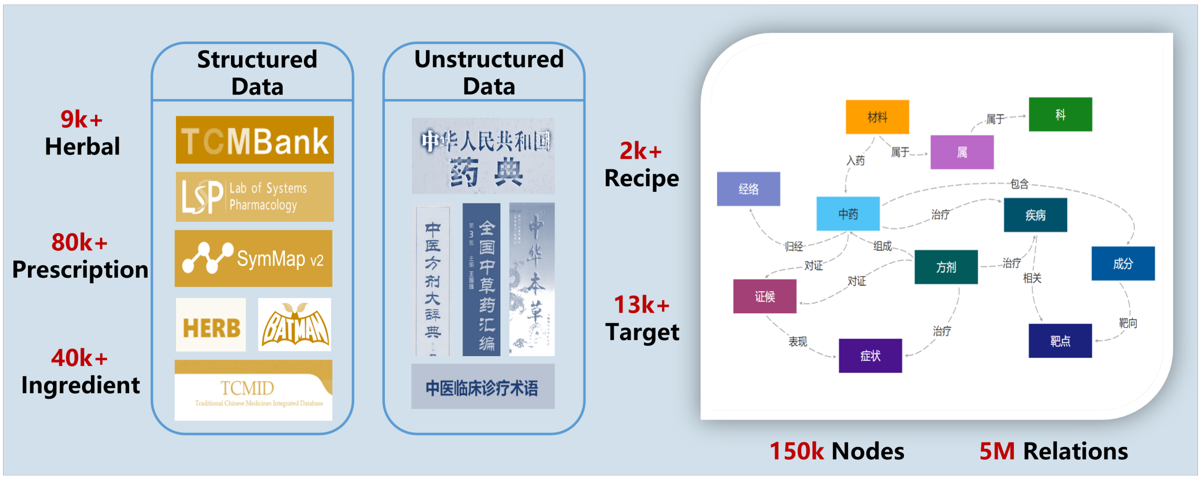

Figure 2: Chinese Medicine Instruction Data Sources.

the building blocks for these RAG applications, as illustrated in the figure3.

3.5 Tools-calling

Tool-augmented LLMs(Naveed et al., 2024) leverage LLMs' reasoning abilities to plan tasks iteratively by breaking them into sub-tasks, selecting necessary tools, and executing actions. This enhances model capabilities and expands their application scenarios. Integrating external tools improves performance in knowledge acquisition, professional skill enhancement, automation, efficiency, and interaction. For instance, limited knowledge base text can be supplemented using search engine APIs. Additionally, TCM models can call command-line APIs to visualize chemical expressions and names mentioned during interactions, enriching the user experience. In this study, we implemented two functionalities through tool invocation: Firstly, we have developed a command-line interactive tool designed to visualize chemical formulas and names mentioned during the interaction between users and the model, thereby enriching the user experience. For detailed implementation and demonstration of the effects, please refer to the Appendix A. Specifically, we followed the approach of Toolformer(Schick et al., 2023) to enhance the awareness and tool-using capabilities of LLMs (Large Language Models), utilizing the RDKit(Bento et al., 2020) small molecule Python package installed in the command-line interaction tool to ensure that the model autonomously triggers function calls when encountering SMILES expressions in dialogues. Initially, we designed the text format for function calls, specifying the input parameters as SMILES expressions and setting the function return value as the properties of chemical elements. Subsequently, we designed 1,000 dialogue data entries to fine-tune the model, ensuring that the model generates tool invocation annotations promptly when producing SMILES expressions during output. After the model completes its output, we further extract and execute the tool invocation annotations generated by the model, thereby completing the visualization of SMILES. The second functionality involves using search engine invocation tools, where we can call various types of search engine APIs to supplement knowledge not included in the knowledge base, enhancing the model's question-answering capabilities.

3.6 Deployment of the Traditional Chinese Medicine QA System

To enhance the accessibility and usability of the Traditional Chinese Medicine (TCM) QA system, we have implemented two distinct deployment methods: a web application and a WeChat Mini Program. These deployment strategies ensure that users can conveniently access the system from various platforms, thereby broadening its reach and utility. The web application provides a comprehensive interface with multiple conversation modes, while the WeChat Mini Program offers streamlined intelligent QA functionality directly on users' mobile devices. Additionally, we have incorporated vLLM(Kwon et al., 2023) inference acceleration technology to improve response speed.

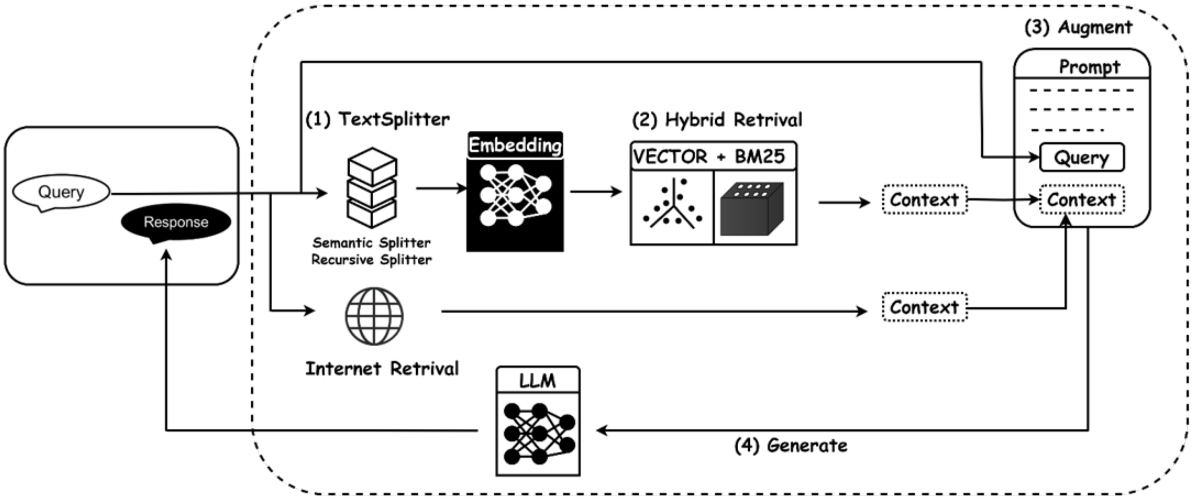

Figure 3: Retrieval-augmented Generation implementation Flowchart.

3.6.1 Web-page deployment

We deployed the Traditional Chinese Medicine (TCM) QA system as a web application with a user-friendly interface,as shown in the Appendix A. The main page includes key features for optimal user interaction and information retrieval:

1.**Intelligent QA** - This mode utilizes a language model fine-tuned with data from the TCM field to provide automated, contextually relevant answers to user queries about TCM.

2.**Knowledge Base QA**- This mode uses the Retrieval-Augmented Generation (RAG) framework to retrieve information from external sources and integrate it with the language model for comprehensive answers.

Additional features include exporting conversation records and clearing chat history. A disclaimer informs users that the system does not replace professional medical advice. The chat interface allows users to input queries and receive real-time responses, utilizing the Bing search engine to retrieve and process the latest information for accurate answers.

3.6.2 WeChat Mini Program deployment

In addition to the web application, we deployed the Traditional Chinese Medicine (TCM) QA system as a WeChat Mini Program, offering users a convenient platform for TCM-related information,as shown in the Appendix A. The main interface emphasizes Intelligent QA functionality, allowing users to ask questions about Chinese herbal medicine, TCM prescriptions, and syndrome differentiation. Powered by a language model fine-tuned with TCM data, the intelligent QA feature ensures accurate, contextually relevant responses, enhancing user experience and accessibility on mobile devices.

4 Experiments

4.1 Training Details

For our application, we utilized Qwen1.5-14B(Bai et al., 2023a) as the base model. Qwen1.5-14B is an open-source, large-scale pre-trained language model built upon the Transformer architecture. Boasting 14 billion parameters and supporting multiple languages, it features a remarkable 32K context length, enabling it to handle complex and lengthy text inputs with ease. This approach was implemented through the transformers and peft libraries. The enhancement pre-training and instruction fine-tuning stages both employed the low-rank adaptation (LoRA) parameter-efficient fine-tuning method(Hu et al., 2021), with lora-rank = 16 and lora-alpha = 32. During the fine-tuning process, several cutting-edge training techniques were strategically employed to optimize efficiency and performance. These included training with bf16 precision, leveraging the Deep-Speed ZeRO-2(Rajbhandari et al., 2020) stage method for parallelization, and utilizing gradient accumulation methods to maximize computational resources. Additionally, we utilized the AdamW optimizer(Loshchilov and Hutter, 2017), a dropout rate of 0.1, and a cosine learning rate scheduler. To maintain training stability, we judiciously halved the loss during gradient bursts and learning rate decay. In the pre-training stage, a high learning rate, large batch size, and fewer epochs were employed to effectively capture broad language patterns from vast amounts of unlabeled data, allowing the model to build a robust foundation. Conversely, in the fine-tuning stage, a low learning rate, small batch size, and numerous epochs were utilized to make precise adjustments to the pre-trained model, enhancing its performance on specific tasks with labeled data without over-fitting. The hyper-parameters for the two fine-tuning stages are meticulously detailed in the accompanying table. Notably, the losses for all training stages successfully converged within an effective range.

4.2 Baselines

In this study, we evaluate the performance of various baseline models for medical dialogue systems. Our selection includes a range of state-of-the-art models developed by leading institutions, each with unique characteristics and strengths in handling medical data and generating relevant responses. The models included in our baseline comparison are:

1.ChatGPT(Gunasekar et al., 2023): Developed by OpenAI, these models represent advanced iterations of the renowned GPT family, lauded for their superior natural language processing capabilities. ChatGPT-3.5-turbo is a powerful general-purpose conversational agent capable of handling a wide range of topics, including medical consultations, while ChatGPT-4 builds upon this foundation with enhanced contextual understanding and response generation, providing more accurate and contextual answers. Given their exceptional performance in general conversational tasks, they collectively form a robust baseline for evaluating medical conversational systems. We compare them both as baseline models.

2.HuatuoGPT(Zhang et al., 2023):Developed by the Big Data Research Institute at The Chinese University of Hong Kong, HuatuoGPT leverages a sophisticated two-stage training approach to optimize its performance in medical consultations. Through supervised fine-tuning with a blend of distilled and real-world data, the model acquires both doctor-like expertise and patient-friendly conversational abilities. This is complemented by reinforcement learning techniques that enhance the generation of precise and contextually appropriate responses, ensuring high-quality interactions in diverse medical scenarios. The model boasts a substantial parameter count of 13 billion, allowing for nuanced understanding and generation of medical dialogue.

3.Bianque(Chen et al., 2023): Developed by the Academy of Future Technology at South China University of Technology, Bianque is a specialized model tailored for Chinese medical dialogues. Integrating extensive medical knowledge with advanced machine learning techniques, Bianque enhances dialogue quality and relevance. It utilizes the BianqueCorpus, a refined multi-turn health conversation dataset comprising 2,437,190 samples. The model is fine-tuned with a WarmupDecayLR learning rate scheduler and features 6 billion parameters, enabling it to handle complex medical conversations effectively.

4.BenTsao(Du et al., 2023b):Developed by Harbin Institute of Technology, BenTsao is a specialized Chinese medical dialogue model engineered to deliver precise and contextually relevant responses in medical interactions. Built upon the LLaMA-7B base model with 7 billion parameters, BenTsao integrates extensive medical literature and clinical data, leveraging structured knowledge from the Chinese medical knowledge graph CMeKG. This approach ensures comprehensive coverage of medical topics such as diseases, drugs, and symptoms. The model's training strategy emphasizes the generation of over 8,000 instruction instances tailored to enhance the model's ability to provide accurate responses aligned with medical facts and guidelines.

5.Zhongjing(Kang et al., 2023): Developed by Zhengzhou University, Zhongjing is a pioneering Chinese medical dialogue model featuring end-to-end training from pre-training to reinforcement learning from human feedback (RLHF). With 13 billion parameters, it significantly enhances understanding and generating relevant medical dialogues. The model utilizes the Chinese multi-turn medical dialogue dataset CMtMedQA, containing 70,000 authentic doctor-patient dialogues, to improve its capability for complex dialogue and proactive inquiry. The comprehensive training pipeline and refined annotation rules enable Zhongjing to outperform baselines and match ChatGPT's performance in some areas, demonstrating the effectiveness of its medical knowledge and instruction-following abilities.

Table 2: Hyperparameters for Continuous Pre-training and Supervised Instruction Fine-Tuning

Parameters	**CPT**	**SFT**
Batch_size	32	8
learning_rate	1e-4	1e-5
Epoch	1	2
Gradient_accumulation_steps	3	3
lr_scheduler_type	cosine	cosine
Precision	bf16	bf16

4.3 The performance of hybrid retrieval

The retrieval process adopts a sophisticated hybrid search strategy(Omrani et al., 2024), which synergistically combines vector similarity and BM25 text similarity. This approach harnesses the efficiency of vector search and the flexibility of keyword search, enabling faster, more accurate, and diverse retrieval results. The model employed for text vectorization is the state-of-the-art bge-large-zh(Chen et al., 2024), while the BM25(Robertson et al., 2009) algorithm leverages the robust capabilities of Elasticsearch.Initially, the query and document are vectorized using a dual-path recall mechanism, simultaneously performing similarity calculations and tokenized full-text search. These complementary queries are executed in parallel, with the results subsequently merged and ranked according to a predefined logic, such as weighted average or rank fusion. This hybrid search approach elegantly overcomes the inherent limitations of both vector and keyword search methodologies, significantly improving retrieval accuracy and reliability.Subsequently, the user query is seamlessly integrated into a structured prompt template, along with the retrieved context information, enriching the prompt with domain-specific insights. Finally, this enhanced prompt is fed into a large language model (LLM), which generates a response leveraging the augmented context. This meticulous process ensures that the generated output is not only contextually appropriate but also incorporates precise and relevant TCM knowledge.

To comprehensively evaluate the retrieval performance, we have randomly extracted data from the knowledge base to construct 100 knowledge documents in the JSON format. These documents are integrated to serve as a dedicated retrieval test dataset. Each knowledge document encapsulates specific Traditional Chinese Medicine (TCM) knowledge points, questions related to these knowledge points, and the corresponding ID numbers of the knowledge points. During the retrieval process, we utilize the questions within the knowledge documents as retrieval queries to obtain the ID and score of the retrieval results. Subsequently, we employ three widely-adopted evaluation metrics: HitRatio@k, Precision@k, and Mean Reciprocal Rank (MRR) to conduct a comparative analysis of the three retrieval methods' efficacy. The HitRatio@k metric assesses whether at least one correct answer is present within the top k results. Its calculation formula is as follows:

$$\text{Hit Ratio@k} = \frac{\text{Number of at least hitting } k}{\text{Total number of queries}}$$

Precision@k evaluates the proportion of relevant results among the top k retrieved items. The calculation formula is:

$$\text{Precision@k} = \frac{\text{Number of relevant docs in top } k}{k}$$

Mean Reciprocal Rank (MRR) is a metric that considers the position of the first correct answer within the results. Its calculation formula is:

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$$

By calculating the aforementioned metrics, we obtained the comparative results of the three retrieval methods, presented in the following table 2. The hybrid retrieval approach achieved the optimal performance across all three metrics. With high Precision@k, Hit Ratio@k, and MRR scores, the hybrid retrieval can more accurately locate relevant documents, ensuring that a greater number of correct answers are included in the top results, and that the most pertinent documents are retrieved more promptly. This method not only enhances the diversity and accuracy of the retrieval results but also improves user satisfaction, enabling users to find the information they need more efficiently.

Metric	Top-1	Top-3	Top-5	Top-10
BM25-HitRatio@K	0.44	0.55	0.63	0.73
Vector-HitRatio@K	0.84	0.95	0.96	0.96
Hybrid-HitRatio@K	**0.84**	**0.96**	**0.97**	**0.98**
BM25-Precision@K	0.44	0.18	0.13	0.07
Vector-Precision@K	0.84	0.32	0.19	0.10
Hybrid-Precision@K	**0.84**	**0.46**	**0.29**	**0.15**
BM25-MRR	-	-	-	0.52
Vector-MRR	-	-	-	0.88
Hybrid-MRR	-	-	-	**0.90**

Table 3: Comparison of BM25, Vector, and Hybrid methods on different metrics

4.4 Evaluation Strategy

We evaluate the TCM knowledge understanding and reasoning performance of ShuzhiQihuang based on the currently publicly available benchmark TCMBench (Yue et al., 2024). We use all multiple-choice questions from the TCM-ED dataset provided by TCMBench as the evaluation dataset and then adopt a unified evaluation process to guide baselines to provide answers and analyses to the questions simultaneously, ensuring the fairness of our evaluation. Finally, we compare the results using the accuracy of answering questions.

4.4.1 Evaluation Datasets and Pipelines

TCM-ED is an LLMs evaluation dataset collected based on the TCM Licensing Exam (TCMLE) in collaboration with TCM experts, covering 5,473 multiple-choice questions. This dataset contains three types of questions, including A1/A2, B1, and A3. The data statistic information is shown in Table 4. In addition, we independently evaluate the sub-datasets under 16 knowledge points in TCM-EB to achieve comprehensive comparison and analysis.

Table 4: The statistical information of TCM-ED.

Question Type	#Qustions	#Sub-Question	#All
A1/A2	1,600	1,600	
A3	198	642	5,473
B1	1,481	3,231	

We follow the evaluation pipeline given in TCM-Bench under the zero-shot strategy, where A1/A2 type questions consist of one question and five options. Therefore, we set up a single-round dialogue to guide all LLMs evaluated in producing answers. Multiple questions of B1 type will share a set of options. Similarly, questions of A3 type set a clinical case as the background, and two or more sub-questions set questions and options around this background, so the knowledge points in these two types of questions have a logical correlation. We use a multi-round dialogue approach to evaluate these two question types, where the output parsing of the previous question is the input of the next question.

4.4.2 Evaluation Results

Overall Result We first evaluate the accuracy of various LLMs on different question types, as shown in Table 5. From Table 5, we can observe that:

- **ShuzhiQihuang achieves the highest performance in TCM.** ShuzhiQihuang achieves the highest accuracy in all three types of questions, surpassing the currently best-performing ChatGPT-4. Still, the parameter scale of ShuzhiQihuang is much smaller than ChatGPT-4, which proves the superiority of ShuzhiQihuang in the TCM domain. Specifically, among the A1/A2 type of questions, ShuzhiQihuang improved by **36.67%** compared to ChatGPT-4.

- **Integrating domain knowledge can make it more effective.** HuatuoGPT outperforms ChatGPT-3.5-turbo on A1/A2 and A3 types of questions, indicating the effectiveness of incorporating Chinese medical knowledge. However, it is far lower than ChatGPT-3.5 turbo on the B1 type of questions, indicating that this model still lacks the logical correlation of TCM knowledge. ShuzhiQihuang is generally superior to other Chinese medical models, indicating that it is necessary to

Table 5: The accuracy on LLMs in three question types of TCM-Bench. The best performing model is bold.

LLMs	A1/A2	A3	B1	Total
Zhongjing	0.19	0.12	0.13	0.15
Bianque	0.21	0.14	0.13	0.16
BenTsao	0.23	0.150.	14	0.17
HuatuoGPT	0.48	0.49	0.22	0.40
ChatGPT-3.5-turbo	0.44	0.47	0.44	0.45
ChatGPT-4	0.60	0.62	60	0.61
ShuzhiQihuang (Ours)	**0.82**	**0.67**	**0.63**	**0.71**
% Improve.	**36.67**	**8.06**	**5**	**16.40**

combine knowledge that is more in line with domain cognition in the field of TCM with domain characteristics. It also indirectly reflects that the knowledge deviation between TCM and Western medicine is still significant even if they belong to the same language field.

Results in different medical domains. We further evaluate LLMs on 16 different TCM subjects in A1/A2 type of questions, and then organize the results according to the various medical branches examined by TCMLE. The results are shown in Figure 4.

The left side of Figure 5 shows the performance of different LLMs in three medical branches: TCM basis, TCM clinical medicine, and Western medicine and its clinical. ShuzhiQihuang far outperforms other models in each branch, indicating that ShuzhiQihuang not only performs well in TCM but also in the field of Western medicine. Specifically, the advantages of ShuzhiQihuang on the TCM basis and clinical medicine are more prominent.

Through the 16 different TCMLE examination subjects on the right side of Figure 5, we can further clearly observe that ShuzhiQihuang has shown significant advantages in traditional Chinese pharmacology, diagnostics of TCM, gynecology of TCM, as well as science of acupuncture and moxibustion. The accuracy of answering questions in traditional Chinese pharmacology is 69.81% higher than that of ChatGPT-4. Surprisingly, ShuzhiQihuang has also shown significant advantages in Western medicine-related subjects such as internal medicine and health regulations.

According to statistics, ShuzhiQihuang achieved an accuracy rate of over 90% in 5 subjects, over 80% in 7 subjects, over 70% in 3 subjects, and only one subject, classical works of Chinese medicine, did not reach 60%. Its accuracy in A1/A2 type of questions in diagnostics of TCM has reached 94%. However, we have also found that all models of classical works of Chinese medicine perform poorly, which may be because TCM classics are the most domain-specific knowledge in the TCM domain. Most of it is recorded in ancient Chinese, lacking a rich corpus. Additionally, LLMs cannot currently process and understand ancient Chinese.

For the four medical LLMs, BenTsao and HuatuoGPT performed better in TCM clinical medicine than in TCM basis, which may be due to the inclusion of medical dialogue data similar to clinical practice or actual clinical data in these models.

4.5 Case Study

In the practical case study, we conducted a comprehensive analysis of the response results generated by the ShuZhiQiHuang model alongside several other state-of-the-art models, including Zhongjing, BianQue, HuatuoGPT, BenTsao, ChatGPT-turbo-3.5, and ChatGPT-4. The answers from these baseline models are listed in the appendix B. Our observations revealed that the Zhongjing, BianQue, and BenTsao models did not perform optimally. Zhongjing's responses were overly general and simplistic, while BenTsao and BianQue exhibited significant hallucinations, rendering their answers incomprehensible. Meanwhile, ChatGPT-turbo-3.5 and ChatGPT-4 provided responses that were logically coherent but lacked factual grounding, leaving uncertainty about their truthfulness and accuracy. In contrast, ShuZhiQiHuang achieved relatively ideal response outcomes across several questions. It not only answered the questions in a detailed and systematic manner but also cited authoritative sources to substantiate its answers.

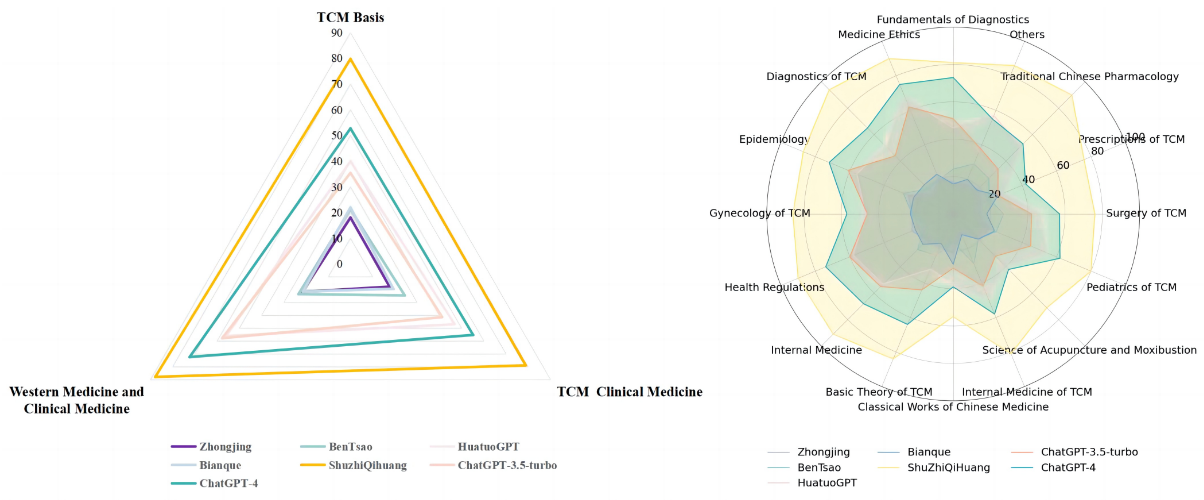

Figure 4: The performance of LLMs in three medical branches and sixteen medical subjects of TCMLE.

For instance, when explaining the location and function of the Shang Yingxiang acupoint, it referenced relevant literature in the response, thereby enhancing the credibility and trustworthiness of the information provided.

5 Limitations

Despite advancements in TCM question answering using LLMs with fine-tuning and RAG technology, several limitations persist:

1.**Data Quality and Availability:** The success of LLMs in TCM relies heavily on the quality and availability of domain-specific data. TCM knowledge primarily comes from classical literature and experiential learning, posing challenges for data acquisition. Limited large-scale and diverse datasets restrict the model's ability to generalize and accurately represent TCM knowledge.

2.**Evaluation Metrics:** Current metrics focus on answering TCM-related questions but overlook other critical aspects such as understanding complex medical texts, handling multi-turn dialogues, and providing context-aware responses. More comprehensive evaluation strategies are needed.

3.**Model Complexity and Interpretability:** The complexity of LLMs and TCM knowledge makes these models difficult to interpret, potentially hindering their acceptance among medical professionals who need understandable decision-making processes.

4.**Clinical Application Challenges:** Implementing LLMs in clinical settings is challenging due to issues like erroneous outputs and hallucinations, which pose significant risks in medical applications where accuracy is crucial.

5.**Ethical and Legal Concerns:** Deploying LLMs in healthcare raises ethical and legal concerns, including patient privacy, informed consent, and potential bias. Ensuring adherence to ethical guidelines and regulatory standards is essential for safe and effective use in clinical practice.

6 Conclusion

In this study, we present ShuZhiQiHuang, an intelligent and comprehensive question-answering platform specifically tailored for the field of Traditional Chinese Medicine (TCM). This platform integrates a sophisticated training workflow that seamlessly combines pre-training and supervised fine-tuning (SFT) methodologies. By strategically leveraging Retrieval-Augmented Generation (RAG) technology, in conjunction with domain-specific expertise extracted from various authoritative databases, molecular visualization techniques, and other advanced approaches, ShuZhiQiHuang effectively addresses the prevalent issue of knowledge hallucinations that plagues specialized LLM question-answering systems. Through rigorous evaluation, ShuZhiQiHuang has demonstrated superior performance in TCM knowledge questioning and answering compared to other mainstream open-source TCM models, as well as the current leading general LLM, ChatGPT-4. To ensure wide accessibility and usability, the platform has been seamlessly deployed across both web pages and WeChat mini-programs, catering to diverse user preferences and technological landscapes. Despite these significant advancements, challenges such as data quality, model interpretability, and ethical considerations persist. Future efforts should be directed towards overcoming these challenges, paving the way for ShuZhiQiHuang to fully realize its potential in enhancing decision-making processes, promoting TCM education and knowledge dissemination, and driving innovative research endeavors in this field.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023b. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

A Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J Bellis, Marleen De Veij, and Andrew R Leach. 2020. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12:1–16.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Odma Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. 2019. Preliminary study on the construction of chinese medical knowledge graph. *Journal of Chinese Information Processing*, 33(10):1–9.

Yihan Cao, Yanbin Kang, and Lichao Sun. 2023. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 2006. An empirical study of chinese chunking. In *Proceedings of the COLING/ACL 2006 main conference poster sessions*, pages 97–104.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, huimin zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023a. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Yanrui Du, Sendong Zhao, Muzhen Cai, Jianyu Chen, Haochun Wang, Yuhan Chen, Haoqiang Guo, and Bing Qin. 2023b. The calla dataset: Probing llms' interactive knowledge acquisition from chinese medical literature.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

BV Elasticsearch. 2018. Elasticsearch. *software], version*, 6(1).

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Sendong Zhao Bing Qin Ting Liu Haochun Wang, Chi Liu. 2023. Chatglm-med: chatglm. https://github.com/SCIR-HI/Med-ChatGLM.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhongzhen Huang, Kui Xue, Yongqi Fan, Linjie Mu, Ruoyu Liu, Tong Ruan, Shaoting Zhang, and Xiaofan Zhang. 2024. Tool calling: Enhancing medication consultation via retrieval-augmented large language models. *arXiv preprint arXiv:2404.17897*.

Yanlan Kang, Yang Chang, Jiyuan Fu, Yan Wang, Haofen Wang, and Wenqiang Zhang. 2023. Cmlm-zhongjing: Large language model is good story listener. https://github.com/pariskang/CMLM-ZhongJing.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2024. A comprehensive overview of large language models.

Pouria Omrani, Alireza Hosseini, Kiana Hooshanfar, Zahra Ebrahimian, Ramin Toosi, and Mohammad Ali Akhaee. 2024. Hybrid retrieval-augmented generation approach for llms query response enhancement. In *2024 10th International Conference on Web Research (ICWR)*, pages 22–26. IEEE.

Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Pei Chu, Yuan Qu, Runyu Peng, et al. 2024. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset. *arXiv preprint arXiv:2402.19282*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Weijieying Ren, Xinlong Li, Lei Wang, Tianxiang Zhao, and Wei Qin. 2024. Analyzing and reducing catastrophic forgetting in parameter efficient tuning. *arXiv preprint arXiv:2402.18865*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Omid Shahmirzadi, Adam Lugowski, and Kenneth Younge. 2019. Text similarity in vector space models: a comparative study. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 659–666. IEEE.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Oguzhan Topsakal and Tahir Cetin Akinci. 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International Conference on Applied Engineering and Natural Sciences*, volume 1, pages 1050–1056.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Guoxing Yang, Jianyu Shi, Zan Wang, Xiaohong Liu, and Guangyu Wang. 2023b. Tcm-gpt: Efficient pretraining of large language models for domain adaptation in traditional chinese medicine. *arXiv preprint arXiv:2311.01786*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. 2022. Gatortron: a large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Wenjing Yue, Xiaoling Wang, Wei Zhu, Ming Guan, Huanran Zheng, Pengfei Wang, Changzhi Sun, and Xin Ma. 2024. Tcmbench: A comprehensive benchmark for evaluating large language models in traditional chinese medicine. *arXiv preprint arXiv:2406.01126*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Heyi Zhang, Xin Wang, Zhaopeng Meng, Yongzhe Jia, and Dawei Xu. 2024. Qibo: A large language model for traditional chinese medicine. *arXiv preprint arXiv:2403.16056*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. Huatuogpt, towards taming language models to be a doctor. *arXiv preprint arXiv:2305.15075*.

Qinglin Zhang, Qian Chen, Yali Li, Jiaqing Liu, and Wen Wang. 2021. Sequence model with self-adaptive sliding window for efficient spoken document segmentation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 411–418. IEEE.

A Interface display

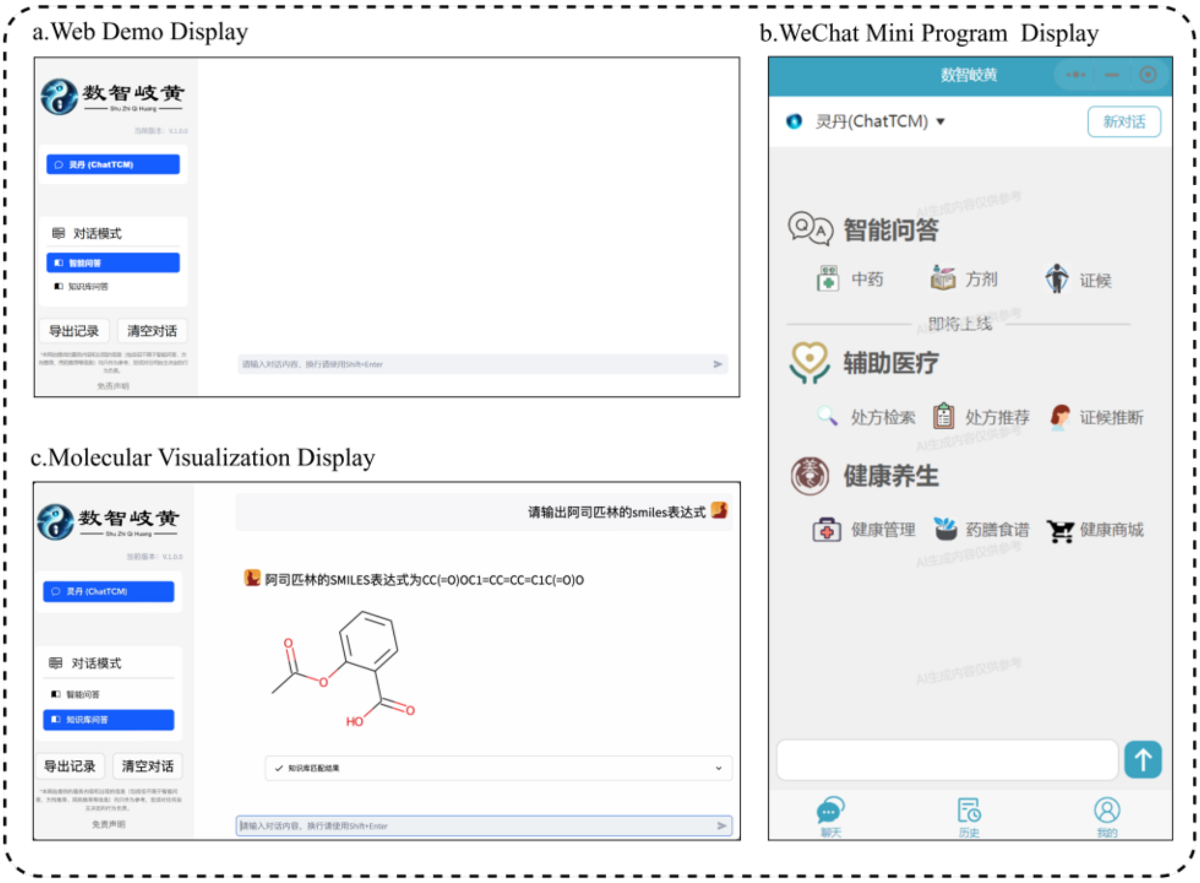

Figure 5: Interface display

B Dialogue Cases

User:

迎香穴的位置和主要功效是什么?

What is the location and primary function of the Yingxiang acupoint?

ShuZhiQiHuang:

迎香穴 (LI-20) 的位置，《针灸甲乙经》中记载：“在禾窖上，鼻下孔傍”；其主要功效，根据古籍《针灸甲乙经》的描述，该穴位可以治疗“鼻塞不通、不闻香臭”等问题。说明迎香穴位于人体的面部，具体位置是在鼻翼外缘中点旁，鼻唇沟上。迎香穴属于手阳明大肠经，是经络上的一个重要穴位，其主要功效治疗鼻部问题，如“鼻塞不通、不闻香臭”。

The Yingxiang acupoint is located on the face, specifically at the midpoint of the outer edge of the nostril, along the nasolabial fold. According to the ancient text "Classical Needling and Moxibustion," this acupoint is beneficial for treating issues such as nasal congestion and loss of smell. It belongs to the Large Intestine Meridian of Hand Yangming and is crucial for nasal health.

Zhongjing:

迎香穴是属于足阳明胃经的腧穴之一，位于人体的胸部，主要功效包括宽胸理气、调节胃肠功能等。此外，迎香穴还可以缓解心绞痛、胸闷等症状。

The Yingxiang acupoint is part of the Stomach Meridian of Foot Yangming and is situated in the chest area. Its primary functions include broadening the chest, regulating gastrointestinal function, and alleviating symptoms like angina pectoris and chest oppression.

HuatuoGPT:

迎香穴位于鼻翼外缘中点旁，即鼻唇沟中。其主要功效包括通鼻窍、祛风邪、宣肺气、调中焦、和胃降逆。

Yingxiang acupoint is located beside the midpoint of the lateral border of the nostril, in the nasolabial groove. It is effective for clearing nasal passages, dispelling wind evils, ventilating the lungs, regulating the middle burner, and harmonizing the stomach to descend counterflow.

BenTsao:

迎香穴的主要功效是抗感染，抑制炎症和维持身体健康。

The primary functions of Yingxiang acupoint are to combat infections, suppress inflammation, and maintain overall health.

Figure 6: Comparison of Model Response Effectiveness

User:

迎香穴的位置和主要功效是什么?

What is the location and primary function of the Yingxiang acupoint?

BianQue:

迎香穴的位置和主要功效：位于人体背部的凹陷处。迎香穴的主要功效是促进血液循环，缓解疼痛和疲劳，增强免疫力，调节血压等。迎香穴的使用方法如下：将一块小棉球或者纱布，用酒精或碘伏消毒后沾上迎香穴的粉末，然后轻轻按摩该部位，每次 2-3 分钟，每天 3-4 次。

The location and main functions of Yingxiang acupoint: Located in the depression on the human back. The main functions of Yingxiang acupoint include promoting blood circulation, relieving pain and fatigue, enhancing immunity, and regulating blood pressure. The method of using Yingxiang acupoint is as follows: Take a small cotton ball or gauze, disinfect it with alcohol or iodophor, then dip it in the powder of Yingxiang acupoint, and gently massage the area for 2-3 minutes each time, 3-4 times a day.

ChatGPT-turbo-3.5:

迎香穴位于鼻翼外缘旁，是鼻孔外缘凹陷的位置。主要功效包括缓解鼻塞, 清热解毒，舒缓头痛, 改善眼部疲劳需要注意的是，按摩穴位时要适度，避免用力过猛，以免引起不适。如有严重不适或疾病，请及时就医。

Yingxiang acupoint is situated beside the lateral edge of the nostril, in the depression of the nostril's outer edge. It helps relieve nasal congestion, clear heat and toxins, soothe headaches, and alleviate eye fatigue. Caution should be taken to massage the point gently to avoid discomfort. Seek medical attention for severe symptoms or illnesses.

ChatGPT-4:

迎香穴位于人体面部鼻翼外缘中点旁开约 0.5 寸处，属大肠经。主要功效有：通鼻窍：缓解鼻塞、流涕、嗅觉减退等鼻部不适。清热：用于治疗面部热症，如面赤、面部红肿热痛等。止痛：缓解面部疼痛，尤其是三叉神经痛。活血：促进面部血液循环，有助于美容和改善面色。你可以通过轻柔按揉迎香穴来刺激这些功效。

Yingxiang acupoint is located approximately 0.5 cun (a traditional Chinese measurement) lateral to the midpoint of the lateral border of the ala nasi on the human face, and it belongs to the Large Intestine Meridian. Its main functions include: Clearing the nasal passages: Alleviating nasal congestion, runny nose, and reduced sense of smell to relieve nasal discomfort. Clearing heat: Used for treating facial heat symptoms, such as flushed face, facial redness, swelling, and painful heat. Pain relief: Relieving facial pain, especially trigeminal neuralgia. Promoting blood circulation: Enhancing facial blood circulation, which can aid in beauty and improve complexion. You can stimulate these effects by gently massaging and kneading the Yingxiang acupoint.

Figure 7: Comparison of Model Response Effectivenes