Contents lists available at ScienceDirect

Digital Chinese Medicine

journal homepage: http://www.keaipublishing.com/dcmed

TCMLCM: an intelligent question-answering model for traditional Chinese medicine lung cancer based on the KG2TRAG method

Chunfang ZHOU[a], Qingyue GONG[a, b*], Wendong ZHAN[c], Jinyang ZHU[a], Huidan LUAN[a]

a. School of Artificial Intelligence and Information Technology, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210023, China

b. Jiangsu Province Engineering Research Center of TCM Intelligence Health Service, Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210023, China

c. School of Life Science, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Article history
Received 24 October 2024
Accepted 20 February 2025
Available online 25 March 2025

Keywords
Traditional Chinese medicine (TCM)
Lung cancer
Question-answering
Large language model
Fine-tuning
Knowledge graph
KG2TRAG method

ABSTRACT

Objective To improve the accuracy and professionalism of question-answering (QA) model in traditional Chinese medicine (TCM) lung cancer by integrating large language models with structured knowledge graphs using the knowledge graph (KG) to text-enhanced retrieval-augmented generation (KG2TRAG) method.

Methods The TCM lung cancer model (TCMLCM) was constructed by fine-tuning ChatGLM2-6B on the specialized datasets Tianchi TCM, HuangDi, and ShenNong-TCM-Dataset, as well as a TCM lung cancer KG. The KG2TRAG method was applied to enhance the knowledge retrieval, which can convert KG triples into natural language text via ChatGPT-aided linearization, leveraging large language models (LLMs) for context-aware reasoning. For a comprehensive comparison, MedicalGPT, HuatuoGPT, and BenTsao were selected as the baseline models. Performance was evaluated using bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), accuracy, and the domain-specific TCM-LCEval metrics, with validation from TCM oncology experts assessing answer accuracy, professionalism, and usability.

Results The TCMLCM model achieved the optimal performance across all metrics, including a BLEU score of 32.15%, ROUGE-L of 59.08%, and an accuracy rate of 79.68%. Notably, in the TCM-LCEval assessment specific to the field of TCM, its performance was 3% – 12% higher than that of the baseline model. Expert evaluations highlighted superior performance in accuracy and professionalism.

Conclusion TCMLCM can provide an innovative solution for TCM lung cancer QA, demonstrating the feasibility of integrating structured KGs with LLMs. This work advances intelligent TCM healthcare tools and lays a foundation for future AI-driven applications in traditional medicine.

*Corresponding author: Qingyue GONG, E-mail: qygong@126.com.

Peer review under the responsibility of Hunan University of Chinese Medicine.

DOI: 10.1016/j.dcmed.2025.03.011

Citation: ZHOU CF, GONG QY, ZHAN WD, et al. TCMLCM: an intelligent question-answering model for traditional Chinese medicine lung cancer based on the KG2TRAG method. Digital Chinese Medicine, 2025, 8(1): 36-45.

1 Introduction

Lung cancer, being one of the malignancies with the highest incidence and mortality rates globally, underscored the critical importance of early diagnosis and precise treatment for enhancing patient survival rates [1]. Traditional Chinese medicine (TCM), with a long history, offers a unique theoretical framework and therapeutic methods that may present new perspectives on lung cancer treatment, such as syndrome differentiation and herbal prescriptions [2]. However, the complexity of TCM knowledge, including intricate disease-syndrome relationships, dynamic treatment principles, and personalized herbal combinations, posed significant challenges for intelligent question-answering (QA) systems [3]. Conventional TCM QA models, which often struggled with the limited domain-specific knowledge integration, static datasets, and the inability to handle nuanced clinical reasoning, failed to provide comprehensive, professional, and personalized answers, thereby restricting the applicability of TCM lung cancer QA systems [4].

Large language models (LLMs), as advanced artificial intelligence (AI) technologies centered around natural language understanding and generation, have revolutionized the field of natural language processing (NLP) through their robust contextual comprehension and content generation capabilities [5]. There were notable examples, including the GPT series [6], ChatGLM series from Tsinghua University [7], and Meta's LLaMA series [8]. These models, equipped with self-attention mechanisms and large-scale unsupervised pre-training, had foundational grammatical structure and semantic feature understanding, thereby significantly enhancing the capacity for deep language comprehension and generation. However, challenges in several specialized domains like TCM lung cancer still exsit. First, LLMs often generate hallucinated or outdated content due to their reliance on general pre-training data and lack of domain-specific medical knowledge [9]. Second, while methods like adaptive pre-training (e.g., TCMDA [10]) and supervised fine-tuning (e.g., LLM-Qibo [11]) can enhance domain relevance, they fail to dynamically integrate structured knowledge from sources such as knowledge graphs (KGs). For instance, although MedChatZH [12] improved dialogue quality through corpus filtering, it also neglected the semantic gap between KG triples and textual representations, making it incapable of retrieving contextually relevant knowledge. Additionally, traditional retrieval-augmented generation (RAG) methods focus on unstructured text retrieval, overlooking the structured relational knowledge inherent in KGs [13]. These shortcomings underscore the need for a hybrid approach that can bridge LLMs and KGs to ensure both linguistic fluency and clinical validity.

To address these gaps, this study proposed the TCM lung cancer model (TCMLCM), an intelligent QA framework that integrates LLMs with KGs through the KG to text-enhanced RAG (KG2TRAG) method. Unlike prior approaches, KG2TRAG converts KG triples into natural language descriptions via ChatGPT-aided linearization, enabling LLMs to leverage structured knowledge dynamically. This method not only enhances the interpretability of retrieved knowledge but also mitigates hallucination risks by grounding responses on verified medical entities and relationships. By fine-tuning ChatGLM2-6B on curated TCM lung cancer datasets and constructing a comprehensive KG spanning diseases, syndromes, formulas, and herbs, TCMLCM aimed to achieve: (i) the improvement of response accuracy through domain-adaptive knowledge retrieval; (ii) the enhancement of clinical applicability via expert-validated knowledge integration; (iii) the establishment of a benchmark for the TCM-specific QA systems. This work provides a foundational step toward bridging the gap between TCM theory and AI-driven clinical support, with potential implications for personalized medicine and interdisciplinary healthcare integration.

2 Data and methods

The TCMLCM model was designed for TCM lung cancer diagnosis and treatment that integrated LLMs with KGs [14] (Figure 1). Initially, LoRA technology was employed to fine-tune ChatGLM2-6B on a specialized dataset for TCM lung cancer, resulting in the fine-tuned model named traditional Chinese medicine lung cancer general language model (TCMLCGLM). A TCM lung cancer KG dataset was collected, and a TCM lung cancer KG was constructed using entity alignment methods. The Beijing Academy of Artificial Intelligence (BAAI) general embedding (BGE) model was applied to vectorize the KG data, which was then stored in a vector database. The KG2TRAG method was applied to ensure that the TCMLCGLM can retrieve the most relevant information from the database with users' queries. This method invoked the ChatGPT application programming interface (API) after the initial retrieval round and combined prompt strategies to textualize the retrieved triplet results [15], generating informative textual statements to ensure more

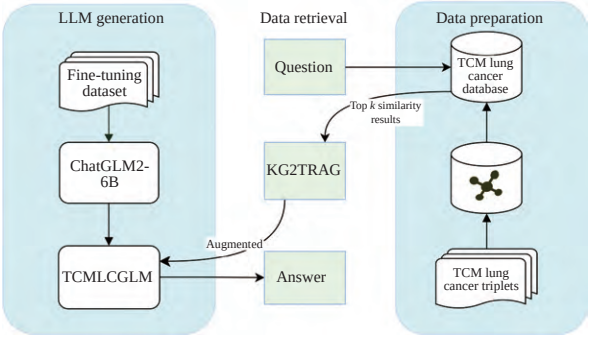

Figure 1 Technical route of the TCM lung cancer intelligent QA model

relevant outcomes in the second retrieval round. Finally, the most similar text blocks retrieved were concatenated with the original users' queries and combined with predefined prompt templates to convert the textual representation of triplets and the question into knowledge-enhanced prompts. These prompts were then input into the TCMLCGLM model for more reliable answers.

2.1 Fine-tuning LLMs

Regarding the baseline model, several Chinese open-source LLMs with comparable parameter scales were compared in relevant benchmark evaluations. This study adopted ChatGLM2-6B as the baseline model with its best performance and stability. Subsequently, a fine-tuning strategy was implemented for the ChatGLM2-6B model. Resources with extremely high memory were required in the full-precision fine-tuning of LLMs. To improve the efficiency of model fine-tuning, previous study indicated that the fine-tuning process of LLMs was at low rank and proposed a low-resource fine-tuning method known as low rank adaptation (LoRA) [16]. The LoRA fine-tuning method achieved this by freezing all parameters of the pre-trained model and injecting trainable low rank decomposition matrices into each weight of the transformer layers, reducing the number of trainable parameters needed for downstream tasks. An efficient parameter fine-tuning framework was provided. Building upon LoRA, previous study has introduced quantized LoRA (QLoRA), which applied k bit quantization to the parameters prior to fine-tuning, significantly reducing the cost of fine-tuning LLMs [17]. As illustrated in Figure 2, QLoRA markedly decreased the memory requirements during the fine-tuning process.

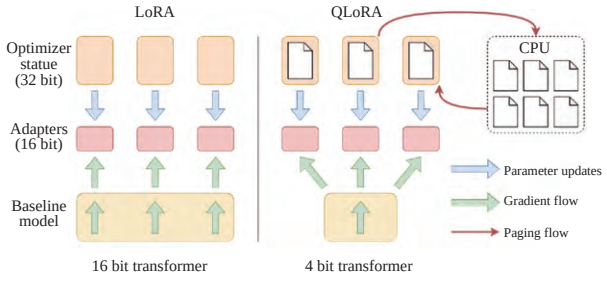

Figure 2 LoRA and QLoRA fine-tuning methods
CPU, central processing unit.

This study implemented 8 bit quantization on the ChatGLM2-6B model for the collected and organized TCM lung cancer datasets. The quantization process of the model was represented by Equation (1).

$$X^{\text{int8}} = \text{round}\left(\frac{127}{\text{abs max}(\text{X}^{\text{FP32}})}\text{X}^{\text{FP32}}\right) = \text{round}\left(\text{c}^{\text{FP32}} \times \text{X}^{\text{FP32}}\right) \quad (1)$$

Here, X represents the input tensor, FP32 represents floating point, with c serving as a constant denoting the quantization scale. The inverse process for the quantized model was given by Equation (2).

$$\text{dequant}\ (c^{\text{FP32}}, X^{\text{int8}}) = \frac{X^{\text{int8}}}{c^{\text{FP32}}} \approx X^{\text{FP32}} \quad (2)$$

QLoRA injects the parameter matrix from Equation (2) into the quantized model, yielding Equation (3).

$$Y = XW + sXL_1L_2 \quad (3)$$

Here, W is the original weight matrix of the model, L_1 and L_2 are two low rank matrices in LoRA, and s is a scalar. The fine-tuned model was referred to as TCM-LCGLM in this study. This model was capable of generating information that aligned with the user's subjective QA task requirements and demonstrated improved understanding and QA capabilities in the domain of TCM lung cancer.

2.2 Construction of the KG

A notable limitation of fine-tuning large models was their failure to adapt to new data and standards, a gap that can be bridged by KGs [18]. To address this issue, this study sourced extensive open-source TCM medical KGs to construct a TCM lung cancer KG, including KGQA_TCM (https://github.com/dreams-flying/KGQA_TCM), Herb KG (https://github.com/FeiYee/HerbKG), and TCM_KG (https://github.com/owlet0605/TCM_KG). TCM lung cancer-related data were filtered out, and entities including diseases, syndromes, formulas, and herbs related to TCM lung cancer were selected, establishing an overall structure centered around disease-syndrome-formula-herb as the ontology. In addition, in order to ensure the comprehensiveness and representativeness of the KG, we incorporated the graph resources of existing research results, covering the classic medical books of *Huangdi Neijing* (《黄帝内经》, *Inner Cannon of Huangdi*), *Shennong Bencao Jing* (《神农本草经》, *Divine Farmer's Materia Medica*), and the clinical medical cases of Zhongying ZHOU, a master of TCM. Through strict data cleaning and standardization processes, a KG containing 11 723 independent nodes (including 1 294 disease entities, 506 syndromes, 678 formulas, and 8 365 herbs) was finally formed. In addition, 15 912 edges were established between these nodes, representing different types of relationships, including but not limited to treatment relationships, composition relationships, etiology, and pathogenesis relationships.

Then, an alignment method based on semantic and text similarity was applied to calculate entity similarity and integrate data from different sources, as shown in Algorithm 1. Entity alignment first computes text similarity; if the text similarity exceeds a threshold, semantic similarity is further calculated to see if it also surpasses a threshold; otherwise, the process returns immediately.

This two-step approach avoids excessive model computation due to a few duplicate entities and prevents mismatches between similar texts with different semantics, thus minimizing misjudgment probability while enhancing computational efficiency.

Algorithm 1: alignment method based on semantic and textual similarity

Input: Entity $x \in G$ that needs to be matched and the candidate entity set N

Output: Duplicate entity set $R \subseteq G$

1. $R \leftarrow \{\}$ // Initialize the duplicate entity set as empty
2. for each n in G do // For each entity n in G
3. $s_1 \leftarrow$ Jaccard(x, n); // Calculate the Jaccard similarity between x and n
4. if $s_1 > t_1$ then // If the Jaccard similarity is greater than threshold t_1
5. $a \leftarrow$ SBERT(x); // Encode x using SBERT
6. $b \leftarrow$ SBERT(n); // Encode n using SBERT
7. $s_2 \leftarrow \cos\langle a, b\rangle$; // Calculate the cosine similarity between a and b
8. if $s_2 > t_2$ then // If the cosine similarity is greater than threshold t_2
9. Add n to R; // Add n to the duplicate entity set R
10. endif
11. endif
12. end

Among them, Jaccard(x, n) represents the Jaccard similarity between entities, and $\cos\langle a, b\rangle$ represents the cosine similarity between vectors a and b. In this study, $t_1 = 0.85$ and $t_2 = 0.85$.

Using the aforementioned method, several sets of duplicate entities were obtained. In this study, the shortest name among identical entities was chosen as the standard name. To ensure the accuracy of the data integrated into the fused KG, the aligned data was reviewed with quality assessments [19]. Finally, BGE was adopted as the vector model. BGE was a Chinese embedding model developed domestically, demonstrating superior semantic retrieval precision and overall semantic representation capabilities compared to models of similar parameter sizes. Consequently, the vectorized KG was stored in a database, thereby constructing a knowledge vector database for TCM lung cancer diagnosis and treatment. The construction process is shown in Figure 3.

In summary, through systematic collation and in-depth analysis of data related to lung cancer in TCM, we have constructed a KG with a clear structure and robust performance, providing a solid foundation for TCM lung cancer research and treatment. This knowledge graph not only covers the core elements in the theoretical system of TCM, but also integrates modern medicine's understanding of lung cancer, providing a bridge for interdisciplinary research.

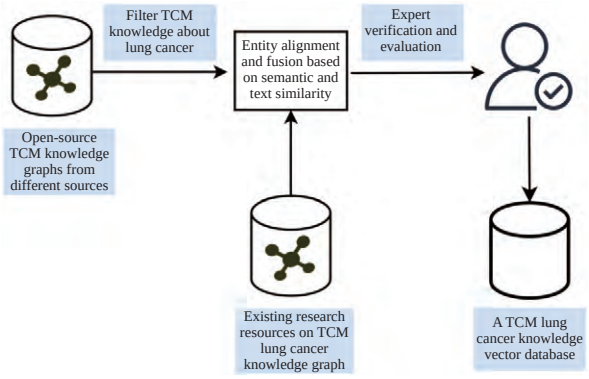

Figure 3 Construction process of TCM lung cancer knowledge vector database

2.3 Construction of the QA model

The KG2TRAG method was employed to integrate LLMs with KGs for building an advanced QA model, thereby infusing knowledge into the LLMs. The primary challenge was to accurately connect the question to the most relevant knowledge within the KG, enabling precise inference of the correct answer [20]. The KG2TRAG method was designed to address the previous issues, which involved converting triplets into appropriate textual descriptions during RAG retrieval tasks. This process included triplet linearization, achieved by invoking the ChatGPT API alongside prompt strategies. The head and tail entities of the structured triplets were treated as subject and object, respectively. The triplet form of subject, relation, and object was transformed into free-form text.

The KG2TRAG method was designed to strive for retrieved answers that closely align with the user queries for the accuracy of the model's final output, thereby potentially enhancing the effect of knowledge augmentation. The intended approach is detailed in Algorithm 2.

Algorithm 2: KG2TRAG method

Input: Subgraph $G' = \{(s, r, o) \mid s, o \in E, r \in R\}$

Output: Text sequence $X = (x_1, x_2, \ldots, x_n)$

1. $G' \leftarrow \{(s, r, o) \mid s, o \in E, r \in R\}$ // Input subgraph G'
2. $X \leftarrow [\,]$ // Initialize an empty text sequence X
3. for each (s, r, o) in G' do // For each triple $(s, r, o) \in G'$
4. triplet_text $\leftarrow s +$ " " $+ r +$ " " $+ o$ // Convert the triple into natural language form
5. prompt $\leftarrow$ P(triplet_text) // Use a predefined template to enhance the expression of triplet_text, e.g., if r is "has capital", then triplet_text could be transformed into "The capital of s is o."
6. response $\leftarrow$ ChatGPT(chatgpt_connection, prompt) // Call the GPT API and provide the optimized template as the prompt
7. $x \leftarrow$ abstract(response) // Extract the generated text snippet x from the GPT API
8. $X \leftarrow$ append(X, x) // Append x to the text sequence X

9. end for // End loop
10. return X // Return text sequence X

G' is the input subgraph of triples. X is the output text sequence. P(triplet_text) is a function that constructs a prompt string, which may either directly use the triplet text as the prompt or construct a more complex prompt string. ChatGPT_connection is a pre-established connection to ChatGPT. ChatGPT(chatgpt_connection, prompt) is a function that invokes the ChatGPT API and passes in the prompt string prompt to receive a response. Abstract(response) is a function that extracts natural language text from the ChatGPT response. Append(X, x) is an operation that adds the generated text x to the text sequence X.

3 Experiments

3.1 Evaluation methodology

During experiments, with bilingual evaluation understudy (BLEU), recall-oriented understudy for gisting evaluation (ROUGE), and accuracy as the evaluation metrics, the TCM-LCEval metric was utilized to objectively compare the performance of the models in the TCM lung cancer domain. Additionally, a subjective method involving ratings from human experts was employed to comparatively analyze the responses generated by the models.

3.1.1 BLEU and ROUGE metric BLEU was proposed as a bilingual translation quality evaluation metric based on text similarity [21]. This metric compared machine-translated output against several reference translations to compute an overall score. ROUGE metric is used for evaluating the performance of text summarization, which automatically compared the generated summaries or translations with a set of reference summaries by counting the number of overlapping basic units, thus calculating the similarity between the automatically generated summary or translation and the reference summaries [22]. While BLEU focused on measuring the accuracy and exact match degree of the generated text, ROUGE emphasized the completeness and coverage of the information.

3.1.2 Accuracy metric Accuracy is one of the commonly used evaluation metrics in the field of deep learning, used to measure the proportion of correct predictions made by a model. The formula for calculating the accuracy metric is shown in Equation (4), where true positives (TP) represents the number of positive samples correctly identified as positive. True negatives (TN) is the number of negative samples correctly identified as negative. False positives (FP) is the number of negative samples incorrectly classified as positive. False negatives (FN) is the number of positive samples incorrectly classified as negative.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

3.1.3 TCM-LCEval metric Due to the inherent randomness in the generation results of LLMs, replicating results is challenging. These models currently faced several issues, such as the lack of domain-specific evaluation metrics. To better assess the professional capabilities of the models, a TCM lung cancer model evaluation method, termed TCM-LCEval, was designed to provide a performance reference standard for LLMs applications. A total of 100 TCM lung cancer-related test questions from medical knowledge competitions and introductory clinical medicine exams were collected. Referencing CEval, a TCM lung cancer-specific LLMs evaluation scheme called TCM-LCEval was constructed. The accuracy rate of the LLMs answering these 100 questions was normalized and recorded as the TCM-LCEval score. An example of a test question from the TCM-LCEval assessment is shown in Table 1.

Table 1 TCM-LCEval question example

Field	Text
Question	"All the five zang organs and six fu organs can cause coughing", but the most closely related is ( ). "五脏六腑皆令人咳", 但关系最密切的是 ( ) 。
Options	A. Heart and Lung B. Lung and Kidney C. Lung and Spleen D. Lung and Stomach E. Lung and Large Intestine A. 心肺 B. 肺肾 C. 肺脾 D. 肺胃 E. 肺大肠
Answer	D

The actual performance of the model in generating answers was highly related to the quality of the prompt templates used. The adopted template is shown in Table 2.

Table 2 TCM-LCEval prompt template

Field	0-shot prompt	*N*-shot prompt
Template	Here is a multiple-choice question from the {subject_name} exam on TCM lung cancer. Please select the correct answer. \n Question: \n {question} \n\n Answer: \n	Here is a multiple-choice question from the {subject_name} exam on TCM lung cancer. Please refer to the examples below to select the correct answer. \n Question: \n {question} \n\n Answer: \n Example(s): \n {examples} \n\n

Here, N is the number of learning samples. The text of {examples} represents the learning samples provided by the N-shot method. Text{subject_name} indicates the type of the current problem, and text{question} is the main text of the question. In the experiment, N takes values of 1 and 3, and the scoring calculation method is shown in Equation (5).

$$\text{TCM} - \text{LCEval score} = \frac{\sum_{k \in S} |C_k|}{|S|} \quad (5)$$

In the formula, $|C_K|$represents the number of questions answered correctly in the kth category, and $|S|$ represents the total number of all questions.

3.2 Experimental datasets

The experimental data consisted of two parts. (i) Retrieval of past medical students' examination papers. From these, TCM lung cancer-related subjective QA items were selected. For multiple-choice questions, the queries and answer options were semantically concatenated using an LLM, followed by textualization through prompt strategies. (ii) Selection of public QA data from the internet. This study opted for the Tianchi TCM dataset [23], HuangDi dataset [24], and ShenNong-TCM-Dataset [25]. Relevant TCM lung cancer data were extracted from these sources, and duplicate entries were removed. An overview of the three TCM data sources is shown in Table 3.

Table 3 Description of datasets for TCM lung cancer knowledge integration

Dataset	Data source description
Tianchi TCM dataset	Including translations of "*Huangdi Neijing* (《黄帝内经》)" "*Famous Doctors Encyclopedia-Traditional Chinese Medicine Chapter* (《名医百科全书——中医药章节》)" "*Chinese Patent Medicine Usage Volume* (《中成药使用手册》)" and "*Chronic Disease Health Care and Popular Science Knowledge* (《慢性病保健与科普知识》)" as the four main sources
HuangDi dataset	Data from TCM textbooks: collected all TCM textbooks from the "13th Five-Year Plan," totaling 22 books. Data from online TCM websites: crawled online TCM websites and knowledge bases such as the "Chinese Medicine Family" and "Folk Medicine Network"
ShenNong-TCM-Dataset	Using an entity-centric self-instruct approach, ChatGPT was utilized to generate over 110 000 instructions focused on TCM

The two types of data mentioned above were processed to construct a unified format fine-tuning instruction dataset, and optimize the data quality for obtaining 50 672 fine-tuning data pairs. The data were divided into three parts: 70% for training, 15% for validation, and 15% for testing.

3.3 Experimental process and analysis

3.3.1 Experimental environment and hyperparameters The following configuration included: a GPU, RTX 4090 (24GB); Python v3.10; CUDA v12.1. The main parameters used in the fine-tuning process of the ChatGLM2-6B model included: after 8 bit quantization of the baseline model, the batch size was set to 64, and the model was continuously optimized for 3 000 steps with a learning rate of 2×10^{-3} during the fine-tuning stage.

3.3.2 Objective evaluation To evaluate the advantages of the QA model, it was compared against baseline medical QA models. The selected medical LLMs for comparison were MedicalGPT [26], HuatuoGPT [27], and BenTsao [28]. These models were introduced into the KG and compared with the TCMLCM model. The comparison performance metrics for the models are presented in Table 4.

Table 4 Performance metrics comparison of TCMLCM with the baseline medical QA models (%)

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Accurary
MedicalGPT-KG	65.58	40.45	53.08	29.57	76.85
HuatuoGPT-KG	56.37	42.20	52.90	27.83	70.38
BenTsao-KG	52.91	41.52	57.18	26.95	68.21
TCMLCM	68.32	45.88	59.08	32.15	79.68

ROUGE-1 focuses on unigram (single-word) matching and is suitable for evaluating lexical coverage. ROUGE-2 measures bigram (two consecutive words) matching and is effective for assessing phrase-level similarity. ROUGE-L is based on the longest common subsequence (LCS), considering both word matching and word order, and is ideal for evaluating sentence-level structural similarity.

The TCMLCM model, based on the KG2TRAG method, outperformed the baseline medical models in terms of performance indicators, indicating its excellent performance in generating accurate and coherent responses to TCM-related lung cancer queries. Compared with the MedicalGPT, HuatuoGPT, and BenTsao models which have been introduced to KG, the TCMLCM model had higher scores in ROUGE-1, ROUGE-2, and ROUGE-L. These indicators measured the overlap between the generated text by the model and the reference text in terms of single words, two-word phrases, and the longest common subsequence, respectively. Similarly, the TCMLCM model also performed excellently in terms of BLEU score compared to a set of reference translations. Its BLEU score was 32.15%, surpassing other models, indicating that it did better in aligning word sequences with reference answers. Finally, the TCMLCM model achieved the highest accuracy score of 79.68%, which was an indicator measuring the overall correctness of the answers, further confirming the ability of the TCMLCM model to provide more accurate TCM lung cancer answers.

In addition, the baseline RAG method was compared with both GraphRAG [29] and MedGraphRAG [30] to evaluate the efficiency of KG2TRAG. The performance metrics of these comparisons are presented in Table 5.

TCMLCM and MedGraphRAG performed excellently in all evaluation metrics, especially showing obvious advantages in ROUGE-L and accuracy. Moreover,

Table 5 Comparison of performance metrics for the different knowledge retrieval methods (%)

Method	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Accurary
RAG	54.85	38.65	49.78	29.58	60.22
GraphRAG	56.38	40.28	50.96	30.13	65.63
MedGraphRAG	67.38	45.52	60.31	32.08	78.23
TCMLCM	68.32	45.88	59.08	32.15	79.68

TCMLCM performed better in accuracy than MedGraphRAG. TCMLCM's advantages lay not only in higher accuracy and better generation quality but also in a deep understanding of TCM theories and efficient utilization of data. KG2TRAG converted KG triples into natural language descriptions, while GraphRAG relied on different forms of knowledge representation. By converting KG triples into natural language descriptions, KG2TRAG fully leveraged the powerful language understanding capabilities of LLMs, enabling the model to better interpret complex contextual questions and improve the effectiveness of knowledge retrieval.

Furthermore, to verify the effectiveness and superiority of the model, an ablation study was conducted, with the results shown in Table 6.

Table 6 Comparison of performance metrics for TCMLCM model ablation experiments (%)

Model	ROUGE-L	BLEU	Accurary
ChatGLM2-6B	50.89	18.35	43.22
TCMLCGLM	50.62	22.33	58.79
TCMLCM	59.08	32.15	79.68

The accuracy of the TCMLCM model has enhanced significantly from 43.22% in ChatGLM2-6B model to 79.68%. Additionally, there were notable increases in the BLEU and ROUGE metrics, indicating significant advancements in the model's ability to understand and generate more accurate and realistic conversational responses. Moreover, the model showed improvements over those that did not incorporate KG2TRAG method. The results demonstrated that KG2TRAG could identify more relevant triplet information within the KG when converted into text, which aided the LLMs in generating more reliable responses. Consequently, the TCMLCM model held significant practical value in the TCM lung cancer medical domain. To further validate the enhancement effects of fine-tuning and KG2TRAG method, examples of the model's responses before and after the ablation study were presented (Figure 4).

The examples illustrated that the responses from ChatGLM2-6B lacked specificity due to the lack of specific herbs or dosages, which was elusive for practical application. Furthermore, the reliance on further adjustments by a professional TCM practitioner might limit its usefulness in areas lacking TCM resources. Although TCMLCGLM has learned a substantial amount of TCM

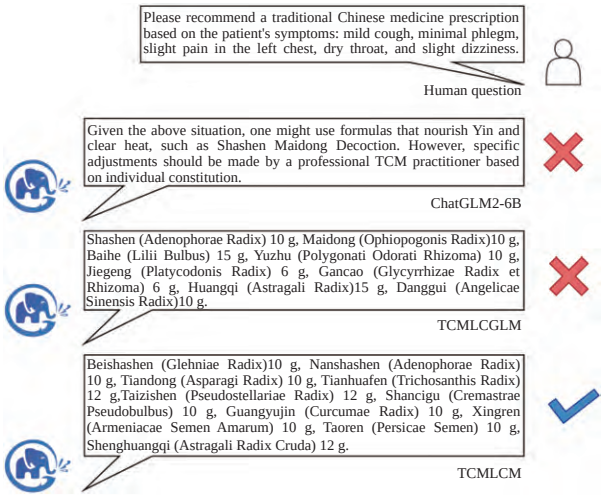

Figure 4 Comparative examples of responses generated by different models

lung cancer-related knowledge, its responses suggested that the specific formulas might be more suitable for common symptoms, like coughing. For complex or severe conditions, such responses might be inadequate, and failed to consider the importance of individual differences and potentially overlooked unique constitutional characteristics of patients. In contrast, the responses generated by TCMLCM were closer to those of medical experts, with detailed information on TCM herbs and dosages, such as the specific mention of Northern and Southern (Scrophularia Radix, 10 g each), which were more applicable for physicians. Overall, the professionalism of the responses was significantly enhanced [31].

In addition to evaluating the model's performance using BLEU and ROUGE metrics, the comparative experiments were conducted to measure the 0-shot and *N*-shot performance of the TCM-LCEval metric across various models. The TCM-LCEval scores are presented in Table 7.

Table 7 Comparison of TCM-LCEval performance metrics of TCMLCM with the baseline medical QA models (%)

Model	0-shot	1-shot	3-shot	Average
MedicalGPT-KG	29.18	28.83	32.80	30.27
HuatuoGPT-KG	33.26	35.90	36.05	35.07
BenTsao-KG	25.77	26.12	28.46	26.78
TCMLCM	37.23	38.65	39.02	38.30

The experimental results demonstrated the advantages of the TCMLCM model within the TCM lung cancer domain. The model's average metric showed a significant improvement compared with other mainstream models. This indicated that LLMs could learn domain-specific knowledge from fine-tuned TCM lung cancer medical dialogue data and KGs. The effectiveness of the TCM-LCEval evaluation method was also validated.

3.3.3 Subjective evaluation To further ensure the professionalism and credibility of the TCMLCM model, two

experienced oncology experts were invited to evaluate responses generated by different models, thereby verifying the model's effectiveness. Fifty questions were input into different models to generate answers, and the results from each model were rated by the experts. The expert scoring criteria covered three core standards: accuracy, professionalism, and usability. For each scoring criterion, a comparison was conducted. For the same question, the expert chose the most satisfactory response, awarding the selected model one point. The total score for each model across all criteria was 100 points. Finally, the voting results from both experts were tallied. The results are presented in Table 8.

Table 8 Expert evaluation scores of different TCM lung cancer QA models

Model	Accuracy	Professionalism	Usability
ChatGLM2-6B	12	15	10
TCMLCGLM	36	39	42
TCMLCM	52	46	48

The data in Table 8 showed that in terms of accuracy, the score of the TCMLCM model surpassed the combined scores of ChatGLM2-6B and TCMLCGLM, indicating that the model's knowledge expression was more precise with fewer errors or inaccuracies. In terms of professionalism, TCMLCM model outperformed ChatGLM2-6B and performed better than TCMLCGLM model, suggesting that its medical knowledge was closely related to TCM lung cancer. Regarding usability, the model had higher scores than ChatGLM2-6B and TCMLCGLM, implying its better adaptability and service efficacy in TCM lung cancer application scenarios, which was helpful in clinical decisions and patients' management.

4 Discussion

TCM integrated with AI presented significant opportunities for extending this model beyond the specific application to lung cancer. By leveraging the principles, similar intelligent systems could be developed within TCM, such as acupuncture, or TCM-based preventive care. These systems could assist practitioners in disease diagnosis, treatments, and personalized healthcare based on patients' profiles and historical data. Furthermore, the potential for integrating TCM practices with modern medical practices through AI is immense. An intelligent QA system can bridge the gap between TCM and western medicine, which could facilitate better communication among multidisciplinary healthcare teams and improve the prognosis in patients. Such integration could lead to the development of hybrid treatment protocols that combined the strengths of both traditional and contemporary medical approaches. Additionally, AI-driven analysis of TCM practices could contribute valuable insights to evidence-based medicine, promoting a more holistic view of people's health and disease treatment.

Although the current model demonstrated promising performances, its internal decision-making process may be a "black box" to users. Future research can focus on more transparent output and better responses in models using visualization techniques and KG information. This approach not only fosters users' confidence but also empowers TCM practitioners to gain a deeper understanding of the model's mechanism, thereby optimizing the diagnosis and treatment protocols. In addition, during the process of KG fusion, it relies on expert experience, which may lead to subjectivity and potential annotation errors. When generating answers, LLMs tend to follow high-frequency patterns and may overlook rare prescriptions or special therapies in TCM ancient books. It can be considered to further introduce an attention mechanism for optimization. The future development of these technologies should also consider ethical and regulatory issues, ensuring that the integration of TCM and AI adheres to best practices in both fields. Collaboration among AI researchers, TCM experts, and regulators will be crucial in addressing challenges related to data privacy, clinical validation, and standardization of TCM knowledge representation.

In summary, the current study not only contributes significantly to TCM lung cancer QA but also paves the way for exploring the intersection of AI and TCM. The continued research and development in this area hold promise in advancing personalized medicine and healthcare delivery for patients worldwide.

5 Conclusion

This study presents TCMLCM, an intelligent QA model tailored for TCM lung cancer that integrates LLMs with KGs through the novel KG2TRAG method. By bridging structured medical knowledge with contextual language understanding, TCMLCM significantly enhances the accuracy and reliability of its responses in complex TCM clinical scenarios. The proposed framework not only advances intelligent TCM healthcare tools by providing actionable clinical insights but also establishes a benchmark for knowledge-driven AI applications in specialized medical domains. Future research could extend this approach to other TCM specialties, such as acupuncture or syndrome differentiation, fostering interdisciplinary integration of AI and traditional medicine.

Fundings

Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX24_2145).

Competing interests

The authors declare no conflict of interest.

References

[1] ZHI XY, SHI JG, TIAN YT, et al. Interpretation of the key points of the 2022 white paper on the quality of life of Chinese lung cancer patients. Chinese Journal of Clinical Thoracic and Cardiovascular Surgery, 2023, 30(8): 1083–1088.

[2] WEI ZC, CHEN J, ZUO F, et al. Traditional Chinese medicine has great potential as candidate drugs for lung cancer: a review. Journal of Ethnopharmacology, 2023, 300: 115748.

[3] YU P, SONG KT, HE FC, et al. TCMD: a traditional Chinese medicine QA dataset for evaluating large language models. arXiv, 2024. Available from: https://arxiv.org/abs/2406.04941v1.

[4] HUANG XF, ZHANG JX, XU ZS, et al. A knowledge graph based question answering method for medical domain. PeerJ Computer Science, 2021, 7: e667.

[5] ZHAO WX, ZHOU K, LI J, et al. A survey of large language models. arXiv, 2023. doi: 10.48550/arXiv.2303.18223.

[6] LIU X, ZHENG YN, DU ZX, et al. GPT understands, too. AI Open, 2024, 5: 208–215.

[7] DU Z, QIAN Y, LIU X, et al. GLM: general language model pretraining with autoregressive blank infilling. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 320–335.

[8] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models. arXiv, 2023. doi: 10.48550/arXiv.2302.13971.

[9] CHANG YP, WANG X, WANG JD, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1–45.

[10] YANG GX, LIU XH, SHI JY, et al. TCM-GPT: efficient pre-training of large language models for domain adaptation in traditional Chinese medicine. Computer Methods and Programs in Biomedicine Update, 2024, 6: 100158.

[11] ZHANG HY, WANG X, MENG ZP, et al. Qibo: a large language model for traditional Chinese medicine. arXiv, 2024. Available from: https://arxiv.org/abs/2403.16056v3 .

[12] TAN Y, ZHANG ZX, LI MC, et al. MedChatZH: a tuning LLM for traditional Chinese medicine consultations. Computers in Biology and Medicine, 2024, 172: 108290.

[13] WU YK, HU N, BI S, et al. Retrieve-rewrite-answer: a KG-to-text enhanced LLMs framework for knowledge graph question answering. arXiv, 2023. doi: 10.48550/arXiv.2309.11206.

[14] JEONG C. Generative AI service implementation using LLM application architecture: based on RAG model and LangChain framework. Journal of Intelligence and Information Systems, 2023, 29(4): 129–164.

[15] WHITE J, FU Q, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv, 2023. doi: 10.48550/arXiv.2302.11382.

[16] HU EJ, SHEN Y, WALLIS P, et al. LoRA: low-rank adaptation of large language models. arXiv, 2021. Available from: https://doi.org/10.48550/arXiv.2106.09685.

[17] DETTMERS T, PAGNONI A, HOLTZMAN A, et al. Qlora: efficient finetuning of quantized llms. arXiv, 2024. Available from: https://doi.org/10.48550/arXiv.2305.14314.

[18] LV X, LIN YK, CAO YX, et al. Do pre-trained models benefit knowledge graph completion? A reliable evaluation and a reasonable approach. ACL Anthology, 2022: 3570–3581.

[19] XUE BC, ZOU L. Knowledge graph quality management: a comprehensive survey. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(5): 4969–4988.

[20] ANDRUS BR, NASIRI Y, CUI SL, et al. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10436–10444.

[21] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002: 311–318.

[22] LIN CY. ROUGE: a package for automatic evaluation of summaries. Text summarization branches out. ACL Anthology, 2004: 74–81.

[23] Aliyun [Internet]. “Wanchuang cup” traditional Chinese medicine tianchi big data competition challenge of traditional Chinese medicine literature problem generation. Available from: https://tianchi.aliyun.com/competition/entrance/531824/rankingList.

[24] ZHANG J, YANG S. HuangDi: research on the construction of a generative large language model for traditional Chinese medicine classics. 2023. Available from: https://github.com/Zlasejd/HuangDi.

[25] ZHU, W, YUE W, WANG X. ShenNong-TCM: a traditional Chinese medicine large language model. GitHub repository, 2023. Available from: https://github.com/michael-wzhu/ShenNong-TCM-LLM.

[26] XU M. MedicalGPT: training medical GPT model. GitHub, 2023. Available from: https://github.com/shibing624/MedicalGPT.

[27] ZHANG HB, CHEN JY, JIANG F, et al. HuatuoGPT, towards taming language model to be a doctor. 2023. Available from: https://arxiv.org/abs/2305.15075v1.

[28] WANG HC, LIU C, XI NW, et al. HuaTuo: tuning LLaMA model with Chinese medical knowledge. arXiv, 2023. Available from: https://arxiv.org/abs/2304.06975v1.

[29] EDGE D, TRINH H, CHENG N, et al. From local to global: a graph RAG approach to query-focused summarization. arXiv, 2024. Available from: https://arxiv.org/abs/2404.16130v2.

[30] WU JD, ZHU JY, QI YL, et al. Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. arXiv, 2024. Available from: https://arxiv.org/abs/2408.04187v2.

[31] HUA R, DONG X, WEI Y, 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, 31(9): 2019–2029.

TCMLCM：基于KG2TRAG方法的中医肺癌智能问答模型

周春芳[a], 龚庆悦[a,b*], 詹文栋[c], 朱金阳[a], 栾慧丹[a]

a. 南京中医药大学人工智能与信息技术学院, 江苏 南京 210023, 中国
b. 南京中医药大学江苏省智慧中医药健康服务工程研究中心, 江苏 南京 210023, 中国
c. 北京理工大学生命学院, 北京 100081, 中国

【摘要】目的 利用从知识图谱到文本增强的检索增强生成（KG2TRAG）的方法将大型语言模型与结构化知识图相结合，提高中医肺癌问答模型的准确性和专业性。**方法** 通过在Tianchi TCM、HuangDi和Shen-Nong-TCM-Dataset数据集以及中医肺癌知识图谱上对ChatGLM2-6B进行微调，构建了中医肺癌模型（TCMLCM）。为增强知识检索能力，引入KG2TRAG方法借助ChatGPT辅助线性化将知识图谱三元组转换为自然语言文本，并利用大型语言模型进行上下文感知推理。为了进行全面比较，选择MedicalGPT、HuatuoGPT和BenTsao作为基线模型。使用双语评估替代（BLEU）、面向召回的自动摘要评估（ROUGE）、准确率以及领域特定的TCM-LCEval指标对性能进行评估，并由中医肿瘤学专家对答案的准确性、专业性和可用性进行验证。**结果** TCMLCM模型在所有指标中均取得最优性能，其中BLEU得分为32.15%，ROUGE-L为59.08%，准确率为79.68%。值得注意的是，在针对中医领域的TCM-LCEval评估中，其性能比基线模型高3%～12%。专家评估显示，在准确性和专业性方面表现卓越。**结论** TCMLCM为中医肺癌问答提供了一种创新的解决方案，证明了将结构化的知识图谱与大型语言模型相结合的可行性。这项工作推动了中医智能医疗工具的发展，并为传统医学中未来的AI驱动应用奠定了基础。

【关键词】中医；肺癌；问答；大语言模型；微调；知识图谱；KG2TRAG方法