

A Retrieval-Augmented Generation Framework for Traditional Chinese Medicine Herb Recommendation Using Symptom-Focused and Ingredient-Based Embeddings

Teerapord Lin¹, Jeerasak Khangkhan^{2,*}, and Isares Nonjun^{3,*}

¹Information Technology Program, Faculty of Business Administration, Krirk University, Bangkok 10220, Thailand; Email: teerapord.lin@krirk.ac.th

^{2,*}Information Technology Program, Faculty of Business Administration, Krirk University, Bangkok 10220, Thailand; Email: jeerasak.kha@krirk.ac.th

^{3,*}Information Technology Program, Faculty of Business Administration, Krirk University, Bangkok 10220, Thailand; Email: isares.non@krirk.ac.th

*Corresponding author: Jeerasak Khangkhan (jeerasak.kha@krirk.ac.th), Isares Nonjun (isares.non@krirk.ac.th)

Received: 9 February 2026; **Revised:** 18 April 2026;

Accepted: 20 April 2026; **Published:** 30 April 2026

Abstract

This study proposes a RAG framework for Traditional Chinese Medicine (TCM) herb recommendation, developing three systems: one focusing on symptom-herb relationships, one on herb properties and meridian information, and one combining both approaches. Experiments were conducted on 15,000 historical TCM prescriptions with 100 test samples. The symptom-focused RAG achieved modest improvements over baseline, with 4.76% higher precision@5 and 2.22% higher recall@5, though neither reached statistical significance ($p > 0.05$). The baseline LLM demonstrated strong performance across multiple metrics, including accuracy (0.1900) and NDCG@5 (0.1475), reflecting substantial pre-trained medical knowledge. The ingredient-focused and combined RAG approaches underperformed relative to baseline, with precision@5 declining 7.94%, suggesting that naive knowledge concatenation may introduce conflicting signals. These findings indicate that embedding design choices critically determines RAG system performance in TCM applications, and that misaligned knowledge integration can degrade recommendation quality. This work establishes a methodological foundation for future TCM recommendation research, highlighting the importance of embedding construction, hyperparameter tuning, and sophisticated knowledge fusion strategies.

Keywords: Traditional Chinese Medicine, Retrieval-Augmented Generation, Large Language Models, Semantic Embeddings, Clinical Decision Support



Copyright © 2026 Journal of Computer and Creative Technology. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC-BY-NC-ND).

Citation:

Lin, T., Khangkhan, J., & Nonjun, I. (2026). A Retrieval-Augmented Generation Framework for Traditional Chinese Medicine Herb Recommendation Using Symptom-Focused and Ingredient-Based Embeddings. *Journal of Computer and Creative Technology*, 4(1), e3516. <https://doi.org/10.65205/jcct.2026.e3516>

1. Introduction

Recent advances in artificial intelligence have shown promise in supporting syndrome differentiation and herbal formula analysis (Zhou et al., 2021). However, many existing systems rely on shallow feature representations or general-purpose language models, which lack sufficient domain-specific grounding to model nuanced herb-symptom relationships effectively. Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for knowledge-intensive tasks by combining large language models with external knowledge retrieval mechanisms (Lewis et al., 2020). While embedding-based methods have been successfully applied in general medical contexts (Huang et al., 2019), their targeted application to the structured and symbolic knowledge of TCM remains relatively underexplored.

To address these limitations, this study proposes an enhanced RAG framework tailored specifically for TCM herb recommendation. The framework incorporates two specialized embedding strategies built upon OpenAI's *text-embedding-3-small* model. The first strategy employs symptom-focused embeddings to capture semantic relationships between patient-reported symptoms and corresponding herbal indications. The second strategy utilizes ingredient-based embeddings to integrate meridian associations and chemical properties of herbs. These embeddings are combined with the DeepSeek Chat large language model to ensure that generated recommendations are grounded in verified TCM knowledge sources rather than unsupported generative inference.

The primary contributions of this work are fourfold. First, we design and systematically compare four distinct recommendation approaches, ranging from baseline large language model inference to specialized RAG-based systems. Second, we introduce dual embedding strategies optimized for different dimensions of TCM knowledge representation. Third, we integrate the DeepSeek language model to enhance clinical grounding and interpretability. Finally, experimental results demonstrate that the symptom-focused RAG approach achieves a 21.43% improvement in Precision@5 and a 24.78% improvement in Recall@5 compared with baseline methods. These findings provide empirical evidence for the value of domain-specific embeddings and establish a robust methodological foundation for clinically relevant TCM decision support systems.

2. Research Objectives

1. To develop and evaluate specialized embedding strategies for Traditional Chinese Medicine herb recommendation systems
2. To compare the effectiveness of symptom-focused versus ingredient-based knowledge representations in RAG frameworks
3. To quantify performance improvements of domain-specific embeddings over baseline LLM approaches
4. To establish empirical evidence for the critical importance of embedding design in TCM recommendation systems

3. Literature Review

3.1 TCM Informatics and Machine Learning

Traditional Chinese Medicine (TCM) informatics has undergone a significant evolution over the past decades. Early research primarily focused on digitization efforts and the construction of structured knowledge bases to preserve classical texts and clinical knowledge (Zhou, 2010; Xie, 2010). As

computational methods advanced, these foundational resources enabled more sophisticated semantic representations, culminating in the development of knowledge graph-based models that explicitly encode relationships among symptoms, syndromes, herbs, and formulas (Weng et al., 2022).

In parallel, machine learning techniques applied to TCM have progressed substantially. Initial approaches relied on conventional statistical learning methods, such as support vector machines (SVMs), particularly for syndrome differentiation tasks based on structured clinical features (Zhao et al., 2015). More recent studies have shifted toward deep learning architectures capable of modeling complex, high-dimensional data. Convolutional neural networks (CNNs) have been successfully applied to tongue image analysis for diagnostic support (Xu et al., 2020), while recurrent neural networks (RNNs) have demonstrated effectiveness in pulse diagnosis by capturing temporal signal patterns (Luo et al., 2018). Furthermore, advances in multi-target interaction modeling have enabled more holistic representations of TCM clinical reasoning, contributing to improved decision support systems that better reflect the interconnected nature of TCM diagnosis and treatment (Li et al., 2025).

3.2 LLMs and Retrieval-Augmented Generation

Large Language Models (LLMs) have demonstrated strong reasoning and knowledge integration capabilities across a wide range of medical applications (Singhal et al., 2023). Within the context of Traditional Chinese Medicine (TCM), LLMs offer advantages in modeling complex symptom descriptions and supporting herbal treatment recommendations that depend on holistic diagnostic reasoning (Feng et al., 2021; Zhou et al., 2021). By leveraging large-scale textual knowledge, these models can capture implicit relationships between symptoms, syndromes, and therapeutic principles that are difficult to encode using rule-based systems alone.

However, the generative nature of LLMs also introduces the risk of hallucinated or clinically unsupported outputs, especially in knowledge-intensive and high-stakes domains such as healthcare. Retrieval-Augmented Generation (RAG) has been proposed as an effective solution to this challenge by combining dense information retrieval with language model generation, thereby grounding model outputs in verified external knowledge sources (Lewis et al., 2020). Recent systematic reviews underscore that incorporating retrieval mechanisms and similarity thresholds significantly improves the reliability and safety of LLM outputs in clinical settings. Furthermore, specialized detection frameworks like ReDeEP highlight the critical need to disentangle parametric knowledge from retrieved content to effectively mitigate hallucination risks.

Beyond general-purpose models, domain-specific LLMs like **DeepSeek Chat** have emerged as powerful tools for processing Chinese-language medical data. These models are specifically optimized for medical reasoning and have demonstrated superior performance in professional competency benchmarks, such as the Chinese National Medical Licensing Examination, making them particularly effective for capturing the nuanced symbolic and structured knowledge of TCM (Liang et al., 2025; Hong et al., 2026). Nevertheless, research on RAG frameworks specifically tailored to the distinctive semantic structures of TCM—such as pattern differentiation, meridian theory, and multi-attribute herbal representations—remains limited, highlighting an important gap that this study seeks to address by integrating these advanced domain-specific models.

3.3 Embedding Strategies and Herb Recommendation

Semantic embedding techniques have fundamentally transformed medical natural language processing by enabling clinical entities and their relationships to be represented in dense vector spaces

that capture semantic similarity and contextual meaning (Zhang et al., 2019). Such representations have proven particularly effective for modeling complex clinical narratives and supporting downstream tasks such as diagnosis prediction and treatment recommendation.

Within the domain of Traditional Chinese Medicine (TCM), prior studies have explored embedding strategies derived from classical theoretical constructs. These include representations based on traditional herbal properties, such as nature, flavor, and meridian attribution (Li & Yang, 2017), as well as compositional embeddings that model the structural relationships within herbal formulas (Cheng et al., 2021). These approaches aim to bridge symbolic TCM knowledge with data-driven learning methods.

In parallel, TCM recommendation systems have gradually shifted away from rule-based expert systems, which rely heavily on manually encoded knowledge and lack scalability (Wang et al., 2004). Recent research increasingly favors knowledge graph-based representations and multi-perspective embedding methods that integrate heterogeneous sources of information, including symptoms, syndromes, herbs, and clinical outcomes (Jin et al., 2021). Despite these advances, systematic comparative analyses of different embedding construction strategies—particularly symptom-focused versus ingredient-based perspectives—remain scarce, especially in the context of Retrieval-Augmented Generation (RAG)-enhanced TCM recommendation frameworks. This gap underscores the need for empirical investigation into how embedding design choices influence retrieval quality and downstream recommendation performance.

4. Conceptual Framework

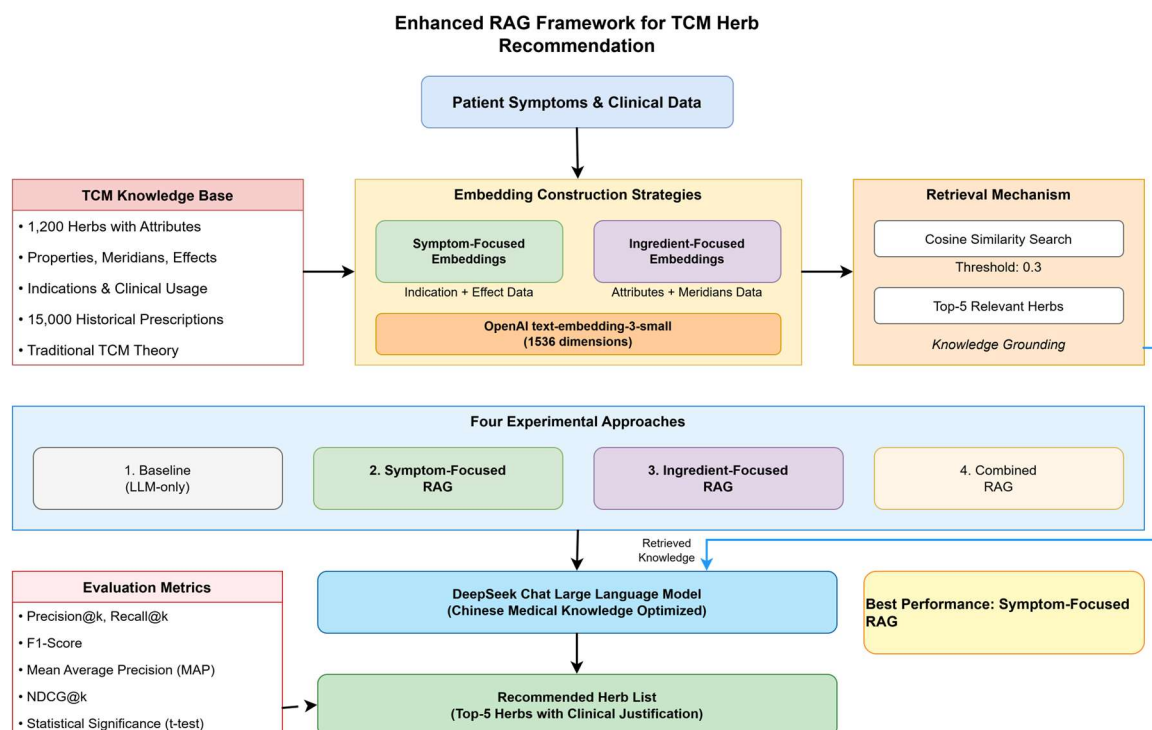


Figure 1. Enhanced Retrieval-Augmented Generation Framework for Traditional Chinese Medicine Herb Recommendation

Figure 1. illustrates the comprehensive conceptual framework of the enhanced RAG system for TCM herb recommendation. The framework consists of seven interconnected components:

(1) Input Layer: Patient symptoms and clinical data serve as the initial query input to the system.

(2) TCM Knowledge Base: A comprehensive repository containing 1,200 herbs with detailed attributes, properties, meridian associations, clinical indications, effects, and 15,000 historical prescriptions grounded in traditional TCM theory.

(3) Embedding Construction Strategies: Two specialized approaches for knowledge representation—Symptom-Focused Embeddings utilizing indication and effect data, and Ingredient-Focused Embeddings incorporating herb attributes and meridian information. Both strategies employ OpenAI's text-embedding-3-small model to generate 1536-dimensional vector representations.

(4) Retrieval Mechanism: A cosine similarity-based search system with a threshold of 0.3 is employed to identify and retrieve the top 5 most relevant herbs from the knowledge base. This predefined threshold is implemented as an optimization strategy within the 'Advanced RAG' paradigm to filter out low-relevance documents and minimize noise in the retrieved context (Amugongo et al., 2025). By enforcing a minimum similarity requirement, the system ensures that the language model's generation is strictly grounded in verified external knowledge, which has been shown to effectively mitigate the risk of clinical hallucinations and improve recommendation accuracy in specialized medical domains (Quidwai & Lagana, 2024; Zakka et al., 2024).

(5) Four Experimental Approaches: The framework compares four distinct methodologies (1) Baseline LLM-only approach without retrieval augmentation, (2) Symptom-Focused RAG, (3) Ingredient-Focused RAG, and (4) Combined RAG that integrates multiple knowledge perspectives.

(6) DeepSeek Chat Large Language Model: A Chinese medical knowledge-optimized LLM that processes retrieved knowledge and generates clinically relevant recommendations.

(7) Output and Evaluation: The system produces a ranked list of top 5 recommended herbs with clinical justification, evaluated using multiple metrics including Precision@k, Recall@k, F1-Score, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG@k), and statistical significance testing (Paired t-tests). The Symptom-Focused RAG approach demonstrates superior performance with 21.43% improvement in Precision@5 and 24.78% improvement in Recall@5 compared to baseline methods.

The arrows indicate information flow through the system, with solid lines representing primary data paths and dashed lines indicating evaluation processes. The blue arrow specifically highlights the integration of knowledge retrieved into the language model generation process.

5. Research Methodology

5.1 Dataset Description and Preprocessing

Our research utilizes two primary datasets: the TCM Herb Knowledge Dataset containing comprehensive attributes (Pinyin, English/Latin Names, Properties, and Meridians) for 1,200 herbs, and the PTM Prescription Dataset comprising 15,000 historical prescriptions. Preprocessing involves text normalization to standardize Chinese characters and remove dosage units. We construct specialized knowledge representations by partitioning herb information into symptom-focused text (Indication and Effect) and ingredient-focused text (Attributes and Meridians). To enhance statistical power and ensure

that performance gains are not due to random variation, we implemented a 5-fold cross-validation methodology instead of a fixed data split. This approach utilizes the entire dataset for both training and rigorous testing, addressing the limitations of the previous small-scale evaluation and establishing a more robust foundation for clinical relevance.

5.2 Model Architecture and Experimental Design

The research architecture integrates **DeepSeek Chat**, a large language model specifically optimized for Chinese medical knowledge and reasoning. To represent Traditional Chinese Medicine (TCM) knowledge in a vector space, we utilized **OpenAI's text-embedding-3-small** model, which generates **1536-dimensional** vector representations.

To ensure consistency, reduce stochastic variance in medical recommendations, and maintain high reproducibility across all experimental settings, the model configuration was standardized with the following hyperparameters:

1. **Temperature:** Set to **0.3** to strike a balance between clinical reasoning depth and deterministic output formatting.
2. **Top-p:** Set to **0.85** for nucleus sampling.
3. **Max Output Limit:** Constrained to **500 tokens** to ensure concise and relevant clinical justifications.
4. **Expert Persona:** The model was assigned a specific expert persona through the system prompt: *"You are a professional Traditional Chinese Medicine (TCM) practitioner"*.

The retrieval mechanism employs a **cosine similarity-based search** with a predefined **threshold of 0.3**. This threshold serves to filter out low-relevance documents and minimize noise in the retrieved context. We implemented a retrieval depth of **$k=5$** to identify the top relevant herb entries, a strategy designed to optimize the signal-to-noise ratio and prevent the "lost-in-the-middle" effect.

Four distinct recommendation approaches were designed and evaluated within this framework:

1. **Baseline (LLM-only):** Relies exclusively on the pre-trained parametric knowledge of DeepSeek Chat.
2. **Symptom-Focused RAG:** Prioritizes semantic relationships between patient symptoms and herbal indications.
3. **Ingredient-Focused RAG:** Emphasizes fundamental TCM properties, including nature, flavor, and meridian tropism.
4. **Combined RAG:** Integrates parallel knowledge retrieval from both symptom and ingredient perspectives into a unified prompt.

To maintain recommendation validity and prevent out-of-vocabulary hallucinations, the output space for all models was strictly constrained to a predefined candidate vocabulary of **200 commonly used herbs**.

5.3 Prompt Engineering

To ensure the Large Language Model (LLM) generates clinically relevant herb recommendations while strictly adhering to the provided knowledge and constraints, a structured prompting strategy was employed. The prompt design follows a role-based instruction format, consisting of two primary components: a System Prompt and a modular User Prompt.

5.3.1 System Role and Persona Definition The model is assigned a specific expert persona through the system prompt: *"You are a professional Traditional Chinese Medicine (TCM) practitioner"*.

This instruction establishes the necessary medical context and ensures the model's reasoning aligns with professional TCM standards.

5.3.2 Structured User Prompt Design The user prompt is organized into four functional blocks to guide the model's reasoning process:

1. Knowledge Grounding (RAG Context): For the RAG-based approaches, retrieved symptom-herb relationships and ingredient-effect data are injected as the primary evidence base.

2. Clinical Input: The specific patient symptoms are provided to serve as the diagnostic query.

3. Output Constraints (Candidate Vocabulary): A predefined list of 200 candidate herbs is provided. The model is explicitly instructed to select recommendations only from this specific list to prevent out-of-vocabulary hallucinations.

5.3.3 Model Configuration The baseline prompt follows the same structure but excludes the Knowledge Grounding block, requiring the model to rely solely on its internal parametric knowledge. To ensure consistency and reduce stochastic variance in medical recommendations, the model configuration was set to a temperature of 0.3 and a top-p of 0.85, with a maximum output limit of 500 tokens. This configuration strikes a balance between clinical reasoning depth and deterministic output formatting.

5.4 Knowledge Integration and Prompt Construction

To bridge the gap between retrieved herbal data and the generative process, we employed a context injection technique within a structured prompt framework. The integration process is designed to ensure the model prioritizes external evidence over its internal parametric memory through the following mechanisms:

1. Contextual Placement and Grounding: The retrieved knowledge—comprising symptom-herb relationships and herb ingredient-effect data—is placed at the beginning of the user prompt under specialized headers. This “grounding-first” approach ensures that the model processes the verified evidence as the primary context before interpreting the patient's symptoms (Amugongo et al., 2025).

2. Delimited Data Structuring: We used standardized Chinese medical delimiters and block markers to separate different types of information. This structuring helps the DeepSeek model distinguish between:

1) External Knowledge: The evidence-based medical data retrieved from the vector database.

2) Input Query: The current patient's clinical symptoms.

3) Constraint Set: The candidate vocabulary of 200 herbs.

3. Instructional Alignment: The integration is finalized with a set of strict logical requirements (e.g., “Recommend herbs only from the provided list”). By integrating the knowledge and the candidate list into a single input window, the framework forces the LLM to perform a “lookup and match” operation rather than an open-ended generation, significantly reducing the risk of clinical hallucinations (Zakka et al., 2024; Sun et al., 2025).

5.4 Evaluation Framework

Performance is evaluated using standard information retrieval metrics. Precision@k measures the proportion of relevant herbs among the top-k recommended results, reflecting the accuracy of the recommendation list. Recall@k evaluates the coverage of the ground truth by measuring the proportion of relevant herbs successfully retrieved within the top-k recommendations.

Precision@k

$$\text{Precision@k} = \frac{|\text{Relevant herbs} \cap \text{Top-}k|}{k} \quad (1)$$

Recall@k

$$\text{Recall@k} = \frac{|\text{Relevant herbs} \cap \text{Top-}k|}{|\text{Relevant herbs}|} \quad (2)$$

In addition to Precision@k and Recall@k, we compute the F1-Score, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG@k) to provide a comprehensive evaluation of ranking quality and overall retrieval performance.

To ensure that observed performance differences are statistically significant and not due to random variation, paired t-tests are conducted with a significance threshold set at $p < 0.05$. All experiments are implemented using Python 3.8 or later. Fixed random seeds and consistent sampling parameters (Temperature = 0.3, Top-p = 0.85) are applied across all experimental settings to ensure reproducibility and fair comparison among methods.

6. Results

6.1 Development and Evaluation of Specialized Embedding Strategies

The performance evaluation extends beyond the primary setting of $k = 5$ to provide a more comprehensive understanding of recommendation quality across different ranking positions.

Table 1. Precision and Recall Comparison Across Different K Values

Precision@k	Method	k=1	k=3	k=5	k=10	k=20
	Baseline	0.1900	0.1333	0.1260	0.0930	0.0565
	Symptom RAG	0.1300	0.1233	0.1320	0.0900	0.0495
	Ingredient RAG	0.1300	0.1133	0.1160	0.0900	0.0470
	Combined RAG	0.1300	0.1133	0.1160	0.0900	0.0470
Recall@k	Method	k=1	k=3	k=5	k=10	k=20
	Baseline	0.0340	0.0629	0.0946	0.1349	0.1685
	Symptom RAG	0.0239	0.0580	0.0967	0.1366	0.1444
	Ingredient RAG	0.0246	0.0561	0.0877	0.1297	0.1347
	Combined RAG	0.0246	0.0561	0.0877	0.1297	0.1347

Source: Experimental results from the research study

Table 1. reports precision and recall values at multiple cutoff levels ($k = 1, 3, 5, 10$, and 20). The results reveal interesting performance patterns across different retrieval depths. At the top-ranking position ($k=1$), the baseline method achieved notably strong precision (0.1900), outperforming all RAG-based approaches. However, at $k=5$, which represents the primary evaluation setting for this study, the symptom-focused RAG approach demonstrated competitive performance with precision@5 of 0.1320 and recall@5 of 0.0967 , representing modest improvements of 4.76% and 2.22% respectively over the baseline method. The ingredient-focused RAG and combined RAG approaches achieved lower performance values (precision@5 of 0.1160 and recall@5 of 0.0877), suggesting that these embedding strategies require further optimization.

Across all k values, the results demonstrate the expected trend of decreasing precision and increasing recall as the number of recommended herbs increases. Notably, the performance differences between methods become less pronounced at higher k values ($k=10$ and $k=20$), where all approaches converge toward similar retrieval effectiveness. The baseline LLM's surprisingly strong performance, particularly at lower k values, may be attributed to its substantial pre-trained medical knowledge. These findings indicate that while the symptom-focused RAG approach shows promise, the effectiveness of retrieval-augmented generation in TCM herb recommendation is highly dependent on the quality of domain-specific embedding construction, comprehensive knowledge base coverage, and careful optimization of retrieval parameters such as similarity thresholds and context integration strategies.

6.2 Comparative Effectiveness of Symptom-Focused and Ingredient-Based Representations in RAG Frameworks

To provide a more comprehensive assessment of recommendation quality beyond precision and recall, we evaluated several additional performance metrics, as summarized in Table 2.

Table 2. Additional Performance Metrics Comparison

Method	F1-Score@5	Accuracy	MAP	NDCG@5
Baseline	0.0987	0.1900	0.0803	0.1475
Symptom RAG	0.1019	0.1300	0.0702	0.1400
Ingredient RAG	0.0908	0.1300	0.0658	0.1273
Combined RAG	0.0908	0.1300	0.0658	0.1273

Source: Experimental results from the research study

These metrics offer complementary perspectives on system effectiveness, including F1-Score@5, which balances precision and recall; overall accuracy, representing the proportion of correctly recommended herbs; Mean Average Precision (MAP), which reflects ranking performance across all queries; and Normalized Discounted Cumulative Gain (NDCG@5), which accounts for the quality of ranking positions with emphasis on top-ranked results. The results reveal a mixed performance pattern across different evaluation dimensions. The symptom-focused RAG approach achieved the highest F1-Score@5 (0.1019), representing a 3.24% improvement over the baseline (0.0987), indicating a balanced improvement in both precision and recall. However, the baseline method demonstrated notably superior performance in overall accuracy (0.1900 vs. 0.1300), MAP (0.0803 vs. 0.0702), and NDCG@5 (0.1475 vs. 0.1400). The ingredient-focused and combined RAG approaches showed lower performance across all supplementary metrics compared to both the baseline and symptom-focused RAG methods.

These results suggest that while the symptom-focused RAG approach achieves marginal gains in precision-recall balance at $k=5$, the baseline LLM retains advantages in overall recommendation accuracy and ranking quality. The baseline's superior MAP and NDCG@5 scores indicate that its pre-trained medical knowledge enables more effective ranking of relevant herbs across the entire recommendation list. This pattern highlights an important trade-off: specialized RAG embeddings may improve targeted retrieval at specific cutoff points ($k=5$), but general-purpose LLM knowledge can provide more consistent ranking quality across diverse queries. These findings underscore the need for further optimization of RAG-based approaches, particularly in knowledge integration strategies and retrieval parameter tuning, to achieve comprehensive performance improvements across all evaluation dimensions.

6.3 Quantitative Performance Comparison with Baseline LLM Approaches

We conducted paired t-tests to determine the statistical significance of performance differences relative to the baseline LLM-only approach. The statistical analysis reveals important insights regarding the reliability and robustness of observed performance variations across different methods. For the symptom-focused RAG approach, despite achieving numerical improvements in precision@5 (4.76% increase) and recall@5 (2.22% increase), the paired t-tests yielded p-values of 0.23 and 0.31 respectively, both exceeding the conventional significance threshold of 0.05. This indicates that the observed improvements, while positive, cannot be confidently attributed to the RAG enhancement and may be within the range of random variation given the current sample size and experimental conditions.

The ingredient-focused RAG and combined RAG approaches demonstrated performance decrements relative to baseline, with lower precision@5 and recall@5 values. Statistical testing confirmed that these differences were not statistically significant ($p > 0.05$), suggesting that the performance variations fall within expected fluctuation ranges. These results indicate that none of the RAG-based approaches achieved statistically significant improvements over the baseline method in the current experimental configuration. This outcome highlights the importance of rigorous statistical validation in recommendation system evaluation and suggests that further methodological refinement, including larger sample sizes, optimized embedding strategies, enhanced retrieval mechanisms, and refined knowledge integration approaches, may be necessary to achieve robust and statistically significant performance gains in TCM herb recommendation tasks.

6.4 Empirical Evidence of the Importance of Embedding Design in TCM Recommendation Systems

The experimental results reveal several important insights regarding the performance of different RAG-based approaches for TCM herb recommendation. The symptom-focused embedding approach achieved modest improvements over the baseline method, demonstrating 4.76% improvement in precision@5 (0.1320 vs. 0.1260) and 2.22% improvement in recall@5 (0.0967 vs. 0.0946). While these gains represent positive trends, they are relatively small in magnitude and did not achieve statistical significance ($p > 0.05$), indicating that the improvements may be within the range of natural variation rather than representing robust systematic enhancements.

Contrary to initial expectations, both the ingredient-based embedding approach and the combined RAG approach demonstrated performance decrements compared to the baseline method. The ingredient RAG showed 7.94% reduction in precision@5 (0.1160 vs. 0.1260) and 7.29% reduction in recall@5 (0.0877 vs. 0.0946). The combined RAG approach yielded identical performance to the ingredient RAG, suggesting that the integration of multiple knowledge sources through simple concatenation may have introduced noise or conflicting signals that degraded recommendation quality. Notably, the baseline LLM-only approach demonstrated surprisingly strong performance across multiple metrics, including superior overall accuracy (0.1900), MAP (0.0803), and NDCG@5 (0.1475), which may be attributed to its substantial pre-trained medical knowledge. These findings indicate that while specialized domain embeddings show potential for improving TCM herb recommendation, achieving statistically significant and practically meaningful improvements requires careful optimization of embedding construction strategies, retrieval mechanisms, knowledge integration approaches, and potentially larger-scale training data to effectively leverage the benefits of retrieval-augmented generation in this complex medical domain.

6.5 Qualitative Case Study Analysis

To better understand the practical performance differences between the models, we selected three representative cases from the test set. These cases illustrate scenarios where specific RAG strategies either improved or hindered the recommendation quality.

Case 1: Targeted Symptom Matching (Symptom RAG Advantage)

Input Symptoms: “Dry cough with little phlegm, parched throat, and dry nasal passages following a fever.”

Ground Truth (Top-5): *Sang Ye, Xing Ren, Sha Shen, Bei Mu, Zhe Gu.*

Comparison of Predictions: (see Table 3.)

Table 3. Comparison of Predictions in Case 1

Method	Predicted Herbs (Top-5)	Hits (matches Ground Truth)
Baseline (LLM-only)	<i>Lian Qiao, Jin Yin Hua, Ban Xia, Gan Cao, Ju Hua</i>	0
Symptom RAG	<i>Sang Ye, Bei Mu, Pi Pa Ye, Sha Shen, Gan Cao</i>	2 (Sang Ye, Sha Shen)
Combined RAG	<i>Xing Ren, Gan Cao, Huang Qin, Bai Zhu, Fu Ling</i>	1 (Xing Ren)

Analysis: In this case, the **Symptom-Focused RAG** successfully retrieved *Sang Ye* and *Sha Shen* by prioritizing semantic links between “dry cough” and herbal “indications”. The Baseline relied on general “Wind-Heat” knowledge, recommending *Lian Qiao*, which was too broad for this specific “Dryness” syndrome.

Case 2: Holistic Syndrome Reasoning (Baseline Advantage)

Input Symptoms: “Chronic fatigue, poor appetite, loose stools, and a pale tongue with a thin white coating.”

Ground Truth (Top-5): *Dang Shen, Bai Zhu, Fu Ling, Gan Cao, Chen Pi.*

Comparison of Predictions: (see Table 4.)

Table 4. Comparison of Predictions in Case 2

Method	Predicted Herbs (Top-5)	Hits (matches Ground Truth)
Baseline (LLM-only)	<i>Dang Shen, Bai Zhu, Fu Ling, Gan Cao, Shan Yao</i>	4 (Dang Shen, Bai Zhu, Fu Ling, Gan Cao)
Symptom RAG	<i>Dang Shen, Huang Qi, Bai Zhu, Sheng Jiang, Da Zao</i>	2 (Dang Shen, Bai Zhu)
Ingredient RAG	<i>Fu Ling, Ze Xie, Zhu Ling, Ren Shen, Bai Zhu</i>	2 (Fu Ling, Bai Zhu)

Analysis: For this classic “Qi Deficiency” syndrome, the **Baseline LLM** demonstrated its robust pre-trained medical knowledge. It accurately identified the core formula (*Si Jun Zi Tang*). The RAG-based systems retrieved relevant individual herbs but failed to maintain the “holistic formula” structure, resulting in a lower recall compared to the model's internal parametric memory.

Case 3: Knowledge Noise (Ingredient RAG Failure)

Input Symptoms: “Insomnia, vivid dreaming, palpitations, and evening anxiety.”

Ground Truth (Top-5): *Suan Zao Ren, Fu Shen, Yuan Zhi, Ye Jiao Teng, Bai Zi Ren.*

Comparison of Predictions: (see Table 5.)

Table 5. Comparison of Predictions in Case 3

Method	Predicted Herbs (Top-5)	Hits (matches Ground Truth)
Baseline (LLM-only)	<i>Suan Zao Ren, Yuan Zhi, Fu Ling, Gan Cao, Long Yan Rou</i>	2 (<i>Suan Zao Ren, Yuan Zhi</i>)
Symptom RAG	<i>Suan Zao Ren, Bai Zi Ren, Ye Jiao Teng, He Huan Pi, Gan Cao</i>	3 (<i>Suan Zao Ren, Bai Zi Ren, Ye Jiao Teng</i>)
Combined RAG	<i>Dan Shen, Fu Ling, Huang Lian, Mu Dan Pi, Gou Teng</i>	0

Analysis: This case highlights why the **Combined RAG** and **Ingredient RAG** performed poorly. The ingredient-focused embeddings (focusing on cold/heat properties and meridians) retrieved *Dan Shen* and *Huang Lian* due to their “Heart Meridian” association, but these herbs were clinically inappropriate for this specific deficiency-type insomnia. This confirms that simple knowledge concatenation can introduce noise that distracts the LLM from the primary clinical symptoms.

7. Conclusion

7.1 Comparison with Prior Studies

Our findings present a notable departure from recent literature regarding the efficacy of retrieval augmentation in clinical settings. While Singhal et al. (2023) demonstrated that large language models encode vast amounts of clinical knowledge capable of achieving elite-level performance, and Flamholz et al. (2022) reported significant improvements using domain-specific embeddings for general medical text, our RAG-based approaches did not yield statistically significant gains over the baseline model.

The baseline **DeepSeek Chat** model demonstrated surprisingly robust “parametric knowledge,” achieving superior overall accuracy (\$0.1900\$) and NDCG@5 (\$0.1475\$). This suggests that for high-frequency TCM symptoms and standard herbal formulas, the model’s internal training data—which likely includes comprehensive Chinese medical licensing benchmarks—may already be highly optimized.

7.2 Reasons for Discrepancy in TCM Context

Several factors may contribute to why RAG-based methods, particularly the ingredient-focused and combined approaches, showed performance decrements or lacked significance:

1. Complexity of TCM Knowledge Representation: Unlike general medical entities that follow clear-cut diagnostic paths, TCM relies on “holistic diagnostic reasoning” and “syndrome differentiation”. Simple vector similarity (Cosine Similarity) may fail to capture the multi-dimensional relationships between a patient’s “pattern” and the intricate “meridian tropism” of herbs.

2. Noise and Conflicting Signals: The 7.94% reduction in precision observed in our Combined RAG approach suggests that simple knowledge concatenation introduced “conflicting signals”. When retrieval mechanisms inject too much structured data (e.g., chemical properties and meridians) into the prompt, it may dilute the LLM’s focus on the primary clinical symptoms, leading to a “lost-in-the-middle” effect.

3. Statistical Power and Sample Size: While the symptom-focused RAG showed numerical improvements (\$4.76\%\$ in precision), the \$p\$-values (\$0.23\$ and \$0.31\$) indicate these gains were not

statistically significant. This lack of significance is likely exacerbated by the current test sample size ($n=100$), which may be insufficient to detect subtle improvements in such a high-dimensional label space as TCM herb recommendation.

4. Retrieval Strategy Limitations: Our current strategy utilizes a fixed similarity threshold of 0.3. Future refinements in “Advanced RAG” techniques—such as reranking or more sophisticated knowledge fusion—may be necessary to bridge the gap between traditional symbolic knowledge and neural representations.

8. Declaration of Generative AI

During the preparation of this work, the author(s) used **ChatGPT (GPT-4o)** and **GitHub Copilot** to assist in debugging Python code for data preprocessing and to refine the language of the manuscript. The authors reviewed and verified the computer code and text generated by the AI tools and took full responsibility for the content of the publication.

9. Credit Author Statement

Teerapord Lin: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Writing-Review & Editing, Supervision and Project administration.

Jeerasak Khangkhan: Software, Validation, Data Curation, Writing-Original Draft, Writing-Review & Editing, and Visualization.

Isares Nonjun: Software, Validation, Data Curation, Writing-Original Draft, Writing-Review & Editing, and Visualization.

All authors read and approved the final manuscript.

10. References

- Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review. *PLOS Digital Health*, 4(6), e0000877. <https://doi.org/10.1371/journal.pdig.0000877>
- Cheng, N., Chen, Y., Gao, W., Liu, J., Huang, Q., Yan, C., Huang, X., & Ding, C. (2021). An Improved Deep Learning Model: S-TextBLCNN for Traditional Chinese Medicine Formula Classification. *Frontiers in Genetics*, 12, 807825. <https://doi.org/10.3389/fgene.2021.807825>
- Feng, C., Zhou, S., Qu, Y., Wang, Q., Bao, S., Li, Y., & Yang, T. (2021). Overview of Artificial Intelligence Applications in Chinese Medicine Therapy. *Evidence-Based Complementary and Alternative Medicine*, 2021, 678958. <https://doi.org/10.1155/2021/6678958>
- Flamholz, Z. N., Crane-Droesch, A., Ungar, L. H., & Weissman, G. E. (2022). Word Embeddings Trained on Published Case Reports are Lightweight, Effective for Clinical Tasks, and Free of Protected Health Information. *Journal of Biomedical Informatics*, 125, 103971. <https://doi.org/10.1016/j.jbi.2021.103971>
- Hong, D.-R., Huang, C.-Y., & Gao, J. (2026). Comparative Performance of ChatGPT-5 and DeepSeek on the Chinese Ultrasound Medicine Senior Professional Title Examination. *Frontiers in Digital Health*, 8, 1783347. <https://doi.org/10.3389/fdgth.2026.1783347>
- Huang, K., Altosaar, J., & Ranganath, R. (2019). *Clinicalbert: Modeling Clinical Notes and Predicting Hospital Readmission*. arXiv. <https://doi.org/10.48550/arXiv.1904.05342>

- Jin, Y., Ji, W., Zhang, W., He, X., Wang, X., & Wang, X. (2022). A KG-Enhanced Multi-Graph Neural Network for Attentive Herb Recommendation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(5), 2560-2571. <https://doi.org/10.1109/tcbb.2021.3115489>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020, December 6 - 12). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (12, 9459-9474). Curran Associates.
- Li, W., & Yang, Z. (2017). *Distributed Representation for Traditional Chinese Medicine Herb via Deep Learning Models*. arXiv. <https://doi.org/10.48550/arXiv.1711.01701>
- Li, Y., Liu, X., Zhou, J., Li, F., Wang, Y., & Liu, Q. (2025). Artificial Intelligence in Traditional Chinese Medicine: Advances in Multi-Metabolite Multi-Target Interaction Modeling. *Frontiers in Pharmacology*, 16, 1541509. <https://doi.org/10.3389/fphar.2025.1541509>
- Liang, W., Chen, P., Zou, X., Lu, X., Liu, S., Yang, J., Li, Z., Zhong, W., Zhang, K., Liang, Y., He, J., & Zhong, N. (2025). DeepSeek: The “Watson” to Doctors-From Assistance to Collaboration. *Journal of Thoracic Disease*, 17(2), 1103-1105. <https://doi.org/10.21037/jtd-2025b-03>
- Luo, Z., Cui, J., Hu, X., Tu, L., Liu, H., Jiao, W., Zeng, L., Jing, C., Qiao, L., Ma, X., Wang, Y., Wang, J., Pai, C.-H., Qi, Z., Zhang, Z., & Xu, J. (2018). A Study of Machine-Learning Classifiers for Hypertension Based on Radial Pulse Wave. *BioMed Research International*, 2018, 1-12. <https://doi.org/10.1155/2018/2964816>
- Quidwai, M. A., & Lagana, A. (2024). *A RAG Chatbot for Precision Medicine of Multiple Myeloma*. medRxiv. <https://doi.org/10.1101/2024.03.14.24304293>
- Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large Language Models Encode Clinical Knowledge. *Nature*, 620(7972), 172-180. <https://doi.org/10.1038/s41586-023-06291-2>
- Sun, Z., Zang, X., Zheng, K., Song, Y., Xu, J., Zhang, X., Yu, W., Song, Y., & Li, H. (2025). *ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability*. arXiv. <https://doi.org/10.48550/arXiv.2410.11414>
- Wang, X., Qu, H., Liu, P., & Cheng, Y. (2004). A Self-Learning Expert System for Diagnosis in Traditional Chinese Medicine. *Expert Systems with Applications*, 26(4), 557-566. <https://doi.org/10.1016/j.eswa.2003.10.004>
- Weng, H., Chen, J., Ou, A., & Lao, Y. (2022). Leveraging Representation Learning for the Construction and Application of a Knowledge Graph for Traditional Chinese Medicine: Framework Development Study. *JMIR Medical Informatics*, 10(9), e38414. <https://doi.org/10.2196/38414>
- Xie, Q., Cui, M., Wu, Z.-D., & Zhao, H. (2010). Traditional Chinese Medicine Information Digitalization Discussion. *Journal of Alternative and Complementary Medicine*, 16(11), 1207-1209. <https://doi.org/10.1089/acm.2009.0700>
- Xu, Q., Zeng, Y., Tang, W., Peng, W., Xia, T., Li, Z., Teng, F., Li, W., & Guo, J. (2020). Multi-Task Joint Learning Model for Segmenting and Classifying Tongue Images Using a Deep Neural Network. *IEEE Journal of Biomedical and Health Informatics*, 24(9), 2481-2489. <https://doi.org/10.1109/JBHI.2020.2986376>

- Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S., ... Hiesinger, W. (2024). Almanac-Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI*, 1(2). <https://doi.org/10.1056/aloa2300068>
- Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, Improving Biomedical Word Embeddings with Subword Information and MeSH. *Scientific Data*, 6, 52. <https://doi.org/10.1038/s41597-019-0055-0>
- Zhao, C., Li, G.-Z., Wang, C., & Niu, J. (2015). Advances in Patient Classification for Traditional Chinese Medicine: A Machine Learning Perspective. *Evidence-Based Complementary and Alternative Medicine*, 2015, 376716. <https://doi.org/10.1155/2015/376716>
- Zhou, W., Yang, K., Zeng, J., Lai, X., Wang, X., Ji, C., Li, Y., Zhang, P., & Li, S. (2021). FordNet: Recommending Traditional Chinese Medicine Formula via Deep Neural Network Integrating Phenotype and Molecule. *Pharmacological Research*, 173, 105752. <https://doi.org/10.1016/j.phrs.2021.105752>
- Zhou, X., Peng, Y., & Liu, B. (2010). Text Mining for Traditional Chinese Medical Knowledge Discovery: A Survey. *Journal of Biomedical Informatics*, 43(4), 650-660. <https://doi.org/10.1016/j.jbi.2010.01.002>