面向針灸穴位可解釋推薦的 LLM－RAG 協同方法研究

Research on LLM-RAG Collaborative Method for Interpretable Recommendation of Acupuncture Points

劉 傳霞[*]・魏璐璐[**]・瞿汶境[**]・吳顺平[**]・顧 沁[**]・甘 泉[**,†]

LIU Chuanxia[*], WEI Lulu[**], QU Wenjing[**], WU Shunping[**], GU Qin[**],GAN Quan[**,†]

（摘要）

針對傳統針灸穴位決策高度依賴經驗、標準化與可追溯機制不足等問題，本文提出一種融合大型語言模型（LLM）與檢索增強生成（RAG）的可解釋推薦方法。以《黃帝內經》為核心知識來源，構建章節級向量數據庫；採用 LoRA 對 Qwen1.5-0.5B-Chat 進行針灸領域微調，設計「症狀—證型—主穴—配穴」的結構化生成流程，並引入邏輯推理模組以實現證據溯源與一致性校核。在系統實作上，前端採用 Vue.js 構建互動介面，後端基於 FastAPI 對接模型與向量數據庫，形成可部署的 Web 決策系統。測試結果表明，系統可穩定輸出涵蓋主穴、配穴及其理論依據的結構化答案，支援經典文獻追溯，具備較高的可解釋性與專業性，整體表現優於通用型 LLM。本研究構建了一套面向針灸從業者與醫學學習者，兼顧公眾健康教育需求的可解釋性推薦系統，驗證了 LLM–RAG 協同機制在中醫針灸智能化與標準化決策中的可行性與推廣潛力。

關鍵詞：針灸；穴位決策；大型語言模型；檢索增強生成；可解釋性

（Abstract）

To address the heavy reliance on practitioner experience and the lack of standardization and traceability in traditional acupoint decision-making, this study proposes an interpretable recommendation approach that couples a large language model (LLM) with retrieval-augmented generation (RAG). Using the Huangdi Neijing as the core knowledge source, we build a chapter-level vector knowledge base; apply LoRA to fine-tune Qwen1.5-0.5B-Chat for the acupuncture domain; design a structured generation pipeline (“symptoms → pattern (zheng) → primary points → adjunct points”); and integrate a logic-reasoning module for evidence tracing and consistency checking. System-wise, we implement a deployable web decision-support system with a Vue.js front end and a FastAPI back end interfacing the model and vector database. Experimental results show that the system consistently produces structured outputs covering primary points, adjunct points, and their theoretical rationales, supports trace-back to classical texts, and outperforms general-purpose LLMs in interpretability and domain competency. This study develops an interpretable decision-support system for acupuncture practitioners and medical learners, while also addressing public health education needs. The findings demonstrate the feasibility and scalability of the LLM–RAG collaborative mechanism for intelligent and standardized decision-making in traditional Chinese medicine acupuncture.

Keywords: Acupuncture; Acupuncture Point Decision-Making; Large Language Model (LLM); Retrieval-Augmented Generation (RAG); Interpretability

* School of Foreign Languages, Jiangsu Ocean University
** School of Computer Engineer, Jiangsu Ocean University
† Corresponding Author

Journal of East Asian Identities Vol. 11 March 2026 (pp. 31-39)

1. 引言

針灸是傳統中醫的重要組成部分，透過刺激特定腧穴以調節氣血運行與臟腑功能，於疾病治療與預防中發揮獨特作用。穴位決策為針灸療效的關鍵環節，其理論源流可上溯《黃帝內經》——現存最早且體系完備的針灸理論典籍，建構了經絡循行、腧穴主治與臟腑關聯的基本框架，為後世臨床取穴提供了核心依據[1–2]。

1979 年，世界衛生組織首次公布 43 種針灸適應證，標誌針灸逐步納入全球醫學體系。然而，臨床實踐仍面臨兩項關鍵挑戰：其一，傳統穴位決策高度依賴個體經驗，缺乏可追溯、可標準化的決策機制；其二，既有病譜總結仍不完備，部分適宜病種未被充分覆蓋，致使決策效率與應用範圍受限[3]。

近年來，人工智慧（Artificial Intelligence，AI；又稱「人工智能」）的快速發展推動傳統中醫資訊化加速演進。大型語言模型（Large Language Model，LLM）在醫學文本問答與病因分析等任務上展現潛力；已有研究顯示，融入辨證體系的 LLM 能支援中醫知識問答、典籍解析與輔助診斷等功能[4–6]，部分模型（如 Qwen、GPT）於西醫真實病例推理的表現亦接近臨床醫師水準[7]。同時，邏輯推理模組與檢索增強生成（Retrieval-Augmented Generation，RAG）的成熟，賦予模型基於經典文獻與知識庫進行語義級知識調用與因果鏈構建之能力，與中醫「由因溯果」的推理機制高度契合[8]。

儘管如此，現有研究多停留於知識檢索與問答層面，尚未有效跨越至面向臨床的高精度、可解釋決策支持，難以滿足針灸在真實場景中的規範化需求 [9–10]。為此，本文面向穴位智慧推薦場景，提出一種融合 LLM、RAG 與邏輯推理模組 的協同方法：以《黃帝內經》為核心語料構建章節級向量數據庫；採用 LoRA（Low-Rank Adaptation，低秩適配）對 Qwen1.5-0.5B-Chat 進行針灸領域微調；規範「症狀—證型—主穴—配穴」之生成路徑，並以邏輯規則校驗實現理論依據的可溯源；最終搭建 Web 系統以支援臨床端部署與驗證。

本研究旨在回應傳統中醫智慧化轉型的實際需求，並探索文獻結構化利用與 AI 深度融合的可行路徑與可驗證範式。所構建系統定位為服務於針灸從業者與醫學學習者的決策支持工具，重點提供具可解釋性的穴位推薦與經典理論依據追溯功能，用以輔助臨床決策與教學實踐，而非作為自動化診療之替代方案。

2. 理論基礎與關鍵技術

2.1 針灸治療

針灸是中醫學中重要的醫療手段，透過針刺或艾灸刺激體表特定部位（穴位，學名「腧穴」），以調節經絡氣血與臟腑功能，達到防治疾病之目的[11]。

依來源、分布與功能特點，穴位大致分為三類：經穴、奇穴（經外奇穴）與阿是穴。其一，經穴分布於十二經脈及任、督二脈上，名稱、定位與所屬經脈明確，並與臟腑功能密切對應；其二，奇穴（經外奇穴）不屬十四經系統，然具備固定名稱與定位，且有特定主治，多源於長期臨床實踐之經驗總結；其三，阿是穴無固定名稱與定位，而以患者疼痛或不適處為取穴依據，臨證取之應手 [12]。

在中醫理論中，穴位—經絡—臟腑並非孤立：臟腑生理／病理變化可反映於相應經脈與腧穴；刺激相關經脈與腧穴，亦可反向調節臟腑功能 [13]。臟腑化生氣血，為人體功能之樞紐 [14]；《靈樞·經脈》曰：「內屬於臟腑，外絡肢節……谷入於胃，脈道以通，血氣乃行」，顯示經絡為聯貫內外、運輸氣血之網絡系統[15–16]。此種「三位一體」的關係構成中醫認識生理與病理的核心理論，亦為針灸等傳統療法的底層邏輯，至今仍是臨床實踐的重要指導思想。

2.2 大語言模型

大型語言模型是以海量文本語料進行預訓練的深度學習模型，多採用 Transformer 架構與自注意力機制，配合「預訓練－微調」範式學習語言規律；其參數規模可達數十億量級，具備強大的自然語言理解與生成能力[17]。可用於問答互動、摘要撰寫、文本生成與翻譯等多種場景，是當前人工智慧（AI）領域實現通用語言能力的核心技術之一。

在針灸領域，大型語言模型可從經典論著與臨床案例等文本中學習經絡—腧穴—主治—操作等知識，並對患者的症狀敘述、既往史與治療反應等自然語言資訊進行語義解析與結構化抽取，據以輔助完成證候辨識與穴位決策。具體而言，模型能將「症狀—證型—主穴—配穴」關係鏈條化，提升決策的效率、規範性與一致性；同時在臨床溝通與教學情境中，提供可讀性良好的知識整合與示例生成，凸顯其在針灸知識體系建構與高效應用方面的價值。

2.3 邏輯推理與 RAG 增強檢索

邏輯推理係指模型在充分理解使用者所述之具體情境後，依循因果關係與論證規則組織語言，生成結構清晰、論據可依、直指問題核心的回答；其輸出既需符合

經典文獻的記載，亦須貼合當下情境之需求 [18]。

檢索增強生成（Retrieval-Augmented Generation，RAG）旨在從外部專業知識庫檢索與問題高度相關的資訊，並於生成過程中動態引入檢索結果，以強化答案的事實基礎與可追溯性 [19]。其關鍵在於：接收問題後，能精準定位至《黃帝內經》等經典文獻所承載的對應內容，作為回應的依據來源。

在此基礎上，採用「RAG + 大型語言模型」的協同架構：由 RAG 保證知識來源充分且可追溯，由大型語言模型完成語義整合與邏輯推理。於針灸決策場景中，二者協同可將「症狀—證型—主穴／配穴—依據標註」的鏈條有機銜接，輸出兼具專業性、可解釋性與一致性的高品質回答，有效支援使用者在針灸知識與決策上的需求。

3. 系統設計與實現

3.1 系統框架

系統運行流程如下：首先，使用者在前端互動介面輸入待查之具體病症並提交請求。系統採用 FastAPI 框架解析前端 POST 請求並完成校驗，確保數據結構完整、欄位合法 [20]。其後，對接收之文本進行基礎預處理，包括去除雜訊字元、中文分詞與術語表達統一等，以保障輸入內容之語義明確。處理完成的數據被封裝為標準化請求格式並由後端接收；後端隨即將該問題提交至大型語言模型。經 LoRA 微調訓練後，模型具備中醫術語的基本理解能力，可對病症歸屬作初步判別，並生成候選之穴位推薦方案 [21]。

隨後，將上述初步方案輸入 RAG 推理模組。該模組以《黃帝內經》為核心知識來源，追溯相關理論依據，並對初步結果進行合理性與正確性檢驗，必要時予以修訂與調整，以提升方案之嚴謹度與一致性。最終，經知識增強與校核後之結果回傳前端呈現，從而使針灸決策輸出更具可靠性、合理性與科學依據。系統架構如圖 1 所示。

3.2 數據集

數據集貫穿大型語言模型之訓練、推理優化與結果驗證全流程，為系統開發之核心環節；其專業性直接影響答案的正確性與使用者對系統之信任度。本專案的數據集構建服務於兩項目標：其一，用於 LoRA 輕量級微調，訓練出精通針灸穴位與辨證表達的大型語言模型，使其能辨識並理解經絡、腧穴與病症之間的複雜關係；其二，為 RAG 模型之語義檢索與（反繹）推理模組提供底層支撐，即基於《黃帝內經》構建向量數據集，以供依據追溯與內容校核之用。

3.2.1 模型微調數據集

本研究採用 Qwen1.5-0.5B-Chat 大型語言模型。雖其預訓練語料覆蓋面廣，但在針灸等專業領域之理解深度仍有限，易出現錯誤資訊（「幻覺」）。為提升系統之專業性與實用價值，我們引入 GitHub 開源的高品質中醫數據集，並藉由 LoRA 輕量化微調完成領域適配 [22]。該數據集涵蓋臨床病例、專家問答與中醫經典文獻等多源文本，資料型態豐富且品質可控。

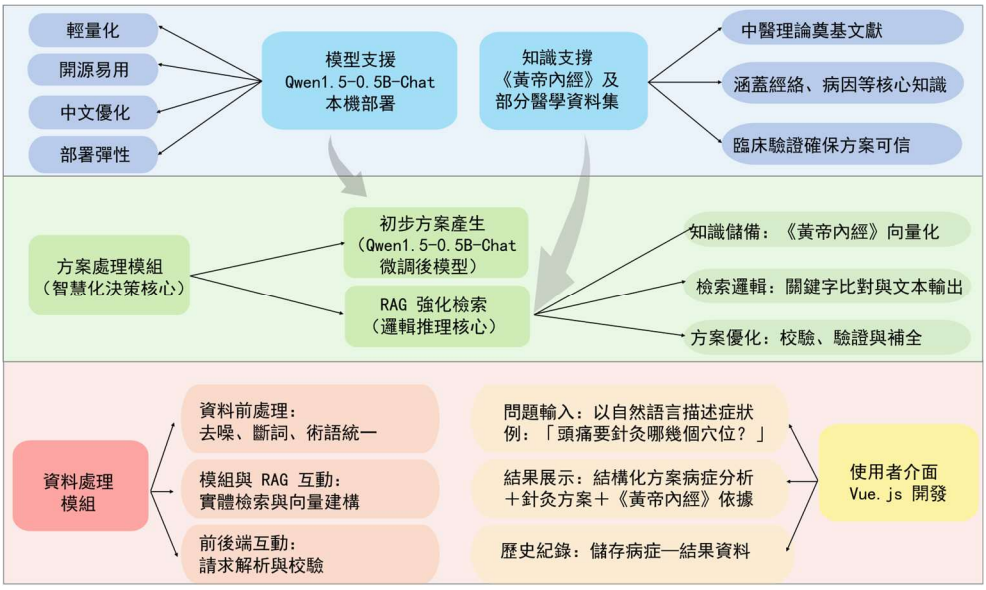

圖 1 系統架構

鑑於原始數據體系龐雜且涵蓋面過廣，本研究對其進行嚴格篩選與主題收斂，聚焦（1）經絡系統分類，（2）穴位核心資訊，（3）病症—穴位對應關係三類知識條目，共遴選約 2500 條。為便於監督式微調（SFT），將資料整理為標準化問答對：「{'input': 問題，'output': 答案}」隨後以該數據集結合 LoRA 完成領域微調，增強模型對經絡—腧穴—病症關係的辨識與表達能力。

3.2.2 向量數據集

RAG 的目的在於從外部知識庫檢索高相關資訊，以輔助大型語言模型生成更準確且更豐富的文本內容，從而增強其處理知識密集型任務的能力 [22]。因此，需依託權威且可追溯的中醫典籍作為知識來源。中醫經典是中醫理論體系的重要組成部分，其核心思想奠定了理論內涵與學術規範 [1]。《黃帝內經》作為我國現存最早的醫學典籍，亦列入傳統醫學四大經典之一，不僅與現代科學範式具有一定契合，亦為生命科學與醫學研究提供方向與方法論啟示 [23]。基於此，本研究選取《黃帝內經》作為構建原始數據集之依據。

具體流程如下：首先，依章節結構與語義單元對《黃帝內經》原文進行分段整理；其次，圍繞原文標註具有醫學知識價值的語句或段落，並以 “query + retrieved_docs + answer” 的「問答—文檔關聯」結構構建原始數據庫，為 RAG 系統優化提供語義關聯標註，以便後續建立向量數據集。

RAG 檢索的本質是語義相似度匹配，故需將文本轉化為可計算之向量表示 [22]。在原始數據集基礎上，本研究利用 Qwen 模型生成 embeddings，構建 “id + text + embedding” 形式的結構化向量數據集，以供 RAG 進行檢索。生成步驟包括：

（1）清洗與格式化：去除雜訊、正規化標點與段落；

（2）向量提取：採用 token embedding 技術獲取各 token 的向量表示；

（3）聚合表示：通過平均池化將 token 級向量聚合為句子級語義向量，以完整表徵文本語義；

（4）數據存儲：按 JSONL 格式保存，每條記錄包含三個欄位：

- id（如 chapter_1，表示章節 / 片段編號）；
- text（如「上古天真論」，表示章節標題或內容摘要）；
- embedding（經嵌入與池化後之語義向量，用於後續向量檢索與匹配）。

上述向量數據集為 RAG 的語義相似度檢索提供了可計算依據，構成本系統在中醫領域實現精準語義檢索與依據追溯的關鍵基礎。

3.3 部署模型

本研究選用 Qwen1.5-0.5B-Chat 作為基礎大型語言模型。該模型為阿里雲研發之 Qwen1.5 系列的輕量版本，參數規模 0.5B（約 5 億）。相較於參數量達數十億乃至上千億的大型模型，此配置在性能—資源折衷上具備顯著優勢：對計算與記憶體的需求更低，可大幅降低硬體門檻，尤契合本課題對系統輕量化運行的要求。

在語言理解層面，該模型經大規模中文語料預訓練，與針灸學這一傳統中醫領域的研究需求高度契合。針灸相關知識多載於中文古籍、文言文與專業文獻，含大量中醫術語與特定表達；Qwen1.5-0.5B-Chat 能較為精準地識別並理解腧穴名稱、定位、經脈歸屬與主治等術語，並把握其在中醫理論體系中的準確概念，為準確處理針灸知識奠定基礎。

面對使用者關於針灸的多樣化提問—無論是穴位定位、主治病症等基礎問題，抑或針灸理論原理等複雜問題—該模型均能把握問題要點並作出較為規範的解析。作為對話優化版本，Qwen1.5-0.5B-Chat 支援多輪交互與上下文維持，並具備良好的指令跟隨能力，可依使用者要求（如「闡述合谷穴的臨床應用」「以易懂語言說明艾灸操作要點」）調整輸出風格。雖屬輕量規模，但結合 RAG 技術可有效補足其在特定領域知識儲備上的不足，適用於針灸學此類需精準取證與依據標註的場景。

為強化 Qwen1.5-0.5B-Chat 對針灸領域知識的理解與適配性，本課題採用 LoRA 進行參數高效微調。LoRA 為面向大規模預訓練模型的參數適配與快速調優方法，其核心思路為凍結絕大部分基座權重，僅新增少量低秩矩陣對部分權重作增量調整 [24–25]。此法不僅能顯著降低運算資源消耗，亦可有效提升模型在特定領域之任務表現，尤適用於 Qwen1.5-0.5B-Chat 等輕量模型的領域適配。此外，LoRA 採用簡潔的線性設計，於部署階段可將可訓練矩陣與凍結權重合併，從而避免額外推理延遲 [26]。

LoRA 微調主要包括三個環節：

（1）結合模型結構與任務需求設計低秩矩陣，作為基座權重的增量參數；

（2）將低秩矩陣注入注意力機制或前饋網路等線性映射層，構成可訓練分支；

（3）以特定任務數據集進行微調，僅更新低秩參數，其餘權重保持凍結不變。

本研究構建以《黃帝內經》及相關針灸典籍為語料

的問答式訓練集，採用 LoRA 進行領域適配，從而增強模型在針灸任務中的領域表達能力與問題針對性。

3.4 大模型推理過程及 RAG 優化

在 Qwen1.5-0.5B-Chat 進行推理的初始階段，基於檢索增強生成（Retrieval-Augmented Generation，RAG）的「語義匹配推理」發揮關鍵的輔助決策作用。RAG 檢索增強演算法透過動態檢索外部醫學知識庫，為大模型提供關鍵領域知識，從而提升生成內容之準確性 [25]；其核心在於，能從《黃帝內經》之向量庫中精準定位與用戶問題相關的知識片段。

當用戶提出問題（例如：「大腸經的 36 個經穴是什麼？」）時，RAG 系統首先調用 Qwen 模型將該查詢轉換為向量表示，再在《黃帝內經》向量庫中計算查詢向量與各章節向量之餘弦相似度，選取與問題最相關的 Top-k 章節（如經脈相關章節等）。計算公式如下：

$$sim(e_q, e_{c_j}) = \frac{e_q \cdot e_{c_j}}{\|e_q\|_2 \cdot \|e_{c_j}\|_2} \quad (1)$$

式中：e_q為用戶提問經編碼得到的嵌入向量；e_{c_j}為《黃帝內經》各章節原文（或切分後片段）經嵌入模型編碼得到的向量。此步驟可準確鎖定回答所需的知識範圍，降低因基座訓練語料侷限而導致的知識偏差，為後續準確作答奠定基礎。

完成語義匹配後，模型進入回答生成的後半段，即「知識整合—邏輯組織推理」流程：

（1）知識整合：依託基於約 2500 條醫療問答對的微調能力與預訓練語言規則，對檢索得到的碎片化文本進行去噪、歸納與對齊，保證內容聚焦於問題核心（如定位至〈脈要精微論〉等關鍵章節，剔除與任務無關的資訊）。

（2）邏輯組織推理：按臨床決策的專業序列將整合後知識串聯為結構化回答——例如針對「失眠如何調理？」先基於病因與嚴重度分層提出生活方式與中醫療法建議，再補充就醫與藥物/心理干預指引，最後提出恢復期管理要點。

同時，模型依提示規則在回答末尾標註知識來源（如具體章節或片段識別符），以增強結果的專業性與可核查性。

4. 系統測試

4.1 系統測試

4.1.1 評估維度與裁定機制

為驗證本針刺決策系統的實際效能，研究團隊圍繞「穴位查詢」與「疾病 / 症狀對應取穴」兩類核心任務設計案例測試。針對系統輸出結果，並與權威知識來源（如百度百科）及主流 LLM（如「豆包」）之回答進行對照分析，本研究採用定性與定量相結合的方法，從正確性、專業性、結構化與可解釋性四個維度進行綜合評估。其中，各評估維度之內涵如下：

（1）正確性：檢驗穴位定位、歸經屬性及主治功能是否符合權威醫學資料與經典文獻記載。

（2）專業性：評估回答是否體現辨證論治思維深度，是否遵循「理—法—方—穴」的中醫決策邏輯，並能合理闡釋穴效關係。

（3）結構化：衡量資訊呈現是否條理清晰、分類明確，是否完整包含「症狀—證型—主穴—配穴」等關鍵決策要素。

（4）可解釋性：審查系統能否明確說明針刺決策的理論依據（如配穴原理、經絡循行關聯與經典文獻來源），而非僅給出結論性建議。

當本系統輸出結果與對照對象出現不一致時，本研究不以通用搜尋結果或其他 LLM 生成內容作為判定依據，而是採用多層裁定機制：

（1）權威文獻一致性判定：以國家統編教材、WHO 標準穴位定位及相關國家標準資料為首要依據，判斷穴位定位、歸經及主治內容是否符合經典與標準記載；

（2）專家評審：邀請 2 名具臨床經驗之針灸專業人員進行獨立評估；若兩者意見一致，則作為最終判定結果；

（3）臨床共識參考：對於存在多種合理取穴方案的情形，依據常見臨床實踐與辨證思路判斷其合理性，而非將差異性簡化為錯誤。

綜上所述，本研究以「權威文獻一致性—專家評審—臨床共識」三層判定機制作為結果評估依據，並以正確性、專業性、結構化與可解釋性四項指標作為核心評估維度。對於系統輸出與其他來源結論不一致之情形，均納入比較與討論範圍，重點分析其理論依據與臨床合理性，以確保評估過程之客觀性、可追溯性與醫學有效性。

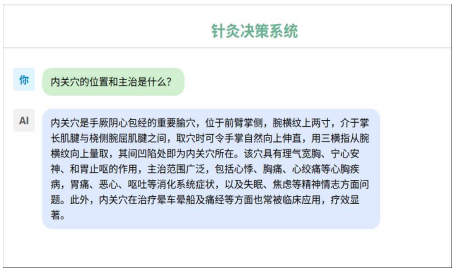

圖 2 案例一：內關穴位置與主治（系統輸出）

圖 3 百度百科對照（內關穴）

4.1.2 案例測試與結果分析

基於上述評估標準，我們進行了以下案例對比分析：

（1）案例一：內關穴的位置與主治

用戶問：內關穴的位置和主治是什麼？

系統回答（見圖 2）：內關穴是手厥陰心包經的重要腧穴，位於前臂掌側，腕橫紋上兩寸，介於掌長肌腱與橈側腕屈肌腱之間，取穴時可令手掌自然向上伸直，用三橫指從腕橫紋向上量取，其間凹陷處即為內關穴所在。該穴具有理氣寬胸、寧心安神、和胃止嘔的作用，主治範圍廣泛，包括心悸、胸痛、心絞痛等心胸疾病，胃痛、噁心、嘔吐等消化系統症狀，以及失眠、焦慮等精神情志方面問題。此外，內關穴在治療暈車暈船及痛經等方面也常被臨床應用，療效顯著。

圖 3 為百度百科「內關」詞條之位置與主治摘要[27]。系統輸出在定位敘述、功效—主治對應與臨床適應證方面均達到與權威來源一致的正確性；且在結構化維度上，系統以高度條理化的段落呈現，相較於網頁詞條冗長的敘述，更便於使用者快速檢索關鍵資訊與操作要點。

（2） 案例二：治療失眠的常用針刺穴位

用戶問：治療失眠的常用針灸穴位有哪些？

系統回答（見圖 4）：治療失眠的常用針灸穴位主要包括百會穴（DU20，位於頭頂正中，可升陽安神）、神門穴（HT7，腕橫紋尺側端，寧心安神）、三陰交（SP6，內踝尖上 3 寸，調理肝脾腎）以及安眠穴（經

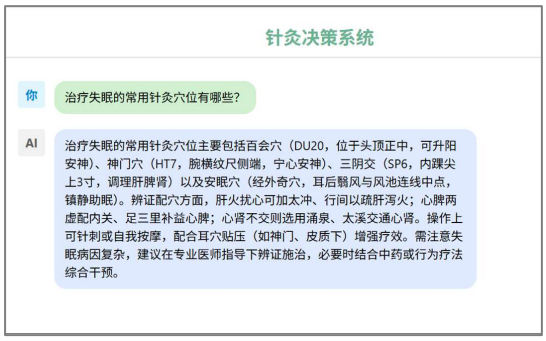

圖 4 案例二：失眠常用穴位（系統輸出）

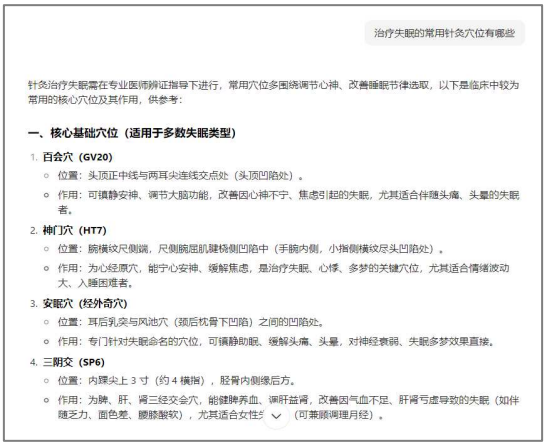

圖 5 豆包輸出：主穴與基本說明

圖 6 豆包輸出：配穴與處置建議

外奇穴，耳後翳風與風池連線中點，鎮靜助眠）。辨證配穴方面，肝火擾心可加太沖、行間以疏肝瀉火；心脾兩虛配內關、足三里補益心脾；心腎不交則選用湧泉、太溪交通心腎。操作上可針刺或自我按摩，配合耳穴壓（如神門、皮質下）增強療效。需注意失眠病因複雜，建議在專業醫師指導下辨證施治，必要時結合中藥或行為療法綜合干預。

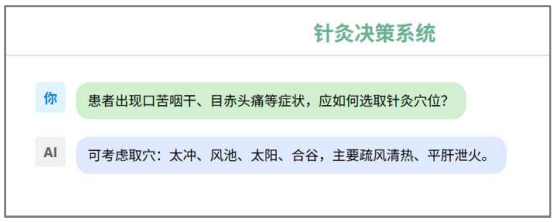

圖 7 未啟用 RAG 的輸出

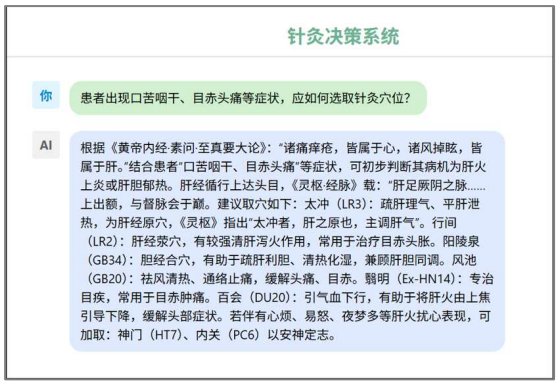

圖 8 啟用 RAG 後的輸出

在相同問題下，本研究選取主流人 AI 工具「豆包」進行提問，以便與同屬 LLM 架構的系統作橫向比較；其生成之回答見圖 5、圖 6 [28]。

從上述結果可以看出，本系統不僅在正確性上明確列出主穴，在專業度與可解釋性上更清晰說明了「穴—效關係」，且內嵌了中醫核心的「辨證配穴」邏輯（如明確指出心火亢盛需加太沖等）。整體而言，系統能提供涵蓋主穴、功效、配穴原則及安全建議的可執行針刺決策，更契合真實臨床與教學情境之需求。

4.2 RAG 強化搜尋測試

在不啟用RAG技術時，系統僅依賴微調後之 LLM。雖可產生看似合理的文字，但常侷限於零散的基礎穴位資訊：缺乏傳統中醫辨證思維的整合（如不同證型之配穴規則）、難以主動調用經典文獻與專業知識庫。圖 7 的輸出即僅為穴位清單，未提供機理解釋與關聯脈絡，難以滿足從業者與學習者對知識深度與決策可解釋性的需求。

將針刺垂直領域知識庫與 RAG 結合後，系統形成「精準檢索—深度整合—專業輸出」的閉環。生成過程中引入《黃帝內經》等經典知識庫進行論據支撐與內容校核。如圖 8 所示，啟用 RAG 後之回答能以可追溯來源支撐關鍵結論，提升可信度；依問題語境動態拼接多段知識，補足辨證與配穴邏輯；隨知識庫持續更新

圖 9 系統介面佈局

而迭代優化輸出深度與廣度。相較「僅微調 LLM」，「RAG + 專業知識庫」顯著緩解資訊碎片化與辨證邏輯缺失之痛點，為針刺從業者與學習者提供更高效、可解釋且可追溯的專業資訊取得途徑。

5. Web 系統實現

本系統前端採用 Vue.js 框架開發，並以元件化與回應式設計為核心理念，確保在不同裝置與螢幕尺寸下皆能提供一致且流暢的互動體驗。介面佈局如圖 9 所示：左側提供對話記錄管理（新建對話、刪除當前對話、清空全部記錄等），右側為互動區域，支援輸入與發送訊息，便於使用者就針刺相關問題與系統進行高效率的溝通與操作。

為緩解可能出現的回應延遲對體驗的影響，前端加入動態回饋機制（如載入提示與過場動效），即時傳達系統狀態，降低等待期間的不確定感與焦慮，並提升整體操作的流暢度與可用性。

在狀態管理方面，前端新增歷史記錄持久化功能，將使用者的對話記錄進行本地儲存；即使重新整理或關閉頁面，既有對話仍可保留與恢復，進一步優化了長時段、跨回合的使用體驗。同時，系統具備前後問題關聯（上下文維持）能力，可分析連續提問之間的語義與邏輯承接，從而更準確地理解使用者意圖，並在後續回答中自然承接上下文，提升互動的一致性、針對性與準確性。

6. 結論與展望

本研究圍繞傳統中醫針刺領域中「穴位決策的可解釋性與智能化」問題，提出一種由 LLM（Large Language Model，大型語言模型）與 RAG（Retrieval-Augmented Generation，檢索增強生成）協同驅動的推薦方法。主要工作包括：以《黃帝內經》等經典文獻構建章節級向量數據庫；採用 LoRA（Low-Rank Adaptation，低秩適配）對 Qwen1.5-0.5B-Chat 進行

領域微調；建構「症狀 — 證型 — 主穴 — 配穴」的結構化推薦流程；引入邏輯推理模組以增強理論依據的可溯源性；並完成前後端 Web 平台部署，形成兼具專業性、交互性與可追溯性的針刺決策輔助系統。

系統測試顯示，模型能穩定輸出涵蓋主穴、配穴與理論依據的結構化建議；在內容專業度與語義完整性方面，整體表現優於通用型大型語言模型。結合 RAG 後，系統可動態調用與驗證經典知識，顯著提升解答的可靠性與臨床適用性，為針刺智慧決策提供了可行路徑。

儘管已取得初步成果，系統仍有待完善之處：其一，當前知識庫以《黃帝內經》為核心，尚未充分整合更多經典文獻與現代臨床案例，知識覆蓋仍需拓展；其二，所用輕量化模型在面對複雜證候、歧義語義與多步推理時仍受限；其三，現階段推理機制以相似度檢索與關鍵詞規則為主，尚不足以支撐符合辨證思維的深層因果推理。

後續研究將致力於：擴展並融合《針灸甲乙經》《針灸大成》等經典與真實病例數據，完善知識體系；探索更高參數量或專門化的語言模型，以強化對中醫語義結構與邏輯鏈條的理解；引入中醫知識圖譜與因果推理技術，構建兼具理論一致性與推理深度的智能推斷鏈，推動系統由輔助問答走向高階推理與個人化診療支持。

總結而言，本研究在技術上拓展了中醫 AI 系統的邊界，亦在方法與實證層面為傳統醫學的現代化轉化與智能化應用提供了可驗證的路徑，具有重要的理論意義與實際價值。

致謝

本研究得到了江蘇省高校哲學社會科學研究一般項目（2023SJYB1829 號）、連雲港人力資源與社會保障局海燕計畫-2023，江蘇海洋大學 2025 年校級本科教育教學改革項目（JGX2025061 號），以及江蘇海洋大學啟動基金（KQ23032 號）的支持，在此謹表謝忱。

文　　獻

[1] D.H.Tian, “Huangdi Neijing”, People’s Medical Publishing House, Beijing (2005) (in Chinese).

[2] J.He, H.Y.Huo, “Lingshu Jing”, China Press of Traditional Chinese Medicine, Beijing (2019) (in Chinese).

[3] Y.H.Du, J.Li, D.W.Sun, et al., “Study on the spectrum of diseases treated by acupuncture in modern China”, Chinese Acupuncture & Moxibustion, vol.27, no.5, pp.373-378 (2007) (in Chinese).

[4] Z.Liu, T.Yang, J.Wang, et al., “Tianyi: A traditional Chinese medicine all-rounder language model and its real-world clinical practice”, Information Fusion (2025).

[5] Y.M.Zhang, H.Y.Li, X.F.Lang, et al., “Construction of a Traditional Chinese Medicine Question-Answering Large Language Model Based on Retrieval-Augmented Generation Technology”, Journal of Nanjing University of Chinese Medicine, vol.40, no.12, pp.1375-1382 (2024) (in Chinese).

[6] S.Y.Qin, Y.F.Wang, T.M.Cui, et al., “Intelligent Chinese patent medicine recommendation framework: Integrating large language models, retrieval-augmented generation, and the largest CPM dataset”, Pharmacological Research, vol.207, 106769 (2025).

[7] Y.Liu, Y.S.Yuan, K.M.Yan, et al., “Evaluating the role of large language models in traditional Chinese medicine diagnosis and treatment recommendations”, NPJ Digital Medicine, vol.8, no.1, pp.1-12 (2025).

[8] H.F.Guo, “Intelligent auxiliary diagnosis and treatment system of traditional Chinese medicine based on large language model and RAG”, China Computer & Communication, vol.37, no.7 (2025) (in Chinese).

[9] T.Yang, X.Y.Wang, Y.Zhu, et al., “Research ideas and methods of intelligent diagnosis and treatment of traditional Chinese medicine driven by large language models”, Journal of Nanjing University of Chinese Medicine, vol.39, no.10, pp.967-971 (2023) (in Chinese).

[10] L.J.Lin, F.Q.Chen, Y.X.Li, “Health medical knowledge question answering system based on large language model RAG framework”, Journal of Ningde Normal University (Natural Science Edition), vol.37, no.2, pp.144-151 (2025) (in Chinese).

[11] J.Z.Yang, “Compendium of Acupuncture and Moxibustion”, People’s Medical Publishing House, Beijing (2006) (in Chinese).

[12] National Standard of PRC, “Acupoint Locations”, China Standards Press, Beijing (2018) (in Chinese).

[13] National Standard of PRC, “Acupoint Names and Locations”, China Standards Press, Beijing (2021) (in Chinese).

[14] J.Guan, J.J.Meng, X.Pu, “Observation on the curative effect of meridian scraping based on the meridian dredging method on stomachache of liver qi invading stomach type”, Xinjiang Journal of Traditional Chinese Medicine, vol.42, no.6, pp.82-84 (2024) (in Chinese).

[15] Z.S.Zhao, S.Yang, Z.T.Yang, et al., “Discussion on the connotation of regulating and harmonizing qi and blood based on the coupling paradigm of zang-fu-xuanfu-yin collaterals”, China Journal of Traditional Chinese Medicine and Pharmacy, vol.40, no.5, pp.2395-2400 (2025) (in Chinese).

[16] T.Wang, “Analysis and comparison of the theories of qi and blood circulation in ancient Eastern and Western medicine”, Chinese Journal of Integrated Traditional and Western Medicine, vol.34, no.9, pp.1035-1041 (2014) (in Chinese).

[17] Y.X.Zeng, Q.Zhao, C.C.Xi, et al., “Technologies and research applications of large language models in the field of traditional Chinese medicine”, Chinese Journal of Experimental Traditional Medical Formulae, vol.31, no.1, pp. 229-239 (2025) (in Chinese).

[18] S.J.Sun, T.H.Ma, K.Huang, “Implementation method of NL2SQL for cultural relics and art auction data based on RAG”, Journal of Computer Applications, vol.45, no.4, pp.1061-1068. (2025) (in Chinese).

[19] C.H.Li, L.P.Zang, H.L.Shi, “Clinical efficacy analysis of

acupuncture and moxibustion in the treatment of acute exacerbation of chronic bronchitis", Psychological Monthly, vol.14, no.3, p.166 (2019) (in Chinese).

[20] Z.D.Li, "Remote loading and implementation of MATLAB functions based on FastAPI", China Computer & Communication (Theory Edition), vol.36, no.3, pp.111-113 (2024) (in Chinese).

[21] Y.Sun, L.X.Zhou, L.J.Sun, et al., "Research on the fusion application of knowledge graph and large language model in aerospace measurement and control question-answering system", Shanghai Aerospace, vol.41, no.5, pp.20-30 (2024) (in Chinese).

[22] X.Wu, T.Fu, "Review on retrieval-augmented generation technology", Computer Engineering and Applications, vol.61, no.2, pp.1-20(2025-09-07）(in Chinese）.

[23] R.S.Chen, "Contemporary value of Huangdi Neijing", Journal of Nanjing University of Chinese Medicine, vol.40, no.10, pp.993-998 (2024) (in Chinese).

[24] X.C.Huang, "Fine-tuning method of large language model based on adaptive quantization", Information Technology and Informatization, no.9, pp.9-12 (2024) (in Chinese).

[25] S.Z.Yang, Y.L.Zhou, X.J.Jie, "Optimization research on medical question answering system integrating RAG retrieval enhancement and LoRA fine-tuning", Jiangxi Science, vol.43, no.3, pp.1-8(2025-06-03）(in Chinese）.

[26] X.L.Han, X.Zeng, K.Liu, et al., "Text stance detection based on LoRA efficient fine-tuning of general language large models", Computer and Modernization, no.1, pp.1-6 (2025) (in Chinese).

[27] Baidu Encyclopedia, "Neiguan (PC6)" (Online), [2024-05-20], https://baike.baidu.com/item/%E5%86%85%E5%85%B3.

[28] Doubao, "Acupuncture for Insomnia" (Online), [2024-09-10], https://www.doubao.com/chat/21145188782254594 .

〈作者略歷〉

劉傳霞

2021 年獲得日本山口大學博士學位。自 2022 年起，在江蘇海洋大學擔任講師。研究興趣包括數據分析與處理、中医药语言学及認知心理學等。

魏璐璐

2023 年至今就讀於江蘇海洋大學本科。其研究興趣包括模型訓練、系統開發等。

瞿汶境

2023 年至今就讀於江蘇海洋大學本科。其研究興趣包括圖像處理、系統前端設計等。

吳順平

2023 年至今就讀於江蘇海洋大學本科。其研究興趣包括演算法設計、系統開發等。

顧　沁

2023 年至今就讀於江蘇海洋大學本科。其研究興趣包括數據集處理、系統前端設計等。

甘　泉

2022 年獲得日本山口大學博士學位。2022 年至 2024 年期間，在山口大學擔任共同研究員。自 2022 年起，在江蘇海洋大學擔任講師。研究興趣包括 Petri 網、系統建模等。