

• 数据挖掘与循证医学 •

# 名医诊治阿尔茨海默病智能体构建方法初探

张玲艳<sup>1</sup>, 倪佩薇<sup>1</sup>, 刘 琴<sup>2</sup>, 张曦丹<sup>1</sup>, 林骏翔<sup>3</sup>, 老膺荣<sup>2\*</sup>, 刘文琛<sup>2\*</sup>

1. 广州中医药大学第二临床医学院, 广东 广州 510405

2. 广州中医药大学第二附属医院(广东省中医院), 广东 广州 510120

3. 华南理工大学电子与信息学院, 广东 广州 510641

**摘要:** **目的** 采用思维链、检索增强生成、提示词等人工智能技术构建基于名医知识的阿尔茨海默病(Alzheimer's disease, AD)诊治智能体, 以应用名医知识辅助临床决策。**方法** 检索名医治疗 AD 的文献、书籍, 选定田金洲、周仲瑛、沈宝藩等 16 位名医, 对其医案、经验和学术理论进行挖掘, 建立名医知识库。通过证素分类解决多位名医辨证方法不统一的问题, 以大语言模型和知识库为核心, 于 Dify 1.9.2 智能体开发平台构建名医诊治 AD 智能体, 智能体包含个体决策方案与群体决策方案。将 108 例真实世界医案的症候群作为输入内容, 智能体输出证型判断与建议方药, 形成测评集。随机抽取其中 30 例采用单盲法进行专家问卷调查, 专家对原医案、智能体个体决策方案、智能体群体决策方案的辨证和处方结果进行合理性打分。**结果** 在专家合理性评分中, 智能体群体决策方案的估算边际平均得分( $3.93 \pm 0.14$ )最高, 高于智能体个体决策方案( $3.47 \pm 0.14$ )和原医案( $3.16 \pm 0.14$ ); 智能体个体决策方案得分亦高于原医案。**结论** 以大模型技术结合证素分类方法, 整合多位名医的辨证经验, 构建了融合群体名医经验的 AD 诊治智能体。专家合理性评分显示群体决策方案得分较高, 提示该方式可发挥多位名医共同决策优势。该智能体的构建为应用名医知识辅助临床决策提供了可复用的方法及路线, 为发挥群体决策优势构建名医诊治专病智能体提供了范例, 为后续前瞻性验证平台奠定基础。

**关键词:** 阿尔茨海默病; 名老中医; 智能体; 检索增强生成; 证素辨证; 群体决策

**中图分类号:** TP18; R285.64 **文献标志码:** A **文章编号:** 0253-2670(2026)14-5601-12

**DOI:** 10.7501/j.issn.0253-2670.2026.14.021

## Preliminary study on construction methods of intelligent agent for famous doctors diagnosis and treatment of Alzheimer's disease

ZHANG Lingyan<sup>1</sup>, NI Peiwei<sup>1</sup>, LIU Qin<sup>2</sup>, ZHANG Xidan<sup>1</sup>, LIN Junxiang<sup>3</sup>, LAO Yingrong<sup>2</sup>, LIU Wenchen<sup>2</sup>

1. The Second Clinical Medical College of Guangzhou University of Chinese Medicine, Guangzhou 510405, China

2. The Second Affiliated Hospital of Guangzhou University of Chinese Medicine (Guangdong Provincial Hospital of Traditional Chinese Medicine), Guangzhou 510120, China

3. School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China

**Abstract:** **Objective** To construct an intelligent agent for the diagnosis and treatment of Alzheimer's disease (AD) based on knowledge of renowned veteran traditional Chinese medicine (TCM) physicians' knowledge by adopting artificial intelligence techniques including chain-of-thought reasoning, retrieval-augmented generation (RAG), and prompt engineering, so as to support clinical decision-making with expert TCM experience. **Methods** Relevant literature and monographs regarding AD treatment by prestigious TCM physicians were retrieved, and a total of 16 renowned veteran TCM physicians, including Tian Jinzhou, Zhou Zhongying, and Shen Baofan, were selected. Their medical records, clinical experience, and academic theories were systematically mined to establish a knowledge base of renowned veteran TCM physicians. To address inconsistencies in syndrome differentiation approaches among different physicians, a syndrome-element classification method was adopted to standardize syndrome differentiation. Based on a large language model and the constructed knowledge base, an intelligent agent for the diagnosis and treatment of AD was developed on the Dify 1.9.2 platform, incorporating both an agent-based individual decision-making scheme and an agent-

收稿日期: 2026-04-10

基金项目: 国家重点研发计划: 中医药典籍智能挖掘与古今融合知识体系构建共性关键技术研究及应用(2023YFC3502900)

作者简介: 张玲艳, 硕士研究生, 从事中医脑病学研究。E-mail: 2537581016@qq.com

\*通信作者: 老膺荣, 中医学博士, 主任中医师, 硕士研究生导师, 从事典籍、五运六气相关研究。E-mail: laoyingrong@qq.com

刘文琛, 中医学博士, 主治医师, 从事中医脑病学、人工智能辅助诊疗研究。E-mail: wenchen20@foxmail.com

based group decision-making scheme. Symptom clusters extracted from 108 real-world medical records were used as inputs to the agent, which generated syndrome differentiation results and recommended prescriptions, thereby creating an evaluation dataset. A total of 30 cases were randomly selected for a single-blind expert questionnaire evaluation. Experts evaluated the rationality of syndrome differentiation and prescription recommendations generated by three approaches: the original medical records, the agent-based individual decision-making scheme, and the agent-based group decision-making scheme. **Results** The agent-based group decision-making scheme achieved the highest estimated marginal mean score ( $3.93 \pm 0.14$ ), outperforming both the agent-based individual decision-making scheme ( $3.47 \pm 0.14$ ) and the original medical records ( $3.16 \pm 0.14$ ). The agent-based individual decision-making scheme also outperformed the original medical records. **Conclusion** By integrating large language model technology with a syndrome-element classification method, this study incorporated the syndrome differentiation experience of multiple renowned veteran TCM physicians and developed an intelligent agent for the diagnosis and treatment of AD that embodies their collective expertise. The agent-based group decision-making scheme achieved higher scores in expert evaluations, suggesting that this approach can leverage the advantages of collaborative decision-making among multiple renowned physicians. The development of this intelligent agent provides a reusable methodological framework and implementation pathway for applying renowned physicians' knowledge to support clinical decision-making, and offers a model for developing disease-specific intelligent agents that leverage the advantages of group decision-making, thereby laying the foundation for a subsequent prospective validation platform.

**Key words:** Alzheimer's disease; renowned veteran traditional Chinese medicine physicians; intelligent agent; retrieval-augmented generation; syndrome-element differentiation; group decision-making

阿尔茨海默病 (Alzheimer's disease, AD) 是一种与年龄相关的、可致严重智力残疾的神经系统退行性疾病, 是全球最常见的痴呆类型, 目前尚无有效的治愈方法<sup>[1-4]</sup>。随着人口老龄化的日益加剧, AD 发病率与死亡率持续上升, 给社会和家庭带来沉重经济负担<sup>[5-6]</sup>。AD 可归属于中医的“痴呆”“呆病”“健忘”“善忘”“愚笨”等范畴, 中医身心同治、标本兼顾、辨证施治的诊疗理念在 AD 防治中提高患者生活质量、延长患者生存期方面有其独特优势<sup>[7-8]</sup>。名老中医 (以下简称名医) 是中医药传承和发展的优秀代表, 在治疗 AD 方面各有所长, 因而挖掘名医群体诊治 AD 的经验, 为 AD 患者提供个性化、有针对性、精准、简便的诊疗方案具有重要意义。然而, 当前名医经验的挖掘大多只针对名医个人的经验, 缺乏对专病领域内名医群体经验的综合运用。群体决策涉及多个决策者, 通过分析不同决策者的客观偏好信息获得一个理想的群体决策结果<sup>[9]</sup>, 能够避免个体决策的局限, 故如何发挥名医群体决策作用, 提升诊疗效果是亟待解决的问题。

随着自然语言处理技术 (natural language processing, NLP)、大语言模型 (large language models, LLM) 的发展, 人工智能 (artificial intelligence, AI) 技术赋能中医药领域已成为大趋势<sup>[10]</sup>。智能体是一种以 LLM 为核心驱动的智能系统, 其利用 LLM 的推理与理解能力来实现任务拆解和策略生成, 具备一定的自主性, 能够与环境交互、独立执行任务、作出决策并持续学习<sup>[11]</sup>。智能

体可以根据用户输入的指令, 自主调用模型、知识库、插件等工具来完成任务。本研究尝试利用提示词技术、思维链 (chain-of-thought, CoT)、检索增强生成 (retrieval-augmented generation, RAG), 构建基于名医知识的 AD 诊治智能体, 为临床决策提供辅助支持, 提高 AD 的诊疗水平。

## 1 资料与方法

为实现用户输入 AD 相关症状后智能体自动输出名医诊疗方案的设计目标, 本研究按以下流程 (图 1) 构建智能体: ①数据收集及清洗; ② workflow、智能体搭建及测试; ③智能体评测。

### 1.1 数据收集、清洗与前期准备

**1.1.1 建立 AD 名医的医案和经验原始数据集** 以“阿尔茨海默”“阿尔茨海默病”“阿尔兹海默病”“痴呆”“痴呆症”“老年痴呆”“认知减退”“认知损害”“智能减退”“失智”“智力减退”等以及“案”“案例”“经验”“思想”“病例”为检索词, 于中国知网 (CNKI)、中国生物医学文献数据库 (CBM)、维普 (VIP)、万方 (Wanfang Data)、中华中医药学会医案成果库、读秀等数据库检索, 检索 2024 年 6 月 30 日之前公开发表的文献, 根据制定的纳排标准, 纳入名医诊治 AD 的医案与经验。

纳入标准: (1) 有明确的 AD 西医诊断; (2) 文献内容含有 AD 相关医案、按语、经验; (3) 医案、经验出自国医大师、全国名中医及痴呆相关领域专家。

排除标准: (1) 无具体有效医案的综述、系统

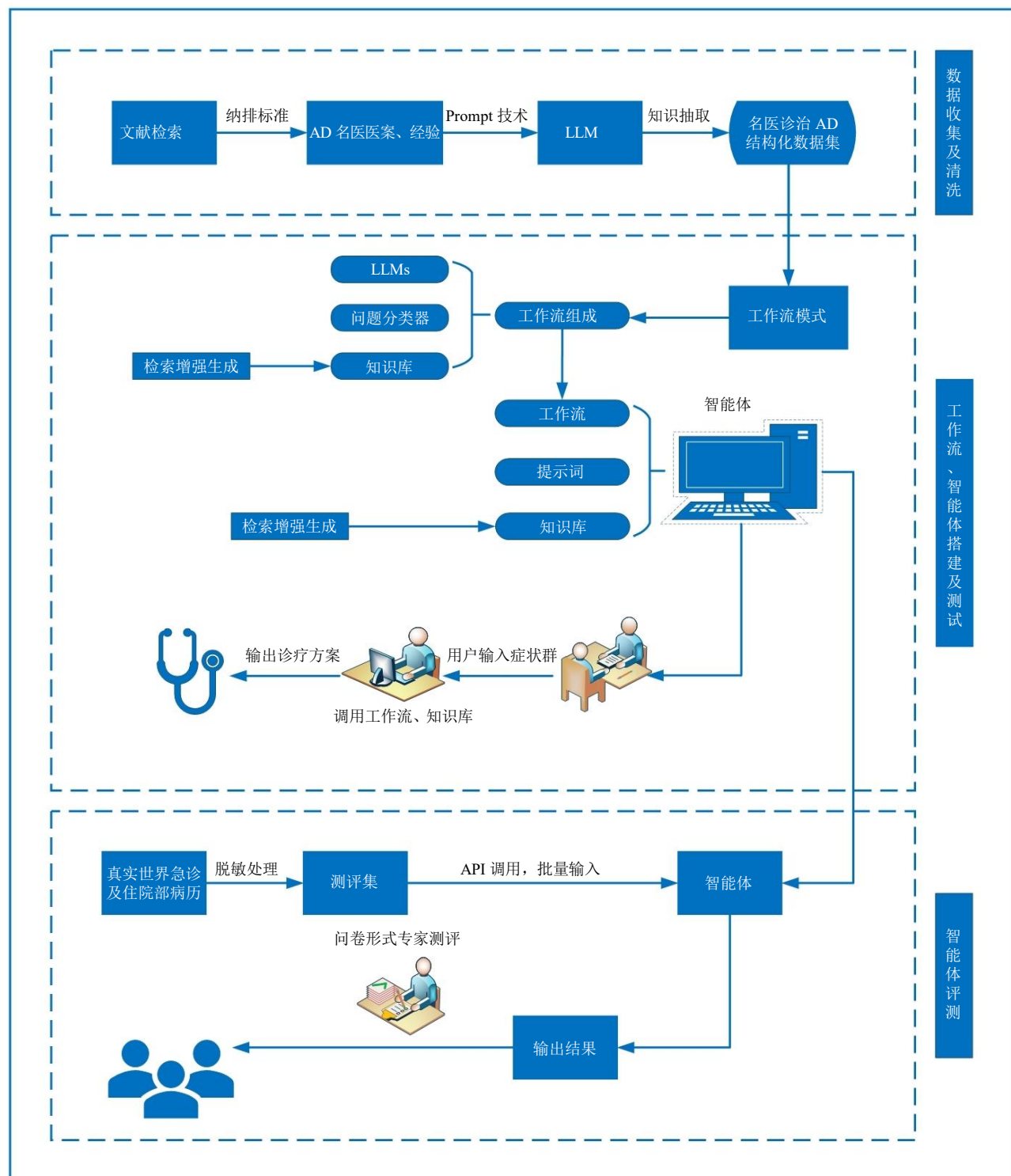


图 1 智能体技术路线图

Fig. 1 Technology roadmap of intelligent agent

评价、Meta 分析等文献；(2) 动物实验、细胞实验、体外实验等非临床类研究文献；(3) 不同篇名但医案内容重复的，纳入信息最为翔实的 1 篇；(4) 医案信息记录不全、有误或前后矛盾者。

根据上述检索策略，于公开发表的文献及书籍内收集到医案 353 例，经验理论 32 条，其中具有明确 AD 诊断的医案 93 例，最终纳入的名医有 16 位，(1) 国医大师：周仲瑛、张琪、郭子光、刘祖贻、沈

宝藩、张志远、王新陆；（2）全国名中医：张发荣、陈宝贵、刘茂才、袁今奇、崔应麟；（3）AD 领域相关专家：田金洲、夏翔、颜乾麟、仝小林。

**1.1.2 数据清洗** 使用大模型提示词，从医家姓名、文献来源、原文、症状与体征、舌脉、病机、

证型、治则治法、方名、组成与剂量等维度对西医诊断明确的原始数据集进行知识抽取，输出为 JSON 格式，人工校对抽取错误之处，共形成规则 70 条，导出至 Excel 形成名医诊治 AD（证型-方药规则）结构化数据集（图 2）。

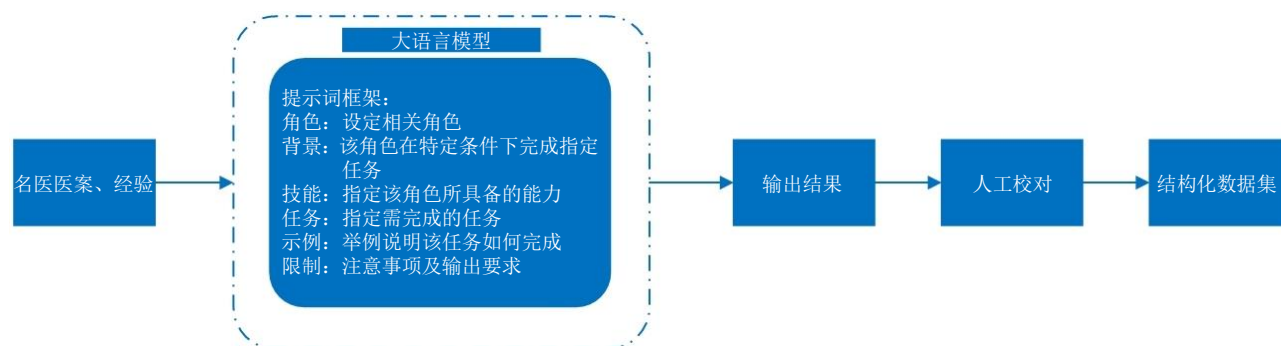


图 2 知识提取流程

Fig. 2 Knowledge extraction process

**1.1.3 基座模型选择** 为选取适合本智能体的基座模型，探索性地对 17 个主流大语言模型（Qwen、DeepSeek、豆包、文心一言、百川、Gemini 等系列）进行了初步测试。以病案为核心设置 5 组提示词，要求各模型输出针对该病案的中医证型及方药建议。由研究团队对各模型输出结果进行评价，采用单盲法，由 3 人独立对证型和方药合理性进行 1~5 分的 Likert 5 级评分（1 分为非常不合理，2 分为不合理，3 分为一般，4 分为合理，5 分为非常合理），汇总评分结果后用 SPSS 25.0 软件进行数据分析。

（1）评分员一致性：采用双向随机、绝对一致性的组内相关系数（intraclass correlation coefficient, ICC）评估 3 位评分员的信度。结果（表 1）显示，证型与方药维度的单个测量 ICC 分别为 0.548 与 0.523（中等一致性），平均测量 ICC 分别为 0.784 与 0.767（良好水平），故后续分析采用 3 位评分员的平均分。

（2）模型间差异及最佳模型表现：以医案为区组进行 Friedman 检验（表 2），证型与方药维度均

表 2 17 个模型在证型与方药维度上的差异比较

Table 2 Comparison of 17 models in dimensions of syndrome type and prescription

| 维度 | 个案数（医案） | $\chi^2$ | 自由度（df） | P 值    |
|----|---------|----------|---------|--------|
| 证型 | 5       | 57.599   | 16      | <0.001 |
| 方药 | 5       | 50.940   | 16      | <0.001 |

显示模型间差异极显著（均  $P < 0.001$ ）。在 17 个模型中，Qwen3-Max-Thinking 20260123 表现最佳，其在证型维度（平均秩 17.00，实际平均分 4.93）和方药维度（平均秩 16.30，实际平均分 3.93）均排名第 1（表 3）。

观察发现，推理模型（如 Qwen3-Max-Thinking、DeepSeek-R1）能够呈现辨证推理步骤，有利于理解模型的中医思维逻辑；但多个推理模型同时测试时，部分模型产生了较长的冗余推理链，可能干扰关键信息的分析；部分推理模型在某些测试案例中出现了与症状无关的“幻觉”内容。结合相关文献报道<sup>[12]</sup>，综合比较中医语境理解能力、输出稳定性及平台兼容性后，Qwen3-Max-Thinking 在本探索性测试中表现相对较优，故选定其作为后续智能体构建的基座模型。

## 1.2 workflow、智能体搭建及测试

**1.2.1 搭建平台选择与设计思路** 利用智能体搭建平台（Dify 1.9.2）搭建智能体，以 Qwen3-Max-Thinking 作为主要推理模型，框架流程采用 LLMs+Workflow 的形式，于 LLMs 内编写设计提示词，运

表 1 评分员信度 (ICC)

Table 1 Inter-rater reliability (ICC)

| 维度 | 单个测量<br>ICC | 95%置信<br>区间    | 平均测量<br>ICC | 95%置信<br>区间    | P 值    |
|----|-------------|----------------|-------------|----------------|--------|
| 证型 | 0.548       | [0.428, 0.659] | 0.784       | [0.691, 0.853] | <0.001 |
| 方药 | 0.523       | [0.399, 0.639] | 0.767       | [0.665, 0.841] | <0.001 |

表 3 各模型在证型与方药合理性评分上的表现

Table 3 Performance of each model in rationality scores of syndrome types and prescriptions

| 模型                                | 证型平均秩 | 证型平均分 | 方药平均秩 | 方药平均分 |
|-----------------------------------|-------|-------|-------|-------|
| Qwen3-Max-Thinking<br>20260123    | 17.00 | 4.93  | 16.30 | 3.93  |
| ERNIE-X1-Turbo-32K-<br>Preview    | 15.40 | 4.27  | 9.40  | 3.27  |
| Qwen3-Max                         | 13.30 | 4.00  | 13.70 | 3.67  |
| MiniMax-M2.1                      | 12.60 | 3.93  | 11.20 | 3.47  |
| DeepSeek-V3                       | 11.30 | 3.73  | 12.40 | 3.53  |
| Kimi-K2-Thinking                  | 10.90 | 3.73  | 11.70 | 3.40  |
| GLM-4.6                           | 9.70  | 3.60  | 10.10 | 3.20  |
| GLM-4.5V                          | 9.50  | 3.53  | 8.40  | 3.13  |
| Baichuan-M3-Plus                  | 8.90  | 3.53  | 3.60  | 2.53  |
| DeepSeek-V3.2                     | 8.10  | 3.47  | 10.60 | 3.33  |
| Gemini 3 Flash                    | 7.60  | 3.40  | 8.70  | 3.13  |
| Doubao-Seed-Character<br>20251128 | 7.30  | 3.40  | 6.90  | 2.93  |
| Seed-OSS-36B-Instruct             | 7.00  | 3.33  | 9.60  | 3.20  |
| DeepSeek-R1                       | 5.80  | 3.20  | 9.10  | 3.20  |
| Ling-flash-2.0                    | 3.70  | 2.87  | 6.40  | 2.87  |
| Hunyuan-A13B-Instruct             | 3.40  | 2.87  | 1.40  | 1.93  |
| Qianfan-IPCharacter               | 1.50  | 2.27  | 3.50  | 2.20  |

表 4 3 级评分标准

Table 4 Three-level scoring criteria

| 分数  | 等级              | 具体描述                                          | 示例（原医案证型为肾虚血瘀证）                                      |
|-----|-----------------|-----------------------------------------------|------------------------------------------------------|
| 3   | 一致              | 输出证型与原医案证型完全一致或表述相同                           | 肾虚血瘀/肾虚兼有血瘀/肾虚夹瘀                                     |
| 1~2 | 部分一致或部<br>分表述相同 | 输出证型与原医案证型部分一致（2 分），或病位与病<br>性要素单一证素部分一致（1 分） | 肾虚血瘀兼脾虚证（2 分）/肾虚血瘀，痰浊蒙窍（2 分）/<br>肾精亏虚（1 分）/肾虚髓减（1 分） |
| 0   | 完全不一致           | 输出证型与原医案证型完全不一致                               | 肝气郁结/风痰阻络，痰热瘀结/痰热瘀火互结，扰动心神                           |

表 5 分数汇总

Table 5 Score summary

| 阶段     | 模式序号                               | 得分  |
|--------|------------------------------------|-----|
| 第 1 阶段 | 模式①：名医结构化数据按名医姓名放<br>置于 1 个 LLM 节点 | 60  |
|        | 模式②：名医结构化数据按相近证型放<br>置于 1 个 LLM 节点 | 11  |
|        | 模式③：名医结构化数据统一放置于 1<br>个 LLM 节点     | 75  |
| 第 2 阶段 | a. 增补名医结构化数据证型缺失信息                 | 122 |
|        | b. 提示词内的名医结构化数据分类放置                | 121 |
|        | c. 提示词中加入证素分类内容                    | 118 |

别按名医姓名放置于独立的 LLM 节点，每位名医的辨证与用药经验直接写入系统提示词中，每个独

行 workflow，将运行稳定的 workflow 与知识库结合搭建智能体。以实现让智能体根据用户输入的症状群，调用名医知识库，输出名医诊疗方案，包括推导步骤、理法方药等信息，体现思维过程，形成名医诊治 AD 思维链。

提示词驱动智能体执行任务，根据智能体预期目标设计大模型提示词，以实现用户输入症状群后可输出名医诊疗方案的功能，提示词内容包括角色、背景、技能、限制、名医结构化数据等模块。提示词的结构与内容直接影响智能体的输出质量。

**1.2.2 预试验** 为系统评估部署方式与提示词策略对智能体输出质量的影响，本研究设计 2 阶段对比实验，分步验证各因素对证型输出准确性的影响。以前期收集的 49 例 AD 中医医案症状及四诊信息作为测试输入，将各模式输出证型与原医案证型进行比较，研究者依据 3 级评分标准（表 4）对输出结果评分并汇总（表 5），尝试对各模式效果进行评价，以此作为 workflow 模式选择依据。

（1）第 1 阶段：名医结构化数据部署方式比较，改变名医结构化数据的部署方式，比较不同部署方式对证型输出结果的影响（图 3）。

模式①：按名医独立部署，16 位名医的经验分

立 LLM 节点根据输入症状输出该名医的证型与方药建议。随后，设置一个“匹配 LLM 节点”，该 LLM 节点接收所有名医的输出结果，结合输入症状进行综合评判，得出最优的诊疗方案。

模式②：按证型聚合部署，将名医经验按相近证型（如肾虚、脾虚、血瘀、火热等）进行归类，每个证型组内多位名医的经验合并写入一个 LLM 节点的提示词中，共形成 6 个证型组 LLM 节点。每个节点根据输入症状输出该组推荐的证型与方药。随后，同样由匹配 LLM 节点对所有证型组 LLM 节点的输出进行综合评判，输出最终方案。

模式③：统一部署，将所有 16 位名医的经验合并写入同一个 LLM 节点的提示词中，该 LLM 直接



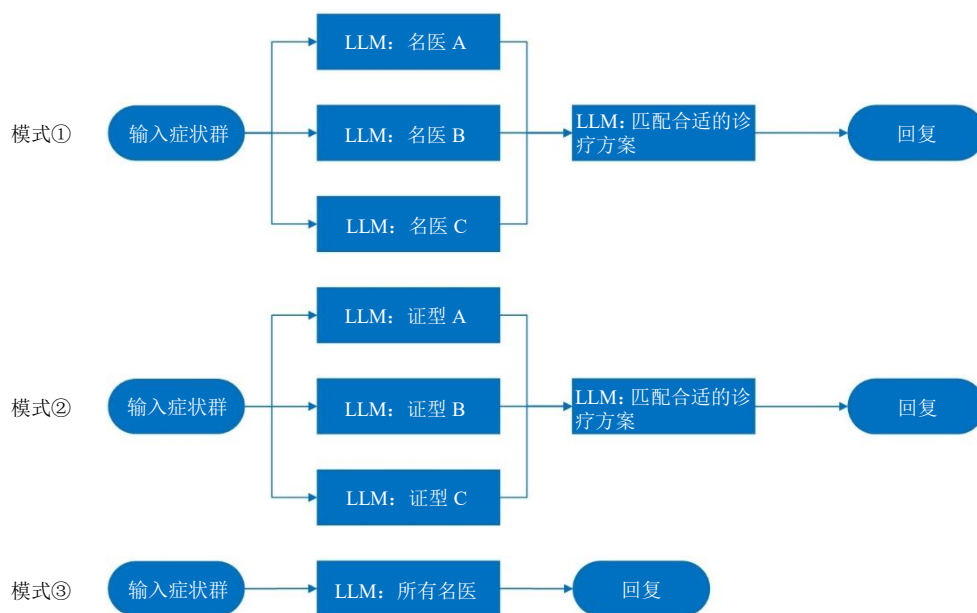


图 3 不同部署方式模式①、②、③流程图

Fig. 3 Flowcharts for different deployment modes ①, ② and ③

根据输入症状判断输出最优的证型与方药，不再设置独立的匹配 LLM 节点。

第 1 阶段对比结果显示，模式③得分（75 分）高于模式①（60 分）与模式②（11 分），提示将所有名医结构化数据统一部署的方式优于按名医或按证型分散部署的方案。故以模式③作为基础模式进入第 2 阶段优化。

（2）第 2 阶段：提示词策略优化。在模式③的基础上保持部署框架不变，使用不同提示词优化策略。每一模式均在前一模式基础上改变单一条件，逐步验证各策略对输出结果的影响。

a. 在模式③的基础上，重新对名医结构化数据进行核查校对，并补齐证型缺失信息，修订后嵌入提示词中。

b. 在 a 的基础上，将嵌入提示词中的名医结构化数据按相近证型进行分类放置。

c. 在 b 的基础上于提示词中加入证候证素分类内容。参考朱文锋所著《证素辨证学》<sup>[13]</sup>及《中医诊断学 新世纪第五版》<sup>[14]</sup>提取名医结构化数据证候的病位和病性证素，进行证候证素分类，引导模型进行证素判断。

第 2 阶段对比结果显示，名医结构化数据的完整性对智能体输出结果影响最为显著，模式③至 a 得分由 75 分升至 122 分，提示名医结构化数据的质量是影响辨证准确性的关键因素。名医结构化数

据的分类放置对输出结果影响有限，a 与 b 得分相差仅 1 分（122 分 vs 121 分）。c 得分较 b 略低（118 分 vs 121 分），但加入证候证素分类后，可使辨证推理过程清晰可追溯，有助于提升结果的可解释性，从而减少“黑箱”效应<sup>[15]</sup>。基于辨证准确性与可解释性，最终选取包含证素辨证分类方法的模式作为智能体基础模式。

1.2.3 智能体结构优化 为进一步优化智能体，在工作流中依次加入 AD 诊断节点、证素分类节点、诊疗方案选择节点，将复杂任务逐步分解为多个子任务，各节点共同构成链式流程<sup>[16]</sup>。在 Dify 1.9.2 平台“知识库”版块创建知识库，采用“关键词全文检索”方式，将 AD 临床表现文本、名医诊治 AD 经验理论文本导入至知识库中，Dify 1.9.2 平台自动生成关键词，人工校对查验关键词后形成 2 个知识库。利用检索增强生成技术，根据实际需求在节点中增加知识检索模块，将 AD 临床表现知识库、名医经验知识库嵌入到相应节点中，将上一节点输出内容在知识库内进行检索并输出，从而使工作流结构更完整、提示词更简明扼要。

用户输入症状群，按工作流顺序执行任务：诊断节点将症状群在 AD 临床表现知识库中进行知识检索，将检索到的结果作为上下文输入模型中，引导模型判断该症状群是否可诊断为 AD，如果是则进入下一节点，否则回复“非 AD”结束对话。证素

判断节点将证素分类知识嵌入提示词中，引导模型推导出该症状群所具备的证素。随后根据所得证素在名医经验知识库进行检索，生成诊疗方案。本智能体可同时输出 2 种决策方案：（1）模式④（个体决策）实现将检索到的知识作为上下文，引导模型进行方案匹配，并给出匹配度高的前 3 个方案，将

这 3 个方案输出到下一个节点进行再次匹配，从 3 个中选择与输入的症状群最为契合的诊疗方案，形成个体决策方案（图 4）。（2）模式⑤（群体决策）调用名医经验知识库，将各医家的经验与理论作为补充知识，通过提示词引导模型综合各家所长，给出 1 个最优的诊疗方案，形成群体决策（图 5）。

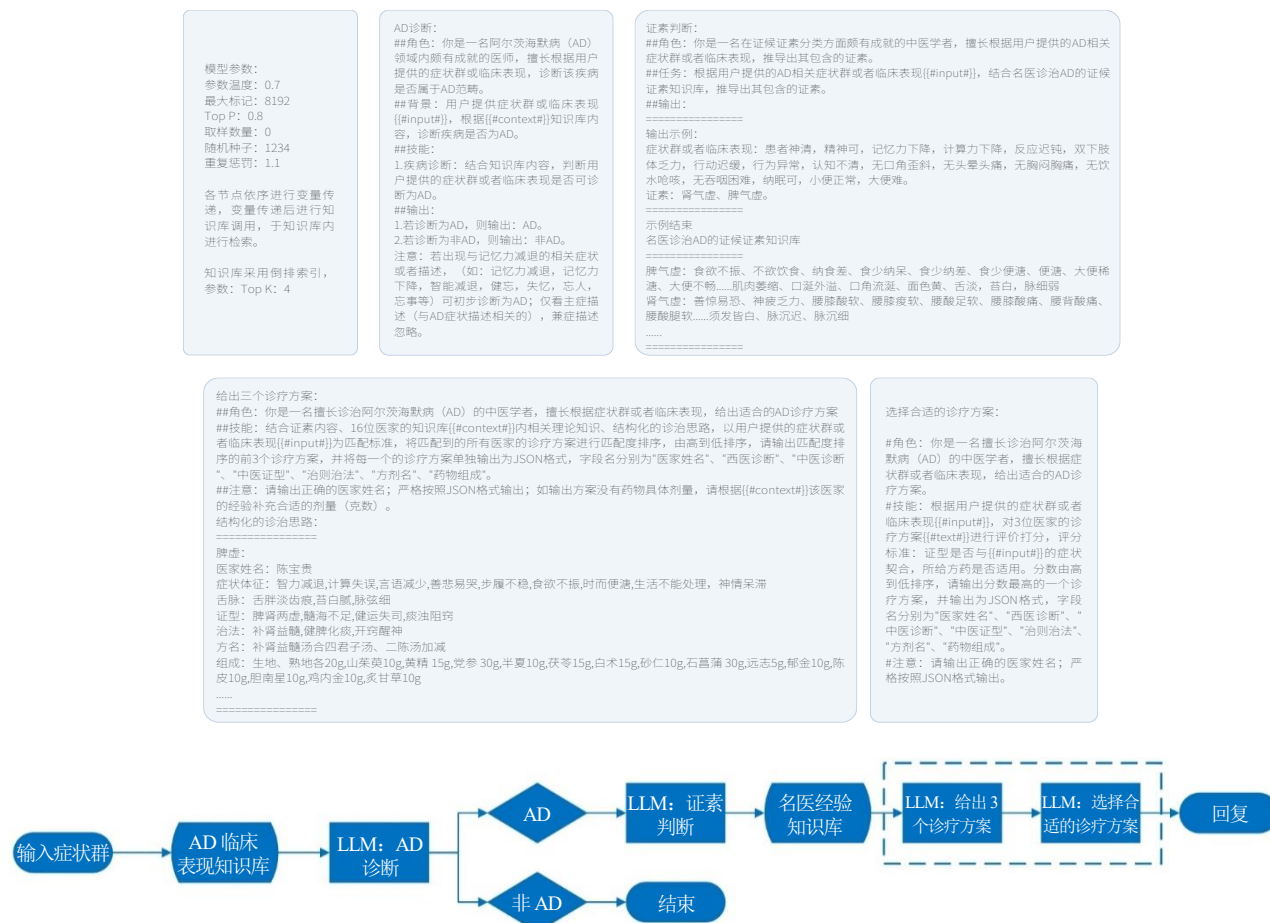


图 4 workflow模式④流程图  
Fig. 4 Flowchart for workflow mode ④

为保证智能体的稳定性和可复现性，本研究对知识库检索、模型参数及各节点提示词进行了统一规范设置。

（1）知识库检索参数：2 个知识库均采用倒排索引（关键词全文检索），检索时以当前节点的输入文本作为查询语句，返回最相关的 4 个知识片段（Top K=4）。

（2）模型参数：所有节点采用相同的模型参数（基座模型为 Qwen3-Max-Thinking），具体为温度=0.7，最大标记=8 192，Top P=0.8，取样数量=0，随机种子=1 234（固定种子以确保结果可复现），

重复惩罚=1.1。

（3）节点提示词框架：各核心节点均采用“角色-背景/任务-技能-输出/ 示例-约束”的结构化模板。以下简述 3 个关键节点的提示词设计。

①诊断节点：定义模型角色为“AD 领域医师”，要求其结合 AD 临床表现知识库的检索结果判断输入症状群是否属于 AD。特别设定规则：“若出现与记忆力减退相关症状（如健忘、智能减退等）可初步诊断为 AD，仅关注主症描述”。输出仅为“AD”或“非 AD”。

②证素判断节点：定义模型角色为“证候证素

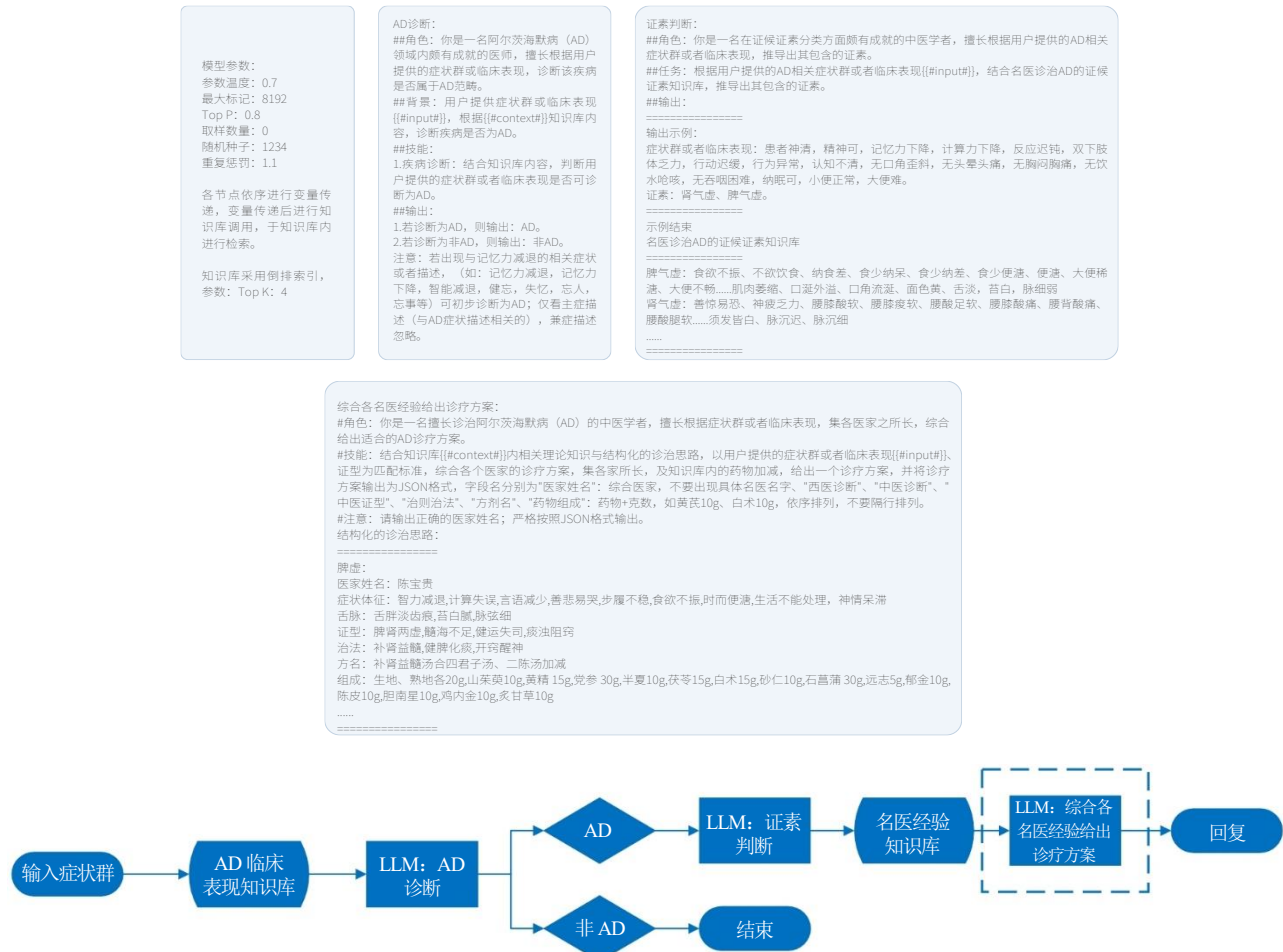


图5 工作流模式⑤流程图

Fig. 5 Flowchart for workflow mode ⑤

分类学者”, 提示词中内嵌了结构化的证素-症状对应知识 (如脾气虚对应的症状列表、肾气虚对应的症状列表等), 并给出输出示例 (如“证素: 肾气虚、脾气虚”), 引导模型进行思维链推理, 输出明确的证素列表。

③诊疗方案节点: 分为群体决策与个体决策 2 种提示词。群体决策提示词要求模型综合各医家经验, 输出 1 个综合方案 (JSON 格式, 字段包括医家姓名固定为“综合医家”、西医诊断、中医诊断、证型、治则治法、方剂名、药物组成)。个体决策提示词则分 2 步: 先让模型匹配出相似度最高的前 3 个医家方案 (JSON 格式), 再由另 1 个模型节点对这 3 个方案进行评分, 选出最优的 1 个输出。所有提示词均强制要求输出规范的 JSON 格式, 并明确约束了字段名称。

(4) 节点衔接方式: 各节点之间通过变量传递方式衔接, 即上一节点的输出文本直接作为下一节

点的部分输入 (如证素判断节点的输出作为诊疗方案节点的输入之一)。

最终, 智能体可根据用户输入的症状群, 自主完成辨证分析与方药推荐, 同时输出个体决策方案与群体决策方案, 为临床诊疗决策提供参考。

### 1.3 智能体评测

为探索性验证智能体辨证处方的水平, 比较其与真实世界医师的差异, 提取广东省中医院 2019 年 1 月—2024 年 10 月第 1 诊断为 AD 的急诊及住院部医案共 108 例。对医案进行脱敏处理后提取其中的症状群 (包含年龄、性别, 刻下症及舌脉象) 作为输入内容。利用 Python 程序调用智能体 API, 批量运行模式④ (个体决策)、模式⑤ (群体决策), 汇总输出结果形成测评集。

从证型和方药 2 个维度出发, 对测评集原医案与智能体输出方案进行专家测评, 采用 Likert 5 分量表法<sup>[17]</sup>设计调查问卷 (表 6)。邀请 10 位中医脑



表 6 Likert 5 分量表  
Table 6 Five-point likert scale

| 维度   | 问题                                                                                                    | 合理性判断         |             |           |              |                |
|------|-------------------------------------------------------------------------------------------------------|---------------|-------------|-----------|--------------|----------------|
|      |                                                                                                       | 非常合理<br>(5 分) | 合理<br>(4 分) | 一般<br>3 分 | 不合理<br>(2 分) | 非常不合理<br>(1 分) |
| 证型维度 | 1. 结论 A/B/C 的辨证证型是否与原始案例的证候表现相符, 对其合理性打分<br>请根据原始案例的证候表现 (主症、兼症、舌象、脉象等), 判断结论 A/B/C 的证型是否与原始案例的证候表现相符 |               |             |           |              |                |
| 方药维度 | 2. 结论 A/B/C 的方药是否适用于原始案例<br>根据原始案例的证候表现, 结论 A/B/C 的方药是否合理                                             |               |             |           |              |                |

病相关领域的临床医生 (1 位主任医师、2 位副主任医师、5 位主治医师、2 位高年资住院医师) 进行问卷咨询。问卷咨询采用单盲法, 为控制专家评阅工作量, 参照相关验证研究的样本量建议<sup>[18-19]</sup>, 从测评集中随机抽取个体决策、群体决策输出结果各 30 份, 将原医案以及个体决策、群体决策的证型和方药<sup>[20]</sup>随机设为结论 A、结论 B、结论 C。专家依据相同的症状群, 对 3 个结论所示证型和方药是否适用于该症状群进行合理性打分。

#### 1.4 数据分析

专家从证型、方药 2 个维度对原医案、智能体个体决策、智能体群体决策 3 种方案进行合理性打分。采用 Excel 软件进行数据录入工作, 导入 SPSS 25.0 软件进行统计分析。对 30 个案例进行均值化处理, 对其进行描述性分析, 一致性检验。因数据存在案例-专家的嵌套与重复评分结构, 故采用线性混合模型对 30 个案例的合理性评分进行分析, 置信区间为 95%, 检验水平  $\alpha=0.05$ , 将合理性评分作为因变量, 以方案 (原医案、智能体个体决策方案、智能体群体决策方案) 和评分维度 (证型、方药) 为固定效应, 案例和专家为随机效应。随机效应部分设定了案例和专家水平的随机截距, 以控制案例间基线评分差异及专家间评分尺度的系统偏移。模型采用受限最大似然法 (restricted maximum likelihood, REML) 进行参数估计, 并通过 -2 倍受限对数似然 (-2 log-likelihood, -2LL)、赤池信息准则 (Akaike information criterion, AIC)、施瓦兹贝叶斯准则 (Bayesian information criterion, BIC) 评估模型拟合优度。针对 3 个方案之间的比较, 采用 Bonferroni 法校正多重比较, 校正后显著性阈值设为  $\alpha'=0.05/3\approx 0.0167$ 。所有检验均为双侧检验,

$P<0.05$  为差异有统计学意义。此外, 为验证线性混合模型结论的稳健性, 以每个案例下各方案的平均评分为分析单元, 采用非参数 Friedman 检验进行敏感性分析。

## 2 结果

共发放并回收院内外专家问卷 10 份。问卷结果显示智能体群体决策方案在证型和方药维度的总分最高, 智能体个体决策方案次之, 二者总分均高于原医案 (表 7)。

表 7 30 个案例各方案总分  
Table 7 Total scores of each scheme for 30 cases

| 方案        | 评分    |       |
|-----------|-------|-------|
|           | 证型维度  | 方药维度  |
| 原医案       | 931   | 965   |
| 智能体个体决策方案 | 1 053 | 1 029 |
| 智能体群体决策方案 | 1 181 | 1 178 |

### 2.1 各方案描述性统计结果

证型与方药维度上, 智能体群体决策方案的专家合理性评分均最高, 智能体个体决策方案次之, 原医案相对最低 (表 8)。

### 2.2 评分一致性结果

肯德尔和谐系数 (Kendall's W) 分析结果如表 9 所示。所有方案在 2 个维度上的 W 值范围为 0.259~

表 8 描述性统计  
Table 8 Descriptive statistics

| 方案        | 证型维度            |      | 方药维度            |      |
|-----------|-----------------|------|-----------------|------|
|           | $\bar{x} \pm s$ | 中位数  | $\bar{x} \pm s$ | 中位数  |
| 原医案       | 3.10±0.34       | 3.00 | 3.22±0.32       | 3.20 |
| 智能体个体决策方案 | 3.51±0.41       | 3.50 | 3.43±0.31       | 3.40 |
| 智能体群体决策方案 | 3.94±0.37       | 4.00 | 3.93±0.27       | 3.90 |

表 9 专家评分肯德尔和谐系数 (Kendall's W) 检验结果  
Table 9 Kendall's W test results for expert ratings

| 方案        | 证型维度  |        | 方药维度  |        |
|-----------|-------|--------|-------|--------|
|           | W 值   | P 值    | W 值   | P 值    |
| 原医案       | 0.386 | <0.001 | 0.473 | <0.001 |
| 智能体个体决策方案 | 0.259 | <0.001 | 0.279 | <0.001 |
| 智能体群体决策方案 | 0.371 | <0.001 | 0.417 | <0.001 |

0.473, 且均具有统计学显著性 ( $P<0.001$ )。其中, 原医案和智能体群体决策方案表现出中等一致性 ( $W=0.371\sim0.473$ ), 而智能体个体决策方案一致性较弱 ( $W=0.259\sim0.279$ )。尽管后者的一致性指标相对较低, 但其  $P<0.001$  仍表明专家对各案例评分排序的共识依然显著。

2.3 线性混合模型分析结果

2.3.1 模型拟合与随机效应 线性混合模型收敛良好, 信息准则显示:  $-2LL=4\ 514.171$ ,  $AIC=4\ 520.171$ ,  $BIC=4\ 536.651$ 。随机效应方差估计显示, 专家水平的随机截距方差为 0.176 ( $P=0.038$ ), 表明不同专家的评分尺度存在显著系统差异, 为中医脑病不同学科背景专家所致; 案例水平的随机截距方差为 0, 提示案例间基线差异可忽略。残差方差为 0.699 ( $P<0.001$ )。上述结果表明模型对数据层次结构的控制基本合理。

表 11 3 个方案在证型与方药维度下的合理性评分 ( $\bar{x} \pm s$ )

Table 11 Rationality scores of three schemes in syndrome type and prescription dimensions ( $\bar{x} \pm s$ )

| 方案/维度     | 合理性评分       |             |             | F 值     | P 值       |
|-----------|-------------|-------------|-------------|---------|-----------|
|           | 证型          | 方药          | 综合评分        |         |           |
| 原医案       | 3.156±0.138 | 3.164±0.138 | 3.160±0.137 | 129.378 | $P<0.001$ |
| 智能体个体决策方案 | 3.466±0.138 | 3.474±0.138 | 3.470±0.137 |         |           |
| 智能体群体决策方案 | 3.928±0.138 | 3.936±0.138 | 3.932±0.137 |         |           |
| 维度        | 3.517±0.135 | 3.524±0.135 | —           | 0.039   | $P=0.844$ |

表 12 3 种方案之间比较

Table 12 Pairwise comparisons among three schemes

| 两方案比较                      | 均值<br>差值 | P 值       |
|----------------------------|----------|-----------|
| 智能体群体决策方案 vs 原医案           | 0.772    | $P<0.001$ |
| 智能体群体决策方案 vs 智能体个体<br>决策方案 | 0.462    | $P<0.001$ |
| 智能体个体决策方案 vs 原医案           | 0.310    | $P<0.001$ |

2.4 敏感性分析: 非参数 Friedman 检验

Friedman 检验结果显示, 在证型维度 ( $\chi^2=34.175$ ,  $df=2$ ,  $P<0.001$ ) 和方药维度 ( $\chi^2=36.643$ ,  $df=2$ ,  $P<0.001$ ) 上, 3 种方案的评分均存在极显

著差异, 秩平均值排序均为智能体群体决策方案>智能体个体决策方案>原医案 (表 13)。该结果与线性混合模型的结论完全一致, 表明本研究的核心发现不依赖于正态分布假设, 具有较高的稳健性。

上述结果提示在本研究设定的评价标准下, 智能体输出方案与专家认知的契合度更高。

表 10 不同方案与评估维度对合理性评分影响的线性混合模型分析

Table 10 Linear mixed model analysis of effects of different schemes and evaluation dimensions on rationality scores

| 变异源 | 分子自由度<br>(df1) | 分母自由度<br>(df2) | F 值     | P 值       |
|-----|----------------|----------------|---------|-----------|
| 截距  | 1              | 9.000          | 690.783 | $P<0.001$ |
| 方案  | 2              | 1 787.000      | 129.378 | $P<0.001$ |
| 维度  | 1              | 1 787.000      | 0.039   | $P=0.844$ |

著差异, 秩平均值排序均为智能体群体决策方案>智能体个体决策方案>原医案 (表 13)。该结果与线性混合模型的结论完全一致, 表明本研究的核心发现不依赖于正态分布假设, 具有较高的稳健性。

上述结果提示在本研究设定的评价标准下, 智能体输出方案与专家认知的契合度更高。

3 讨论

本研究采用大模型思维链、检索增强生成技术, 利用名医经验知识库, 构建了“CoT 任务分解-RAG 补充知识-证素框架统一辨证”的协同智能体, 并引入群体决策理论<sup>[21]</sup>模拟多专家诊治阿尔茨海默病的决策过程。评测结果提示, 智能体群体决

表 13 3 种方案证型与方药合理性评分的 Friedman 检验结果

Table 13 Friedman test results for syndrome type and prescription rationality scores under three schemes

| 维度 | 方案        | 秩平均值 | $\chi^2$ | 自由度 (df) | P 值    |
|----|-----------|------|----------|----------|--------|
| 证型 | 原医案       | 1.23 | 34.175   | 2        | <0.001 |
|    | 智能体个体决策方案 | 2.07 |          |          |        |
|    | 智能体群体决策方案 | 2.70 |          |          |        |
| 方药 | 原医案       | 1.37 | 36.643   | 2        | <0.001 |
|    | 智能体个体决策方案 | 1.78 |          |          |        |
|    | 智能体群体决策方案 | 2.85 |          |          |        |

策方案在辨证与处方方面的合理性得分均高于原始医案及个体决策方案，提示群体决策方式在专家评价中具有一定优势，表明该方式可发挥多位名医共同决策优势。群体决策方案优势在于通过检索增强生成技术，使智能体检索并补充名医知识，再运用提示工程综合各医家经验给出诊疗方案。此过程模拟了多专家会诊，有助于规避单一经验可能存在的局限性与偏倚。

本智能体通过对 16 位医家的医案、经验结构化数据进行证素提取，将其放置于工作流中的证候证素分类节点，让大语言模型以证素为核心进行辨证，为不同名医的经验提供了统一的、机器可理解的表现形式。“证素”是按照病位、病性将证候（症状、体征等）分解成的若干要素，“证素辨证体系”根据证候，辨别证素，组成证名，形成“三阶双网”结构，有助于解决各医家、各流派的多种辨证方法错杂重复的问题<sup>[13,15,22]</sup>。多专家多种辨证方法的整合使智能体能够跨越流派限制，降低不同辨证体系之间的差异性，为实现真正的“群体智能”奠定了基础。同时证素辨证结合人工智能技术，也为中医证素辨证迈向智能化发展提供了新的思路和方法<sup>[23]</sup>。

本智能体通过 CoT、RAG 与证素辨证协同作用，形成了从症状输入到诊疗方案输出的可解释、可验证的思维推理链条。思维链技术可使大语言模型模拟人类推理过程，将辨证处方的过程逐步拆解，外显思维过程，能显著提升大语言模型的复杂推理能力，引导模型将复杂任务拆解为多个子任务，逐步逻辑推理完成目标任务<sup>[24-25]</sup>。RAG<sup>[26]</sup>结合大语言模型的外部知识库，能有效减少大语言模型的幻觉，从而增强模型的语义理解和生成能力，提高模型回答的可解释性<sup>[27-28]</sup>。在工作流中嵌入知识库，以关键词检索知识库全文并进行生成<sup>[29]</sup>，以此减少大语言模型幻觉，提升回答的准确度。本智能

体分别在 AD 诊断、方案匹配 2 个步骤中加入知识库，依序进行知识检索，实现“诊断→证素→方案”分步骤推理，让中间推理过程可视化，使输出结果可溯源。相较于单个大语言模型的单次推理，智能体通过链式结构设计进行任务分解，组合大语言模型与知识库等工具构建工作流，从 AD 的诊断、证素分类、方案匹配、最后输出名医诊疗方案，各环节环环相扣，使推理过程更具透明性、逻辑性、可解释性<sup>[30]</sup>，可为临床工作者提供具有潜在应用价值的个体化中医药临床辅助决策系统（clinical decision support system, CDSS）<sup>[31-33]</sup>，为之提供临床决策参考，从而辅助临床决策。

本研究仍存在一定的局限性，由于大语言模型的普适性问题，其本身可能出现的知识偏倚会影响输出结果，致使智能体输出的诊疗方案与实际临床应用存在一定偏差；公开发表的 AD 名医医案及经验数据量相对较少，导致在知识检索时原始数据不足，从而影响输出的诊疗方案质量；本研究评测规模较小，后续应扩大样本量及专家人数。未来智能体可以基于疗效的反馈进行调整优化，模拟真实的临床情境，使诊治方案更适用于临床实际<sup>[34]</sup>。此外，需要明确的是，本智能体当前仅作为临床辅助决策工具，其输出方案不能替代临床医师的诊断和处方决策。临床医师应结合患者具体情况，审慎参考智能体提供的建议。后续研究需进一步开展更大样本量、多中心、前瞻性的临床验证，以评估本智能体在实际临床应用中的有效性和安全性。可以预期，随着更多高质量数据的不断注入与相关人工智能技术的日益成熟，智能体将不断迭代优化，更好地发挥名医的群体智慧，助力临床实践。

志谢：广东省中医院温泽淮与孙景波研究员，冯梅副主任医师对本研究的指导与支持！

利益冲突 所有作者均声明不存在利益冲突

## 参考文献

- [1] Scheltens P, De Strooper B, Kivipelto M, *et al.* Alzheimer's disease [J]. *Lancet*, 2021, 397(10284): 1577-1590.
- [2] Grøntvedt G R, Schröder T N, Sando S B, *et al.* Alzheimer's disease [J]. *Curr Biol*, 2018, 28(11): R645-R649.
- [3] 林璐, 马辛, 王刚, 等. 中国阿尔茨海默病早期预防指南 (2024) [J]. 阿尔茨海默病及相关病, 2024, 7(3): 168-175.
- [4] 高雁, 渠硕, 李漾, 等. 《痴呆的行为和精神症状的评估和管理临床实践指南(2025)》解读 [J]. 临床荟萃, 2025, 40(5): 389-393.
- [5] 王刚, 齐金蕾, 刘馨雅, 等. 中国阿尔茨海默病报告 2024 [J]. 诊断学理论与实践, 2024, 23(3): 219-256.
- [6] Nichols E, Szeoke C E I, Vollset S E, *et al.* Global, regional, and national burden of Alzheimer's disease and other dementias, 1990—2016: A systematic analysis for the Global Burden of Disease Study 2016 [J]. *Lancet Neurol*, 2019, 18(1): 88-106.
- [7] 路漫漫, 谷峰, 刘伟, 等. 基于《黄帝内经》“五脏藏神”理论探析阿尔茨海默病与五脏的关系 [J]. 辽宁中医杂志, (2024-09-04) [2026-01-06]. <https://link.cnki.net/urlid/21.1128.R.20240904.0909.004>.
- [8] 金博文, 王豫骞, 樊薛津, 等. 基于 Preferential Attachment-VIKOR 算法的阿尔茨海默病中药处方“方-证”关系预测与处方综合评价 [J]. 中草药, 2025, 56(22): 8273-8282.
- [9] 刘亚婷, 徐维军, 陈霞, 等. 群体决策中基于最小成本的个体偏好策略操纵模型研究 [J]. 控制与决策, 2026, 41(2): 445-454.
- [10] 曾玉秀, 赵琼, 席崇程, 等. 大语言模型在中医领域使用的技术及研究应用 [J]. 中国实验方剂学杂志, 2026, 32(14): 50-59.
- [11] 陈亚楠, 姜姗, 李瑞锋, 等. 大模型智能体在中医药教育应用中的 SWOT 分析 [J]. 中医教育, 2026, 45(2): 7-12.
- [12] Kong S F, Yang X R, Wei Y Y, *et al.* MTCMB: A multi-task benchmark framework for evaluating LLMs on knowledge, reasoning, and safety in traditional Chinese medicine [J]. *ArXiv*, 2025, doi: 10.48550/arXiv.2506.01252.
- [13] 朱文锋. 证素辨证学 [M]. 北京: 人民卫生出版社, 2008: 91-290.
- [14] 李灿东, 方朝义. 中医诊断学 [M]. 第 5 版. 北京: 中国中医药出版社, 2021: 138-187.
- [15] 朱文锋. 创立以证素为核心的辨证新体系 [J]. 湖南中医学院学报, 2004(6): 38-39.
- [16] 王阶, 宋逸杰, 惠小珊, 等. 链式思维驱动的冠心病中医证候要素辨证大语言模型 [J]. 中国实验方剂学杂志, 2025, 31(19): 1-10.
- [17] 何怡潇, 陈敦金, 胡淼, 等. 大语言模型对胎盘植入性疾病患者咨询响应的准确性与全面性研究 [J]. 中华医学杂志, 2025, 105(40): 3650-3656.
- [18] Liu W W, Miao S C, Ma Q, *et al.* Digitally assisted clinical decision-making in traditional Chinese medicine: Comparative study of 5 large language models [J]. *JMIR Form Res*, 2026, 10: e80167.
- [19] Li K N, Guo J B, Shang Z Y, *et al.* A benchmark dataset for evaluating syndrome differentiation and treatment in large language models [J]. *ArXiv*, 2025, doi: 10.48550/arXiv.2512.02816.
- [20] Wang B F, Lu Y W, Wang Z, *et al.* TCMEval-PA: A question-answering benchmark dataset for the prescription audit of traditional Chinese medicine [J]. *Sci Data*, 2026, 13: 79.
- [21] 刘敏, 张露祥, 平卫英, 等. 基于大语言模型的多阶段网络舆情驱动群体共识决策方法研究 [J]. 管理学报, 2025, 22(4): 750-759.
- [22] 朱文锋, 张华敏. “证素”的基本特征 [J]. 中国中医基础医学杂志, 2005, 11(1): 17-18.
- [23] 孙肇阳, 汪洋, 马铭泽, 等. 基于大语言模型与文本嵌入计算的中医证素辨证自动化方法研究 [J]. 北京中医药大学学报, 2025, 48(8): 1176-1184.
- [24] 燕天. 基于大语言模型的思维链技术应用展望 [J]. 百科知识, 2025(33): 12-13.
- [25] Wei J, Wang X Z, Schuurmans D, *et al.* Chain-of-thought prompting elicits reasoning in large language models [J]. *ArXiv*, 2023, doi: 10.48550/arXiv.2201.11903.
- [26] Lewis P, Perez E, Piktus A, *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks [J]. *ArXiv*, 2021, doi: 10.48550/arXiv.2005.11401.
- [27] 吴璇, 付涛. 检索增强生成技术研究综述 [J]. 计算机工程与应用, 2025, 61(20): 19-35.
- [28] 王星臣, 田夫蓉, 孙琼巍, 等. 检索增强生成的概念、核心技术与面临的挑战 [J]. 计算机应用, 2025, 45(S2): 7-13.
- [29] 吴恙, 于彤, 于琦, 等. 基于检索增强生成技术的脾胃病大语言模型问答研究 [J]. 世界中医药, 2025, 20(19): 3516-3523.
- [30] 张鸿波, 李京. 大模型仿人思维链优化的主观题评分方法 [J]. 小型微型计算机系统, 2025.
- [31] 郑婉婷, 李敬华, 田少磊, 等. 中医临床辅助决策系统功能与算法的发展评述 [J]. 中国数字医学, 2022, 17(2): 86-94.
- [32] 庞晓燕, 尹思艺, 蔡秀军, 等. 构建基于大数据的人工智能临床辅助决策系统方法与效果研究 [J]. 中国数字医学, 2020, 15(9): 49-52.
- [33] 彭颖, 杨超然, 于广军. 基于人工智能的临床辅助决策支持系统问责制度探索 [J]. 海军军医大学学报, 2025, 46(8): 963-969.
- [34] 倪敬年, 曹天雨, 陈雅静, 等. 名老中医经验数字化传承的思考 [J]. 中医杂志, 2023, 64(17): 1754-1758.

[责任编辑 潘明佳]