

• 论著 •

基于首次住院病程资料的生成式人工智能 与医师制定鉴别诊断及诊疗计划的比较

杨万杰¹ 傅浩² 孟祥飞² 倪广健³ 李长松² 庞晓磊⁴ 万振⁵ 李莉⁶ 崔超¹ 刘军¹
刘金婷⁷ 范爽⁸ 韩东⁹ 史丽霞⁶ 韩晶¹⁰ 杨晔¹¹ 李雪¹² 赵卫康⁵ 林宝禄⁷ 邢维荣¹³
张俊全¹³

¹ 天津大学海河医院院办, 天津 300350; ² 国家超级计算天津中心, 天津 300457; ³ 天津大学医学工程与转化医学研究院, 天津 300072; ⁴ 天津智临天河科技有限公司, 天津 300192; ⁵ 天津大学海河医院信息科, 天津 300350; ⁶ 天津大学海河医院呼吸与危重症医学科, 天津 300350; ⁷ 天津市医疗服务评价和指导中心, 天津 300131; ⁸ 天津大学海河医院中医科, 天津 300350; ⁹ 天津大学海河医院胸外科, 天津 300350; ¹⁰ 天津大学海河医院医务科, 天津 300350; ¹¹ 天津市云山道学校, 天津 300450; ¹² 天津大学海河医院基础医学实验部, 天津 300350; ¹³ 天津大学海河医院病案科, 天津 300350

通信作者: 杨万杰, Email: yangwanjie0709@126.com

【摘要】 目的 比较生成式人工智能(GenAI)与专科西医医师和专科中医医师在住院患者首次病程记录阶段疾病诊断及诊疗计划制定中的质量差异。**方法** 采用基于同一病历资料的双路径对照研究方法。选择 2025 年 4 月至 11 月天津市海河医院呼吸与危重症医学科收治的住院患者, 所有入选患者均在入院后由专科医师按照规范流程完成问诊、体格检查及相关辅助检查, 形成病历文本资料。按照诊疗方案制定方式不同分为医师组(分为西医医师组和中医医师组)和 GenAI 组(分为西医 GenAI 组和中医 GenAI 组), 其中医师组由专科医师独立完成诊疗方案制定, GenAI 组由研究人员调用本地化定制的 DeepSeek 模型服务生成诊疗建议。鉴别诊断评价指标包括诊断冗余度、诊断遗漏度、诊断优先级和诊断适配度; 诊疗计划评价指标包括方案完整度和方案精准度。由西医或中医专家采用 Likert 5 点量表分别对医师组与 GenAI 组的鉴别诊断和诊疗计划进行评分。**结果** 共纳入 96 例病例资料, 其中肺部感染患者 28 例, 罕见病患者 9 例。① 鉴别诊断质量: 在诊断遗漏度方面, 西医 GenAI 组评分高于西医医师组[分: 4.00(4.00, 4.50)比 4.00(3.75, 4.25), $P<0.05$], 中医 GenAI 组评分高于中医医师组[分: 5.00(4.00, 5.00)比 3.00(2.00, 3.00), $P<0.05$]; 在诊断适配度方面, 西医 GenAI 组评分高于西医医师组[分: 4.25(4.00, 4.50)比 4.00(4.00, 4.25), $P<0.05$], 但中医 GenAI 组与中医医师组评分差异无统计学意义; ③ 在诊断冗余度和诊断优先级方面, 西医 GenAI 组与西医医师组之间以及中医 GenAI 组与中医医师组之间的评分差异均无统计学意义。② 诊疗计划质量: 在方案完整度方面, 西医 GenAI 组评分高于西医医师组[分: 4.25(4.00, 4.50)比 4.00(3.75, 4.25), $P<0.05$], 中医 GenAI 组评分高于中医医师组[分: 5.00(4.00, 5.00)比 3.00(2.00, 3.00), 均 $P<0.05$]; 在方案精准度方面, 西医 GenAI 组与西医医师组之间以及中医 GenAI 组与中医医师组之间的评分差异均无统计学意义。**结论** 在住院患者首次病程记录阶段, GenAI 在西医与中医的诊断遗漏度和诊疗方案完整度方面, 以及在西医的诊断适配度方面可能具有潜在优势, 可作为专科医师诊疗决策的辅助工具。

【关键词】 鉴别诊断; 诊疗计划; 生成式人工智能; 深度求索; 中西医结合

基金项目: 天津市科技重大专项与工程项目(24ZXKJGX00070); 天津市科技计划项目(23JCZDJC00970)

DOI: 10.3760/cma.j.cn121430-20260313-00118

Comparison of generative artificial intelligence and physicians in differential diagnosis and treatment planning based on initial hospital medical records

Yang Wanjie¹, Fu Hao², Meng Xiangfei², Ni Guangjian³, Li Changsong², Pang Xiaolei⁴, Wan Zhen⁵, Li Li⁶, Cui Chao¹, Liu Jun¹, Liu Jintong⁷, Fan Shuang⁸, Han Dong⁹, Shi Lixia⁶, Han Jing¹⁰, Yang Mou¹¹, Li Xue¹², Zhao Weikang⁵, Lin Baolu⁷, Xing Weirong¹³, Zhang Junquan¹³

¹Hospital Administration Office, Haihe Hospital, Tianjin University, Tianjin 300350, China; ²National Super Computer Center in Tianjin, Tianjin 300457, China; ³Academy of Medical Engineering and Translational Medicine, School of Medicine, Tianjin University, Tianjin 300072, China; ⁴Tianjin Zhilin Tianhe Technology Co., Ltd., Tianjin 300192, China; ⁵Department of Information, Haihe Hospital, Tianjin University, Tianjin 300350, China; ⁶Department of Respiratory and Critical Care Medicine, Haihe Hospital, Tianjin University, Tianjin 300350, China; ⁷Tianjin Medical Service Evaluation and Guidance Center, Tianjin 300131, China; ⁸Department of Traditional Chinese Medicine, Haihe Hospital, Tianjin University, Tianjin 300350, China; ⁹Department of Thoracic Surgery, Haihe Hospital, Tianjin University,

Tianjin 300350, China; ¹⁰Department of Medical Administration, Haihe Hospital, Tianjin University, Tianjin 300350, China; ¹¹Tianjin Yunshan Road School, Tianjin 300450, China; ¹²Department of Basic Medicine, Haihe Hospital, Tianjin University, Tianjin 300350, China; ¹³Department of Medical Records, Haihe Hospital, Tianjin University, Tianjin 300350, China

Corresponding author: Yang Wanjie, Email: yangwanjie0709@126.com

【Abstract】 Objective To compare the quality of differential diagnoses and diagnostic and treatment plans formulated by generative artificial intelligence (GenAI) and by specialist physicians practicing western medicine or traditional Chinese medicine at the stage of preparing the initial progress note for hospitalized patients. **Methods** A dual-path controlled comparative study was conducted using the same medical record data. Hospitalized patients admitted to the Department of Respiratory and Critical Care Medicine at Haihe Hospital, Tianjin University, from April to November 2025 were selected. After admission, standardized history taking, physical examination, and relevant ancillary examinations were performed by specialist physicians, and textual medical record data were compiled. According to the method used to formulate the diagnostic and treatment plans, two groups were established: a physician group, comprising a western medicine physician subgroup and a traditional Chinese medicine physician subgroup and a GenAI group, comprising a western medicine GenAI subgroup and a traditional Chinese medicine GenAI subgroup. In the physician group, diagnostic and treatment plans were independently formulated by specialist physicians. In the GenAI group, diagnostic and treatment recommendations were generated by investigators using a locally deployed, customized DeepSeek model service. Evaluation indices for differential diagnosis included diagnostic redundancy, diagnostic omission, diagnostic prioritization, and diagnostic appropriateness. Evaluation indices for diagnostic and treatment plans included plan completeness and plan precision. The differential diagnoses and diagnostic and treatment plans generated by the physician and GenAI groups were separately rated by experts in western medicine or traditional Chinese medicine using a 5-point scale. **Results** A total of 96 medical records were included, including records from 28 patients with pulmonary infections and 9 patients with rare diseases. (1) Quality of differential diagnosis: For diagnostic omission, scores were higher in the western medicine GenAI subgroup than in the western medicine physician subgroup [score: 4.00 (4.00, 4.50) vs. 4.00 (3.75, 4.25), $P<0.05$], and scores were higher in the traditional Chinese medicine GenAI subgroup than in the traditional Chinese medicine physician subgroup [score: 5.00 (4.00, 5.00) vs. 3.00 (2.00, 3.00), $P<0.05$]. For diagnostic appropriateness, scores were higher in the western medicine GenAI subgroup than in the western medicine physician subgroup [score: 4.25 (4.00, 4.50) vs. 4.00 (4.00, 4.25), $P<0.05$], whereas no statistically significant difference was found between the traditional Chinese medicine GenAI and traditional Chinese medicine physician subgroups. For diagnostic redundancy and diagnostic prioritization, no statistically significant differences were found between the corresponding GenAI and physician subgroups in either western medicine or traditional Chinese medicine. (2) Quality of diagnostic and treatment plans: For plan completeness, scores were higher in the western medicine GenAI subgroup scored higher than in the western medicine physician subgroup [score: 4.25 (4.00, 4.50) vs. 4.00 (3.75, 4.25)], and were also higher in the traditional Chinese medicine GenAI subgroup than in the traditional Chinese medicine physician subgroup [score: 5.00 (4.00, 5.00) vs. 3.00 (2.00, 3.00); both $P<0.05$]. For plan precision, no statistically significant differences were found between the corresponding GenAI and physician subgroups in either western medicine or traditional Chinese medicine. **Conclusions** At the stage of preparing the initial progress note for hospitalized patients, GenAI may offer potential advantages in reducing diagnostic omissions and traditional Chinese medicine, as well as in improving diagnostic appropriateness in western medicine. GenAI may serve as an adjunctive tool to support clinical decision-making by specialist physicians.

【Key words】 Differential diagnosis; Diagnostic and treatment plan; Generative artificial intelligence; DeepSeek; Integrated traditional Chinese and Western medicine

Fund program: Tianjin Major Science and Technology Projects and Engineering Projects (24ZXKJGX00070); Tianjin Natural Science Foundation Project (23JCZDJC00970)

DOI: 10.3760/cma.j.cn121430-20260313-00118

疾病诊断与治疗是临床医疗活动中最核心、最基础的环节,其准确性和及时性直接关系到患者预后与医疗安全。任何诊断偏差或决策延误均可能对患者造成不同程度的不良影响,甚至引发严重医疗事件。因此,该过程高度依赖医务人员长期积累的专业知识、临床经验及综合判断能力。然而,当前临床实践中仍普遍存在医师培养周期长、诊疗经验差异较大及诊疗同质化水平不均等现实问题。生成式人工智能(generative artificial intelligence, GenAI)及大语言模型的快速发展,为优化临床诊疗流程,

提升传统诊疗的质量、效率和安全提供了新的技术路径^[1]。天津大学海河医院在“全流程疾病诊疗智能辅助系统”研发基础上,完成了 DeepSeek 模型的本地化定制部署,并将其作为临床决策支持系统(clinical decision support system, CDSS)的重要组成部分嵌入现有诊疗流程^[2]。本研究采用基于同一病历资料的双路径对照设计,作为探索性分析,旨在初步比较 GenAI 与医师在首次病程记录中生成诊疗文本的质量差异,为后续设计严格的验证性研究提供基础数据。

1 资料与方法

1.1 病例资料来源:选择 2025 年 4 月至 11 月天津市海河医院呼吸与危重症医学科收治的住院患者作为研究对象。所有入选患者均在入院后由专科医师按照规范流程完成问诊、体格检查及相关辅助检查,形成病历文本资料。样本量基于研究周期内符合纳排标准的连续入院病例确定。

1.1.1 纳入标准:① 年龄 ≥ 18 周岁;② 具备完整的一般资料、病史、体格检查和辅助检查等病历资料;③ 入院时尚未明确最终诊断及完整诊疗计划。

1.1.2 排除标准:根据现有病历资料即可明确诊断和诊疗方案。

1.2 伦理学:本研究符合医学伦理学要求,并经天津市海河医院医学伦理委员会审批通过(审批号:2025HHKT-0029)。

1.3 研究对象及分组:采用双路径对照研究方法,基于相同病历文本资料,分别由医师和 GenAI 模型独立生成鉴别诊断和诊疗方案进行比较。按照诊疗方案制定方式不同分为医师组(由专科医师独立完成诊疗方案制定,包括西医医师组和中医医师组)和 GenAI 组(由研究人员调用本地化定制的 DeepSeek 模型服务生成诊疗建议,包括西医 GenAI 组和中医 GenAI 组)。参与人工诊疗的西医医师和中医医师均具有医师及以上资格,且从事本专业临床工作 3 年以上。

1.4 GenAI 的技术支撑条件

1.4.1 医院基础信息系统:医院基础信息系统以 WiNEX 6.0 临床信息系统(上海卫宁健康科技集团股份有限公司)为核心,涵盖医院信息系统(hospital information system, HIS)、实验室信息管理系统(laboratory information management system, LIS)、医学影像存档与通信系统(picture archiving and communication system, PACS)、电子病历(electronic medical record, EMR)等主要业务模块,为临床信息的规范化采集和整合提供基础支撑。

1.4.2 智算资源配置:医院完成本地化智算资源部署,具备满足临床场景下模型实时调用与稳定运行的计算和存储能力,为 GenAI 系统在院内安全运行提供硬件保障。算力为 1.2 PFlops,数据存储 448 TB, 32 块通用 GPU 计算卡(美国 NVIDIA 公司, A6000)。

1.4.3 模型部署与调用方式:GenAI 组采用本地化定制的 DeepSeek R1 蒸馏版模型(杭州深度求索人工智能基础技术研究有限公司),模型参数规模为

32 B。模型基于医院场景下进行深度的定制优化,构建了一套完整的技术体系,涵盖针对专业场景的模型微调与对齐、能力评估、检索增强技术优化及领域知识库建设。研究以首次病程记录形成的“病历文本资料”作为输入内容,通过预设的结构化提示词规则(角色、目标、要求、输出格式和限制等),对“鉴别诊断”“诊疗计划”等核心字段进行嵌入式调用,通过输入“//”符号调取模型并输出相应诊疗建议内容。

1.4.4 数据安全与网络防护:算力资源与模型均采用本地化部署方式,未涉及患者数据外传;同时医院已建立完善的网络安全服务系统(项目编号:TJBD-2024-C-295),实现器防、人防安全保障机制的常态化运行,确保临床数据与模型应用过程中的信息安全。

1.5 观察指标:采用 Likert 5 点量表对鉴别诊断及诊疗计划进行质量评价,其中 1 分为非常差,5 分为优秀。

1.5.1 鉴别诊断评价指标:① 诊断冗余度:是否存在与核心病情明显不相关的诊断;不相关诊断越多,评分越低。② 诊断遗漏度:是否遗漏重要或关键诊断;遗漏越多,评分越低。③ 诊断优先级:鉴别诊断的排序是否遵循“先重病,后轻病”“先常见病,后罕见病”的原则;排序越不符合优先级,评分越低。④ 诊断适配度:诊断是否与患者个体化特征(如年龄、性别、职业及流行病学背景等)相契合;契合度越差,评分越低。

1.5.2 诊疗计划评价指标:① 方案完整度:诊疗方案是否完整、无明显遗漏;遗漏项目越多,评分越低。② 方案精准度:诊疗方案是否合理、不过度;不必要或过度干预项目越多,评分越低。

1.6 评价流程:由西医或中医专家分别对医师组与 GenAI 组的鉴别诊断和诊疗计划进行评分。西医专家来自本院呼吸与危重症医学科;中医专家来自本院中医相关科室。所有专家均为副高及以上职称,具有 5 年以上本专业临床工作经验,且在评分前均接受统一培训,并通过评价标准一致性考核。

具体过程:在入院后,每个患者分别由 1 名西医医师和 1 名中医医师进行问诊、体格检查,并收集已有的检验、检查资料。对获得的病史、体格检查和已有的辅助检查等信息进行收集,生成《首次病程记录》的病历文本资料。专科西医医师根据病历文本资料,运用自身掌握的知识和理论,得出西医医师组的鉴别诊断和诊疗计划;专科中医医师得

出中医医师组的中医辨证和中医治疗计划；由二线医师（主治及以上职称）在 24 h 内完成审核与修订。医师组在整个诊疗过程中不使用任何 AI 工具辅助；由课题组研究人员对同一份病历文本资料调用嵌入的 DeepSeek 定制模型，作为 GenAI 组的鉴别诊断和诊疗计划，GenAI 生成的诊断与治疗方案仅供研究分析使用，并未反馈给临床团队，也未对患者的实际诊疗过程及预后产生任何影响。整理病例资料及两组鉴别诊断和诊疗计划内容，采用去标识化与随机呈现顺序提供给专家，由 4 名西医专家和 4 名中医专家使用 Likert 5 点量表，对 6 项指标分别进行独立评分。西医专家仅评价西医相关内容；中医专家仅评价中医相关内容。每项指标取 4 名专家评分的平均值，作为该病例本指标的最终得分。

1.7 统计学处理：使用 SPSS 26.0 统计软件进行数据分析。非正态分布的计量资料以中位数（四分位数） $[M(Q_1, Q_3)]$ 表示，组间比较采用 Wilcoxon 符号秩检验。由于评分数据呈偏态分布，故采用肯德尔和谐系数（Kendall's W）评估专家评分的一致性。Kendall's W 系数取值范围 0 ~ 1，越接近 1 表示一致性越强。检验水准 α 值取双侧 0.05。

2 结 果

2.1 患者资料：共纳入 96 例病例资料，其中肺部感染患者 28 例，罕见病患者 9 例，其他主诊断包括慢性气道疾病、间质性肺病、肺部肿瘤、肺血管疾病、胸膜疾病等；罕见病病例包括抗中性粒细胞胞质抗体（antineutrophil cytoplasmic antibody, ANCA）相关性血管炎、特发性肺纤维化、恶性胸膜间皮瘤。

2.2 专家评分一致性：所有指标的专家评分均具有一致性（ $P < 0.01$ ）。各指标的 Kendall's W 系数为 0.378 ~ 0.546，表明专家组的整体一致性处于中等水平。专家们对各项指标的理解和评价标准整体较为一致，基于专家评分得出的诊断质量评价具有较好的可靠性。

2.3 鉴别诊断质量比较（表 1 ~ 2）：在诊断遗漏度方面，与西医医师组比较，西医 GenAI 组表现出优势（ $P < 0.05$ ）；与中医医师组比较，中医 GenAI 组亦表现出优势（ $P < 0.05$ ）。在诊断适配度方面，与西医医师组比较，西医 GenAI 组表现出优势（ $P < 0.05$ ）；与中医医师组比较，中医 GenAI 组差异无统计学意义（ $P > 0.05$ ）。在诊断冗余度和诊断优先级方面，与西医医师组比较，西医 GenAI 组差异无统计学意义（ $P > 0.05$ ）；与中医医师组比较，中医 GenAI 组差异亦无统计学意义（ $P > 0.05$ ）。

2.4 诊疗计划质量比较（表 1 ~ 2）：在方案完整度方面，与西医医师组比较，西医 GenAI 组表现出优势（ $P < 0.05$ ）；与中医医师组比较，中医 GenAI 组表现出优势（ $P < 0.05$ ）。在方案精准度方面，与西医医师组比较，西医 GenAI 组差异无统计学意义（ $P > 0.05$ ）；与中医医师组比较，中医 GenAI 组差异亦无统计学意义（ $P > 0.05$ ）。

3 讨 论

鉴别诊断与诊疗计划制定环节是患者入院后临床诊疗的起点，也是临床诊疗过程中最重要且易出错的环节，将直接影响后续检查方案和治疗决策^[3-4]。本研究通过对住院病历《首次病程记录》

表 1 西医医师组与西医 GenAI 组鉴别诊断质量和诊疗计划质量比较〔分， $M(Q_1, Q_3)$ 〕

组别	鉴别诊断质量				诊疗计划质量	
	诊断冗余度	诊断遗漏度	诊断优先级	诊断适配度	方案完整度	方案精准度
西医医师组	4.00 (4.00, 4.50)	4.00 (3.75, 4.25)	4.00 (4.00, 4.50)	4.00 (4.00, 4.25)	4.00 (3.75, 4.25)	4.00 (4.00, 4.31)
西医 GenAI 组	4.00 (4.00, 4.44)	4.00 (4.00, 4.50)	4.25 (4.00, 4.50)	4.25 (4.00, 4.50)	4.25 (4.00, 4.50)	4.25 (4.00, 4.25)
Z 值	-0.685	-2.858	-0.312	-4.349	-5.659	-1.062
P 值	0.494	0.004	0.755	<0.001	<0.001	0.288

注：GenAI 为生成式人工智能

表 2 中医医师组与中医 GenAI 组鉴别诊断质量和诊疗计划质量比较〔分， $M(Q_1, Q_3)$ 〕

组别	鉴别诊断质量				诊疗计划质量	
	诊断冗余度	诊断遗漏度	诊断优先级	诊断适配度	方案完整度	方案精准度
中医医师组	4.00 (4.00, 4.00)	3.00 (2.00, 3.00)	4.00 (4.00, 4.00)	4.00 (3.00, 4.00)	3.00 (2.00, 3.00)	4.00 (4.00, 4.00)
中医 GenAI 组	4.00 (3.00, 4.00)	5.00 (4.00, 5.00)	4.00 (3.00, 4.00)	4.00 (3.00, 4.00)	5.00 (4.00, 5.00)	4.00 (4.00, 4.00)
Z 值	-1.077	-8.585	-0.862	-0.041	-8.063	-0.518
P 值	0.281	<0.001	0.389	0.968	<0.001	0.605

注：GenAI 为生成式人工智能

中的鉴别诊断与诊疗计划制定环节的病历质量进行探索性分析,进一步评估人工智能在真实临床场景中的辅助价值。

3.1 GenAI 组的诊断遗漏度评分和方案完整度评分均高于医师组:一是知识储备的“宽度”与实践应用的“深度”。不同医师掌握的基本知识、理论、技能存在差异,将临床指南或共识转化为具体实践的效果也不尽相同,不排除发生遗漏诊断或诊疗计划不足的情形。相比之下,本课题组根据实际需求进行模型本地化部署时进行了定制调整,知识来源包括国际与国内权威教材、临床指南、专家共识、医学期刊论文以及中医学资源等^[5-7]。此外,本研究显示,西医 GenAI 组与西医医师组的诊断遗漏度评分中位数相同,但前者的评分分布更集中于高分端,四分位区间整体高于后者,能够更稳定地覆盖重要诊断,组间差异具有统计学意义。二是对少见或罕见疾病识别的敏感度。GenAI 组基于丰富的知识库以及设置的“提示词”规则,在少见或罕见疾病辅助识别方面展现出逊色于医师组的能力。例如,本研究中 1 例女性患者因“间断咳嗽、胸闷,胸部 CT 提示肺部间质性改变,伴双侧颈部及纵隔多发淋巴结肿大”被收治入院,GenAI 的鉴别诊断中列出了“结节病”这一罕见疾病,并提出了进一步确诊的建议。另有研究显示,不同大语言模型在诊断罕见病与诊断常见病的能力上无显著差距^[8]。三是知识体系的持续更新能力。医学学科的知识体系快速更迭,但受限于学习意愿、能力、时间、环境等因素,不同医师的知识体系更新速度、知识面拓展宽度往往参差不齐。本院定期更新本地化部署模型的知识库,弥补了医师知识体系更新方面的不足。四是工作状态的稳定性。专科医师的实际工作表现可能受到专长学科、知识储备、工作能力、责任心、疲劳程度、情绪波动及工作环境等多重因素影响^[9],而 GenAI 则不受上述因素干扰,始终保持稳定的工作状态^[7]。

3.2 西医 GenAI 组的诊断适配度评分高于西医医师组:一是诊断结论的个体化。GenAI 可以更加全面、高效地整合患者的病历信息,结合年龄、性别、职业、流行病学背景对鉴别诊断进行个性化论证,按可能性排序生成诊断建议,并对可排除的诊断选项提供具有针对性的分析说明。如研究中对 1 例“支气管扩张”的鉴别诊断,GenAI 指出“患者虽有长期咳嗽及偶发咯血表现,需考虑肺结核可能,但因其无午后低热、盗汗等典型结核中毒症状,胸部 CT

未显示结核病灶,且结核抗体检测为阴性,故可排除该诊断”,推理过程增强了排除诊断的逻辑透明度与个体化适配性^[10]。二是 GenAI 对微弱或隐匿信息具有高敏感度。在临床实践中,微弱线索或“沉默”信息易被人工诊断忽略,而 GenAI 则无差别地处理此类信息。如研究中对 1 例普通感染患者进行评估时,GenAI 能够将患者“免疫抑制状态”的潜在风险因素纳入考量,并提出“机会性感染”排查建议。这归功于模型优化的长期情境处理框架,使大语言模型增强了识别临床变量之间微妙相互关系的能力^[11-12]。三是 GenAI 的诊断适配程度更高。GenAI 组的西医诊疗评分显著高于医师组,而二者在中医诊疗方面无显著差异,主要是由于西医知识呈现高度结构化、标准化特征,与 AI 的模式识别能力高度契合,评分标准相对客观,专家一致性较高。

3.3 中医 GenAI 组的诊断适配度评分与中医医师组无显著差异:GenAI 能够基于病历文本生成与西医诊断具有一定映射关系的中医辨证结论。如研究中 1 例西医诊断为“胸腔积液”的病例,GenAI 基于西医与中医病理机制的内在关联,自动映射为“痰瘀水停”的经典证型。既往文献表明,恶性胸腔积液的中医核心实证为“瘀水互结证”和“饮停胸胁证”^[13-14],GenAI 的输出与该类疾病的常见中医证型方向一致。又如 1 例西医诊断为“空洞结节、支气管扩张”的病例,GenAI 整合病灶特征与中医病因辨证逻辑,凝练出“痰热瘀结”的核心病机,与《支气管扩张症中医证候诊断标准(2019 版)》^[15]中将痰热壅肺证作为支气管扩张症的核心证型之一且在合并感染的急性期尤为多见的标准具有一致性。此外,GenAI 对微弱线索或“沉默”信息的敏感性也在中西结合方面有所体现。如研究中 1 例西医缺乏特效干预的“放疗后低热”病症,GenAI 能捕捉治疗后机体的功能状态变化,将其归属到“气虚痰郁”的辨证范畴。有研究指出,气虚发热系癌性发热的常见证型之一^[16]。GenAI 可能具备将部分西医临床信息与中医辨证术语进行关联的能力,另有研究也对 DeepSeek 模型处理中医语料的技术特点进行了探讨^[17]。值得关注的是,本研究对中医辨证结论的评价是基于输入病历文本中可见症状、体征及检查信息的模式匹配与语义映射,而非基于确证诊断。

3.4 本研究的特点及不足:一是研究以鉴别诊断与诊疗计划作为评估对象,以专家评分作为评价标准。疾病诊疗基于患者病情状态、诊疗进行动态调整,

入院初期的方案不能代表患者全流程诊疗方案,后期将以“出院诊断”为标准进行研究。此外,未来会将医师对 GenAI 建议是否采纳整合入临床 workflow,比较在医师决策支持下,GenAI 组与医师组在上述临床结局及过程指标方面的差异。二是排除标准剔除了根据现有病历资料即可明确诊断和诊疗方案的患者,纳入病例主要以疑难、复杂、需鉴别诊断者为主,旨在聚焦更具挑战性的临床任务。另外,研究结论仅适用于呼吸与危重症医学科住院患者中首次病程记录阶段需要进行鉴别诊断的复杂或疑难病例。三是本研究中大语言模型提示词在角色、目标、要求、输出格式和限制等方面相对固定,在临床实践中仍需要动态调整、不断优化。四是本研究的发现为探索性证据,基于研究周期内连续入组病例数确定样本量,因未进行估算,可能对部分指标的统计效能及结果的稳定性产生影响,未来将在设计阶段对前瞻性样本量估算作进一步验证。五是 DeepSeek 模型为通用模型,训练数据主要来源于互联网,受个人信息和商业秘密保护等因素限制,相关关键医疗数据几乎不在互联网公开且来源渠道较窄。因此,在医疗领域实现大语言模型运用的突破,仍寄希望于对专业模型的持续训练。目前,本课题组与国家超级计算天津中心合作的以“天河·天元模型”为基础的 CDSS 相关研究正在进行中。

综上所述,GenAI 在减少诊断遗漏、提高个体化适配程度及完善治疗方案方面具有潜在优势,但其临床价值尚需以出院诊断、治疗结局为金标准的验证性研究加以确认。此外,优秀专科医师的每一句问诊、对每一项检查结果的解读,以及对患者情绪状态等非语言信息的关注,都可能触发医师的逻辑思维“印象”。这种无形“印象”的内容可能多于病案中有形“文字”的内容,进而影响医师诊疗结果。专科医师的直觉、经验和伦理判断在需要极高创造性、复杂情感交互或处理前所未有病例的场景下仍具有不可替代性。

利益冲突 所有作者声明不存在利益冲突

志谢 感谢天津市海河医院(天津大学海河医院)西医专家组(王合荣、李红蔚、华静娜、胡金苗)和中医专家组(马兆润、石铁、张虹、范爽)在本研究中的贡献

作者贡献声明 杨万杰: 酝酿和设计实验、实施研究、采集数据、分析/解释数据、起草文章、对文章的知识性内容作批评性审阅、统计分析、获取研究经费;傅浩: 酝酿和设计实验、实施研究、采集数据、对文章的知识性内容作批评性审阅;孟祥飞: 酝酿和设计实验、实施研究、对文章的知识性内容作批评性审阅、支持性贡献;倪广健: 酝酿和设计实验、对文章的知识性内容作批评性审阅、支持性贡献;

李长松、庞晓磊、刘军、刘金婷、林宝禄、邢维荣、张俊全: 实施研究、对文章的知识性内容作批评性审阅、支持性贡献;万振、崔超、史丽霞、赵卫康: 实施研究、采集数据、对文章的知识性内容作批评性审阅、支持性贡献;李莉: 实施研究、采集数据、分析/解释数据、对文章的知识性内容作批评性审阅、支持性贡献;范爽: 实施研究、采集数据、分析/解释数据、对文章的知识性内容作批评性审阅、统计分析、支持性贡献;韩东、韩晶: 实施研究、采集数据、分析/解释数据、起草文章、对文章的知识性内容作批评性审阅、统计分析、支持性贡献;杨眸: 实施研究、对文章的知识性内容作批评性审阅、统计分析、支持性贡献;李雪: 实施研究、采集数据、分析/解释数据、起草文章、对文章的知识性内容作批评性审阅、支持性贡献

参考文献

- [1] 闫温馨, 刘珏, 梁万年. DeepSeek 赋能全科医学: 潜在应用与展望[J]. 中国全科医学, 2025, 28(17): 2065–2069. DOI: 10.12114/j.issn.1007-9572.2025.0023.
- [2] 杨万杰, 傅浩, 孟祥飞, 等. 基于生成式人工智能的全流程疾病诊疗辅助系统的研发与应用[J]. 中华危重病急救医学, 2025, 37(5): 477–483. DOI: 10.3760/ema.j.cn121430-20250127-00093.
- [3] Hoofman J, Dijkstra AC, Suurmeijer I, et al. Common contributing factors of diagnostic error: a retrospective analysis of 109 serious adverse event reports from Dutch hospitals[J]. BMJ Qual Saf, 2024, 33(10): 642–651. DOI: 10.1136/bmjqs-2022-015876.
- [4] Jackson R, Feder G. Guidelines for clinical guidelines[J]. BMJ, 1998, 317(7156): 427–428. DOI: 10.1136/bmj.317.7156.427.
- [5] 百度智能云. DeepSeek 开源模型深度解析: 技术演进与应用全景 (2024–2025)[EB/OL]. (2025–09–25) [2025–12–21]. <https://cloud.baidu.com/article/3747767>.
- [6] Guo DY, Yang DJ, Zhang HW, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning[J]. Nature, 2025, 645(8081): 633–638. DOI: 10.1038/s41586-025-09422-z.
- [7] 北京卫生学会大数据与互联网人工智能医疗专委会, 中国生物医学工程学会医学人工智能分会法律伦理专家组, 曹艳林, 等. 医疗机构部署 DeepSeek 专家共识[J]. 中国医学伦理学, 2025, 38(5): 674–678. DOI: 10.12026/j.issn.1001-8565.2025.05.19.
- [8] Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making[J]. Nat Med, 2025, 31(8): 2546–2549. DOI: 10.1038/s41591-025-03727-2.
- [9] Brashear J, Mize R, Laposata M, et al. Impact of diagnostic management team on patient time to diagnosis and percent of accurate and clinically actionable diagnoses[J]. Diagnosis (Berl), 2024, 11(2): 132–135. DOI: 10.1515/dx-2023-0175.
- [10] Zhou JS, Cheng YH, He SX, et al. Large language models for transforming healthcare: a perspective on DeepSeek-R1[J]. MedComm Futur Med, 2025, 4(2): e70021. DOI: 10.1002/mef2.70021.
- [11] Afzal M, Alam F, Malik KM, et al. Clinical context-aware biomedical text summarization using deep neural network: model development and validation[J]. J Med Internet Res, 2020, 22(10): e19810. DOI: 10.2196/19810.
- [12] Liu YX, Yu FF, Zhang XF, et al. Assessing the role of large language models between ChatGPT and DeepSeek in asthma education for bilingual individuals: comparative study[J]. JMIR Med Inform, 2025, 13: e65365. DOI: 10.2196/65365.
- [13] 郑磊, 刘猛, 贾立群. 恶性胸腔积液中医证候及用药规律分析[J]. 北京中医药, 2015, 34(7): 529–531. DOI: 10.16025/j.1674-1307.2015.07.007.
- [14] 吕洪清, 罗娟娟, 刘桂荣, 等. 国医大师张志远治疗胸腔积液用药经验[J]. 山东中医杂志, 2024, 43(12): 1414–1417, 1422. DOI: 10.16295/j.cnki.0257-358x.2024.12.016.
- [15] 中华中医药学会内科分会, 中国民族医药学会肺病分会, 中华中医药学会肺系病分会. 支气管扩张症中医证候诊断标准 (2019版)[J]. 中医杂志, 2020, 61(15): 1377–1380. DOI: 10.13288/j.11-2166/r.2020.15.022.
- [16] 王泽文, 王振强. 癌性发热的中医辨证治疗[J]. 中医临床研究, 2024, 16(6): 70–74. DOI: 10.3969/j.issn.1674-7860.2024.06.015.
- [17] Long HL, Deng Y, Guo YG, et al. Large language model evaluation in traditional Chinese medicine for stroke: quantitative benchmarking study[J]. JMIR Form Res, 2025, 9: e81545. DOI: 10.2196/81545.

(收稿日期: 2026–03–13)

(本文编辑: 张耘菲)