·2025 年中山医医学教育发展大会征文·

大语言模型在医学检验技术教育中的能力评估

汪洋　吴嘉豪　张帆　程静　谭红霞　欧阳涓　李俊勋
中山大学附属第一医院医学检验科，广州 510080
通信作者：李俊勋，Email：lijunx5@mail.sysu.edu.cn

【摘要】 目的　评估主流大语言模型（large language model，LLM）在医学检验教育中的专业知识水平，探究其是否具备辅导医学检验技术专业学习的可能性。**方法**　选取 2023 年我国临床医学检验技术（师）资格考试真题共 400 题，分 2 轮测试 5 款 LLM（Copilot、Grok、元宝、豆包及 Kimi）：首轮采用零提示词策略，次轮采用交互优化策略，评估其答题正确率及生成内容质量。通过 Cochran's *Q* 检验比较 LLM 性能差异，并基于 CLEAR 框架（完整性、准确性、循证性、恰当性、相关性）对生成内容进行评分。**结果**　首轮测试中，豆包总正确率最高（375/400），且豆包与元宝的总正确率显著优于 Copilot 与 Kimi（*P*<0.001）。次轮优化交互后，Kimi 的正确率显著提升（*P*<0.05）；其他 LLM 的正确率均略有提升，但差异无统计学意义（*P*>0.05）。豆包的总正确率仍最高（380/400），豆包和元宝的总正确率显著高于 Copilot（*P*<0.005）。基于 CLEAR 框架的评估显示，元宝、豆包、Kimi 在循证性（*P*<0.003）和完整性（*P*<0.05）维度显著优于国外 LLM，能规范引用权威证据，生成内容质量更优。**结论**　主流 LLM 具备丰富的医学检验知识，通过单题输入、明确证据要求及启用高级推理功能可提升答题正确率与生成内容质量。国产 LLM 在正确率上与国外 LLM 相当，且在循证性和完整性维度具有显著优势。研究表明，主流 LLM 可作为辅助工具用于医学检验技术专业知识学习。

【关键词】 大语言模型；　医学检验技术；　临床医学检验技术（师）资格考试；　CLEAR 框架；　豆包；　元宝

【中图分类号】 R-05

基金项目：2023 年广东省基础与应用基础研究基金项目（2023A1515220150）；2023 年吴阶平医学基金会临床科研专项资助基金（H2023053）；中山大学 2025 年校级教学质量与教学改革工程项目

DOI: 10.3760/cma.j.cn116021-20250516-02108

Proficiency evaluation of large language models in medical laboratory technology education

Wang Yang, Wu Jiahao, Zhang Fan, Cheng Jing, Tan Hongxia, Ouyang Juan, Li Junxun
Department of Medical Laboratory, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China
Corresponding author: Li Junxun, Email: lijunx5@mail.sysu.edu.cn

【Abstract】 Objective　To assess the professional knowledge proficiency of mainstream large language models (LLMs) in medical laboratory education and to explore their potential as educational aids for medical laboratory technology students. **Methods**　A comprehensive evaluation was conducted using 400 authentic questions from the 2023 Chinese National Clinical Medical Laboratory Technician Qualification Examination. Five LLMs (Copilot, Grok, Yuanbao, Doubao, and Kimi) were tested through two-round interactions using zero-shot prompting and interaction-optimized prompting strategies. The accuracy of answers and the quality of generated content were evaluated. Performance disparities were analyzed using Cochran's Q test. Content quality was scored through the CLEAR framework (completeness, lack of false information, evidence-based reasoning, appropriateness, relevance). **Results**　In the first-round test, Doubao achieved the highest overall accuracy rate (375/400). The overall accuracy rates of Doubao and Yuanbao significantly outperformed Copilot and Kimi (*P*<0.001). After the second-round interactive optimization, the accuracy rate of Kimi significantly improved (*P*<0.05), whereas other LLMs showed slight improvements (*P*>

0.05). Doubao still had the highest overall accuracy rate (380/400). The overall accuracy rates of Doubao and Yuanbao significantly outperformed Copilot (P<0.005). Evaluation based on the CLEAR framework revealed that Yuanbao, Doubao, and Kimi significantly outperformed foreign models in the dimensions of evidence-based reasoning (P<0.003) and completeness (P<0.05), demonstrating standardized citation of authoritative evidence and superior content quality. **Conclusions** The tested LLMs possess extensive medical laboratory knowledge. The accuracy of their answers and the quality of the generated content can be improved through single-question input, specifying evidence requirements, and enabling advanced reasoning functions. Domestic LLMs are comparable to foreign LLMs in terms of accuracy, and have significant advantages in the dimensions of evidence-based reasoning and completeness. LLMs can serve as auxiliary tools for learning professional knowledge in medical laboratory technology.

【Key words】 Large language model; Medical laboratory technology; Clinical Medical Laboratory Technician Qualification Examination; CLEAR framework; Doubao; Yuanbao

Fund program: 2023 Guangdong Province Basic and Applied Basic Research Fund (2023A1515220150); 2023 Wu Jieping Medical Foundation Clinical Research Special Fund (H2023053); 2025 Sun Yat-sen University Teaching Quality and Reform Project

DOI: 10.3760/cma.j.cn116021-20250516-02108

近年来，以ChatGPT为代表的大语言模型（large language model，LLM）通过海量知识库训练获得的语义理解与推理能力，在医学考试评估中展现出令人瞩目的表现[1]。这种技术突破不仅重塑了医学知识获取方式，更对传统医学教育体系提出了全新挑战。已有研究表明，将ChatGPT整合到医学教育中可获得显著收益[2-3]。但当前研究多聚焦于国外LLM在临床医学领域的表现评估，缺乏对国内LLM的评价，亦缺乏LLM在医学检验这一细分领域的知识评价。

基于此，本研究旨在探讨核心研究问题：LLM是否具备辅导医学检验技术专业学习的能力？具体分解为3个子研究问题：①主流LLM是否具备较丰富的医学检验专业知识？②主流LLM生成的答案解析是否具有较高的质量？③如何提升LLM生成内容的质量？

为回答以上研究问题，本研究选取2023年临床医学检验技术（师）资格考试全题库，以两阶段交互策略分别测试5款主流LLM，通过Cochran's Q检验比较正确率差异，并基于既有的框架评价LLM生成内容的质量，旨在揭示LLM作为新型智能教育工具的潜力边界与优化路径。

1 资料与方法

1.1 测试题目

为检验主流LLM的医学检验专业知识，本研究选取2023年我国临床医学检验技术（师）资格考试全部试题作为测试题目，通过官方指定教材及权威教育平台获取试题文本及标准答案。该考试共设“基础知识”“相关专业知识”“专业知识”及“专业实践能力”4个科目，题型均为客观题，包含单选题（A1/A2型）、共用题干单选题（A3/A4型）及共用备选答案单选题（B型）。每科目题量均为100题，考试时间均为90 min。考试内容覆盖临床检验基础、临床血液学检验、临床化学、临床免疫学和免疫检验、微生物学检验、寄生虫学及检验、医学伦理学等方面，全面考查考生在临床医学检验领域的专业知识和技能水平。

1.2 LLM的选择与测试环境

测试模型选取当前主流的5款LLM，包括Copilot、Grok、元宝（DeepSeek-R1）、豆包及Kimi，均使用公开可用免费网页版本。测试时间为2025年4月15日至18日。各LLM均进行2轮对话：第1轮：将所有选择题每5题为一组，以零提示词的方式将题干及选项发送给各LLM，待其完成内容生成后，再输入下一组题目。本轮对话中5款LLM均不选用高级功能。第2轮：与第1轮同样的题目，每次1题，加入统一的提示词进行对话：“我是一名医学生，请您帮我详细分析这个题目并指出证据出处：副流感嗜血杆菌生长的必要因子是（ ）A.需要X因子；B.需要V因子；C.同时需要X、V因子；D.不需要V因子；E.不需要X、V因子”。第2轮对话中，5款LLM均启用其高级功能：Copilot选择“Think Deeper”功能；Grok选择“DeeperSearch”功能；元宝选择DeepSeek

“R1 深度思考”及“联网搜索”功能；豆包选择“深度思考”功能；Kimi 选择“长思考 k1.5”功能。对含图片的题目，采用直接粘贴原题图片的方式输入，未对图片内容进行人工预处理或 OCR 处理，旨在模拟普通用户直接上传图片的真实使用场景。

1.3　LLM 正确率计算

对每个 LLM 的答案以官方标准答案为基准进行人工核对，记录正确答题数量，计算正确率（正确题数/总题数 × 100%）。

1.4　基于 CLEAR 框架的生成文本质量分析

参照 Malik Sallam 等提出的对人工智能生成的医学相关信息的“CLEAR 评估框架”，对本次研究中各 LLM 在 2 轮对话中生成的答案解释分别进行评分。CLEAR 框架每个维度的评分采用 Likert 5 级量表评分（1=极差，5=优秀）。课题组详细定义了每个维度以及每级评分的内涵，并对 2 名评分员（均为已从事医学检验临床工作及教学工作 15 年以上的资深人员）进行评分开展专题培训，内容包括：①题目错误类型分析及归类；②LLM 医学输出典型偏差识别（知识幻觉、虚假引用、逻辑断层、术语错误等）；③CLEAR 框架的定义及评分标准内涵（表 1）。培训后，2 名评分员根据评分标准预评分 40 题，由另一位高年资医学检验技术从业人员审核，计算 Cohen’s κ 系数超过 0.75，证实一致性较好，方可进行正式评分。2 名评分员评分结果取平均分作为最终得分，得分越高表明信息质量越优。

1.5　统计学方法

使用 IBM SPSS Statistics 26.0 及 Microsoft Excel 2019 进行统计学分析。对于多个 LLM 在相同题目集上的正确率比较，采用 Cochran’s Q 检验。若 Cochran’s Q 检验显著（P<0.05），进一步采用 McNemar-Bowker 检验进行两两比较，并通过 Bonferroni 校正调整显著性阈值。正确率的 95% 置信区间采用 Wilson Score 法计算。评分者间一致性评估：通过 Cohen’s κ 系数检验 2 名评分者对 CLEAR 框架评分的一致性，κ≥0.75 视为一致性良好。

1.6　伦理与局限性

本研究经中山大学附属第一医院医学伦理委

表 1　CLEAR 框架的定义及评分标准内涵

维度	分值	内涵
完整性（completeness）：答案是否涵盖问题所需的全部关键信息，无遗漏或冗余	5 分	解析包含题干分析、全部选项解析及答案依据，逻辑完整
	4 分	解析题干及正确答案，部分错误选项有分析，细节稍缺
	3 分	解析题干和正确答案，无错误选项分析，内容简单
	2 分	仅给出正确答案，无任何解析或解释
	1 分	无答案且无解析，内容空白或无关
准确性（lack of false information）：内容是否准确，无事实性错误或误导性表述	5 分	解析无任何错误，专业知识、逻辑推导完全正确
	4 分	整体正确，偶有非关键细节偏差，不影响答案判断
	3 分	存在部分次要错误，但核心解题逻辑和答案正确
	2 分	较多错误信息，可能导致答案理解偏差或混淆
	1 分	严重错误，解析与正确知识矛盾，答案错误
循证性（evidence）：是否引用可信证据（如教材、指南、权威文献）支持结论	5 分	能详细标出每一句话或每一个知识点的出处，且出处无误，均为较权威的证据
	4 分	能提供部分证据出处，但部分缺失，证据无误，但不够权威
	3 分	能提供证据出处，但不具体，需要手动查阅相关文献验证
	2 分	能提供证据出处但证据难以验证
	1 分	未能提供具体证据出处及引用文献，或证据出处为幻觉
恰当性（appropriateness）：语言是否清晰、简洁、专业，符合医学检验领域的表达规范	5 分	语言简洁易懂，逻辑清晰，术语解释到位，适合目标受众
	4 分	表述较清晰，逻辑连贯，偶有复杂术语但不影响理解
	3 分	内容基本可懂，但存在表述模糊或逻辑衔接不顺畅
	2 分	语言繁琐、逻辑混乱，术语堆砌且未解释，理解困难
	1 分	表述完全混乱，无法提取有效解题思路
相关性（relevance）：内容是否紧扣问题核心，没有无关信息干扰	5 分	内容紧扣题目，无任何无关信息，精准解析考点
	4 分	主体相关，含少量边缘信息，不干扰解题核心
	3 分	部分内容偏离题目，需结合上下文才能理解关联
	2 分	大量无关信息，严重干扰解题思路和答案判断
	1 分	内容与题目无关，无法辅助答题或导致误导

员会批准(批准号〔2023〕403)。

2 结果

2.1 第1轮答题正确率分析

5款LLM在临床医学检验技术(师)资格考试不同科目的答题正确率情况如表2所示。5款LLM在不启用其高级功能的情况下,总体上正确率均较高(均达80%以上),其中豆包正确率最高(总分376),元宝次之(总分366),Grok再次(总分357),Copilot与Kimi正确率相对偏低(总分分别为339与326)。5款LLM在基础知识、相关专业知识、专业实践能力及总正确率方面有显著差异。继而进行两两比较(表3)。结果显示,豆包、元宝和Grok之间及Copilot和Kimi之间在各科目和总正确率上均无显著性差异;豆包和元宝在总正确率上均显著优于Kimi及Copilot(P<0.001)。

2.2 第2轮答题分析

第2轮答题为每次1题,加入统一的提示词并选择高级功能的模式进行再对话。第2轮答题中,Copilot、Grok、元宝及豆包的正确率较第1轮略有提高(均P>0.05),少量第1轮错题在第2轮被修正;而Kimi的表现有显著提升,在基础知识、专业实践能力及总正确率均较第一轮有明显改善(P<0.05)(表4)。5款LLM的正确率均超90%,其中豆包的正确率最高(95%)。正确率两两比较结果(表5)显示,豆包、元宝和Grok之间在各科目及总正确率上均无显著性差异;豆包的基础知识(P<0.001)、相关专业知识(P=0.001)及总正确率(P<0.001)均显著优于Copilot;元宝的总正确率显著优于Copilot(P=0.002);豆包的相关专业知识和总正确率均显著优于Kimi(P=0.004)。

表2 5款LLM第1轮答题正确率比较

科目及子板块(题目数量)	Copilot 正确率(95%*CI*)	Grok 正确率(95%*CI*)	元宝 正确率(95%*CI*)	豆包 正确率(95%*CI*)	Kimi 正确率(95%*CI*)	Cochran's *Q*值	*P*值
基础知识(100)	0.81(0.73~0.89)	0.94(0.89~0.99)	0.90(0.84~0.96)	0.95(0.90~1.00)	0.81(0.73~0.89)	30.13	<0.001
相关专业知识(100)	0.82(0.76~0.90)	0.87(0.80~0.94)	0.93(0.88~0.98)	0.95(0.90~1.00)	0.78(0.70~0.86)	27.10	<0.001
专业知识(100)	0.90(0.84~0.96)	0.91(0.85~0.97)	0.93(0.88~0.98)	0.92(0.86~0.98)	0.88(0.81~0.96)	4.23	0.392
专业实践能力(100)	0.86(0.79~0.93)	0.86(0.79~0.93)	0.90(0.84~0.96)	0.94(0.89~0.99)	0.79(0.71~0.97)	17.97	0.001
总正确率(400)	0.89(0.82~0.96)	0.89(0.83~0.95)	0.92(0.86~0.98)	0.94(0.89~0.99)	0.82(0.76~0.90)	68.01	<0.001

表3 5款LLM第1轮答题正确率两两比较(*P*值)

项目	基础知识	相关专业知识	专业实践能力	总正确率
Kimi-Copilot	1.000	0.305	0.060	0.063
Kimi-元宝	0.011	<0.001[a]	0.003[a]	<0.001[a]
Kimi-Grok	<0.001[a]	0.021	0.060	<0.001[a]
Kimi-豆包	<0.001[a]	<0.001[a]	<0.001[a]	<0.001[a]
Copilot-元宝	0.011	0.005	0.282	<0.001[a]
Copilot-Grok	<0.001[a]	0.200	1.000	0.010
Copilot-豆包	<0.001[a]	0.001[a]	0.031	<0.001[a]
元宝-Grok	0.256	0.124	0.282	0.198
元宝-豆包	0.156	0.608	0.282	0.152
Grok-豆包	0.776	0.040	0.031	0.007

注:a,Bonferroni校正法调整检验水准,P<0.005有显著性差异;"专业知识"科目未检测出LLM正确率间存在显著差异,因而未执行多重比较。

表4 5款LLM第2轮答题正确率比较

科目及子板块(题目数量)	Copilot 正确率(95%*CI*)	Grok 正确率(95%*CI*)	元宝 正确率(95%*CI*)	豆包 正确率(95%*CI*)	Kimi 正确率(95%*CI*)	Cochran's *Q*值	*P*值
基础知识(100)	0.86(0.79~0.93)	0.94(0.89~0.99)	0.93(0.88~0.98)	0.96(0.92~1.00)	0.93(0.88~0.98)[b]	16.82	0.002[a]
相关专业知识(100)	0.87(0.80~0.94)	0.90(0.84~0.96)	0.95(0.90~1.00)	0.97(0.93~1.00)	0.88(0.81~0.96)	15.76	0.003[a]
专业知识(100)	0.93(0.88~0.98)	0.96(0.92~1.00)	0.95(0.90~1.00)	0.93(0.88~0.98)	0.92(0.86~0.98)	3.85	0.663
专业实践能力(100)	0.92(0.86~0.98)	0.88(0.81~0.96)	0.92(0.86~0.98)	0.95(0.90~1.00)	0.92(0.86~0.98)[b]	6.35	0.174
总正确率(400)	0.90(0.84~0.96)	0.92(0.86~0.98)	0.94(0.89~0.99)	0.95(0.90~1.00)	0.91(0.85~0.97)[b]	21.15	<0.001[a]

注:a,P<0.05;b,与第1轮比较,P<0.05。

表 5　5 款 LLM 第 2 轮答题正确率两两比较（*P* 值）

项目	基础知识	相关专业知识	总正确率
Kimi–Copilot	0.007	0.749	0.204
Kimi–元宝	1.000	0.025	0.042
Kimi–Grok	0.701	0.523	0.717
Kimi–豆包	0.250	0.004[a]	0.004[a]
Copilot–元宝	0.007	0.011	0.002[a]
Copilot–Grok	0.002[a]	0.338	0.103
Copilot–豆包	<0.001[a]	0.001[a]	<0.001[a]
元宝–Grok	0.701	0.110	0.147
元宝–豆包	0.250	0.523	0.277
Grok–豆包	0.443	0.025	0.011

注：a，Bonferroni 校正法调整检验水准，*P*<0.005 有显著性差异；“专业知识”与“专业实践能力”科目未检测出 LLM 正确率间存在显著差异，因而未执行多重比较。

2.3　2 轮答题错误原因类型分析

分析 LLM 答题错误原因，可以归纳为 3 个主要类型：无法识别图片关键特征、医学检验专业知识错误及对文本理解不足导致逻辑推理错误。其中占比最大的是对文本理解不足导致的逻辑推理错误，其次是医学检验知识错误。如图 1 所示，交互优化后，Copilot、Grok、元宝及 Kimi 因文本理解不足导致逻辑推理错误所致的错题数量显著减少（*P*<0.05）；Copilot 与 Kimi 因医学检验知识错误所致的错题数量亦显著减少（*P*<0.05），而各 LLM 因无法识别图片关键特征所致的错题数量则无明显减少。

2.4　生成文本质量分析

根据 CLEAR 框架对第 2 轮中 5 款 LLM 的生成内容进行分析，结果如图 2 所示。前后 2 轮的评分中，5 个维度的 Cohen’s κ 系数均介于 0.72~0.85，2 名评分员的评分较为一致。5 个维度的评分中，完整性和循证性的得分在 5 款 LLM 中有显著差异。事后检验（Dunn-Bonferroni）结果显示，豆包、元宝的完整性得分均优于 Copilot（*P*<0.005）；豆包、元宝、Kimi 的循证性得分均较 Grok 和 Copilot 高（*P*<0.005）。分析其得分差异原因主要在于 2 点：一是部分 Copilot 生成的解析不能涵盖所有选项，在完整性上略逊于其他 LLM；二是 Copilot 和 Grok 的生成内容尽管能给出参考文献或证据来源，但并无详细的链接或页码，而元宝、豆包和 Kimi 则能以类似学术论文的参考文献的格式列出证据出处，显著提升了可验证性。

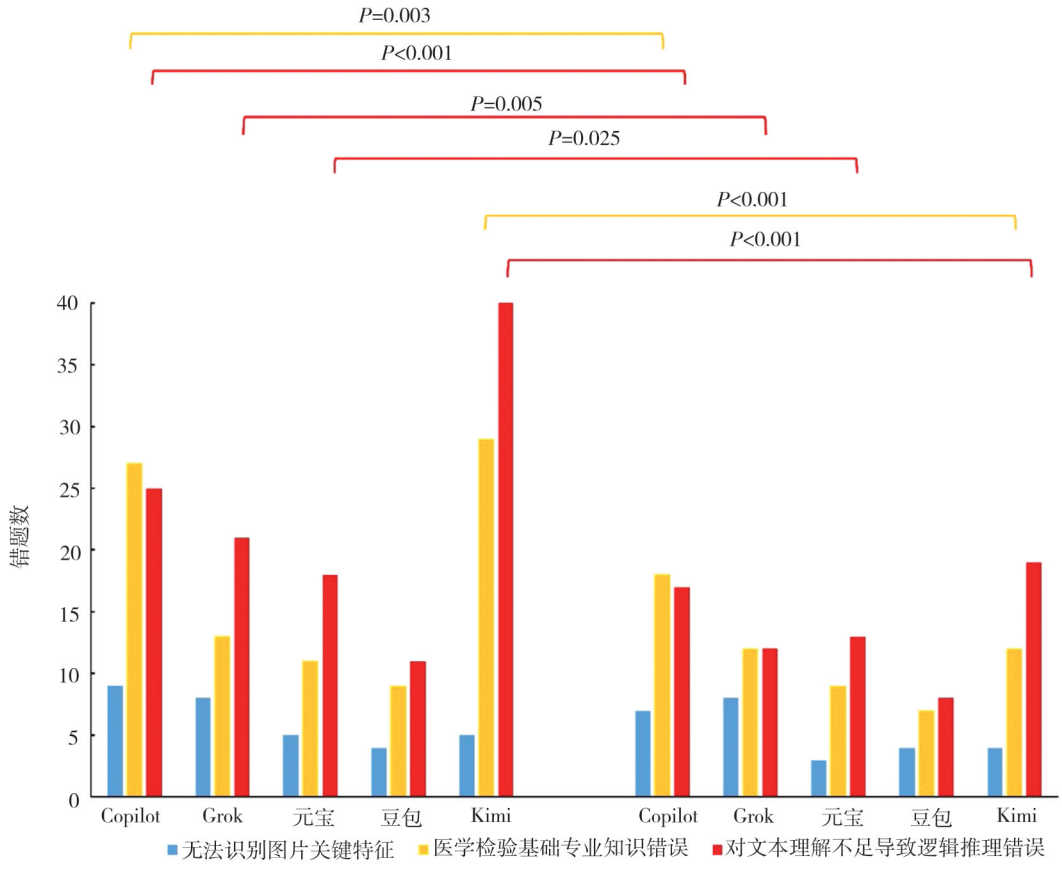


图 1　LLM 错误原因类型分析

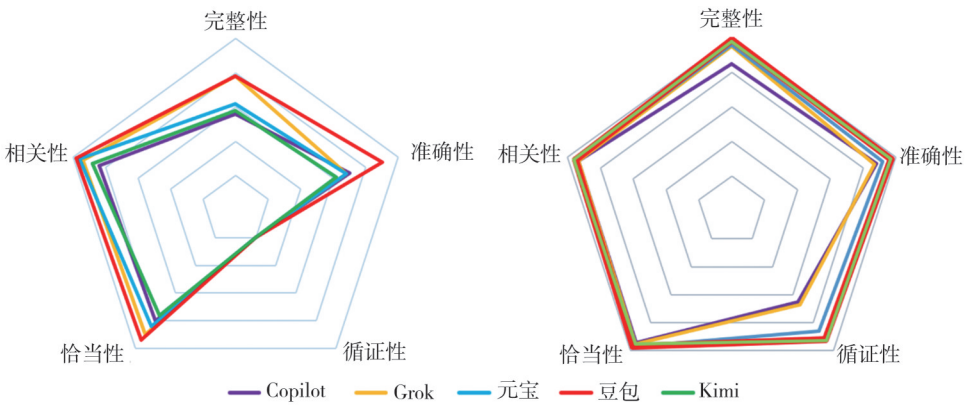


图2　5款LLMCLEAR评分雷达图

3　讨论

3.1　LLM具备丰富的医学检验相关知识

本研究选取2023年临床医学检验资格考试全题库作为评估基准，全面覆盖医学检验核心知识点，具有权威性和时效性，能系统反映LLM的知识水平。本研究结果显示，5款LLM在第1轮答题中仅使用零提示词的情况下均有较高的正确率，说明5款LLM已具备丰富的医学检验相关知识，可作为医学检验知识载体。本研究的结果与国际上其他文献一致：Kung等[4]研究发现GPT4.0在美国医师执照考试中正确率接近或超过及格水平；Long等[5]发现ChatGPT有潜力通过加拿大耳鼻咽喉-头颈外科认证考试；Patel等[6]发现GPT4很有可能通过美国耳鼻喉科（鼻科学）标准化委员会考试；Kung等[7]也观察到GPT4.0在骨科实习考试中的表现优于研究生的水平。从正确率分析，部分国产LLM（豆包、元宝）表现突出。无论是否开启高级功能并优化交互策略，豆包与元宝的总正确率均较Copilot高，与Grok相仿。

从错题的错误原因分析，第1轮答题的错因主要为对文本理解不足导致的逻辑推理错误及医学检验知识错误，而第2轮采用优化交互策略后，5款LLM均能将大部分上述错题纠正。这说明优化交互策略及启用高级功能可以明显改善LLM的逻辑推理能力及医学检验知识的应用水平。

值得注意的是，5款LLM对图片的识别正确率相对较低，优化交互及启用高级功能并不能明显提高图片识别的正确率，推测这与图片内容相对复杂但质量相对较低有关，也与本研究未对图片进行预处理的设计有关。然而，整体偏低的图片识别能力也提示LLM的多模态分析功能仍有待改善。

3.2　交互策略优化激活LLM的认知增强潜能

2轮对话实验表明交互策略显著影响LLM性能。本研究中第1轮的输入策略参考了Miller的认知负荷，即人类工作记忆的容量有限，通常为（7 ± 2）个信息单元[8]，因而采用每5题一组的输入方式。虽然第1轮答题正确率较高，但各款LLM的生成文本内容均较简短，难以支持高质量的学习辅导。在第2轮采用优化交互策略后，5款LLM均对部分第1轮错题进行了修正，其中Kimi的修正比例最高。第2轮采用单题输入的方式，有效降低了LLM的认知负荷，提升了生成文本的质量。更关键的是，观察发现启用LLM的高级功能，如Kimi的“长思考”功能，可显著提升其推理能力。结合提示词中“详细分析这个题目并指出证据出处”的要求，3款国产LLM能系统且规范地标注引用来源（如指南、权威文献等网络资源），方便使用者验证查阅。这种提供结构化解析和证据引用的方式，与Vygotsky提出的“脚手架”学习理论高度契合：当学生处于近端发展区时，LLM可通过逐步提示引导其自主构建知识体系[9]。

3.3　生成内容质量评估揭示国内LLM的优势

在生成内容质量评估中，国产LLM展现出显著优势：元宝、豆包及Kimi在循证性维度得分显著高于Copilot与Grok。3款国产LLM的生成内容中，普遍能在句末直接引用多篇较高质量的参考文献，如指南、权威杂志发表的论著等。相比之下，Copilot与Grok虽然能在生成内容最后列出参考的证据来源，也能包括部分权威文献，但并未提供链接，增加了溯源难度。另一方面，3款国产LLM的完整性评分显著高于Copilot，其主要原因在于Copilot生成内容多聚焦正确答案解析，对错误选项的分析覆盖率较低，在教学场景中信息完整性稍逊，对用户的帮助相对有限。由此可见，国内LLM的生成内容质量具有循证性和完整性方面的优势。

3.4　作为新型知识更丰富者的医学教育赋能路径

根据Vygotsky的近端发展区（the zone of proximal

development, ZPD）理论，知识更丰富者（more knowledgeable other, MKO）是指在学习者就特定任务、过程或概念而言，具有更好理解或更高能力水平的人。ZPD被定义为学习者独立学习所能达到的水平与在MKO指导下所能达到的水平之间的差异[9-12]。在传统的教学模式中，指导教师是主要的MKO，通过社交互动（social interaction）指导学生在ZPD中学习，但指导教师并不是随时可以求教的对象[13]。本研究评估表明，LLM具备丰富的医学检验技术相关知识，并能通过优化提示词生成高质量的内容。这使其能突破传统MKO（如教师）的时空限制，提供7×24 h、类似社交互动（social interaction）的即时反馈与个性化问答。LLM还能详细列出相关学习资源，以“脚手架”的形式引导学生在ZPD中学习。因此，LLM可作为构建动态化学习支持系统的关键工具[10,13]。

3.5　研究局限与未来方向

尽管取得突破，但LLM仍存在局限：一方面，在未预处理的图片识别类题目中错误率偏高，这可能与图片质量及视觉推理能力不足有关。建议后续研究可探索开发整合专业医学知识库和增强视觉理解能力的定制化LLM智能体，以提升其在医学图像识别任务中的性能。另一方面，由于LLM仍有一定的错误率，且其解析能力尚无法比肩经验丰富的人类教师，因而LLM的应用需警惕技术依赖风险，避免仅以获取答案为学习目标，应构建“人工智能辅助基础认知+教师主导临床思维”的混合学习模式：人工智能负责知识点的标准化解析与海量习题训练，教师聚焦难点解析、重点深化及病例分析等高阶临床知识的培养，实现优势互补。

4　结语

5款LLM具备丰富的医学检验专业知识。优化交互策略（如单题输入、明确证据要求及启用高级推理功能），可有效增强LLM的逻辑严谨性与循证性，提升生成内容的质量。部分国产LLM（如豆包、元宝）在正确率上与本研究中的国外LLM表现相当。元宝、豆包及Kimi的生成内容在循证性维度显著优于Copilot及Grok，在完整性维度显著优于Copilot。LLM可作为新型的知识更丰富者，通过类似社交互动的对话模式，随时随地提供高质量的知识解析与学习资源，辅助学生在近端发展区中学习医学检验相关知识。LLM作为辅导学习工具，应与传统教师主导的学习模式紧密结合，构建合理的混合学习模式，以最大化提升学习效率。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　汪洋、李俊勋：项目设计、论文构思、撰写论文；汪洋、吴嘉豪：项目实施；张帆、程静、谭红霞、李俊勋：采集并分析数据；欧阳涓、李俊勋：提出研究命题、总体把关

参考文献

[1] Sallam M, Al-Salahat K, Eid H, et al. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, bard, ChatGPT-3.5 and humans in clinical chemistry multiple-choice questions [J]. Adv Med Educ Pract, 2024, 15: 857-871. DOI: 10.2147/AMEP.S479801.

[2] Triola MM, Rodman A. Integrating generative artificial intelligence into medical education: curriculum, policy, and governance strategies [J]. Acad Med, 2025, 100(4): 413-418. DOI: 10.1097/acm.0000000000005963.

[3] Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: randomized controlled trial [J]. J Med Internet Res, 2024, 26: e57037. DOI: 10.2196/57037.

[4] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models [J]. PLoS Digit Health, 2023, 2(2): e0000198. DOI: 10.1371/journal.pdig.0000198.

[5] Long C, Lowe K, Zhang J, et al. A novel evaluation model for assessing ChatGPT on otolaryngology-head and neck surgery certification examinations: performance study [J]. JMIR Med Educ, 2024, 10: e49970. DOI: 10.2196/49970.

[6] Patel EA, Fleischer L, Filip P, et al. Comparative performance of ChatGPT 3.5 and GPT4 on rhinology standardized board examination questions [J]. OTO Open, 2024, 8(2): e164. DOI: 10.1002/oto2.164.

[7] Kung JE, Marshall C, Gauthier C, et al. Evaluating ChatGPT performance on the orthopaedic in-training examination [J]. JB JS Open Access, 2023, 8(3): e23.00056. DOI: 10.2106/JBJS.OA.23.00056.

[8] van Merriënboer JJG, Sweller J. Cognitive load theory in health professional education: design principles and strategies [J]. Med Educ, 2010, 44(1): 85-93. DOI: 10.1111/j.1365-2923.2009.03498.x.

[9] Vygotsky LS. Mind in society: development of higher psychological processes [M/OL]//Cole M, Jolm-Steiner V, Scribner S, et al. Cambridge, MA, USA: Harvard University Press, 1980 [2023-03-01]. https://www.jstor.org/stable/10.2307/j.ctvjf9vz4.

[10] Li J, Ouyang J, Liu J, et al. Artificial Intelligence-based online platform assists blood cell morphology learning: a mixed-methods sequential explanatory designed research [J]. Med Teach, 2023, 45(6): 596-603. DOI: 10.1080/0142159X.2023.2190483.

[11] Groot F, Jonker G, Rinia M, et al. Simulation at the frontier of the zone of proximal development: a test in acute care for inexperienced learners [J]. Acad Med, 2020, 95(7): 1098-1105. DOI: 10.1097/ACM.0000000000003265.

[12] Lindgren H, Lu MH, Hong Y, et al. Applying the zone of proximal development when evaluating clinical decision support systems: a case study [J]. Stud Health Technol Inform, 2018, 247: 131-135.

[13] 李俊勋，张帆，李润钊，等. 基于人工智能的线上学习平台在医学检验技术专业血细胞形态学实习教学中的应用[J]. 中华医学教育杂志，2025，45(3)：204-209.

（收稿日期：2025-05-16）
（本文编辑：张学颖）