·数智医学教育·

从工具替代到人机协作：深度推理大语言模型评阅医学生反思性写作作业的个案研究

柴桦[1,2]　闫昱江[1,2]　卿平[2,3]

[1]四川大学华西第二医院筑浪学院，成都 610041；[2]国家教材建设重点研究基地（高等学校医学教材研究），成都 610041；[3]四川大学华西医学中心医学教学部，成都 610041

通信作者：闫昱江，Email：1051173813@qq.com

【摘要】 本研究对具有深度推理功能的大语言模型（reasoning-capable large language models，rcLLMs）评阅医学生反思性写作作业的可行性和效果进行了探索。在选择作业样本、建立评阅方法的基础上，用2种国内主流rcLLMs进行评阅，并采用定量与定性相结合的混合研究方法将评阅结果与教师评阅结果及评分标准相比较。研究发现2种rcLLMs和教师三者评分的一致性及两两之间评分的相关性较差，2种rcLLMs各自评分的稳定性也较差；rcLLMs生成的评语结构完整，但关注主题与评分标准不完全一致。本研究揭示出目前rcLLMs尚不能作为评阅医学生反思性写作作业的有效工具，讨论了相关影响因素，并强调了建立人机协作策略及反思AI评阅正当性的重要性。

【关键词】 医学教育；　生成式人工智能；　深度推理大语言模型；　学业评价；　反思性写作

【中图分类号】 R-05

基金项目：四川大学2025年度医学教材建设研究项目（面上项目）（SCUYJ0126）；四川大学高等教育教学改革工程（第十一期）研究项目（SCU11404、SCU11411）；西部妇幼医学研究发展中心2025年研究项目（XBFY-YJXM2025003）

DOI: 10.3760/cma.j.cn116021-20250805-02199

From substitution to human-AI collaboration: a case study on assessment of reflective writing assignments of medical students using reasoning-capable large language models

Chai Hua[1,2], Yan Yujiang[1,2], Qing Ping[2,3]

[1] Zhulang Career Development Center, West China Second University Hospital, Sichuan University, Chengdu 610041, China; [2] National Research Center for Educational Materials, Chengdu 610041, China; [3] Department of Medical Education, West China Medical Center, Sichuan University, Chengdu 610041, China

Corresponding author: Yan Yujiang, Email: 1051173813@qq.com

【Abstract】 This case study explored the feasibility and efficacy of employing reasoning-capable large language models (rcLLMs) to assess the reflective writing assignments of medical students. Following the selection of writing samples and the establishment of an assessment methodology, two prominent domestic rcLLMs were utilized to evaluate the assignments. A method combining quantitative and qualitative analyses was employed to compare the evaluations from rcLLMs with those from instructors in accordance with the established grading rubric. The findings revealed poor inter-rater reliability due to weak correlations among all three raters collectively and pairwise. Both rcLLMs demonstrated significant variability in their own scoring consistency. While the feedback comments generated by the rcLLMs were well-structured, their focuses exhibited a misalignment with the grading rubric. This study suggests that current rcLLMs are not yet suitable for use as effective tools for assessing the reflective writing assignments of medical students. The influencing factors were discussed, and the need to develop human-AI collaborative strategies, as well as to critically reflect on the legitimacy of employing rcLLMs in assessment, was underscored.

【Key words】 Medical education; Generative artificial intelligence; Reasoning-capable large language model; Assessment; Reflective writing

Fund program: 2025 Research Project of Medical Textbook Development of Sichuan University (Major Project) (SCUYJ0126); The 11th Higher Educational Teaching Reform Project of Sichuan University (SCU11404, SCU11411); 2025 Research Project of West China Center for Research and Development of Women and Children's Health (XBFY-YJXM2025003)

DOI: 10.3760/cma.j.cn116021-20250805-02199

近年来，生成式人工智能（generative artificial intelligence，GenAI）在高等医学教育领域的应用已成为医学教育研究的热点[1]。在智能助“教”方面，研究者普遍认为GenAI可以帮助教师处理教学中的各种烦冗事务，例如撰写教案、统计考勤、评阅作业等，从而减轻教师的工作负担[2-12]。其中，关于作业评阅，对于有参考答案的短文本类主观题（如简答题），部分大语言模型（large language models，LLMs）的评分准确性及与教师评分的一致性已能达到较高水平[13]；但对于无参考答案的长文本类主观题（如课程论文、研究报告、反思性写作等），LLMs能否较好地完成评阅（包括评分和生成评语）工作，目前尚缺少研究。

文献计量学研究指出，GenAI评阅长文本类主观题是当前AI赋能学生学业评价的研究热点之一，其中人机评分的一致性作为评阅工具效度评价的重要指标，尤其值得关注[14]。2025年，具有深度推理功能的大语言模型（reasoning-capable LLMs，rcLLMs）的出现，通过逻辑分析、因果推断、知识关联等能力，为长文本类主观题评阅的复杂任务提供了更加有力的工具[15]。本研究模拟高等医学教育中普通教师的真实工作场景，选用具有可及性的国内主流rcLLMs对医学生反思性写作作业进行评阅，深入探究实施过程中遇到的限制与困难，并采用定量与定性相结合的混合研究方法对评阅结果进行分析，以揭示当前条件下GenAI评阅长文本类主观题的可靠性和实用性，为正在探索AI赋能医学教育的同行提供借鉴。

1　设计与实施

1.1　评阅对象与评分标准

选择2017年秋季学期四川大学（以下简称“学校”）临床医学导论课程的反思性写作作业作为评阅对象；由于2022年11月才出现较为成熟且公众可及的LLMs[16]，这一选择确保当时学生的写作和教师的评阅均未受到LLMs的影响。

临床医学导论课程是学校1998年起在全国医学院校中率先开设的新型医学综合课程，目的是使临床医学专业学生早期理解医学模式转变、卫生健康行业发展、医师职业素养要求和医学学习方法等，并安排“早期接触临床”见习活动；课程还鼓励学生利用寒假到家乡基层医疗卫生机构见习。该课程2017年秋季学期的考核方案中，反思性写作作业40分，要求学生在下一学期开学（2018年2月底）前通过在线教学平台提交电子文档；共收到293份反思性写作作业。

反思性写作作业的要求为，主题（四选一）：①医学人才面面观：调查受社会欢迎的医学人才的特征，结合课堂内容进行自我审视与反思；②卫生事件之我见：结合医疗卫生相关热点事件中的经历或感想，以及课程或寒假见习所见，对生命、死亡、健康、医学、伦理、人文及医师职业责任等进行反思；③医路成长在我心：结合课程所学及见习所见，对自身职业生涯成长进行思考与规划；④医学拾贝初探寻：选择课程所学或见习所见中感兴趣的主题，通过社会实践、调查研究等形成调研报告，或撰写其他与本专业有关主题的反思性论文。写作体裁不限、题目自拟，字数不少于2 000字；考核方案对写作格式（如字体、字号、行距、文件命名方式等）提出了具体要求，并提供了评分标准（表1）。

表1　反思性写作作业评分标准

分数段	评分标准
36~40分	原创性的反思，文笔畅达，能从实践出发，有深度，能结合专业和职业特点；或原创性的调查报告（需附上调查原始表格、数据等）、综述、论文等，格式符合学术论文、报告的要求，各部分完整、标准。能反映出思考的深度、广度和认真负责的态度
32~35分	原创性的反思，有自己的见解，结合实践稍欠；综述，摘要、关键词、参考文献等内容完备，格式标准；或一般的调查报告，结合实际问题稍欠。能反映出一定的思考深广度，行文较认真
24~31分	一般的反思或综述，各部分格式不标准，缺乏原创性的见解。文字不能反映深入思考的过程和结果
0~23分	一般的综述，格式不标准，没有原创性的见解，且有明显网上拷贝痕迹，没有注明参考文献出处。思考肤浅，行文敷衍

1.2　评阅方法

选择DeepSeek-R1[15]（杭州深度求索人工智能基础技术研究有限公司，浙B2-20250178）（以下简称“DeepSeek”）和Kimi K1.5长思考[17]（北京月之暗面科技有限公司，京B2–20240852）（以下简称“Kimi”）2个当时具有代表性的主流rcLLMs作为评阅工具；于2025年3月15日至19日将提示词、反思性写作作业、作业要求与评分标准（以下简称“评分标准”）、2017年秋季学期课程教学进度表（以下简称“课表”）提交给rcLLMs，获得并记录评阅结果。

运用提示工程技术并结合rcLLMs的特点，采用“零样本（zero-shot）”策略构建提示词，经过多轮优化调试，确定提示词（表2）。综合2种rcLLMs的功能特点并经反复测试，将任务提交策略确定为：分批提交，每批均新建对话，采用相同的提示词，并上传8份反思性写作作业、1份评分标准、1份课表；重复这一过程，直到所有作业被评阅完。

表2　经优化调试后确定的提示词

文本类型	文本内容
提示词（Prompt）	我是一名高等医学院校的教师，我为临床医学专业大一的学生布置了“临床医学导论”课程的课后“反思性写作”作业，他们利用大一结束后的寒假完成了该作业。请你仔细对照附件上传的《作业要求与评分标准》，并参考该课程的课表，对这些作业进行评阅，包括：①为每份作业评分；②为每份作业写不超过200个字符的评语，总结该文章的优点、缺点（即扣分原因）并提出修改完善的建议。评分完成后，请将学生姓名（即文件名）、分数[只写实际得分（如：34），不用注明满分（如：34/40）]和评语整理成一个表格 附：对评语的要求：①写成完整的一小段话，不需要明确分为“优点、缺点、建议”三部分，也不需要写成要点式的；②如发现某份作业存在抄袭网络文章或抄袭其他同学作业的情况，请在评语中明确指出，并附上被抄袭的网络文章的链接（如链接较长，可以不受评语不超过200个字符的限制），或指明其抄袭了哪位同学的作业；③即使多份作业有共性的问题或改进建议，也把它们写到每份作业的评语中，并整理到表格里，不需要额外对共性问题或建议进行总结提炼

研究者将293份反思性写作作业提交给2种rcLLMs评阅，其中5份作业被DeepSeek以“可能违反使用规范”为由拒绝评阅，故将这5份作业排除；记录2种rcLLMs对其余288份作业的评分和评语，并将其与2017年秋季学期教师的原始评阅结果（以下简称“教师评分”或“教师评语”）一一对应，用于分析评阅的一致性。在第一次评阅的基础上，通过DeepSeek和Kimi提供的“重新生成”功能获取第2次评分，用于分析rcLLMs评分的稳定性。

1.3　评分的定量分析方法

DeepSeek、Kimi和教师三者共对288份反思性写作作业评分864个。使用Excel 2021和SPSS 20.0进行数据处理和统计分析；使用Openbiox Hiplot（ORG）[18]和Excel 2021绘制统计图。采用Kolmogorov-Smirnov检验描述评分分布的正态性，根据是否符合正态分布，采用均数 ± 标准差或中位数（四分位数间距）描述分布情况；采用组内相关系数（intraclass correlation coefficient，ICC）[选用双向随机效应模型和绝对一致性，报告单个测量的ICC（2，1）]和Kendall’s W检验评价三者评分的一致性，采用Spearman秩相关系数评价三者评分两两之间的一致性。分别对2种rcLLMs的第1次评分和第2次评分的分布情况进行比较；采用Spearman秩相关系数分别评价2种rcLLMs两次评分的一致性。检验水准α=0.05。

1.4　评语的定性分析方法

2种rcLLMs均为288份作业撰写了评语，教师为288份作业中的248份撰写了评语，评语长度用中位数（四分位数间距）描述，并对评语的基本结构进行分析。对DeepSeek、Kimi和教师三者的评语及评分标准进行文本分析，由1位研究者对所有文本进行逐句编码，再从中析出若干一级主题、二级主题，析出过程中的疑难问题与另外2位研究者讨论决定；统计每个主题的编码数量占比，从而比较三者评语与评分标准关注主题的一致性，并用调整后的总绝对偏差[=Σ（|评语中一级主题的编码数量占比–评分标准中对应一级主题的编码数量占比|）÷2]半定量描述三者评语与评分标准的差异程度；对评语内容及存在的主要问题进行分析，筛选出有代表性的评语作为示例。使用Excel 2021、Sublime Text（Build 4180）进行文本处理和转换，使用Nvivo 15进行文本分析、词云图绘制。

2　评阅结果

2.1　评分的总体分布情况

Kolmogorov–Smirnov检验显示2种rcLLMs和教师对288份作业的评分均不满足近似正态分布（DeepSeek：D=0.099，P<0.001；Kimi：D=0.160，P<0.001；教师：D=0.220，P<0.001）。DeepSeek评分的中位数（四分位数间距）为34.00（3.00）分，Kimi评分为34.00（2.00）分，教师评分为34.00（1.00）分，三者评分的分布较为接近（图1）。

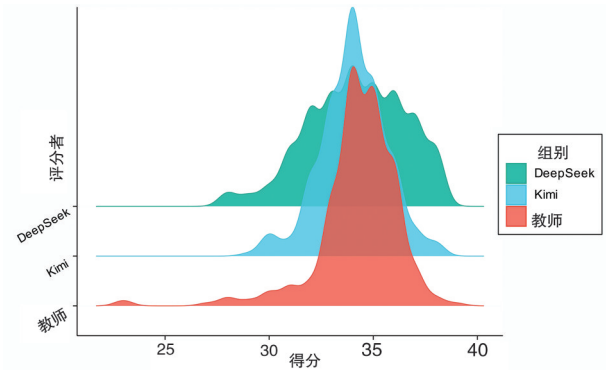


图1　DeepSeek、Kimi 与教师评分的分布情况

2.2　rcLLMs 与教师评分的一致性

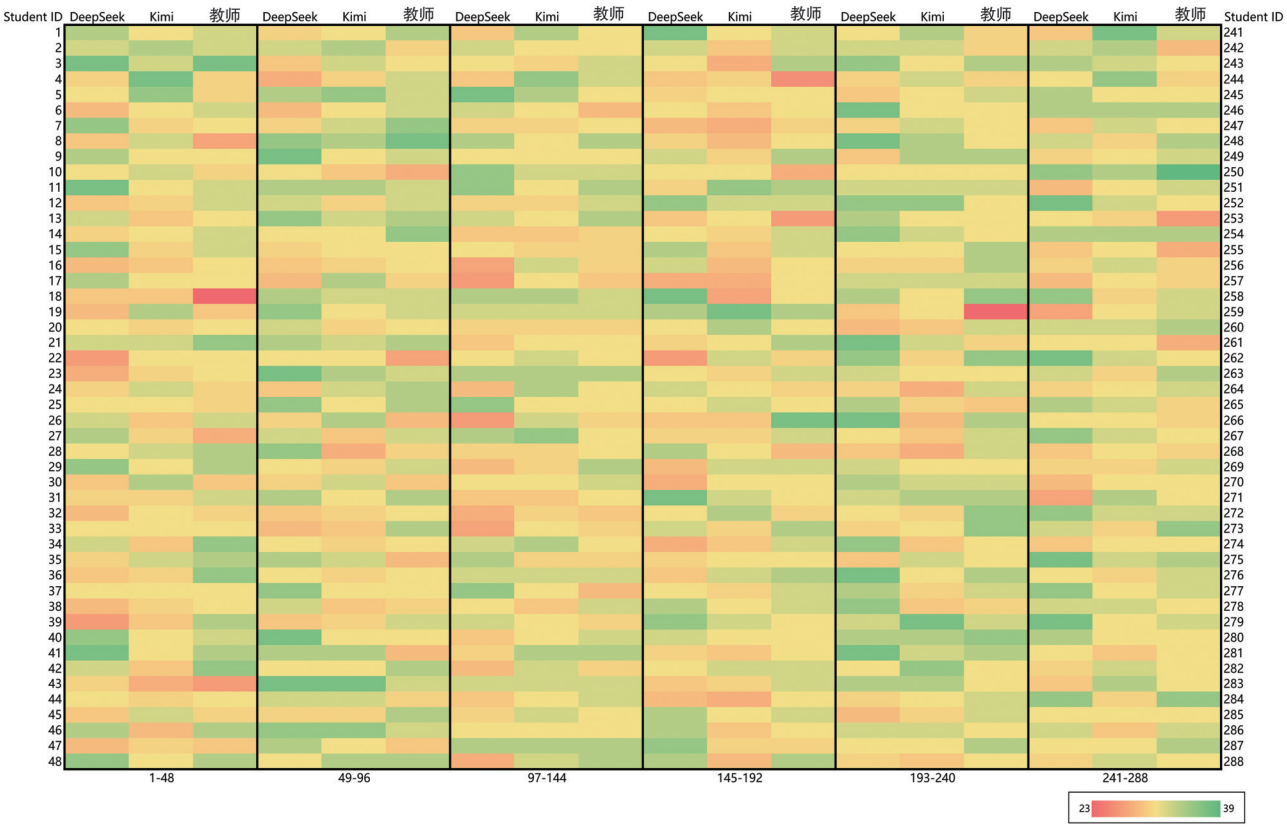
DeepSeek、Kimi 和教师三者评分的一致性分析结果显示，ICC(2,1)=0.126，95% 置信区间(*CI*)为 0.055~0.201；由于 ICC(取值范围 0~1)远小于 0.5，反映出在绝对分数上三者评分的一致性较差(图 2)。Kendall's *W*=0.012，*P*=0.033；由于检验统计量 *W*(取值范围 0~1)非常接近 0，反映出在相对排序上，三者的评分一致性也较差。

DeepSeek 与教师评分的 Spearman 秩相关系数 *r*=0.204(*P*=0.001)，Kimi 与教师评分的 *r*=−0.030(*P*=0.612)，DeepSeek 与 Kimi 评分的 *r*=0.136(*P*=0.021)。由于 Spearman 秩相关系数 *r*(取值范围−1~1)的绝对值均远小于 0.6，可以认为三者评分两两之间的相关性均很弱，其中 Kimi 与教师评分的相关性不具有统计学意义。

2.3　rcLLMs 评分的稳定性

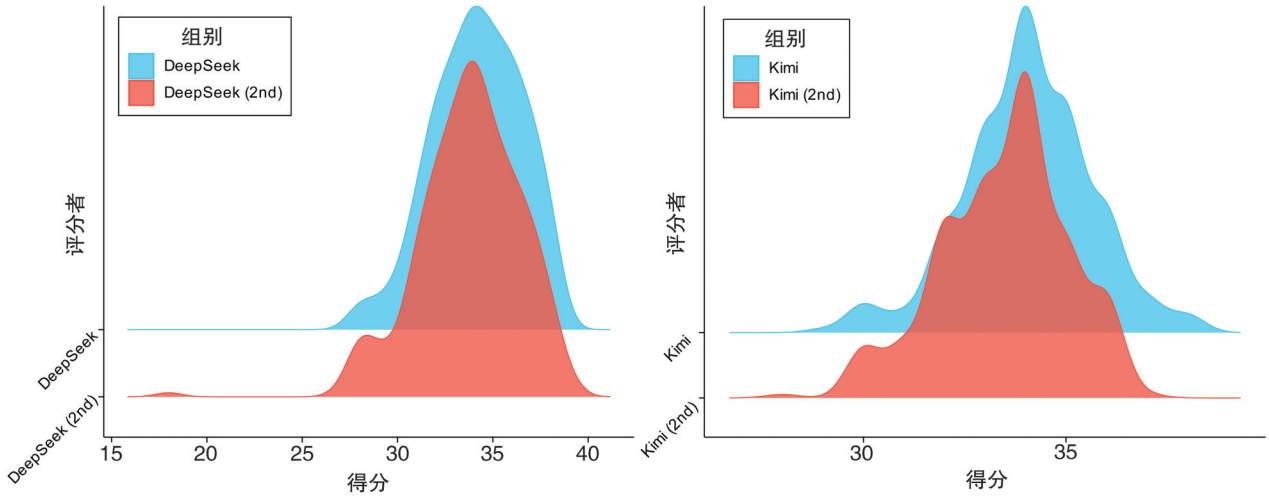
使用 DeepSeek、Kimi 对 288 份作业进行第 2 次评分，Kolmogorov-Smirnov 检验显示二者的第 2 次评分同样均不满足近似正态分布(DeepSeek 2nd：D=0.109，*P*<0.001；Kimi 2nd：D=0.178，*P*<0.001)。DeepSeek 评分的中位数(四分位数间距)为 34.00(4.00)分，Kimi 评分为 34.00(2.00)分；与第 1 次评阅相比，二者评分的中位数均不变，但 DeepSeek 评分的离散程度略增加(图 3)。

DeepSeek 第 1 次与第 2 次评分的 Spearman 秩相关系数 *r*=0.358(*P*<0.001)；Kimi 第 1 次与第 2 次评分的 *r*=0.354(*P*<0.001)；可以认为 DeepSeek 和 Kimi 各自 2 次评分的一致性均较弱，且这种弱相关性具有统计学意义(图 4)。



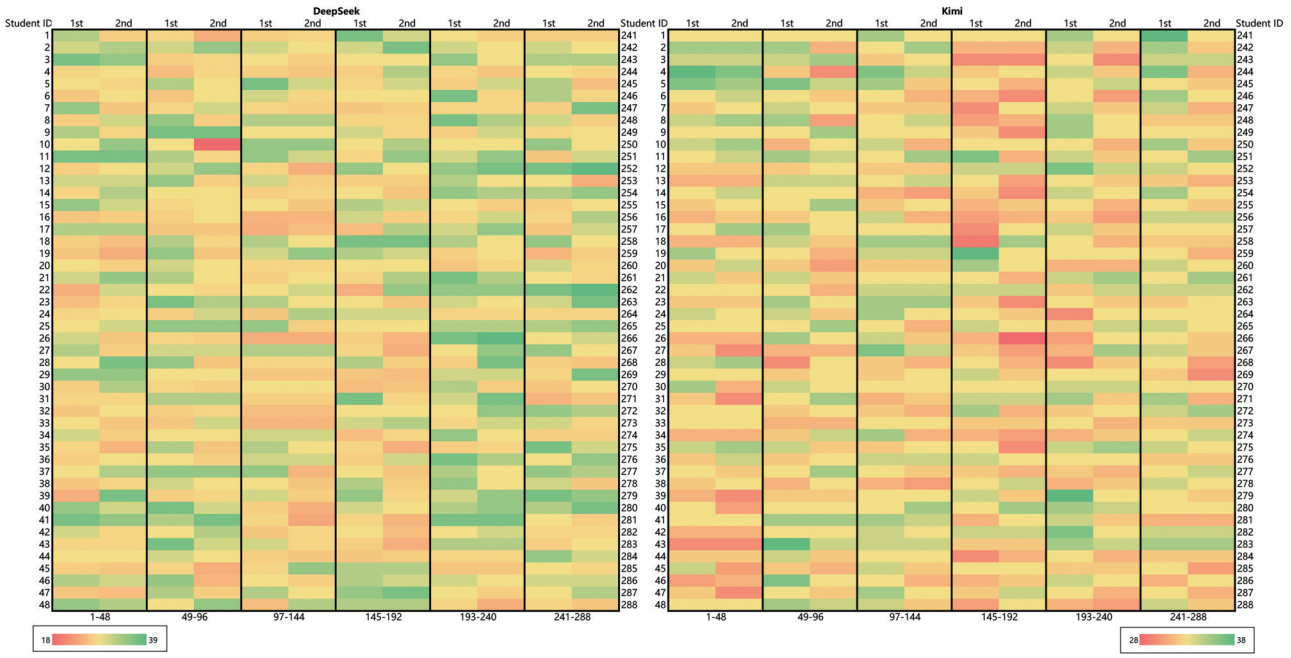
注：热图由 48 行、6 个大列、18 个小列构成，共 864 个小格，每一小格代表一个评分；其中，每一大列中每一行 3 个小格代表 DeepSeek、Kimi 与教师三者对同一份作业的评分，这 3 个相邻小格的颜色越接近，反映三者评分的一致性越好。

图2　DeepSeek、Kimi 与教师评分热图



A. DeepSeek　　B. Kimi

图 3　rcLLMs 两次评分的分布情况对比



注：左为 DeepSeek；右为 Kimi。左右 2 幅热图均由 48 行、6 个大列、12 个小列构成，共 576 个小格，每小格代表 1 个评分；其中，每大列中每 1 行 2 个小格代表该大语言模型对同一份作业的 2 次评分，这 2 个相邻小格的颜色越接近，反映该大语言模型 2 次评分一致性越好。

图 4　rcLLMs 第 1 次评分与第 2 次评分热图

2.4　评语的总体情况

DeepSeek 和 Kimi 的评语内容与作业内容具有对应关系，未发现错评现象；评语内容与评分基本一致，即评语偏向负面的作业得分较低、而评语偏向正面的作业得分较高；评语长度为 DeepSeek 84.00（16.00）个字符，Kimi 87.00（41.25）个字符。教师评语长度为 14.00（12.00）个字符。

rcLLMs 的评语字数相对较多，句式结构完整，通常以概括一份作业的选题或写作思路开头，随后分析作业的主要优缺点，最后提出修改建议（DeepSeek 和 Kimi 分别提出修改建议 243 条、262 条）；而教师的评语字数相对较少，多为两三个短语，较为精炼地指出这篇文章主要优缺点，不直接提出修改建议。

2.5　评语的主题分析及与评分标准的一致性分析

对所有评语及评分标准进行文本分析，共编码 3 295 个，其中 DeepSeek 评语编码 1 529 个、Kimi 评语编码 1 259 个、教师评语编码 459 个、评分标准编码 48 个；从中析出一级主题 4 个（内容质量、写作结构、写作表达、学术规范）、二级主题 18 个。由三者

评语及评分标准中每个一级主题的编码数量占比（图5）可见：教师评语与评分标准对写作的内容质量的关注度相近（41.78% vs. 45.45%），而DeepSeek和Kimi评语对内容质量的关注度均更高（60.18%、75.19%）；DeepSeek评语对学术规范的关注度与评分标准相近（14.21% vs. 15.91%），而教师评语对学术规范的关注度最高（35.16%）。三者评语与评分标准调整后的总绝对偏差分别为24.29%（DeepSeek vs. 评分标准）、36.10%（Kimi vs. 评分标准）、27.48%（教师 vs. 评分标准），即DeepSeek评语和教师评语与评分标准相对接近，而Kimi评语与评分标准的差距相对较大。

根据三者评语及评分标准中涉及每个二级主题的编码数量及占比（表3）可见：DeepSeek评语最关注写作内容是否联系实际（17.46%）、选题及其合理性（14.98%）；Kimi评语最关注写作的深度（23.11%）、选题及其合理性（18.03%）；教师评语最关注写作的深度（25.27%）、引用参考文献的规范性（21.35%）；而评分标准最关注的是写作的基本规范（33.33%）和写作的深度（18.75%）。在析出的18个二级主题中，有6个二级主题（内容的学术性、论据的充分性、写作的逻辑性、情感是否真挚、语言是否简洁、使用术语的规范性）是评分标准中未明确提及的，但2种rcLLMs和教师均给予了不同程度的关注：教师关注了其中3个二级主题，累计关注强度最低（10.24%）；2种rcLLMs均关注了上述6个二级主题，累计关注强度分别为：DeepSeek 28.84%、Kimi 20.02%。

基于自然语言识别技术、以2个字符为最小长度对评语进行分词，在合并同义词的基础上为2种rcLLMs和教师的评语中前200个高频词绘制词云图（图6），结果与对一、二级主题的分析基本一致。

2.6　评语的内容分析及存在的主要问题

2.6.1　关于内容质量的评语

rcLLMs的评语通常会首先归纳作业的选题，但很少评价选题是否符合作业要求（限于文章篇幅，相关示例略去，下同）。rcLLMs比较关注写作是否联系了课程内容，如医学伦理、公共卫生政策等主题或见习经历等，但有时缺乏变通，例如：作业已经结合了课程见习经历，但仍要求增加对课程知识的引用；作业已经结合了医学伦理原则剖析医疗资源分配问题，但仍要求联系分级诊疗制度等内容。

总体来看，rcLLMs的评语能够比较准确地评价作业内容。例如：能准确发现学生在谈及疫苗相关科普问题时，只强调医务工作者的责任，没有结合医学生应该建立的科学素养和批判思维；能准确发现学生对全科医生的理解存在偏差。但也有些评价存在内在矛盾。例如：既肯定了写作具有一定深度，又认为部分观点的展开不够充分；既肯定了写作结合见习经历、内容丰富，又认为结合自身实际和具体案例不足。还有部分评语让学生难以找到改进的平衡点。例如：引用文献太少被评价为论据支撑不足，太多又被评价为缺乏原创性。

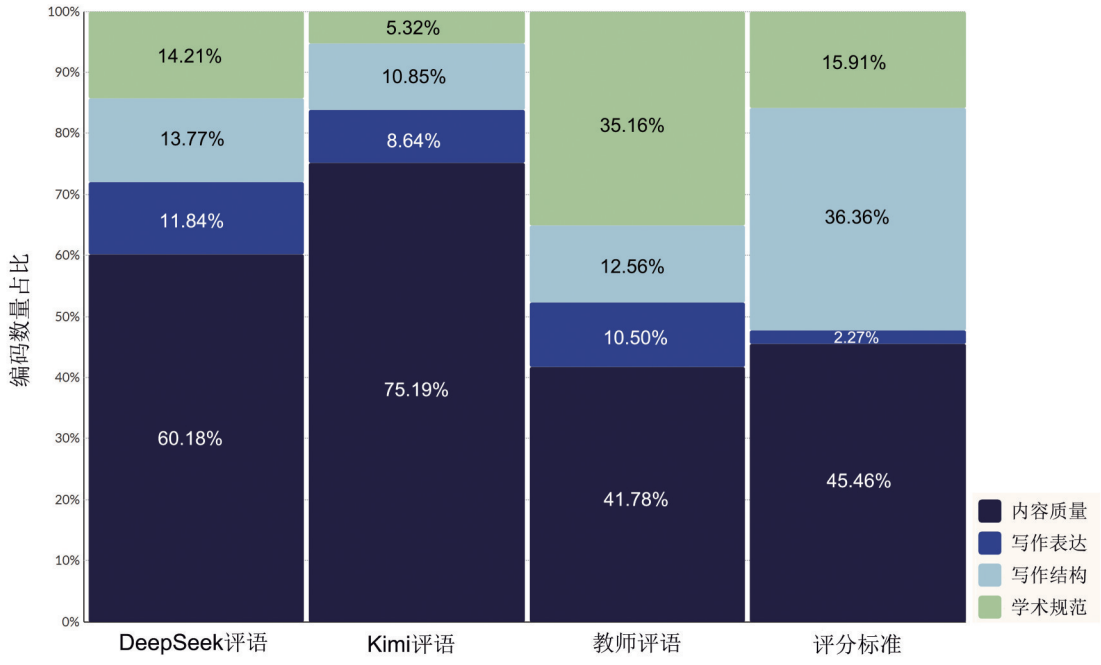


图5　rcLLMs和教师评语及评分标准一级主题编码数量占比的堆积柱状图

表 3　DeepSeek、Kimi 与教师的评语的文本分析及与评分标准的对比[例(%)]

一级主题	二级主题	主题释义	评语或评分标准中涉及该主题的编码数量及占比			
			DeepSeek 评语	Kimi 评语	教师评语	评分标准
内容质量	选题及其合理性	选题是否符合反思性写作要求	229(14.98)	227(18.03)	13(2.83)	1(2.08)
	内容的广度	内容是否全面充实而非简单拼凑	29(1.90)	109(8.66)	23(5.01)	2(4.17)
	内容的深度	是否有自己的分析、反思、感悟，是否有自己的观点并能加以论述，观点是否具有批判性、启发性	198(12.95)	291(23.11)	116(25.27)	9(18.75)
	内容的学术性	内容是否具有专业性、科学性	18(1.18)	11(0.87)	0(0.00)	0(0.00)
	内容的原创性	内容是否具有原创性、创新性、新颖性、趣味性	22(1.44)	24(1.91)	2(0.44)	3(6.25)
	论据的充分性	是否有充足的数据、案例、政策、理论用于支撑论点，论据是否质量高、时效性强	213(13.93)	121(9.61)	12(2.61)	0(0.00)
	内容是否联系实际	内容是否能够结合个人经历、职业规划、课程内容、见习见闻等；内容是否具有实用性和社会价值，能否提出问题的解决方案	267(17.46)	212(16.84)	30(6.54)	7(14.59)
	内容是否涉及特定领域	内容是否能结合医学人文、医学伦理、职业精神或职业素养；是否能结合预防、康复、全科、全生命周期健康观等理念	78(5.10)	34(2.70)	3(0.65)	2(4.17)
写作结构	写作的基本规范	标题与内容是否相符、论文结构是否完整、结构与体裁是否相符、论文格式(排版/字体/字号/行距/图表等)的规范性、是否存在错别字与标点使用问题、字数是否符合要求	53(3.47)	39(3.10)	42(9.15)	16(33.33)
	写作的逻辑性	写作结构是否紧凑、逻辑主线是否清晰、段落过渡衔接是否自然	104(6.80)	61(4.85)	13(2.83)	0(0.00)
写作表达	情感是否真挚	是否有真情实感	49(3.20)	41(3.26)	22(4.79)	0(0.00)
	语言是否简洁	语言是否简洁、重点是否突出	45(2.94)	16(1.27)	0(0.00)	0(0.00)
	语言是否通顺	语言是否优美通顺、文笔是否生动流畅	47(3.07)	23(1.82)	26(5.67)	1(2.08)
学术规范	使用术语的规范性	语言或术语的准确性、专业性、学术性、正式性、可读性、可理解性	12(0.78)	2(0.16)	0(0.00)	0(0.00)
	研究设计的规范性	研究方法是否严谨、结果是否可信、数据是否规范、结论是否科学	65(4.25)	41(3.26)	42(9.15)	1(2.08)
	引用参考文献的规范性	是否规范引用参考文献，参考文献格式是否符合标准、是否统一	62(4.06)	5(0.39)	98(21.35)	2(4.17)
	是否具备学术素养	写作态度是否认真、是否具有学术探索与研究潜质或能力、能否认识到写作的局限性	5(0.33)	2(0.16)	4(0.88)	3(6.25)
	是否存在学术诚信问题	是否存在抄袭网络文章、学术论文的嫌疑；是否与其他同学的作业雷同	33(2.16)	0(0.00)	13(2.83)	1(2.08)



A. DeepSeek　　B. Kimi　　C. 教师

图 6　2 种 rcLLMs 和教师评语的词云图

2.6.2　关于写作结构的评语

rcLLMs 的评语极少指出作业字数上的不足，以及写作标题与内容不符等问题。2 种 rcLLMs 在作业中有图片时，实际均能够读取并识别其中的文本，

也能够读取表格；但无法读取从Word文档中创建的可编辑图表（如：饼图、柱状图等）。

2.6.3　关于写作表达的评语

rcLLMs比较关注写作语言是否简洁生动流畅、情感是否真挚，但DeepSeek有较多评语反映出其不欣赏文学化的叙事，以及比喻、象征等修辞手法。

2.6.4　关于学术规范的评语

对于少数批次上传的作业，rcLLMs的评语中会说明“未发现抄袭”。其他批次上传的作业则无相关表述，是未检测抄袭还是未发现抄袭，尚不明确。个别评语会指出，作业有抄袭网络文章的嫌疑并附链接或文章标题，但这些链接几乎都无法访问（如：链接中部分字符被隐去、链接的文章已删除、需要特定权限才能访问等），多数文章标题也检索不到，无法核实抄袭是否属实。对于同批次上传的作业，rcLLMs偶尔会指出某2篇作业疑似存在雷同，但经核查多为具有相似主题、均不能证实为抄袭；相比之下，教师评语中指出某2位同学的作业存在部分雷同，则通常比较准确。rcLLMs在个别情况下会对判定为严重抄袭的作业评0分，但经核查多为引用内容未规范标注参考文献，其程度尚难以判定为抄袭。此外，还少量存在教师发现了抄袭而rcLLMs未发现的情况。

2.6.5　评语中的其他问题

rcLLMs倾向于在评语中提供对作业进行修改完善的建议，但建议的质量不一。有一些比较具体、有针对性的建议。例如：作业主题为反思临床与科研的关系，评语建议应结合课程中循证医学相关内容；作业主题为调查青少年近视问题，评语建议应进一步探讨近视对学生学习生活的影响，以及对家庭和社会的长远影响；针对家庭医生相关主题的作业，评语建议应结合家庭医生签约服务相关政策。但也有一些建议内容空洞，缺乏针对性或建设性，如“可尝试提出更具前瞻性的观点”“可进一步优化内容，使文章更有深度和条理”等。

2种rcLLMs的评语均存在提出的建议要求过高、不符合大一学生能力水平的问题。例如：对调查设计的规范性、完成问卷的份数要求过高；要求做更深入分析、提出解决方案或独特见解、加强理论深度等。此外，在2种rcLLMs撰写的所有评语中，只有1条评语明确说明了其与评分的关系；另有1条评语建议学生参考同批次上传的另一位同学的作业进行修改。

对比2种rcLLMs与教师对作业的评分和评语，尽管有少量三者一致性较好的情况，但更多情况下三者的一致性并不理想。例如：评语中反映的观点相似、但评分差异较大；或三者对作业的评价存在分歧。

3　讨论与思考

3.1　对评阅结果的分析与讨论

在评分效果方面，DeepSeek、Kimi和教师三者的评分分布总体上比较接近，但三者评分的一致性较差，反映评分者间信度较差；三者两两之间评分的相关性也较差。如果将教师评分视为效标，可以认为2种rcLLMs作为评分工具的效标关联效度较差。类似地，DeepSeek和Kimi各自2次评分的分布总体上比较接近，但2次评分的相关性较差，可以认为2种rcLLMs作为评分工具的复测信度较差。

在评语质量方面，2种rcLLMs都能够撰写结构完整的评语，包括提出修改完善建议，但评语的关注点与评分标准不完全一致。Kimi评语相对于评分标准的偏离程度相对较高，而DeepSeek评语、教师评语的偏离程度相对较低；发生偏离的主题存在差异，2种rcLLMs都更关注内容质量（这正是LLMs擅长的领域），而教师更关注学术规范。在评语内容上，尽管有一些质量较高的评语，但2种rcLLMs均存在对大一学生要求过高、执行评分标准缺乏变通、较少指出写作结构上的问题、不欣赏文学化的叙事和修辞（特别是DeepSeek）、对抄袭的判定不可靠、部分评语内容存在内在矛盾、有些建议比较空泛等问题。因此，还不适合将rcLLMs生成的评语直接用于反馈。

由此可见，现阶段rcLLMs尚不能作为评阅医学生反思性写作作业（无参考答案的长文本类主观题）的有效工具。

3.2　对评阅过程、效果及影响因素的讨论

尽管已有较多文献指出，GenAI可以帮助教师评阅作业[2-12]，乃至实现即时、智能的反馈[2,4,6,8,19]，但实际实施过程并不顺畅。其一，在本研究开始时，多数LLMs尚不提供深度推理功能，对复杂指令的理解和执行（特别是分解工作步骤）效果不够理想；有些LLMs在深度推理模式下不允许上传附件，因而无法用于评阅作业；当时，提供深度推理功能且同时允许上传附件和启用联网搜索功能的模型选择十分有限。其二，rcLLMs一次允许上传的附件数量有限，这与模型上下文窗口（context window）的大小有关，并受到硬件性能、算法架构、运行成本

等因素的限制，因此无法将同一班级乃至同一专业学生的作业进行整体评阅，加之普通教师通常不具备调用rcLLMs应用程序编程接口（application programming interface，API）的能力，不得不将作业分批上传，导致不同批次评阅的同质性存疑。例如：rcLLMs偶尔会指出同批次上传的某2篇作业疑似存在雷同，但对于不同批次上传的作业则无法进行这样的检测，而教师发现此类问题的能力更强。其三，即使rcLLMs允许上传较多数量的附件（如50个），在本研究建立评阅方法的过程中观察到，模型出现明显的“中间丢失”现象（信息召回率下降，导致部分作业漏评）；理论上还可能因“注意力稀释”（关键信息被无关内容淹没）而导致评分不准确和评语质量下降，从而影响评阅效果。其四，个别作业还会被rcLLMs以“可能违反使用规范”为由拒绝评阅，实际上并不存在违规、侵权内容，其具体机制尚不明确。因此，在目前的技术条件下，还不能认为“GenAI帮助教师评阅作业”是完全可行的。

除了rcLLMs的功能限制以外，评分标准也会影响评阅效果。本研究采用的评分标准是由课程负责人基于教学目标并结合自身经验编制的，以描述不同分数段作业特征的形式呈现。这种类型的评分标准为教师提供了评分指南，有助于合理确定作业的分数段，一定程度上约束了评分的随意性。同时，由于没有明确的评分维度，也为教师评分保留了一定的灵活性。然而，这种类型的评分标准也许不太适合rcLLMs，提供更清晰的评分维度、更细致的赋分标准可能效果会更好[20]。这要求教师采用主要特征分析法（primary trait analysis，PTA）、层次分析法（analytic hierarchy process，AHP）等科学方法或基于对过往作业评语的主题分析来构建评分标准，并通过因子分析等方法提前检验评分标准的有效性。此外，本研究提示，教师在执行自己制定的评分标准时也会出现一定的偏差，这反映了评分标准中某些条目的描述不够精准，未准确反映教师的设想。例如：教师对反思性写作作业的设想是不必强求大一新生写出学术标准很高的文章，重点在于鼓励反思、早期培养写作的学术规范，但在评分标准的文字表述中强调不足，因此才会出现教师评语“看似”偏离了评分标准、更多关注学术规范的现象。同样，rcLLMs生成的评语中存在的一些问题也与此有关。例如：教师对于写作要求中“结合课程所学及见习所见”的预期是学生写作时能够有相关结合即可，但评分标准并未说明相关结合的数量要求，这可能导致GenAI在评阅时的理解不同，因此才会出现学生写作时已经结合了一个知识点，rcLLMs还要求结合更多相关知识点或见习见闻的情况。当然，尽管评分标准的描述越精确，GenAI的评阅效果可能越好，但这是否又与反思性写作本身的主观性、灵活性背道而驰，值得思考。

3.3　对评阅结果的进一步思考

本研究结果提示，rcLLMs评阅无参考答案的长文本类主观题，效果尚不满意。从技术角度看，这与LLMs本身处理长文本的能力有限有一定关系，其瓶颈在于高质量长文本预训练数据的稀缺[21-22]。近来有研究者指出，预训练数据质量、长距离信息聚合能力等因素同样重要[23]。因此，相关研究者正在探索合成长文本[24]、负向文档扩展[22]等新策略，来提高LLMs处理长文本的能力。

在基础教育领域，近年来对GenAI评阅作文进行过一些研究，总体上效果略优于本研究评阅医学生反思性写作作业的效果，但仍然存在较明显的局限性[25]。部分研究结果与本研究相仿，包括LLMs评分与教师评分相关性弱、LLMs两次评分稳定性差[26]、LLMs评语与教师评语关注的主题差异较大（LLMs更关注内容与表达、对写作结构的评价精准度较差）[27]等。有研究认为LLMs在评阅中实用性较强的仍是纠正错别字、语病和优化用词等浅层应用[28-30]，但这些研究尚未使用rcLLMs进行评阅，相关研究者认为可以开发专用的作文评阅系统来提高评阅效果[10,25,28-29]。在高等医学教育领域，也有研究者支持开发专用作业评阅系统来提高作业评阅的效率和质量[5,14,19]。应该注意到，中小学作文有较为固定的评价维度且主题相对有限，在开发专用评阅系统方面具有一定的优势。例如：可以针对特定评价维度进行优化训练、根据语料库或知识图谱来评价写作内容或提出写作建议等[27,29]，且开发后可以反复使用；而在高等医学教育领域，诸如反思性写作这样的作业涉猎范围十分广泛、不同课程间要求多变，开发专用作业评阅系统难度较大、性价比较低。

现阶段，如果要使用GenAI帮助教师评阅作业，应当重视人机协作策略的建立[31]。相关研究者指出，应该重视教师评阅、AI评阅和学生自评相结合，使得教师和AI对学生形成双重指导关系，而不是由AI替代教师[28,32]。结合本研究结果，建议教师关注以下人机协作策略[33-37]：①避免不经检视地采信rcLLMs的评阅结果；②对照评分标准逐份检视作业

并修改 rcLLMs 生成的评语，用好评语中概括写作主题或思路、提出修改建议的框架，以减轻教师的书写负担[27,32-33]；③对照修改后的评语进行评分，或调整 rcLLMs 的评分；④使用多种 rcLLMs 进行评阅，同一份作业评分差异较大、评语存在矛盾时要重点复核；⑤涉及抄袭嫌疑的，必须人工复核。未来，随着 rcLLMs 的性能变化，人机协作策略需要动态调整。

此外，教师还需要反思用 GenAI 评阅作业的正当性：①如果认为用 GenAI 评阅作业属于帮助教师应对繁冗事务、减轻工作负担，那么这样的“繁冗”作业对学生有何价值，是否有必要布置？②如果教师批评学生使用 GenAI“应付”写作，那么教师自己是否也不该用 GenAI“应付”评阅？有学者指出，目前教育领域已经出现了一个“教师使用 AI 设计和布置作业→学生使用 AI 完成作业→教师使用 AI 评阅作业→学生因作业缺乏意义而不再重视以后的作业”的“人工智能恶性循环”[37]。这反映出尽管 GenAI 可以帮助教师提高工作效率，但它也会异化教育的核心价值，特别是诱使教师放弃自己的评价主体地位，放任 GenAI 的代具性膨胀为对教师的技术规训，应当引起高度警惕[32-35,38]。因此，当教师准备使用 GenAI 评阅作业时，应当思考：①自己是否处于评阅信息流的中心，是否实施了负责任的评阅[39-43]，保证了评分的可靠性、公平性，确保了反馈的准确性、建设性；②如果作业可以被 GenAI 高质量地批改，是否在某种程度上意味着自己布置的作业所考查的认知层次还不够高、对学生元认知能力的考查还不到位[44-46]。

4 结语

本研究揭示了目前 rcLLMs 尚不能作为评阅医学生反思性写作的有效工具。当然，本研究还存在一些局限性：①选择了一门课一学期的反思性写作作业作为研究样本，样本量稍偏小，代表性较有限；②教师评分采用 2017 年秋季学期的原始评分，由 1 位教师独立完成，作为效标的价值有限；③评语的文本分析由 1 位研究者主要完成，主题析出过程可能带有一定主观性；④未请学生根据评语修改作业，并评价 rcLLMs 生成评语的指导价值；⑤未评价 GenAI 评阅作业的长期效应，如是否会抑制学生的创造性表达、使写作风格趋于同质化等[30]。

未来，在第四次教育革命和第四代医学教育改革的背景下，医学教育工作者可以尝试以个性化教育和促进终身学习能力发展的视角，构建和完善以健康为中心、跨学科交叉融合、促进转化式学习、深度融入 GenAI、促进全方位胜任力发展的新型作业[47]，密切关注 rcLLMs 领域的技术进步和人机协作策略的发展[13-14,48]，将 GenAI 赋能医学教育的视域从帮助教师评阅作业扩展到学生学业评价体系的完善与创新，使学业评价能够更加有效地促进卓越医学人才培养。

利益冲突 所有作者声明不存在利益冲突

作者贡献声明 柴桦：研究选题、搜集整理文献、研究设计与实施、数据分析、论文撰写；闫昱江：研究设计指导、数据分析、修改论文；卿平：政策保障、数据分析、审定论文

参考文献

[1] 黄明芳, 侯青涵, 张韦. 生成式人工智能在医学教育领域的应用现状与未来趋势[J]. 医学与社会, 2025, 38(1): 29-34, 47. DOI: 10.13723/j.yxysh.2025.01.005.

[2] 刘邦奇, 聂小林, 王亚飞, 等. 生成式 AI 赋能教育：技术框架、应用场域及价值：2024 智能教育发展研究报告[J]. 中国电化教育, 2025(3): 61-70. DOI: 10.3969/j.issn.1006-9860.2025.03.009.

[3] 周洪宇, 余江涛. 迈向通用人工智能时代的教师工作：挑战与出路[J]. 电化教育研究, 2025, 46(7): 5-12. DOI: 10.13811/j.cnki.eer.2025.07.001.

[4] 董娜, 陈寞塬, 黄甫全. AI 教师身份何以证成：来自新兴角色理论的回答[J]. 中国电化教育, 2025(5): 42-52. DOI:10.3969/j.issn.1006-9860.2025.05.007.

[5] 王茹, 李毅恒, 孙世仁, 等. 人工智能在医学教育中的应用前景与挑战[J]. 中国医学教育技术, 2025, 39(3): 306-310. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202503003.

[6] 田阳, 王运武, 栾慧敏. DeepSeek 类人工智能变革学习样态：适配逻辑、推动进程与人智协同[J]. 中国医学教育技术, 2025, 39(3): 291-298. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202503001.

[7] 曹伍军. 人工智能时代教师角色的三重转变与应变[J]. 中国医学教育技术, 2025, 39(3): 316-321. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202503005.

[8] 江哲涵, 奉世聪, 王维民. 人工智能生成内容在医学教育中的应用、挑战与展望[J]. 中国教育信息化, 2024, 30(8): 29-40. DOI: 10.3969/j.issn.1673-8454.2024.08.003.

[9] 信斯言, 沈子曰, 吴红斌, 等. 面向未来医学教育 全球健康与人工智能：2024AMEE 年会综述 2[J]. 中华医学教育杂志, 2024, 44(12): 896-900. DOI: 10.3760/cma.j.cn115259-20241006-01031.

[10] 王运武, 周甜, 邬亚楠. 人工智能创作物何以变革教育：基于 ChatGPT 类人工智能的理性思考[J]. 中国医学教育技术, 2024, 38(5): 543-549. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202405001.

[11] 徐傲飞, 刘俭涛, 李嘉, 等. ChatGPT 在医学教育中的应用探讨[J]. 中国医学教育技术, 2024, 38(3): 266-270. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202403002.

[12] 田阳, 王运武, 于燕娟, 等. ChatGPT 类人工智能推动教育数字化变革的失范风险及应对策略[J]. 中国医学教育技术, 2023, 37(3): 247-253, 267. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202303001.

[13] 向巴卓玛, 王珍珍, 畅洪昇, 等. 基于大型语言模型的药理学考试主观题智能评分研究[J]. 中国医学教育技术, 2024, 38(5): 572-579. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202405006.
[14] 赵磊, 田丹. 人工智能何以赋能教育评价: 基于国际研究热点和前沿的分析与启示[J]. 中国医学教育技术, 2024, 38(1): 1-9, 22. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202401001.
[15] DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning [J]. arXiv, 2025: abs/2501.12948. DOI: 10.48550/arXiv.2501.12948.
[16] OPEN AI. ChatGPT: Optimizing Language Models for Dialogue [EB/OL]. (2022-11-30) [2025-07-31]. https://openai.com/index/chatgpt/.
[17] Kimi Team, Du A, Gao B, et al. Kimi k1.5: scaling reinforcement learning with LLMs [J]. arXiv, 2025: abs/2501.12599. DOI: 10.48550/arXiv.2501.12599.
[18] Li J, Miao B, Wang S, et al. Hiplot: a comprehensive and easy-to-use web service for boosting publication-ready biomedical data visualization [J]. Brief Bioinform, 2022, 23(4): bbac261. DOI: 10.1093/bib/bbac261.
[19] 来茂德. AI时代医学教育改革该往何处去[J]. 中国高等医学教育, 2025(4): 1-3. DOI: 10.3969/j.issn.1002-1701.2025.04.001.
[20] 陈曦, 胡中锋. 基于DeepSeek的智能评分: 效度、信度与可行性研究[J]. 高教探索, 2025(3): 62-67. DOI: 10.3969/j. issn.1673-9760.2025.03.014.
[21] 孙维. 文本数据增强:解决NLP数据稀缺问题的5大策略 [EB/OL]. (2024-11-20) [2025-07-31]. https://wenku. csdn. net/column/22vikgq33g.
[22] Gao C, Wu X, Lin Z, et al. NExtLong: toward effective long-context training without long documents [J]. arXiv, 2025: abs/2501.12766. DOI: 10.48550/arXiv.2501.12766.
[23] Xiong W, Liu J, Molybog I, et al. Effective long-context scaling of foundation models [J]. arXiv, 2023: abs/2309.16039v3. DOI: 10.48550/arXiv.2309.16039.
[24] Gao C, Wu X, Fu Q, et al. Quest: query-centric data synthesis approach for long-context scaling of large language model [J]. arXiv, 2024: abs/2405.19846. DOI: 10.48550/arXiv.2405.19846.
[25] 李春明. 基于人工智能技术的智慧校本作业应用研究[J]. 广州开放大学学报, 2024, 24(6): 14-20, 108. DOI: 10.3969/j.issn.1672-0385.2024.06.004.
[26] 魏顺平, 张悦, 冉柔. 国产大语言模型的语文作文评价能力测试[J]. 现代教育技术, 2025, 35(3): 24-33. DOI: 10.3969/j.issn.1009-8097.2025.03.003.
[27] 吴军其, 刘萌, 王嘉桐, 等. AIGC辅助教师作文评价的效果研究: 以九年级语文作文为例[J]. 现代教育技术, 2024, 34(10): 53-64. DOI: 10.3969/j.issn.1009-8097.2024.10.006.
[28] 张泽治. 人工智能融入课堂 创设教育新场景: 以山东青岛崂山区为例[J]. 中小学信息技术教育, 2022(5): 8-10. DOI: 10.3969/j.issn.1671-7384.2022.05.004.
[29] 孙晓莉. 人工智能在小学教育中的伦理与创新实践[J]. 奥秘, 2025(12): 207-209.
[30] 刘梦君, 邹文龙, 宗敏, 等. 生成式人工智能何以赋能教学: 中学生英语写作教学实证研究[J]. 中国电化教育, 2025(6): 70-79. DOI: 10.3969/j.issn.1006-9860.2025.06.010.
[31] 王晶莹, 杜蕾, 张登博. GAI赋能教育变革: 哲学审视、国际比较与未来进路[J]. 中国电化教育, 2025(3): 51-60. DOI: 10.3969/j.issn.1006-9860.2025.03.008.
[32] 郑智勇, 范卿泽, 贾伟. 人工智能技术赋能教师发展的三重幻象及破解之道[J]. 中国电化教育, 2024(7): 28-34, 73. DOI: 10.3969/j.issn.1006-9860.2024.07.005.
[33] 侯浩翔, 王旦. 生成式人工智能时代教师教学创新的风险隐忧及规避路径[J]. 中国电化教育, 2025(3): 20-26. DOI: 10.3969/j.issn.1006-9860.2025.03.004.
[34] 申灵灵, 周一鸣, 卢锋, 等. 真实的虚无: 灰色生态学视域下人工智能与教育的偏离与对齐[J]. 中国电化教育, 2025(2): 29-35. DOI: 10.3969/j.issn.1006-9860.2025.02.006.
[35] 卢秀. 从离身走向具身: 人工智能运用于学习评价的风险及其化解[J]. 中国电化教育, 2025(6): 102-108. DOI: 10.3969/j.issn.1006-9860.2025.06.014.
[36] 姜朝晖. 促进人工智能助力教育变革: 战略任务与政策建议[J]. 中国电化教育, 2025(4): 9-12. DOI: 10.3969/j.issn.1006-9860.2025.04.006.
[37] Scott McLeod. The AI vicious cycle (Revised) [EB/OL]. (2025-04-29) [2025-07-31]. https://dangerouslyirrelevant.org/2025/04/the-ai-vicious-cycle-revised.html.
[38] 焦建利. 高度警惕AI教育应用的恶性循环[EB/OL]. (2025-05-03) [2025-07-31]. https://www.jiaojianli.com/20790.html.
[39] 陈倩倩, 逯行, 张立新, 等. 生成式人工智能教育应用的价值冲突与调适路径[J]. 现代教育技术, 2025, 35(6): 5-13. DOI: 10.3969/j.issn.1009-8097.2025.06.001.
[40] 蒋贵友, 殷文轩. 变革抑或危机: 大语言模型赋能大学教学及其限度: 基于斯坦福大学的案例考察[J]. 电化教育研究, 2025, 46(1): 122-128. DOI: 10.13811/j.cnki.eer.2025.01.017.
[41] 杨满福, 桑新民. 从Deep-Learning到DeepSeek: 人工智能赋能大学功能范式重构的挑战、转型与新生态[J]. 现代教育技术, 2025, 35(4): 5-13. DOI: 10.3969/j.issn.1009-8097.2025.04.001.
[42] 何昌旺, 熊和平. ChatGPT类生成式人工智能教育伦理危机及其应对[J]. 中国教育信息化, 2024, 30(2): 81-90. DOI: 10.3969/j.issn.1673-8454.2024.02.008.
[43] 钟柏昌, 刘晓凡. 生成式人工智能何以、以何生成教育[J]. 电化教育研究, 2024, 45(10): 12-18, 27. DOI: 10.13811/j.cnki.eer.2024.10.002.
[44] 袁欢欢, 全祉悦, 林祎, 等. 当医学课程考核遇上生成式人工智能: 价值、挑战与未来展望[J]. 中国医学教育技术, 2023, 37(3): 254-259. DOI: 10.13566/j.cnki.cmet.cn61-1317/g4.202303002.
[45] 郑庆华. 人工智能赋能创建未来教育新格局[J]. 中国高教研究, 2024(3): 1-7. DOI: 10.16298/j.cnki.1004-3667.2024.03.01.
[46] 李运福. 人工智能赋能高等教育评价改革的国际借鉴[J]. 电化教育研究, 2025, 46(2): 32-40. DOI: 10.13811/j.cnki.eer.2025.02.005.
[47] 巴德年. 从互联网与医学到人工智能与医学: 我的“未来医学”观和医学教育发展观[J]. 中华医学杂志, 2025, 105(29): 2437-2439. DOI: 10.3760/cma.j.cn112137-20250417-00956.
[48] Bearman M, Ajjawi R. When I say … artificial intelligence [J]. Med Educ, 2024, 58(11): 1273-1275. DOI: 10.1111/medu.15408.

（收稿日期:2025-08-05）
（本文编辑:兰明娟）