

基于人工智能的高中地理主观题批阅对比实践研究

李梓萌¹ 姜静² 顾成林^{1*}

1. 佳木斯大学; 2. 佳木斯市第一中学

摘要: 随着生成式人工智能技术的发展, AI 辅助教育评价成为热点。本研究以高中地理主观题为对象, 选取 DeepSeek、豆包、文心一言三款主流 AI 模型, 与教师批阅结果对比, 采用“地理 2025 年辽宁高三 12 月联考”真题及答卷, 覆盖三类题型进行多维度评估。结果显示, AI 显性采分点识别与人工一致性较高 (Kappa 系数 0.58-0.82), 效率提升且模型表现差异显著, 据此提出“分层分类人机协同”模式, 为 AI 赋能地理教学评价提供实践路径。

关键词: 人工智能; 地理主观题; 人机协同; 教学评价

【DOI】10.12252/j.issn.2096-6288.2026.01.131

引言

“双减”推动评价向素养导向转型, 对高中地理教学提出新要求。地理核心素养评价复杂, 主观题需综合空间逻辑与辩证思维, 使教师面临批阅负担重、反馈滞后、评分标准不一、影响信度的多重压力, 负担与一致性的矛盾成为改革难题。目前 AI 虽已应用于部分文科主观题批阅, 但在兼具文理属性的高中地理领域, 系统性对比研究仍较缺乏。人工智能在教育评价领域的应用始于 20 世纪 60 年代, 21 世纪以来, 自然语言处理技术推动作文自动评分系统发展并应用于标准化考试, 近年来大语言模型为开放性试题自动批阅提供新可能^[1-3]。现有地理主观题评价研究多聚焦评分标准与核心素养测评, 实证研究匮乏, 且多关注地理信息技术, 对 AI 批阅效度的检验存在空白^[4]。

本研究以高中地理主观题为对象, 构建数据集并选取多种 AI 模型开展批阅对比实验, 分析 AI 与教师在评分准确性、一致性及效率上的差异, 探究 AI 在不同题型中的优劣, 细化人机协同流程并检验其平衡效果, 旨在客观评估 AI 应用效能, 为智能评价落地提供实证依据, 推动地理教学评价智能化、精准化发展^[5-7]。

一、人工智能批阅的实践

本研究选取佳木斯市第一中学高三学生试卷为样本, 收集描述类、原因类、措施类地理主观题有效答卷各 15 份, 共 45 份, 样本覆盖四个成绩层级。人工批阅由两位具有 10 年以上高考阅卷经验的地理教师独立完成, Kappa 一致性检验系数为 0.89 ($p < 0.01$), 信度极高。实

验参照相关评分细则, 先将手写答卷转为电子文本并经 AI 识别校准, 再结合人工复核完成采分、评分及评语, 既提高批阅效率, 也保障数据真实可靠。

(一) 描述类题目批阅实践

本研究选取以河流水系特征描述 (见图 1) 为核心内容的高中地理综合主观题作为测试题目。



图1

描述晚更新世河套古湖湖水外泄前盆地内河流的水系特征。(3分)

在 AI 批阅任务中, 对模型进行明确指令设定:

您是一位有着 10 年高考阅卷经验、熟悉地理课标与评分细则的高中地理教师。请按照《地理 2025 年辽宁高三 12 月联考》评分细则, 批阅学生地理主观题, 逐点评分、计算总分, 撰写 50-100 字评语 (含问题与建议), 并识别地理术语使用准确性。

得到的结果 (见表 1)

表1 deepseek 评阅描述类题目细则

学生	得分	采分点	评语
同学 1	0 分	无	审题严重偏差。答案完全未涉及河流水系特征, 属于答非所问。建议仔细阅读题目要求, 明确“水系特征”的具体内涵后再作答。
同学 2	0 分	无	概念混淆。将含沙量大、水流速度快等水文特征误答为水系特征, 说明对地理基础概念掌握不牢。建议回归课本, 区分“水系特征”与“水文特征”的差异。
同学 3	2 分	①河流数量较多 ②向心状水系	要点较准但不全。能准确抓住“河流众多”和“向心状”两个核心采分点, 但遗漏了“分布不均”和“流程短”。建议答题时从多个角度全面描述, 争取满分。
同学 4	2 分	①河流数量较多 ②流程较短	要点有得分, 表述需规范。“流域面积较小”可推断流程较短, “梳状”暗示有多条支流, 但“生梳状”表述不规范, 应为“羽状水系”或“支流较多”。建议加强专业术语的准确使用。
同学 5	1 分	①河流数量较多	得分点单一且有错误表述。仅答出“支流较多”一个采分点, “水域范围较大”与正确答案要求的“流程短”相悖, 属于方向性错误。建议仔细审题, 避免与题意相反的表述。

通过对比描述类试题教师批阅与人工智能批阅结果可知，在描述类试题中，AI与教师均能识别核心采分点，对偏离题意作答判定一致，得分趋势相近。但教师对表述规范、概念正误更敏感，模糊赋分更谨慎；AI依赖关键词匹配，对逻辑错误与概念混淆敏感度较低。AI一致性较高，但专业逻辑与术语评判仍需人工校准。

（二）原因类题目批阅实践

表2 豆包评阅原因类题目细则

学生编号	地壳运动	断裂 / 地震	滑坡 / 泥石流	裂隙下渗	冰期冷干	总分	术语准确性	评语
同学 1	0	2	0	2	2	6	规范	抓住冰期、地震、外泄核心点，遗漏构造抬升、滑坡填充
同学 2	2	0	0	0	2	4	准确	答出地壳运动、气候冷干，未涉及地震、下渗等过程
同学 3	2	0	0	0	0	2	偏差	仅认可地壳运动，误答全球变暖，混入人类活动
同学 4	2	2	0	2	0	6	规范	覆盖构造抬升、地震、下渗外泄，遗漏冰期冷干
同学 5	2	2	0	0	0	4	偏差	答出地壳运动、地震，气候背景判断错误

对比人工批阅，AI对地壳抬升等显性关键词采分点识别准确，对答非所问判定与人工一致，但局限性明显：对“滑坡、泥石流填充湖盆”等需因果推理的隐性采分点识别不足，因依赖关键词匹配，无法解读地理过程与隐含逻辑，采分覆盖不全，这是其亟待优化的关键。

（三）措施类题目批阅实践

设问：结合同城化发展经验，为沪杭甬湾区的发展提出建议。（6分）

文心一言给出其批阅结论：在本次答题中，大部分同学表现优异，总分普遍在5分左右。高分答案的亮点集中在对区域发展宏观策略的把握上，如同学1提到的“辐射带动”与“国际合作”，同学2提出的“特色互补”与“信息互联”，以及同学3和5强调的“区域一体化”“创新驱动”等，均切中要点。然而，一个普遍存在的问题是表述较为笼统，缺乏具体的实施细节和路径规划，这也是同学4失分的主要原因。此外，同学5在部分地理术语的准确性上略有欠缺。总体而言，同学们对宏观战略理解到位，但需在结合具体区域、细化实施策

本研究选取原因类典型设问（3）“分析除河流袭夺外，湖泊逐渐消亡的原因。”（8分）作为测试内容，参照上述描述类题目的批阅指令与规范要求，使用AI软件豆包对该题学生答卷开展智能批阅，严格依据联考评分细则完成采分点判定、得分核算、术语准确性识别及针对性评语撰写，具体AI批阅结果如下表2所示。

略方面进一步加强。

文心一言对措施类题目的批阅与教师批阅存在较大差异，文心一言的批阅可能更侧重于答案中提及的要点数量，而教师批阅则更注重答案的完整性、深度和实用性。同学4的答案在教师批阅中明显获得了更高评价，这可能是因为教师认为同学4的答案在细节和政策内容上更为完善，且逻辑清晰、完整。

二、人工智能实践对比与数据分析

本研究以高中地理描述、原因、措施类主观题为对象，用DeepSeek、豆包、文心一言进行模拟批阅。三款AI均以关键词匹配判分，对显性采分点和偏离题意作答的判定与人工一致性较高，能提升阅卷效率，但在地理逻辑推理、隐性知识解读上弱于人工。三者差异明显：DeepSeek术语把控严，适配描述类；豆包识别显性过程采分点较好，适配原因类；文心一言侧重要点覆盖、评分较松，适合措施类快速初评。

整体而言，人工智能可作为地理主观题辅助阅卷工具，但在隐性采分点、地理过程与作答质量深度评判上，仍需结合人工阅卷进行校准与完善，详见表3。

表3 AI与人工批阅评分一致性分析

题型	AI 模型	人工均分 ±SD	AI 均分 ±SD	分差	相关系数 r	Kappa	一致性等级
描述类	DeepSeek	2.1±0.8	1.8±0.9	-0.3	0.85	0.78	高度一致
原因类	豆包	2.3±1.9	4.4±1.7	2.1	0.72	0.65	中度一致
措施类	文心一言	3.5±1.2	4.8±0.4	1.3	0.68	0.58	中度一致

注：p<0.05,p<0.01

三、“人机协同”优化策略构建

（一）当前AI批阅的适用场景

依据与人工阅卷的一致性与可靠性，可将AI地理主

观题批阅划分为不同适用场景。按与人工阅卷的一致性，AI地理主观题批阅可分场景：答案明确的描述类题为高适用场景，AI可高效稳定批阅；常规措施类为中适

用场景, AI能识别要点但需人工复核; 开放创新类题为低适用场景, AI难判隐性逻辑与创新价值, 需以人工主批。

(二) “人机协同”批阅模式设计

结合不同题型与AI适配性差异, 构建分层分类的人机协同批阅模式, 实现效率与质量统一, 详见图2。

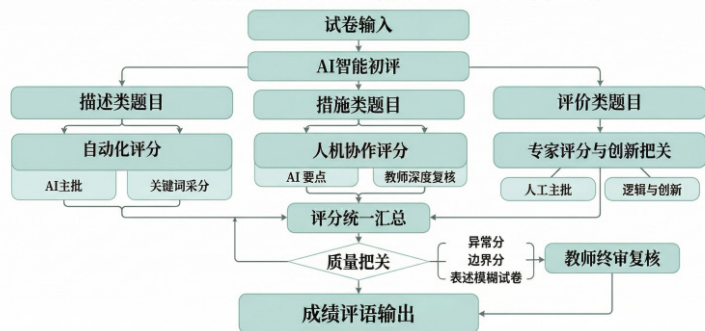


图2 “人机协同”地理主观题批阅流程图

该图呈现的“AI智能初评—分类处理—质量把关”三级人机协同评卷系统, 为教育评价领域人机深度协作提供了可行路径。其核心是按题目类型采用梯度协同策略: 描述类、原因类题目由AI独立评分、人工抽检; 常规措施类题目AI要点评分后教师复核; 开放性题目则人工主导、AI辅助。这种分工打破了全人工或全智能的单一模式, 体现人机各取所长、互补嵌套的核心逻辑。AI承担规模化、标准化任务, 教师专注高阶价值判断, 二者形成良性循环。

此系统所诠释的“人机协同”绝非简单叠加, 而是一种基于能力边界与任务特性的深度协作范式, 最终实现“AI高效辅助与人工精准把关的优势互补”, 对推动教育评价数字化转型具有重要实践价值。

(三) 教师AI素养提升建议

为保障人机协同阅卷落地, 需提升教师AI应用素养。强化批判性使用AI的意识, 明确AI辅助定位, 不盲目依赖机器结果。开展提示词工程基础技能培训, 使学员掌握精准规范的指令编写方法。提升人机评分差异识别与处理能力, 能够修正AI评分误差, 对隐性采分点、创新作答进行合理赋分, 保证阅卷质量。

结语

本研究对比DeepSeek、豆包、文心一言三款AI与教师的批阅效果, 发现AI显性采分点识别效率高, 一致性Kappa值0.58-0.82, 但在隐性地理逻辑、术语规范及创新思维评价上存在不足。三款模型各有特点: DeepSeek术语把控严, 豆包地理过程识别优, 文心一言要点覆盖广但深度不足。据此提出“分层分类、人机互补”协同策略, 以AI处理标准化环节、教师把控高阶评价, 兼顾效率与信度, 为智能工具选用及模型优化指明方向。

总体而言, 人工智能可作为高中地理评价的高效辅助手段, 但必须在人机协同框架下与教师专业判断结合。只有技术效率与人文评判相互支撑, 才能实现素养评价的智能化、精准化与高效化统一, 助力“双减”背

景下地理教学评价改革深入推进。

参考文献

[1] Attali Y, Burstein J. Automated essay scoring with e-rater® V.2[J]. The Journal of Technology, Learning and Assessment, 2006, 4(3): 1-30.

[2] Elliot S. IntelliMetric™: From here to validity[C]. Machine scoring of student essays: Truth and consequences. Utah State University Press, 2006: 147-168.

[3] 祝智庭, 胡姣. 智能教育的发展趋势与关键特征[J]. 开放教育研究, 2022, 28(1): 13-25.

[4] 段玉山, 周义钦. 地理新课程教学方法[M]. 北京: 高等教育出版社, 2017: 156-162.

[5] Holstein K, Aleven V. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms[J]. Artificial Intelligence in Education, 2019: 154-166.

[6] Baker R S. Challenges for the future of educational data mining: The Baker learning analytics prizes[J]. Journal of Learning Analytics, 2019, 6(1): 1-7.

[7] 祝智庭, 彭红超. 智能教育: 人工智能时代的教育变革[J]. 教育研究, 2020, 41(4): 16-22.

作者简介: 李梓萌, 女, 2002年2月出生, 汉族, 黑龙江省双鸭山市人, 硕士, 研究方向为地理综合评价与考试测量。

通讯作者: 顾成林, 男, 1978年5月出生, 汉族, 黑龙江省齐齐哈尔人, 博士, 副教授, 研究方向为地理学教学研究。

基金项目: 黑龙江省高等教育教学改革研究项目“融入信息技术的地方高校地理科学专业野外实践教学模式研究”(编号: SJGYB2024492)。