计算机与智能教育技术

2026年11月 第1卷第4期

智能体赋能教育评价：应用模式与评价效能研究

向惠阳，聂新明

江苏师范大学物理与光电工程学院，江苏徐州，221116

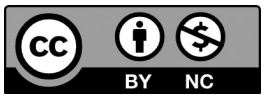
摘　要 | 现有智能体辅助评价研究多聚焦评分一致性与反馈接受度，对评价环节之间的连接关注不足。本文以证据采集—证据解释—反馈生成—后续干预为分析框架，基于科大讯飞“智学网”、江西省“智慧作业”与北京中学英语写作智能体三个公开案例的资料，并以“前智能体平台—LLM智能体”的演进视角加以对照，结合形成性评价理论开展案例资料对照与理论分析。智能体经由证据连续化、诊断即时化、反馈闭环化三条路径连接离散的评价活动；其应用可归纳为单智能体辅助、多智能体协作、角色化反馈与人机协同闭环四种模式，四者在评价连贯性上呈层级递进。评价连贯性以人机合理分工为前提，智能体承担标准化初评与反馈草拟，教师保留价值判断与教育决策权，并需标准与权责制度化、教师判断力维护、场景驱动的数据迭代三方面配套。上述结论属基于公开资料的分析性概化，所提路径为规范性建议，效能尚未经大样本教学试验检验。

关键词 | 教育智能体；教育评价；形成性评价；人机协同

Copyright © 2026 by the author(s).

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

https://creativecommons.org/licenses/by-nc/4.0/



1 研究现状与传统评价的碎片化问题

近年来，随着人工智能技术从内容生成向任务执行演进，智能体开始深度介入教育评价领域。基于大语言模型的智能体具备目标分解、工具调用与多步推理能力，正在重塑评价流程的整体样貌。纵观国内外相关文献，当前学界在智能体辅助评价方面的研究主要集中于两个维度：一是评分自动化的一致性检验，即探究智能体依据既定标准给出的分数或等级与人类教师的一致性程度；二是反馈感知层面的效果评估，即调查使用者对智能体生成内容的接受度[1]。随着生成式大模型介入评价，自动作文评分（AES）研究进一步表明，GPT-4等模型与人工评分具有较高相关性，其反馈的具体性与相关性因评分维度而异[2]；但模型作为评价者仍存在位置、冗长等偏差，评分一致性与构念效度不可混同。这些研究为教育评价的精准化与效率提升提供了重要依据，也证明了智能体在复杂教育场景中具备较高的技术可行性。

然而，在追求单点评价精度与效率的同时，现有研究往往忽视了传统教育评价长期存在的一个结构性困境：评价活动通常是相互孤立与碎片的。无论是在课堂观察、日常作业还是阶段性测验中，评

基金项目：江苏师范大学 “人工智能 +” 应用场景立项项目（基于多模态融合与移动式 VR 平台的师范生教学能力动态诊断与训练应用案例，项目负责人：聂新明）。

作者简介：向惠阳，江苏师范大学物理与光电工程学院，学科教学物理研究生，研究方向为中学物理课程与教学；聂新明，江苏师范大学物理与光电工程学院，硕士生导师，研究方向为物理课程与教学论。

文章引用：向惠阳，聂新明. 智能体赋能教育评价：应用模式与评价效能研究[J]. 计算机与智能教育技术, 2026, 1(4): 166-173.

https://doi.org/10.68070/rs4ps737

价往往在生成分数或简短评语后便宣告结束[1][3]。形成性评价理论指出，评价的本质价值在于为后续的教学干预提供依据，但在现实操作中，从获取评价证据到采取教学调整之间，通常存在明显的信息滞后与流失。教师获取的往往是高度压缩的结果性数据，学生接受到的也多为缺乏操作指引的终结性反馈[4]。这种各自为政的评价状态，使得过程性数据难以转化为系统的教学改进线索。

如果智能体的介入仅仅停留在提升单一环节的评分效率上，其结果不过是让原本孤立的评价活动发生得更加频繁，无法从根本上改变评价体系的碎片化现状。因此，本研究认为，智能体赋能教育评价的核心突破口，不应局限于工具层面的效率提升，而应聚焦于其连接能力的构建，即将零散的评价动作整合为持续追踪、反馈与改进的学习支持过程。

需要特别界定的是，学界常以“目标分解、工具调用与多步推理”刻画以大语言模型为核心的自主智能体的能力特征[5]：目标分解指将总体任务自主规划为子任务，工具调用指依据环境反馈选择并调用外部工具，多步推理指基于中间结果迭代修正行动。但需强调，这三项是判别自主性的参照维度，而非只要出现大模型或多个节点便自动成立，关键在于任务如何拆解、由谁决定下一步动作。本文所称“智能体”即在此严格意义上使用，特指以大语言模型为核心、具备上述能力的自主智能体（LLM-based AI Agent）[6][7]。已有研究亦开始探索GAI赋能学科教学智能体的开发[8]。基于图像识别、点阵笔采集、题库标签匹配与规则推送的智慧教育平台，本质上属于数据驱动型评价平台或个性化推荐系统，并非严格意义上的LLM智能体。考虑到评价链条的演进具有连续性，本文将此类平台作为“前智能体阶段”的技术底座与对照案例保留，并补充一个由一线教师基于通用大模型自主搭建的写作评价智能体作为核心案例，以保证概念边界与实证材料的一致。为落实上述界限，第2.1节将逐项列明各案例中“教师预先编排的步骤”与“模型运行时作出的决策”，并据此区分“大模型支持的工作流”“单智能体”与“多智能体协作”，不把多个工作流节点直接解释为多个智能体协作。

2 研究设计

为避免理论建构与模式归纳流于思辨，本研究选取三个真实落地的应用案例，开展多案例比较研究。分析围绕“证据采集—证据解释—反馈生成—后续干预”这一评价链条展开，为后文的作用机制阐释与应用模式归纳提供经验依据。

2.1 案例选择与数据采集

评价环节之间的连接机制在真实情境中如何运作、评价效能如何形成，属于“如何”与“为什么”的问题，正适合以案例研究作答。本研究据此采用以质性分析为主的多案例比较设计：按照理论抽样逻辑选定案例，先逐一刻画各案例中智能体评价的运作机制与效能表现，再进行跨案例比较，提炼共性规律与差异条件。

案例资料检索于 2024 年 9 月至 2026 年 6 月在中国知网、教育部及江西省教育厅官网、企业公开技术文档与《人民日报》等渠道完成，检索词为‘智学网/智慧作业/写作智能体＋物理/作业/作文评价’，仅纳入可溯源至原始出处的资料，按案例分别编号存档。

据此确定三个案例。案例A是科大讯飞“智学网”在初中物理作业与测验评价中的应用[9][10]。该系统面向学校日常考试与作业场景，学生完成后将作业扫描上传，系统基于图像识别技术解析卷面文字、图形与公式，自动批改客观题及部分主观题，批改结果逐次沉淀为持续更新的个人学情数据库；在此基础上，系统绘制知识点掌握图谱，生成班级与个体两级学情报告，并据此为每位学生推送个性化学习手册，汇集专属错题、变式巩固题与配套讲解资源。

其评价成效已有可查证的学科研究：一线教师围绕智学网阅卷数据在初中物理试卷讲评中的运用开展了研究[9]，另有研究讨论了教育智能体工具为增值评价提供数据支撑的路径[10]。需说明的是，这两项均为篇幅简短的期刊文章，并非学位论文；其中刊于《小学科学》，聚焦增值评价的数据支撑，并非初中物理教学研究，故不宜据此推断该系统已在初中物理情境中经过前后测对照检验；上海市徐汇区一项面向初三物理总复习的实践研究亦报告，依托智学网测评诊断个体薄弱点并推送个性化资料后，实验班平均提分明显高于对照班。在常态化应用层面，南师附中江宁分校初中物理教研组自2018年起将智学网嵌入日常考试与作业评价，依托其逾两百万道试题的标注题库开展组卷、学情图谱分析与个性化推荐，为本研究的模式刻画提供了丰富的运行细节。

案例B是江西省依托“赣教云”平台建设的“智慧作业”工程，属于区域层面的人机协同作业评价实践[11][12]。该工程不改变学生在纸质练习册上作答的习惯：教师以点阵笔在铺有点阵码的练习本上批改，作答信息经笔端摄像头实时上传平台，亦可经高拍仪扫描采集；后台即时统计每道题的作答情况，为每名学生生成专属错题集，同步推送错题讲解微课与举一反三巩固练习，学生在家通过电视机

https://doi.org/10.68070/rs4ps737

顶盒即可免费观看。

工程规模亦有权威数据支撑：截至2022年，全省使用学校达7359所、注册师生479万人，常态化使用教师2万余名、学生100万余名；配套作业微课52.8万余节，由8万余名教师参与录制，覆盖省教育厅教辅目录上全部练习册的作业题目[11]。2021年9月，教育部办公厅发布《关于推广学校落实“双减”典型案例的通知》（教基厅函〔2021〕37号），江西省“智慧作业”以技术赋能案例入选十大典型案例[13]；该通知是对中办、国办“双减”意见的落实，二者并非同一文件。

案例C是北京中学英语教师基于飞书aily平台搭建的“北中英语赵老师智能体”，属于由一线教师自主构建的生成式AI评价智能体[14]。该智能体的流程由教师在搭建阶段预先设定：教师将二十余年中考阅卷经验、四档评分标准、高分句型与逻辑词汇写入提示词，设置条件分支节点分别接收文本与手写图片作文，再交由大语言模型节点生成评分与修改建议，并将交互结果写入多维表格。智能体上线10天完成390次对话，服务70余名学生；学生依据反馈反复修改二至三次，直至达到9—10分（满分10分）。交互记录同步至多维表格形成结构化学情数据，教师据班级70%—80%学生在定语从句上的共性薄弱点开展专项讲解，将机器初评转化为数据驱动的教学决策。

2.2 数据分析流程与效度信度保障

资料分析采用三级编码：先按案例归类材料并撰写内容摘要，再进行开放式编码与主轴编码，围绕评价链条的四个环节归纳各案例中智能体评价的运作机制；在此基础上构建跨案例比较矩阵，从评价环节覆盖范围、技术架构复杂度、评价连贯性与闭环水平等维度比较三个案例，提炼各类模式的形成条件与层级差异。如图1所示，其中所述的四个评价环节，在前文提及的案例A、案例B、案例C均可找到对应证据。

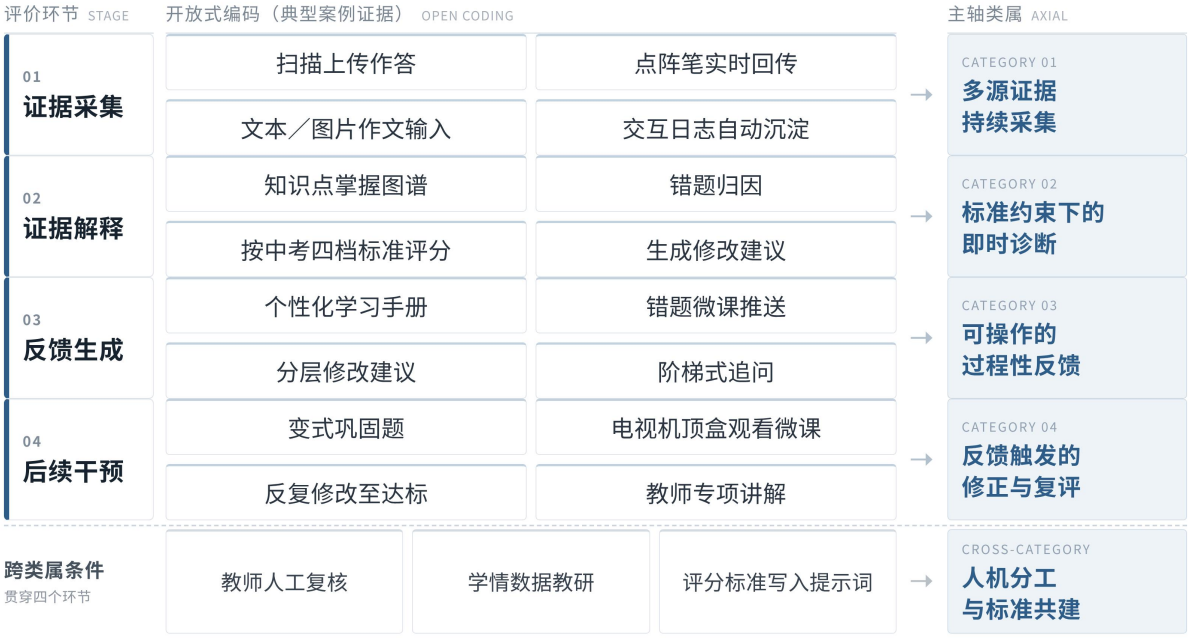


图 1 开放式编码与主轴编码结果

为保障研究质量，分析遵循案例研究的效度规范：观察维度统一锚定评价链条的四个环节，关键结论均以官方报告、期刊论文与媒体报道等多源证据相互印证；机制解释采取模式匹配的策略，并主动排查竞争性解释；结论推广不依赖统计概化，而是借助学校、区域与课堂三级案例之间的复制逻辑，谨慎划定适用边界。资料收集与分析全程依循事先拟定的案例研究草案，编码记录完整留存，研究过程可被重复检验。资料均为公开文本，未开展课堂观察与师生访谈，效能判断依据既有报告，结论属分析性概化。

3 理论依据与智能体连接评价的作用机制

将智能体定位为评价活动的连接者，须有坚实的理论支撑。从形成性评价的视角看，如图2所示评价是由证据采集、证据解释、反馈生成与后续干预构成的闭环，而反馈研究也一再表明，指向学习过程与自我调节的反馈比单纯的成绩判定更能引发学生的深度认知参与。这为智能体的功能设计划定了方向：其价值不在于算出一个准确的分数，而在于能否在恰当的教学节点给出兼具诊断价值与行动指引的过程性反馈。



图 2 智能体连接评价链条的作用机制示意图

3.1 智能体实现评价连接：证据的连续化

传统学业评价长期以终结性测评为主，依托单次测验、作业成果等静态结果作出判定，本质上是抽取学习者某一时刻的表现快照：既分不清短期记忆带来的偶然高分与认知稳步建构形成的稳定能力，也丢掉了学生在试错、调整、反思中的发展信息，评价结论因而片段化、滞后化[15]。智能体则能在学习场景中持续、非侵入地采集作答行为、交互记录、资源访问等多源信息，按时间线索将其清洗、对齐并整合为连贯可追溯的学习轨迹。评价基准由此从某一时间截面的瞬时表现，升级为具有时间纵深的趋势判断——时序分析能够解释结果背后的形成过程，为分层干预与个性化学习支持提供客观依据。

就本文案例而言，案例A以扫描件持续沉淀个人学情数据库，案例B以点阵笔实时回传作答数据，案例C则以多轮对话日志自动留存修改轨迹，三者共同体现了评价证据由静态结果向连续轨迹的转变。

3.2 智能体实现评价连接：诊断的即时化

标准化初步诊断最能体现智能体的技术优势。学生的认知偏差常在练习与问答中即时产生，教师受精力与负荷所限，难以对每一次学习输出都及时识别，错误认知往往在等待反馈的间隙里不断固化，后续纠偏要付出更高的教学成本。智能体可以高频捕捉作答、交互、资源操作等学习信号，按预设规则快速完成初筛与归因，把反馈介入压缩到错误尚未固化之前[16]。这种即时干预并不取代教师的专业判断——它承担的是批量初筛，优先识别共性错误与典型问题；教师由此得以将精力集中于复杂认知误区、情感动机与高阶思维等更需要研判的环节。

案例A的知识点掌握图谱、案例B的即时错题统计与案例C的评分节点，均承担了批量初筛功能；但只有案例C的生成式建议能够依据学生具体作答给出可操作的修改方向，诊断的即时化由此迈向解释的可理解化。

3.3 智能体实现评价连接：反馈的闭环化

反馈要真正促学，就要接入后续环节。反馈的教育价值并不取决于信息的完备程度，而在于它能否改变学生接下来的学习行为：一条再精准的诊断，若只停留在信息告知，不能顺势衔接修改订正、二次作答与复评，学生看过便放下，错误理解依旧留在原处。智能体的意义正在于把反馈从“告知”推进为“触发”，趁学生仍保有解题时的情境记忆，把订正任务与同类变式推送出来，使纠错发生在遗忘之前。依托时序化学习轨迹，系统能够追踪学生基于反馈开展的修正行为，记录订正内容与二次作答的正确率变化，据此判断错误是否真正得到纠正，也区分哪些差错只是偶发笔误、哪些反复出现并指向深层概念偏差。就在这样的往复中形成诊断、反馈、修正、复评相互衔接的循环，复评的结果又成为下一轮诊断的起点，过程性评价的效能便在一轮轮闭合中落到实处。

案例C中学生依据智能体建议反复修改二至三次直至达标，教师再据交互数据开展专项讲解，较为完整地呈现了“反馈—修正—复评—干预”的闭环；案例A、B则分别以变式题和错题微课实现闭环的初级形态。

https://doi.org/10.68070/rs4ps737

4 智能体参与教育评价的典型应用模式

综合国内外技术实践与相关文献，智能体参与教育评价的落地应用可归纳为四种典型模式。四类模式在技术架构复杂度、功能覆盖范围与评价连贯性上呈清晰的层级递进，对应着不同教学场景的应用需求与技术成熟度。



图 3 跨案例比较矩阵与模式对应图

由图3可见，案例A在功能上对应模式一（单智能体辅助评价），并开始向模式四过渡，但其技术基础是图像识别与题库标签匹配，并非LLM智能体；案例B具备角色化反馈与初步人机协同特征，可视为模式三、模式四的初级形态；案例C作为严格意义上的LLM智能体，承载模式一的评分效率与模式四的人机协同闭环，是四种模式层级递进的经验注脚；其内部虽含条件分支、评分与结果写入等多个节点，但均属单智能体内部由教师预设的工作流环节，并非多个智能体之间的协作，不能据节点数量将其归入模式二。需要强调的是，案例A、B之所以纳入比较，是为呈现评价链条从“前智能体平台”到“生成式智能体”的连续演进，而非将其直接等同于LLM智能体。

单智能体辅助评价是最基础的入门形态，定位于单点环节的效率工具，聚焦客观题自动批改、标准化作业评语生成等规则明确的单一任务；技术上多依托检索增强生成框架，将评分细则与参考答案嵌入检索库，以约束模型输出。这一模式在特定任务上评分一致性与可控性较强，落地门槛低，能快速替代人工完成高重复性的标准化评阅，明显缓解大班额与规模化在线课程中的教师负担。但它终究只是对传统评价单点环节的技术替代，不同作业、测验、阶段性评价的结果彼此独立，难以撼动评价碎片化的整体结构。

当评价对象转向开放性论述、项目式成果等主观性较强的复杂任务时，单智能体的短板逐渐凸显，多智能体协作评价由此应运而生。该模式通过角色分工与任务拆解构建协作拓扑：既可让多个同质智能体并行评审、交叉核验以降低个体偏差，也可按评价流程串行衔接不同职能的智能体，分别承担证据提取、维度比对、偏差核查与规范性审核等子任务[17]。这一思路与多智能体辩论（Multi-Agent Debate）及ChatEval等多智能体评审框架的结论一致：多个模型实例经由独立应答、相互质疑与迭代修订，可提升事实性与推理的可靠性，异质角色与多智能体协同机制亦被证明有助于降低单一评价者的偏差。多角色交叉校验显著提升了定性评价的稳健性，能够覆盖知识掌握、逻辑表达、创新思维等多个维度，适配开放性作答与实践类作业等复杂场景。不过其协作大多仍限于单次评价任务内部，评价结果与后续个性化练习、针对性辅导之间并未建立联动，评价与学习过程依然相互割裂。

角色化反馈智能体突破了“智能体只管评分”的

窄化定位，把公平性监控与元认知支持直接嵌入评价系统：独立的公平监控智能体持续监测不同群体间的评分分布差异，及时识别潜在的算法偏见；反思引导角色则以阶梯式提问引导学生归因错误、自我评估与调节学习策略，让评价过程本身成为元认知训练的载体。偏差监测由此从偶发的事后抽查转为评价流程的常规环节，反馈的教育附加值也随之拓展。但这类干预大多发生在评分结果生成之后，仍属事后补救，难以在评价标准设计与任务开发的源头消除结构性偏差，反馈与后续学习任务的联动亦显不足。

人机协同闭环评价是目前最契合“评价—学习深度连接”理念的高阶形态：学生作答后由智能体完成即时初评与诊断反馈，快速定位知识漏洞；教师结合学情开展情境化介入与最终审定，补充机器无法覆盖的情感、价值与情境判断；再引导学生进入修正学习与二次复评。人机互补弥合了机器认知盲区与算法不透明性，保障了评价的教育性与连贯性，但对教师的数字素养、专业能力与时间投入要求较高，缺乏配套制度支撑时，在规模化教学场景中容易退化为形式化审核。

整体而言，四种模式依次递进，勾勒出智能体从单点效率工具走向系统连接者的演进路径。评价效能的提升不仅取决于单点技术的先进性，更取决于模式设计是否把评价后的学习改进纳入系统结构，形成前后衔接、持续迭代的评价与学习闭环。

5 推动评价走向连贯的实施路径

将智能体在学习评价中的连接潜力转化为可持续、可推广的教育实践，无法仅凭技术层面的功能迭代实现，而需要构建覆盖标准、权责、主体与机制的系统性保障体系，推动技术逻辑与教育评价规律深度契合，避免智能评价系统陷入“技术先进但落地失效”的困境。

5.1 评价标准与人机权责的制度化构建

评价连接的质量，根本上取决于评价标准的清晰度与可执行性。传统评价量规多停留在抽象的能力维度描述层面，无法直接支撑智能体完成 “证据识别 — 等级判定 — 反馈生成” 的连贯逻辑。要实现评价全链路的有效衔接，推动评价标准向学科化、颗粒化、可操作化转型：明确界定不同表现水平对应的可观测学习证据，建立判定结论与干预策略的对应映射规则，让每一项评价判断都有明确的行为数据作为支撑。这一过程需要学科教育专家与技术开发人员深度协同，将学科核心素养的评价要求转化为可被系统识别的数据结构与判定逻辑，从源头减少因标准模糊导致的智能体输出随意化、反馈表层化问题，保障评价的构念效度与实践价值。

与此同时，以制度化方式明确人机权责边界，筑牢教师在评价中的主导地位。随着智能体承担的评价任务从简单客观题批改向复杂认知诊断延伸，若缺乏清晰的流程规范，极易出现技术越位、教师专业判断被架空的风险。制度设计需清晰划定智能体可独立输出的任务范围，明确经由教师人工确认、复核的关键评价节点，并建立人机评价分歧的仲裁流程与算法评价后果的追溯机制。制度化的核心并非限制技术应用，而是守住评价的教育属性底线，确保在规模化教学场景下，教师始终保有实质性的评价决策权，避免评价异化为脱离育人目标的纯技术化数值判定。

5.2 教师专业评价判断力的维护与发展

教师的专业评价判断力是智能评价系统长期健康运行的核心支撑，也是技术落地过程中极易被忽视的隐性风险点。随着智能体逐步承接大量标准化初评、反馈起草等重复性工作，教师独立开展完整评价实践的频次会相应降低。若长期依赖智能体的初步诊断结论开展形式化审核，教师对复杂学习表现的辨识能力、对学生认知状态的权衡判断能力将面临逐步退化的风险，最终削弱整个评价体系的专业根基与人文温度。

因此，在系统落地设计与日常教学流程中，同步配套能够维持并发展教师评价能力的保障机制。一方面要设置实质性的深度审核环节，避免审核沦为无意义的“一键确认”，要求教师对典型案例、边界案例、争议案例进行完整的独立研判，保留专业判断的实践场景；另一方面要依托智能体积累的全量评价数据开展常态化教研活动，组织教师对智能体输出结果进行批判性审视与集体研讨，将技术工具转化为教师专业发展的载体，在人机协同的实践中持续提升教师的评价素养与数字教学能力。

5.3 教学场景驱动的数据迭代闭环

智能体评价系统的持续优化，不能仅依靠技术人员在后台进行封闭式的参数调试，而应建立扎根真实教学场景的开放式数据迭代机制。脱离一线教学实践的技术优化，往往会导致系统功能与真实学情、教学需求渐行渐远，最终沦为脱离教育场景的技术展品。

真正可持续的迭代闭环，需要将教学全流程产生的实践数据反向输送至系统优化的前端：具体包括学生对反馈的采纳率与修正效果、后续学习表现的变化轨迹、教师在审核过程中标记的系统性偏差与典型误判案例、不同学情群体的评价适配性数据等。通过在系统架构中内置常态化的数据采集与回流通道，将一线教学的真实反馈转化为评价标准修

https://doi.org/10.68070/rs4ps737

订、反馈策略优化、模型能力迭代的核心依据，形成教学实践、数据沉淀、系统优化、回归实践的持续循环，让智能评价系统始终适配真实的教育教学需求，伴随教学实践的发展共同演进。

需要说明的是，本文的机制阐释与模式归纳建立在三个案例的公开资料与理论推演之上，尚未经大样本教学试验检验，所提实施路径属规范性建议而非实证结论，其有效性有待后续研究检验。

6 结语与研究展望

综上所述，智能体赋能教育评价的核心价值，并非评价精度的单点突破，而是对离散评价活动的连接与重构能力。长期以来，传统教育评价受限于数据采集能力与组织形式，各阶段测评相互独立、学习过程信息难以留存，评价始终局限于终端结果判定的单一职能。智能体通过实现证据连续化、诊断即时化与反馈闭环化，打通了学习数据采集、认知状态诊断、干预反馈落地的完整链路，将原本孤立的结果性评价转化为可持续支持学习改进的过程性评价。这一价值的落地以人机合理分工为基础，既不能完全依赖机器替代教师的专业判断，也无法仅靠人力实现规模化的过程追踪，其效能发挥最终依赖标准建设、制度规范、能力维护与数据迭代多维度的系统性支撑。

围绕评价连贯性这一核心目标，未来研究可从三个方向推进：一是测量对象的转移。现有相关研究多聚焦于单次评价输出的评分信度与结果精度，对评价活动的链式育人效应关注不足。后续可将研究焦点从单点评价质量转向评价链条的整体运作机制，重点考察反馈信息对后续学习行为的长期影响，以及评价干预对学生知识迁移能力、元认知水平的实际作用效果，完整衡量智能评价的综合教育价值。二是模式比较的规范化。当前智能体评价应用模式多样但缺乏统一的效能参照，多数研究仅针对单一模式开展案例验证，难以支撑实践场景中的科学选型。后续可在严格控制任务特征与评价标准的前提下，对不同智能体应用模式的效能开展系统横向比较，明确各类模式的适用场景、成本收益与效能边界，为不同教学场景下的模式选择提供实证依据。三是公平性与可解释性的前置设计。现有研究对评价公平性的关注多集中于输出结果的事后审核与偏差修正，属于被动的补救式治理。未来可将社会价值审视与偏差排查前置到评价标准构建、系统架构设计与模型训练的全流程中，从源头降低算法偏见与结构性偏差的产生风险，而非局限于输出结果的事后审核。

教育评价的智能化是一个循序渐进的长期演进过程，不可能一蹴而就。当前智能体在连接评价环节、支持持续学习方面的技术潜力已逐步显现，但其教育效能的充分释放无法仅靠技术迭代实现，仍有赖于技术、制度与教育实践三者的深度协同，最终形成技术适配教育规律、制度保障育人方向、实践反哺系统优化的良性循环。

参考文献

[1] 郑娅峰, 赵亚宁, 黄璟玥, 等. 教育智能体: 研究现状和发展趋势[J]. 现代远程教育研究, 2025, 37(4): 3-13, 59.

[2] MIZUMOTO A, EGUCHI M.Exploring the potential of using an AI language model for automated essay scoring[J]. Research Methods in Applied Linguistics, 2023, 2(2): 100050.

[3] BLACK P, WILIAM D. Assessment and classroom learning[J]. Assessment in Education: Principles, Policy & Practice, 1998, 5(1): 7-74.

[4] NICOL D J, MACFARLANE-DICK D.Formative assessment and self-regulated learning:a model and seven principles of good feedback practice[J]. Studies in Higher Education, 2006, 31(2): 199-218.

[5] XI Z H, CHEN W, GUO X, et al. The rise and potential of large language model based agents:a survey[J]. Science China Information Sciences, 2025, 68(2): 121101.

[6] 李艳, 陈琳, 马学纯, 等. 智能时代教育的变革、发展与治理——2026世界数字教育大会综述[J]. 远程教育杂志, 2026, 44(3): 28-36.

[7] 教育部等九部门. 关于加快推进教育数字化的意见：教办〔2025〕3号 [EB/OL]. (2025-04-11) [2026-11-11]. https://www.moe.gov.cn/srcsite/A01/s7048/202504/t20250416_1187476.html.

[8] 姚雪梅, 罗骁, 周维国, 等. GAI赋能高中地理启发式提问智能体开发——以三江源冰冻圈教育为例[J]. 冰川冻土: 1-12.

[9] 陈镇华. 核心素养视角下基于考测数据的初中物理试卷讲评策略[J]. 新课程研究, 2025, (17): 13-15.

[10] 吴萍. 教育智能体工具为增值评价提供数据支撑[J]. 小学科学, 2026, (24): 97-99.

[11] 江西省加强教育信息化建设促进优质资源普惠共享 [EB/OL]. 中国教育信息化网, 2022-01-07[2026-11-11]. https://www.ictdedu.cn/html/special/2022/0107/3733.html.

[12] 中共中央办公厅, 国务院办公厅. 关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见：中办发〔2021〕40号[EB/OL]. (2021-07-19)[2026-11-11].https://www.gov.cn/zhengce/2021-

07/24/content_5627132.htm.
[13] 教育部办公厅. 关于推广学校落实“双减”典型案例的通知: 教基厅函〔2021〕37号[EB/OL]. (2021-09-17)[2026-11-11]. https://www.moe.gov.cn/srcsite/A06/s3321/202109/t20210926_567037.html.
[14] 温婧, 黄依璇. 一个北京英语老教师的“数字分身”: 飞书的AI教学实践[N]. 北京青年报, 2026-04-20.
[15] 中共中央, 国务院. 深化新时代教育评价改革总体方案[EB/OL]. (2020-10-13)[2026-11-11]. https://www.gov.cn/zhengce/2020-10/13/content_5551032.htm.
[16] 葛俊如. 人工智能支持下教育评价方法改革研究[J]. 科教文汇, 2026, (20): 99-102.
[17] 林秀瑜, 黄慕雄, 蔡晓桑, 等. 教育智能体特征组态对探究式教学成效影响的实证研究[J]. 电化教育研究, 2026, 47(10): 62-68.

AI Agents Empowering Educational Evaluation: A Study on Application Models and Evaluation Efficacy

Xiang Huiyang , Nie Xinming

School of Physics and Optoelectronic Engineering, Jiangsu Normal University, Xuzhou 221116

Abstract: Research on AI agents in educational assessment has centred on scoring consistency and users' acceptance of automated feedback, leaving the linkages between assessment stages underexamined. Drawing on three publicly documented cases-iFLYTEK Zhixuewang in junior secondary physics, Jiangxi's "Smart Homework" initiative, and a teacher-built LLM writing-assessment agent at Beijing Middle School-and distinguishing data-driven platforms from genuine LLM agents, this paper conducts a comparative reading of case materials and theoretical analysis along the chain of evidence collection, interpretation, feedback generation and follow-up intervention. The analysis indicates that agents bridge discontinuities through continuous evidence collection, immediate diagnosis and closed-loop feedback, and distinguishes four models-single-agent assisted, multi-agent collaborative, role-based feedback and human-AI closed-loop assessment-that form a hierarchy in assessment coherence. Such coherence presupposes a division of labour whereby agents handle standardized tasks while teachers retain value judgement and decision-making, supported by institutionalized standards and accountability, sustained teacher expertise and practice-driven data iteration. As the analysis rests on public materials and theoretical inference, the findings are analytical generalizations and the proposed pathways are normative recommendations not yet tested in large-scale teaching experiments.
Keywords: Educational AI Agents; Educational Evaluation; Formative Assessment; Human-AI Collaboration