基于大语言模型的多模态研究性学习行为评价研究

刘小龙

（上海市教师教育学院/上海市教育委员会教学研究室，上海 200233）

[摘要] 人工智能与大数据技术的迭代升级，推动传统评价模式迈向系统性、深层次变革。本文基于多模态学习分析构建覆盖问题解决能力、合作能力和创新思维三个维度，包含35项外显行为观测点的研究性学习评价指标体系。研究借助视频分析、文本挖掘、系统日志追踪等多模态数据采集技术，实现从行为数据到能力画像的动态映射。基于大语言模型的多模态数据聚类融合，文章识别出四类不同实践创新能力特征学生群体，并提出针对性改进策略。结果表明，多模态数据驱动的评价体系能有效量化学生研究性学习的高阶能力，为综合素质评价提供技术支撑。

[关键词] 多模态学习分析；研究性学习行为评价；大语言模型

[中图分类号] G622 [文献标识码] A [文章编号] 1007-2179(2026)02-0110-10

研究性学习指学生在教师指导下以类似科学研究的方式获取和应用知识的学习方式，它注重培养学生在真实情境中主动建构知识并解决实际问题的能力（张肇丰，2000）。在具体实践中，研究性学习往往面临评价滞后、重结果轻过程、技术赋能不足等问题，学习实效不强，甚至流于形式。因此，识别研究性学习行为特征，并基于行为分析学生能力发展水平，实现及时、精准的教学指导尤为重要。

一、研究现状

引导学生将提出假设、实验观察、数据收集与解析、反思研究结果等研究性学习方法应用于真实问题解决，被广泛认为有助于提升学生问题解决能力、探究能力与思维能力（Belland et al., 2019; Pedaste et al., 2015）。信息技术的发展，为动态评价研究性学习提供了可能，正悄然推动研究性学习范式的深刻变革。最新研究显示，多模态数据支持的个性化指导系统能显著提升探究效率（Zhang et al., 2023）。

多模态学习分析（multimodal learning analytics, MMLA）从最初的单一模态探索逐步发展为多源数据融合（Schneider et al., 2021; Ochoa et al., 2020），已成为能整合多种传感与数据采集技术支持复杂认知与学习过程分析的重要方法。与传统偏重学习结果的评价方式不同，多模态学习分析借助自动化分析与智能反馈机制简化数据处理流程，可追踪学生在活动参与、认知加工与情感体验等多维度的表现，为过程性评价提供支持（Blikstein et al., 2016），

[收稿日期]2026-02-03 [修回日期]2026-03-10 [DOI编码]10.13966/j.cnki.kfjyyj.2026.02.011

[基金项目] 全国教育科学“十四五”规划2021年度教育部青年课题“基于多模态学习分析的研究性学习行为评价体系研究”（ECA210405）。

[作者简介] 刘小龙，助理研究员，上海市教师教育学院、上海市教育委员会教学研究室，研究方向：研究性学习、综合素质评价、课程教学改革、人工智能教育（122377307@qq.com）。

[引用信息] 刘小龙(2026). 基于大语言模型的多模态研究性学习行为评价研究[J]. 开放教育研究，32(2)：110-119.

帮助教师与学习者更好地理解并改进学习表现(Guerrero-Sosa et al.,2025)。多模态行为数据驱动的评价,能够综合考量行为轨迹、交互频次等多维指标(葛岩等, 2023; 王春洁, 2025),通过情感语义库分析学生课堂参与度(蒋艳双等, 2025; 李昂等, 2024)。大语言模型被用于解释多模态分析结果(吴永和等, 2021; 王刚等, 2025),不仅可以实现学习行为量化,还能显著增强结果的可解释性和教育意义(Li et al., 2025)。

然而,研究性学习行为评价研究仍存在局限:一是评价方式多采用纸笔测评或写实记录的方式,评价过程的主观性较强,难以保证评价结果的真实有效;二是多强调定性分析学习结果,缺乏数据支撑,一定程度上忽略了学习过程评价和增值评价,难以指导学生高阶能力培养;三是技术应用多基于单模态数据进行统计分析,将多模态学习分析和大模型等技术应用于研究性学习行为评价的不多,高阶能力量化研究尤少。

本研究通过梳理研究性学习行为的一般过程、目标技能与构成要素,确定不同目标技能的评价行为,并引入多模态学习分析,整合研究性学习全过程中视频、文本、语音和系统行为等多源数据,从中提取信息并量化分析,以期精准评价研究性学习行为。

二、评价模型与指标体系

(一)评价模型

本研究从视频、文本、语音和系统操作等数据源提取信息,利用大语言模型实现多模态数据融合,并与学生能力画像相映射,构建基于多模态数据的研究性学习行为评价分析模型(见图 1)。该模型旨在突破传统评价体系的局限,深入理解学生主观学习行为与学习结果之间的关系,揭示学生学习行为模式的共性与差异,分析师生与生生之间的互动联络拓扑,重构“教—学—评”一体化的教育生态,为建设高质量教学体系提供支持。

(二)指标体系

研究性学习一般包括学生围绕挑战性问题驱动的复杂任务,开展方案设计、问题解决、调查活动或进行决策,并最终呈现或展示研究成果。本研究通过梳理国内外高阶能力思维评价框架及其指标体系,确定评价目标和评价对象。其中,评价目标是学习者实践能力和创新思维,评价对象是能力思维外显于学习过程中的行为表现。因此,本研究参考余明华等(2021)的项目式学习评价指标体系,然后融合问题解决能力、合作能力和创新思维等构成要素,构建研究性学习行为评价指标体系。

本研究编制了《中学生研究性学习调查问卷》。

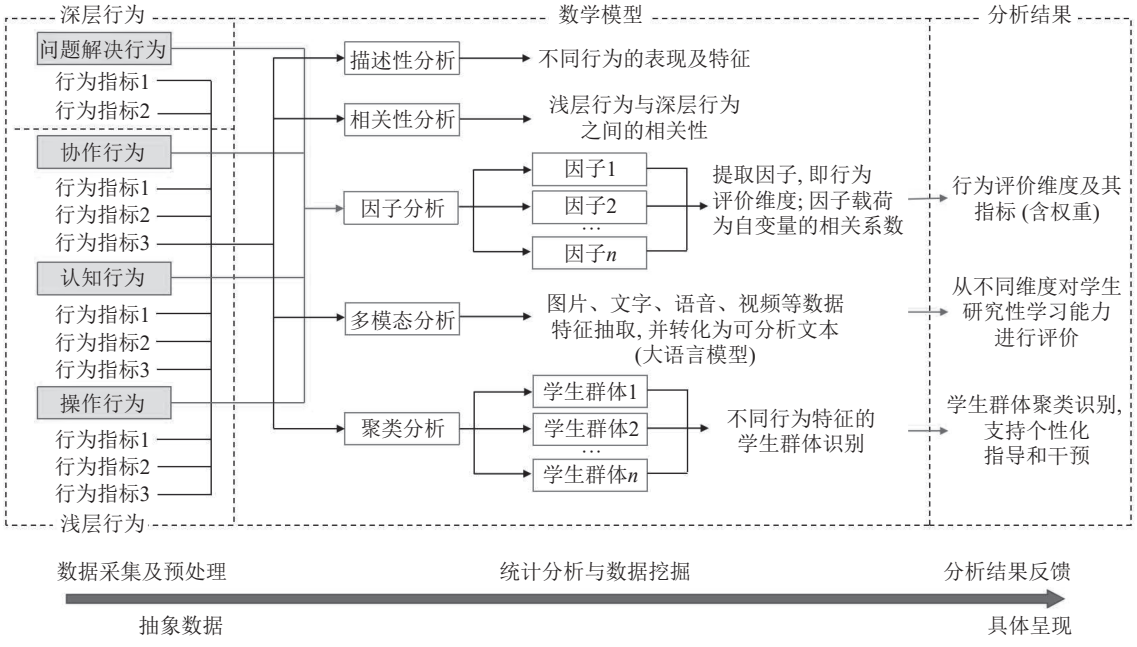

图 1　基于多模态数据的研究性学习行为评价模型

问卷调研采用李克特五点量表形式,分预试和正试阶段。预试阶段的问卷样本用于项目分析和探索性因素分析,旨在检验题项的效度和信度。正试阶段的问卷样本用于验证性因素分析,旨在检验指标建构的适切性和真实性。开始探索性因素分析之前,研究者对问卷题项的 KMO 系数与 Bartlett 球形进行检验。KMO 值>0.9,说明问卷题项适合进行探索性因素分析。

研究者采用主成分分析法进行探索性因素分析,删除因素负荷量低于 0.45 的题项,同时保证两个或以上因素的负荷量超过 0.45,且每个维度不少于 3 道题,可解释方差的累积贡献率为 66.872%(见表 1)。问卷采用随机抽样方式面向上海市各区中学生发放,回收有效问卷 3452 份。

问卷验证性因素分析结果显示,因素载荷在 0.596 至 0.868 之间,说明所有的因素载荷均在合理范围内。信度检验克隆巴赫 Alpha 值 0.96,各维度克隆巴赫 Alpha 值均大于 0.8,说明问卷题项内部一致性较好。模型的 CMIN/DF 值 2.123,其它拟合指数均达到要求。

因素载荷值是潜在变量到观测变量的标准化回归系数,代表观测变量对潜在变量的影响程度。本研究将其归一化为权重系数,形成研究性学习行为评价指标体系(见表 2)。

三、数据采集与融合方法

依托上海市学生研究性学习智能支持系统,研究者采集视频、语音、文本、图像和系统行为等多模态数据,并通过数据清洗、集成、归约和转化等,以多样化的格式和粒度存储数据(见图 2):学业成绩、课程参与、课时学分、成果评价等数据以 json 文件格式存储在文件数据库,并进行归一化处理;行为数据通过平台功能嵌入“埋点”,实现对用户行为数据的标准化采集和处理,并采用多种统计分析方法为此类交互行为数据合理赋值;将语音、视频、图片、文本等多种数据类型存储于关系型数据库及文件服务器内,借助语音识别、视频分析、图像识别及自然语言处理等技术抽取行为指标的特征向量,将其转化为可供分析的文本或统计数据,进而运用大语言模型实现多模态数据融合,开展对学生研究性学习行为的评价和能力分析。

(一)学习行为数据采集

本研究参考网络学习行为 OCCP(observation, classification, coding, profiling, OCCP)层次化模型(彭文辉, 2012),将在线研究性学习行为分为操作行为、认知行为、协作行为和问题解决行为。操作行为指学生操作学习平台行为,包括登录、退出、点击访问、浏览下载、保存、删除等。认知行为指学生开展针对性学习,包括查看优质课题并收藏、点赞和评论,搜索相关资源关键词,查看学习资源,记录笔记等。协作行为包括组员之间、师生之间的讨论等。问题解决行为是最高级、最深层次的学习行为,包括完成研究任务、填写研究报告、提交报告等。学生在上海市中学生研究性学习智能支持系统上开展研究性学习,上述四个层次行为数据以 xAPI 格式被记录到学习记录库(见表 3)。学习记录库类似数据库,用来存储和维护学习者行为记录。

(二)基于大语言模型的多模态数据融合

本研究基于多模态学习行为的时间性和过程性特点,对不同模态、粒度、来源的学习行为数据进行时间轴对齐和因果分析,并将其转化为可分析

表 1 量表解释总变异量

成分	初始特征值			提取载荷平方和			旋转载荷平方和		
	总计	方差百分比	累积百分比(%)	总计	方差百分比	累积百分比(%)	总计	方差百分比	累积百分比(%)
1	23.307	54.203	54.203	23.307	54.203	54.203	10.673	24.821	24.821
2	1.791	4.165	58.368	1.791	4.165	58.368	5.620	13.070	37.891
3	1.412	3.283	61.651	1.412	3.283	61.651	4.292	9.983	47.874
4	1.157	2.690	64.341	1.157	2.690	64.341	4.144	9.638	57.512
5	1.088	2.531	66.872	1.088	2.531	66.872	4.025	9.360	66.872

注:数据采用主成分分析法。

表 2 研究性学习行为评价指标体系

一级指标	二级指标	三级指标	观测行为表现描述
D1 批判性思维与问题解决	E1 提出问题	F1 问题清晰度	S1.问题表述清晰，容易理解，与研究目标相呼应
			S2.问题与研究主题紧密相关，有助于深入理解主题
		F2 问题有效性	S3.问题触及研究核心，能转化为具体解决方案，引发深入思考和讨论
	E2 理解与规划	F3 分析问题	S4.简洁、清晰地呈现研究背景与目的
		F4 制定研究计划	S5.研究方案设计结构严谨、逻辑清晰且可操作，步骤、内容与方法明确
			S6.基于成员能力合理分配研究任务
			S7.研究方案紧扣研究问题，内容与目标一致
	E3 信息采集	F5 信息获取	S8.选择适当的手段，尽可能多地收集研究信息
	E4 分析证明	F6 信息处理	S9.有效设定实验变量与科学测量
			S10.综合评价信息（数据）的有效性、准确性与科学性
			S11.通过推理有效性检验、证据与问题的匹配度及覆盖度，评价推论、观点的可靠性
	E5 表述与呈现	F7 结论阐述	S12.通过图表工具优化数据组织与呈现，提升信息传递的准确性与效率
			S13.对研究结论进行基于实证的深度分析与阐释
D2 合作能力	E6 自我管理	F8 自我意识	S14.自觉按时完成任务（行动力、参与度和坚持力等）
			S15.调查和反思团队提出的观点
		F9 帮助团队	S16.为他人提供具体且有可行性的反馈
			S17.提供他人需要的帮助
	E7 团队管理	F10 组织工作	S18.创建详细的任务列表，并合理划分成员任务
			S19.在规定时间讨论并跟踪进展
		F11 团队协作	S20.每位成员都积极参与讨论
			S21.所有队友参与，并形成观点和研究成果
D3 创新思维	E8 思维开放性	F12 发现探索	S22.通过专家咨询、社区调研、企业合作或专业文献等途径获取信息
			S23.通过群体交流激发多元视角，形成创造性观点
		F13 灵活联想	S24.通过概念衍生与创新联想提出多样化问题解决策略
			S25.识别不同观点的优劣并整合最优观点，增强推论的合理性
			S26.尝试提出新问题，从不同视角阐述和改进观点
			S27.当想法转化为观点时，能发挥创造力和想象力
		F14 多样表达	S28.通过动态演示、交互展示等形式，呈现研究成果的创新性
	E9 创新性表述与构思	F15 成果独创性	S29.研究成果总是新颖、独特、让人眼前一亮
			S30.研究成果能打破常规，或运用新的、灵活的、让人耳目一新的方法处理已有信息（材料或数据）或观点
		F16 成果有效性	S31.研究成果有用且有价值，能解决问题
			S32.研究成果切合实际且可推广
		F17 成果科学性	S33.研究报告能清晰、简明、有逻辑地呈现信息、调查结果和证据
			S34.研究报告能将不同元素组合成连贯的整体
			S35.研究报告呈现出精心设计、引人注目、恰当且鲜明等特点

的文本，以理解抽象的语义；利用大语言模型进行融合分析，最大程度地综合利用多来源数据中的信息，从多角度视角和借助功能多样的深度学习技术实现精确的研究性学习行为评价。研究者选择 ChatGLM 作为研究性学习行为评价分析的核心技术。该技术是专为中文场景优化的大模型。与基于英文训练的模型相比，它不仅在中文通用语言文本理解和生成方面表现更优，在研究性学习课题报

告等学术文本处理方面也展现出独特优势，能准确解析和评估研究报告的内容，以及提供开源的模型参数和代码；与超大规模的闭源模型相比，它能降低使用成本，保障数据的隐私安全，确保技术的可控性，适合在教育领域应用。

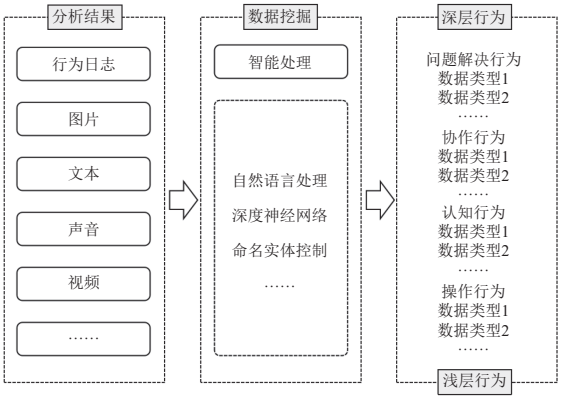

图 2　多模态数据采集与处理

1. 大语言模型的评分细则提示词设计

本研究在上述研制的指标体系基础上，为每个观测行为设计详细的评分准则，并通过 ChatGPT 和 ChatGLM 对比测试，优化评分效果（见图 3）。

研究者结合前述指标体系研究中归一化处理得到的权重系数，为 35 个观测行为设计“有限解决、仍需改进、基本解决、良好尝试、卓越解决”5 个档次评分细则提示词，确保每个观测行为的评分标准清晰、具体，以便大语言模型能够理解并按照标准进行分析和评分。

设计完评分细则后，本研究先用 ChatGPT 进行

表 3　基于 xAPI 的研究性学习行为数据

行为层次	行为指标	基于 xAPI 的研究性学习活动流映射
操作行为 （登录、退出、浏览、点击、访问、选择、关闭等）	登录系统时长、频次 访问不同功能模块的频次	Actor+ logged-in（登录）+ MOORs（上海市中学生研究性学习智能支持系统） Actor+ logged-out（退出）+ MOORs（上海市中学生研究性学习智能支持系统） Actor+ accessed（访问）+ Research（研究室） Actor+ accessed（访问）+ RClass（研课堂） Actor+ accessed（访问）+ RCircle（研讨圈） Actor+ accessed（访问）+ RDiscuss（研究群组） Actor+ accessed（访问）+ RNote（灵感笔记）
认知行为 （创建、修改、搜索、查看、阅读观看、收藏、点赞、评论等）	创建灵感笔记数 搜索课题/资源频次及其关键词 查看课题频次 收藏课题数 课题点赞频次 阅读文本资源/校本文本资源频次及时长 观看视频资源/校本视频资源频次及时长	Actor+ created（创建）+ INote（灵感笔记） Actor+ modified（修改）+ INote（灵感笔记） Actor+ searched（搜索）+ RProject（课题） Actor+ searched（搜索）+ Resource（资源） Actor+ read（阅读）+ RProject（课题） Actor+ read（阅读）+ TextResource（文档资源） Actor+ read（阅读）+ STextResource（校本文档资源） Actor+ watched（观看）+ VideoResource（视频资源） Actor+ watched（观看）+ SVideoResource（校本视频资源） Actor+ collected（收藏）+RProject（课题） Actor+ liked（点赞）+ RProject（课题）
协作行为 （评论、讨论、交流、提问、回复等）	评论课题频次 发送消息频次 讨论贡献率 发送私信频次	Actor+ commented（评论）+ RProject（课题） Actor+ sended（发送）+ Question（问题） Actor+ sended（发送）+ MMessage（消息） Actor+ sended（发送）+ PMessage（私信）
问题解决行为 （创建、填写、修改、完成、提交、共享、上传等）	创建课题数 填写/修改某些研究内容频次 提交课题数 分享课题数	Actor+ created（创建）+ RProject（课题） Actor+ wrote（填写）+ Content（研究内容） Actor+ modified（修改）+ Content（研究内容） Actor+ saved（保存）+ Content（研究内容） Actor+ completed（完成）+ Report（研究报告） Actor+ submitted（提交）+ Report（研究报告） Actor+ submitted（提交）+ Artifact（课题制品） Actor+ shared（分享）+ RProject（课题）

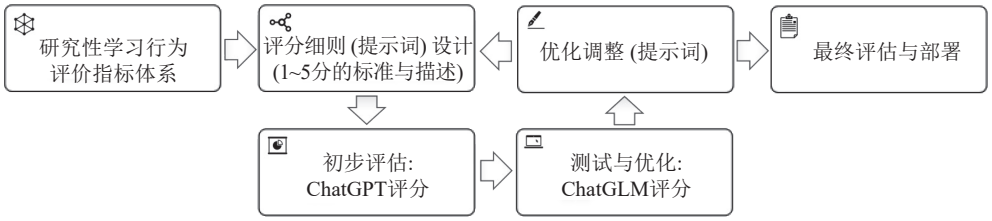

图 3 评分细则提示词设计流程

测试, 用于模型和提示词调优的参考基准。研究者将设计好的评分细则提示词输入 ChatGPT 中, 根据评价指标对学生研究行为和报告进行评分; 记录 ChatGPT 对观测行为的评分, 再与人工评分对比, 检查分数分布是否合理, 以验证评分的准确性和合理性; 再用 ChatGLM 进行测试, 将评分细则提示词输入 ChatGLM, 再次评价学生研究行为和报告, 比较 ChatGLM 与 ChatGPT 的评分, 最终确定评估结果的准确性、相关性和分布情况。

对比结果显示, ChatGPT 与 ChatGLM 评分存在不一致, 原因可能是部分评分维度的描述不够清晰或具体, 研究者进一步调整评分细则提示词, 增加示例或解释, 使模型能准确理解评分标准, 并重新评分和记录, 以提高评分的精确度和一致性。

在第二轮优化中, 研究者细化了评分标准并要求模型解释评分原因, 提高评分的透明化程度。经优化后, 模型的准确率跃升至 55.5%, 斯皮尔曼相关系数达到 0.343。这表明模型评分一致性和准确性均有显著提高(见图 4)。评分输出不仅包括分数, 还附带详细的评分理由, 以增强评分过程的合理性和透明度。

经多轮优化后, 研究者综合评估 ChatGLM 的最终评分效果, 包括对比智能评分与人工评分的计算准确率、精确度、召回率和 F1 分数等指标(见表 4), 以全面评估模型的评分准确性, 检查评分结果分布, 降低评分结果的误差。

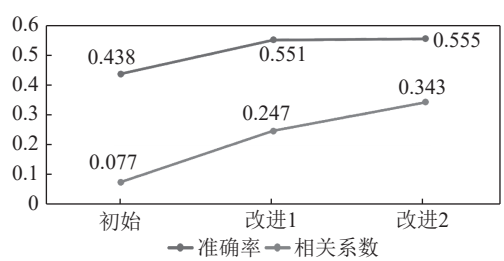

图 4 提示词准确率和秩相关系数变化

2. 多模态数据融合处理流程

首先, 为确保提示词对研究性学习行为评价的准确性、稳定性和保持生成框架的通用性, 避免主观评价和无关干扰, 研究者设计具备重试机制的模型调用流程, 配置 temperature=0.1 和 do_sample=True, 采用无历史上下文的独立交互方式, 提取 0—5 区间的数值型评分。

其次, 采用少样本学习思路嵌入指标评分标准、边界阈值及典型案例, 确保提示词具备明确的指令导向与歧义消解能力。研究者将单指标提示词与对应的学生课题报告文本和学习行为数据进行标准化预处理(包括文本处理、冗余信息过滤、提取关键信息等), 按 JSON Schema 规范封装传入大语言模型交互推理逻辑, 通过批量异步推理机制驱动模型完成 35 个指标的量化, 同步记录评分置信度参数; 按照研究性学习行为评价指标体系的层级架构, 动态校准预设指标权重, 提升权重分配的科学性与客观性; 基于校准后权重, 通过加权求和算法、Min-Max 归一化处理, 完成各层级指标得分的分层加权聚合, 输出三级指标的综合得分, 计算各维度分值。

再次, 研究者对聚合后得到的结果进行清洗和检验。考虑到学生开展研究性学习实践活动可能导致部分行为数据缺失, 研究者通过孤立森林算法检测并剔除评分异常值, 采用插值法补全缺失值, 并利用跨指标逻辑校验消除不一致数据; 分层随机抽取 5%—10% 的样本进行人工检验, 验证模型评分与人工评分的契合度(目标契合度≥0.85), 同步反馈抽验结果用于模型提示词迭代优化, 提升量化结果的准确性、完整性与可靠性。

最后, 研究者对量化数据集开展聚类分析, 按聚类标签对数据集进行分组处理: 剔除无样本的空聚类簇后, 统计分析有效聚类簇内各行为特征的数值分布与离散程度; 基于预设的能力级别阈值判定

表 4　ChatGLM 评分效果

类别	准确率	精确度	召回率	F1 分数	类别	准确率	精确度	召回率	F1 分数
S1	0.8715	0.8512	0.8923	0.8713	S19	0.7710	0.7392	0.8048	0.7706
S2	0.8347	0.8123	0.8578	0.8344	S20	0.7321	0.6903	0.7765	0.7311
S3	0.8031	0.7739	0.8341	0.8029	S21	0.7488	0.7089	0.7910	0.7478
S4	0.7819	0.7536	0.8120	0.7817	S22	0.8664	0.8467	0.8865	0.8661
S5	0.8628	0.8434	0.8829	0.8626	S23	0.8973	0.8820	0.9130	0.8971
S6	0.7423	0.7018	0.7847	0.7411	S24	0.8396	0.8179	0.8618	0.8392
S7	0.9041	0.8912	0.9173	0.9039	S25	0.8254	0.7998	0.8523	0.8250
S8	0.8516	0.8320	0.8719	0.8514	S26	0.8620	0.8425	0.8821	0.8617
S9	0.9120	0.9018	0.9223	0.9119	S27	0.7965	0.7664	0.8283	0.7961
S10	0.7645	0.7341	0.7962	0.7638	S28	0.7579	0.7234	0.7946	0.7573
S11	0.8189	0.7912	0.8481	0.8186	S29	0.7246	0.6810	0.7710	0.7234
S12	0.7932	0.7629	0.8250	0.7928	S30	0.7133	0.6689	0.7605	0.7119
S13	0.8473	0.8265	0.8687	0.8470	S31	0.8081	0.7802	0.8375	0.8077
S14	0.7764	0.7447	0.8101	0.7760	S32	0.8320	0.8091	0.8556	0.8316
S15	0.8295	0.8043	0.8556	0.8291	S33	0.8442	0.8229	0.8660	0.8439
S16	0.8137	0.7850	0.8439	0.8133	S34	0.7869	0.7558	0.8198	0.7865
S17	0.8882	0.8731	0.9036	0.8879	S35	0.8787	0.8603	0.8976	0.8784
S18	0.8551	0.8374	0.8733	0.8549	均值	0.8129	0.7943	0.8346	0.8138

聚类簇的整体能力层级，生成对应维度的语义描述、匹配针对性的教学建议，实现从量化数据到质性结论的转化。

四、研究发现

本研究基于上述研制的指标体系和数据融合方法对上海市学生研究性学习行为数据量化评分，建立指标维度和观测行为的初始常模值。教师可基于指标维度或某一观测行为统计数据反思和研究教学质量，找到改进方向，实现理论与数据的双向驱动。考虑到城郊和校际差异以及数据连续性、完整性等因素，本研究以 X 区（城区）和 Q 区（郊区）3548 名高三学生作为样本开展数据分析。

（一）研究性学习水平分层

研究者采用 K-means 聚类算法分析学生批判性思维与问题解决、协同合作、创新思维等维度的能力，选择 E1—E9 共 9 项二级指标作为聚类分析的特征向量。聚类结果评估显示，设置 k=4 时，分析质量较好（见表 5）。

为了直观展示九维特征空间的聚类结果，研究者使用 t-SNE 降维方法将数据降至 2D 空间，识别四类不同特征学生群体（见图 5 和表 6）：引领型创新者（类别 1）、潜力型学习者（类别 2）、思辨型探索者（类别 3）和资源型协作者（类别 4）。

基于学生答辩视频发现，对比学习特征和行为数据，学校可探寻学生答辩表现不同的原因。例如，部分学生能清晰说明研究方法和过程，但在研究问

表 5　聚类质量评估结果

SSE（Inertia）	轮廓系数	Davies-Bouldin 指数	Calinski-Harabasz 指数
1732.63	0.417	0.672	1703.823

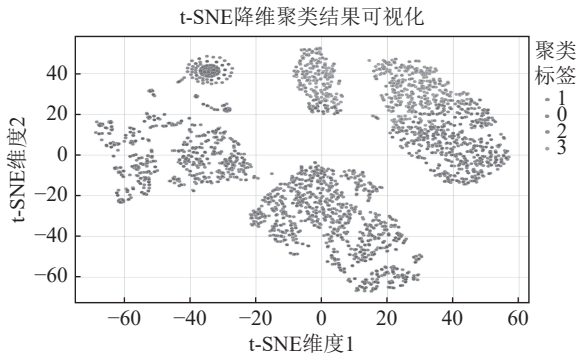

图 5　四类群体降维空间分布

表 6 四类学生群体特征值和特征表现

类别	E1	E2	E3	E4	E5	E6	E7	E8	E9	特征表现
类别 1	3.69	3.65	3.96	3.96	2.10	4.03	3.86	3.65	3.71	综合能力强，在所有指标上表现优异。
类别 2	3.04	3.18	2.98	3.12	1.71	3.30	3.17	2.87	2.91	各项能力均有待提升，整体表现较弱。
类别 3	3.26	3.23	3.00	3.69	2.07	3.81	3.58	3.35	3.38	推理能力较强，但提出问题和资源收集能力一般。
类别 4	3.25	3.31	4.01	3.35	1.89	3.55	3.35	3.22	3.29	资源收集能力突出，但其他能力表现一般。

表 7 X 区和 Q 区常模值对比

指标名称	提出问题	理解与计划	信息采集	分析证明	表述与呈现	自我管理	团队管理	思维开放性	创新性表述与构思
全市均值	63.63	65.42	68.52	71.44	53.24	73.29	67.81	62.56	66.45
X 区	65.21	68.56	70.42	73.23	55.39	78.92	70.31	65.47	68.52
Q 区	63.87	65.18	67.33	72.91	52.85	75.30	68.04	61.43	64.72

题来源、意义和反思方面的表现相对较弱。研究者分析学习行为数据发现，学生查阅同领域研究案例的行为频繁，但几乎没有查阅学术文献和开展小组讨论，主要是借鉴和模仿已有研究，呈现类型 3 的特征。由此，学校可结合识别出的不同类型学生群体开展个性化培养。对于类别 1 学生，教师可提供挑战性的研究课题和项目，鼓励学生参与高水平的学术交流和竞赛活动；建立“导师制”为学生提供专业指导。对于类别 2 学生，教师要加强训练其基础能力，特别是问题探究、表述与呈现能力；提供更多的实践机会和案例示范；建立学习支持小组，促进同伴互助。对于类别 3 学生，教师要强化培养其问题探究和资源收集能力；引导学生将分析推理能力应用于实际问题解决。对于类别 4 学生，教师要指导其有效整合和利用收集到的资源，并提升其分析推理和表述能力。

（二）人才培养水平和能力发展差异

比较 X 区与 Q 区学生指标平均水平可以发现，两者差异不大，且趋于全市常模值，但不同水平学生占比差异明显（见表 7 和图 6），如 X 区“引领型创新者”型学生数占比高于 Q 区。这表明，两区的研究性学习均能落实国家课程要求，但 X 区在拔尖创新人才培养方面优于 Q 区。由此，管理部门和学校可深入分析指标和观测行为表现，完善课

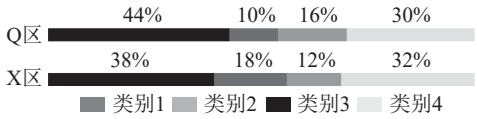

图 6 X 区和 Q 区不同类别学生占比

程建设、改进学生学习过程和优化教育资源投入。

在“表述与呈现”方面，两区学生在“S12:通过图表工具优化数据组织与呈现，提升信息传递的准确性与效率”上的表现均较弱。这表明，两区学生都不擅长运用数据分析和可视化呈现工具，呈现方式多停留在文字描述。对此，教师要指导学生利用统计分析工具组织和呈现研究成果，提升学生应用工具的熟练度。

（三）思维能力培养

相关性分析结果显示（见表 8），大部分指标之间存在较强的正相关关系，相关系数普遍在 0.75 以上，但“思维开放性”与其他指标的相关系数值较低，大多为 0.5—0.6 之间。这表明，学生思维能力发展相对独立，教师教学需关注培养学生的思维品质，从知识传授转向素养导向，注重设计开放式问题、强化跨学科整合、拓展思维工具应用，激发学生深度反思知识与学习过程，促进其认知与元认知的协同发展。

五、总结与反思

研究性学习是科技教育实施的重要载体。信息技术不仅拓展学生开展研究性学习的过程性、多模态数据采集范围，而且能实现研究性学习行为多模态数据融合和分析评价、生成教学策略建议，为学生自主开展研究性学习提供个性化支持。

本研究通过构建学生研究性学习行为评价指标体系，以及多模态数据采集路径和基于大语言模型的数据融合策略，实现对学生研究性学习行为的量化评价，为基于数据循证的教学设计改进、分层

表 8 指标相关性分析

E9	0.866	0.857	0.828	0.847	0.859	0.825	0.865	0.604	1
E8	0.581	0.575	0.559	0.591	0.577	0.588	0.591	1	0.604
E7	0.842	0.836	0.804	0.814	0.837	0.838	1	0.591	0.865
E6	0.798	0.801	0.777	0.811	0.804	1	0.838	0.588	0.825
E5	0.86	0.849	0.816	0.824	1	0.804	0.837	0.577	0.859
E4	0.809	0.809	0.786	1	0.824	0.811	0.814	0.591	0.847
E3	0.842	0.845	1	0.786	0.816	0.777	0.804	0.559	0.828
E2	0.885	1	0.845	0.809	0.849	0.801	0.836	0.575	0.857
E1	1	0.885	0.842	0.809	0.86	0.798	0.842	0.581	0.866
	E1	E2	E3	E4	E5	E6	E7	E8	E9

教学实施、学习方法指导和教育资源投入提供了全新可能。基于对上海市 X 区和 Q 区学生研究性学习数据的分析和调研，本研究发现：

第一，学生统计分析和结果呈现能力相对较弱，缺乏动手实验、整理数据、分析研讨等研究过程的经验。原因可能在于：部分学生多依赖文献查阅和问卷调查等方法，而不是采用实验、访谈、数据分析等方法完成研究性学习任务；在“分数”的重压下，教师常会弱化实验分析等动手环节，采用装置搭建、演示或播放视频等方式替代学生的亲手操作，或以验证型实验提供数据分析结果，导致学生无法在试错中重构认知框架，也无法在认识冲突中迭代观点，最终消解研究性学习“在不确定中创造”的核心价值。第二，部分教师把评价重点放在研究报告的格式规范上，忽视对假设的合理性、数据采集的伦理边界、结论推导的逻辑链条等核心思维能力的培养。教师常同时指导几十甚至上百名学生，受限于学科知识背景和工作负担，难以对每位学生提供个别指导和评价，也是其中的重要原因。但这种评价倾向，可能导致学生思维发展的停滞与真实问题解决能力的萎缩。第三，数据“解读”能力是改进教与学的关键。基于多模态学习分析的研究性学习行为评价的核心是数据驱动与理论引导的辩证统一。教师需具备从“关联”走向“因果”的解释与经验概括能力，能将数据转化为指导教学干预的“证据”。

人工智能的应用为研究性学习行为评价的逻辑、路径与实施提供了全新可能，但也潜藏着不容忽视的认知偏差与价值异化。其一，基于“模型算法”的评价一般以标准化数据为依据，而在实际教学中，学生非常规探究、批判性质疑、创造性试错等往往是非标准化的，教师依赖或盲目应用技术易陷入“数据陷阱”，既弱化自身专业能力，也不利于发现拔尖创新人才。其二，数据驱动的评价易剥离学习的情感性与情境性，无法真正捕捉行为背后的思维深度与情感价值，教师主体性的弱化会导致评价沦为数据拟合而非育人判断。其三，教师要警惕学生借助人工智能完成学习任务的依赖性。多项研究表明，过度使用生成式人工智能技术易导致“认知负债”，学生将逐渐丧失独立思考、逻辑推理与原创表达的能力，陷入“浅层学习”的困境。

本研究后续将依托上海市中学生研究性学习智能支持系统和综合素质评价信息管理系统，整合学生学业成绩与在线学习行为等数据，构建兼顾量化指标与质性表现的研究性学习能力画像，深入挖掘影响学生核心素养与探究能力发展的关键因素；同时，立足评价育人本质，探索基于研究性学习行为评价的个性化教学策略推荐，将学生的研究状态、资源、路径精准匹配，帮助教师及时为学生开展研究性学习提供知识补给、资源支持和情感引导，提升学生开展研究性学习的专注度与效率，满足不同层次学生的个性化学习需求，推动人工智能在评价场景中的应用回归育人本源。

［参考文献］

[1] Belland, B. R., Weiss, D. M., Kim, N. J., Piland, J., & Gu, J. Y.(2019). An examination of credit recovery students' use of computer-based scaffolding in a problem-based, scientific inquiry unit[J]. International journal of science and mathematics education, 17(2): 273-293.

[2] Blikstein, P., &Worsley, M.(2016). Multimodal learning analytics and education data mining: Using computational technologies to meas-

ure complex learning tasks[J]. Journal of Learning Analytics, (2): 220-238.

[3] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., &Olivas, J. A.(2025). A comprehensive review of multimodal analysis in education[J]. Applied Sciences, 15(11): 5896.

[4] 葛岩, 朱若瑜, 崔璐(2023). 面向多模态数据融合的混合式学习参与度分析模型构建 [J]. 中小学电教, (5): 16-19.

[5] 蒋艳双, 许宗嗣, 逯行, 朱立新(2025). 人机协同教学中的学生参与度: 关键特征、分析模型与实践进路 [J]. 现代教育技术, 35(8): 77-86.

[6] 李昂, 唐章蔚(2024). 基于多模态数据分析框架的在线直播课堂学生参与度研究 [J]. 教育传播与技术, (2): 87-96.

[7] Li, X., &Wang, S.(2025). Explainable multimodal learning analytics using large language models[J]. Computers & Education, 186: 104785.

[8] Ochoa, X., &Worsley, M.(2020). Augmenting learning analytics with multimodal sensory data[J]. Journal of Learning Analytics, 7(3): 1-13.

[9] Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E.(2015). Phases of inquiry-based learning: Definitions and the inquiry cycle[J]. Educational Research Review, 14: 47-61.

[10] 彭文辉(2012). 网络学习行为分析及建模 [D]. 武汉: 华中师范大学: 70-73.

[11] Schneider, B., Dowell, N., &Thompson, K.(2021). Multimodal learning analytics for collaborative learning: A systematic review[J]. Computers & Education, 163: 104123.

[12] 王春洁(2025). 基于多模态数据的个性化学习者画像模型的构建与应用 [J]. 电脑知识与技术, 21(22): 30-34.

[13] 吴永和, 郭胜男, 朱丽娟, 马晓玲(2021). 多模态学习融合分析(MLFA)研究: 学理阐述、模型样态与应用路径 [J]. 远程教育杂志, 39(3): 32-41.

[14] 王刚, 孙方, 郭磊(2025). 数智时代教学范式转型: 多模态数据融合驱动的个性化学习路径重构 [J]. 淮南师范学院学报, 27(4): 93-99.

[15] 余明华, 张治, 祝智庭(2021). 基于学生画像的项目式学习评价指标体系研究 [J]. 电化教育研究, 42(3): 89-95.

[16] 张肇丰(2000). 试论研究性学习 [J]. 课程. 教材. 教法, (6): 42-45.

[17] Zhang, L., Chen, X., &Wang, Y.(2023). Personalized guidance system for inquiry-based learning: Enhancing student engagement through adaptive scaffolding[J]. Computers & Education, 185: 104521.

(编辑: 李学书)

Research on the Evaluation of Multimodal Research-based Learning Behavior Based on Large Language Models

LIU Xiaolong

(*Shanghai Teacher Institute/The Teaching Research Section of Shanghai Municipal Education Commission*, *Shanghai* 200233, *China*)

Abstract: *The advancement of artificial intelligence and big data technologies is driving a systematic and profound transformation from traditional evaluation models. This study focuses on the scientific requirements for evaluating research-based learning behaviors, an evaluation indicator system, constructed on multimodal learning analytics, covering three dimensions: problem-solving ability, collaboration ability, and innovative thinking, and with 35 observable behavioral indicators. Through multimodal data collection techniques such as video analysis, text mining, and system log tracking, a dynamic mapping from behavioral data to form competency profiles and based on the multimodal data clustering and fusion of large language models, four types of student groups with different characteristics of practical innovation ability were successfully identified with significant regional differences discovered. The research results show that the multimodal data-driven evaluation system can effectively quantify the high-order abilities of students in research-based learning, providing solid technical support for comprehensive quality evaluation.*

Key words: *multimodal learning analytics*; *assessment of research-based learning behaviors*; *large language model*