中国人工智能学会
Chinese Association for Artificial Intelligence

中国人工智能学会系列白皮书
——教育研究中的AI4S

中国人工智能学会
二〇二五年十一月

中国人工智能学会系列白皮书

——教育研究中的AI4S

中国人工智能学会

二〇二五年十一月

《中国人工智能学会系列白皮书》编委会

主　　任：戴琼海

执行主任：马华东

副 主 任：赵春江　何　友　王恩东　郑庆华　刘成林
周志华　孙富春　庄越挺　胡德文　杜军平
杨　强

委　　员：陈松灿　董振江　付宜利　高新波　公茂果
古天龙　何　清　胡清华　黄河燕　季向阳
蒋田仔　林浩哲　梁吉业　刘奕群　潘　纲
石光明　孙茂松　孙长银　陶建华　王海峰
王熙照　王　轩　王蕴红　吴　飞　于　剑
余有成　张化光　张学工　章　毅　周鸿祎
周　杰　祝烈煌

《中国人工智能学会系列白皮书——教育研究中的AI4S》编写组

主　　编：陈向东

副 主 任：武法提　柯清超　金　慧

编　　者：刘泽民　卢淑怡　褚乐阳　靳旭莹
刘城烨　潘香霖　等

目　录

第 1 章 引言

人类的知识生产体系正经历一场深刻的范式变革，我们正亲历一场足以载入史册的“哥白尼时刻”[1]。2024 年诺贝尔奖的颁布，无疑是这一巨变强有力的印证。物理学奖授予 Geoffrey Hinton 和 John Hopfield，不仅表彰了他们在人工神经网络领域的奠基性贡献，更昭示了硅基智能对我们理解世界方式的重塑。与此同时，化学奖颁给 David Baker、Demis Hassabis 和 John Jumper，以表彰他们运用 AI 在蛋白质结构预测上实现的革命性突破，此举正开启生物学研究与药物发现的全新纪元。这一事件宣告，人工智能已然超越了单纯的辅助工具范畴，它正作为一种前所未有的力量，深度介入并驱动着基础科学前沿的探索与突破。

以深度学习为基础，并以大语言模型（Large Language Models, LLMs）为代表的生成式人工智能（Generative AI），正以前所未有的态势，深刻改变着科学研究的本质与范式。AlphaFold2 成功解决了困扰生物学界 50 年的蛋白质折叠难题[2]，DeepMind 的 FunSearch 系统在纯数学领域发现了上限集问题（Cap Set Problem）令人惊叹的全新解法[3]。这些突破性的成果，不仅昭示 AI 正在重塑科学研究的基础逻辑与实践形态[4]，更预示着它将从根本上颠覆我们定义问题、构建理论、收集证据乃至确立真理的传统方式[5]。

这场变革的深刻性足以并肩历史上任何一次重大的认知革命：印刷术改变了知识传播方式，显微镜开启了微观世界的探索，计算机实现了大规模数据处理，每一次技术革命都从根本上扩展了人类

[1] 亨利·基辛格, 埃里克·施密特, 丹尼尔·胡滕洛赫尔. 人工智能时代与人类未来[M]. 胡利平, 风君, 译. 中信出版社, 2023.

[2] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[3] Romera-Paredes B, Barekatain M, Novikov A, et al. Mathematical discoveries from program search with large language models[J]. Nature, 2024, 625(7995): 468-475.

[4] Agrawal A, Gans J, Goldfarb A. Power and prediction: the disruptive economics of artificial intelligence[M]. Boston, Massachusetts: Harvard Business Review Press, 2022.

[5] Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence[J]. Nature, 2023, 620(7972): 47-60.

的认知边界。而当前的 AI 革命，正以其前所未有的广度和深度，重塑着人类的认知图景。

科学发展的历史表明，知识体系的重大进步往往伴随着研究范式（Paradigm）的转换[1]。文明以降，以观察、记录与归纳为核心的经验主义构成了科学的第一范式。随后，以牛顿、爱因斯坦为代表的、通过公理化与逻辑演绎揭示宇宙底层规律的理论推演，塑造了科学的第二范式。随着计算机的诞生，通过数值计算模拟星系演化、气候变化等复杂现象的计算科学，崛起成为第三范式。而近几十年，互联网与传感器的普及催生了大数据时代，从海量数据中发现隐藏模式与相关性的数据密集型科学，则被公认为第四范式[2]。这四大范式构成了现代科学方法论的基础框架。

此刻，一个由人类智慧与机器智能深度耦合、共生演化的新的研究范式正在浮现。它超越了仅仅将计算机作为数据分析工具的第四范式边界，其核心特征在于将 AI 从一个被动的分析工具，提升为知识创造过程的积极参与者[3]。在这个新范式中，AI 不仅能够高效处理数据，更能够提出科学假说、设计实验方案，在广阔的可能性空间中探索人类研究者未曾设想的解决路径。这种人机协同的知识生产模式正在各个学科领域展现其颠覆性潜力，尤其是在那些传统方法难以企及的复杂领域。教育研究，作为一个涉及人类认知、情感、社会互动等多重复杂因素，具有多维度、高动态、弱规律性特征的复杂领域，正是检验这一新范式独特优势的重要场域。因此，人工智能驱动的科学研究（AI for Science，AI4S)在教育研究中的应用，不仅是技术迁移的尝试，更是对新型知识生产模式及其深层运作机制的前瞻性探索。

[1] 托马斯·库恩, 伊安·哈金. 科学革命的结构[M]. 金吾伦, 胡新和, 译. 北京大学出版社, 2012.
[2] Hey T, Tansley S, Tolle K, 等. 第四范式: 数据密集型科学发现[M]. Bei jing: 科学出版社, 2012.
[3] Wei J, Yang Y, Zhang X, et al. From AI for science to agentic science: a survey on autonomous scientific discovery[EB/OL]. (2025-08-18)[2025-09-14]. http://arxiv.org/abs/2508.14111.

1.1 教育研究的独特性：范式之争与实践鸿沟

人工智能驱动的科学研究（AI4S），这个脱胎于自然科学的强大新范式，当其试图跨越学科界限，向教育研究领域迁移时，正遭遇着一系列深刻而根本的挑战。其核心症结在于自然科学与教育科学在哲学基础上存在深刻差异，这种差异体现在世界观、知识观和方法论等多个层面。理解这种差异对于探讨 AI4S 在教育研究中的应用至关重要。

正是在这种哲学分野的土壤上，教育研究领域长期存在两个显著特征：一是持续不休的“范式之争”，其内部共识难以达成；二是研究成果与教育实践之间难以逾越的“实践鸿沟”。这些固有特征不仅深刻影响着该领域的自身发展，也构成了 AI 技术应用在此落地时必须面对、且亟待解决的复杂现实语境。

1.1.1 范式之争的历史遗产

教育研究领域的第一个显著特征是内部存在深刻的哲学分野与方法论张力，这种现象被学界称为“范式战争”（Paradigm Wars）[1]。与自然科学领域相对统一的实证主义传统形成鲜明对比，教育研究的发展历程始终伴随着关于其科学性质的激烈论辩。

早在 20 世纪上中叶，教育研究在追求科学合法性的驱动下，借鉴自然科学，特别是物理学的方法论，形成了以量化、实验和因果规律探寻为特征的实证主义传统。实证主义基于实在论的本体论立场，认为存在独立于人类意识之外的客观现实世界。其认识论则坚持客观主义，主张研究者应当保持价值中立，通过科学方法发现客观真理。这一时期的代表性研究者如约翰·B·卡罗尔（John B. Carroll），致力于将学习过程分解为可量化的变量体系，构建具有普遍适用性的学校学习模式[2]。李·克隆巴赫（Lee Cronbach）在

[1] Gage N. The paradigm wars and their aftermath a “historical” sketch of research on teaching since 1989[J]. Educational Researcher, 1989, 18(7): 4-10.

[2] Carroll J B. The carroll model: a 25-year retrospective and prospective view[J]. Educational Researcher, 1989, 18(1): 26-31.

1957 年的经典论述中，进一步阐明了实证主义对控制变量和寻求普遍规律的追求（尽管他本人后来也有所反思）[1]。

然而，随着研究者日益认识到人类行为的复杂性、情境性和价值负载性，诠释主义范式应运而生[2]。诠释主义持相对主义的本体论立场，主张现实是多元的，由社会和个人主观建构。在认识论上坚持主观主义，强调知识是研究者与研究参与者互动中共同建构的产物。诠释主义研究者将研究重点从发现客观规律转向理解行动者赋予其行为和经历的主观意义，由此推动了民族志、案例研究等质性研究方法的发展。

随后，批判理论范式的出现进一步丰富了教育研究的理论图景。批判理论将研究目标指向揭示和改变社会权力结构与不平等现象，强调研究的解放性价值[3]。这些不同范式在基础假设上的根本分歧，引发了 20 世纪 70 年代至 90 年代的激烈争论。虽然当前教育研究领域已进入多元范式并存的阶段[4]，但这种哲学多元性对任何试图引入的新方法或技术都提出了严峻挑战。任何单一的、技术驱动的研究模式（如 AI4S 内含的实证主义倾向）都必然会与教育研究的复杂传统产生摩擦与碰撞。

1.1.2 理论与实践的断裂

第二个核心特征是理论与实践之间存在的巨大鸿沟，即“研究-实践鸿沟”（Research-Practice Gap）[5]。这一问题比范式之争更为根本和持久，直接影响着教育研究的社会价值实现。学术界产生的理论、模型和实证发现，往往难以被一线教育工作者有效理解、采纳和应用到真实的课堂环境中，一项新的循证实践平均需要长达 17 年

[1] Cronbach L J. The two disciplines of scientific psychology[J]. American Psychologist, 1957, 12(11): 671-684.
[2] Nickerson C. Interpretivism paradigm & research philosophy[EB/OL]. (2024-02-13)[2025-09-14]. https://www.simplypsychology.org/interpretivism-paradigm.html.
[3] Denzin N K, Lincoln Y S. The SAGE handbook of qualitative research[M]. Fifth edition. Los Angeles London New Delhi Singapore Washington DC Melbourne: SAGE, 2018.
[4] Taylor P C, Medina M. Educational research paradigms: from positivism to pluralism[J]. College Research Journal, 2011, 1(1): 1-16.
[5] Rycroft-Smith L. Knowledge brokering to bridge the research-practice gap in education: where are we now?[J]. Review of Education, 2022, 10(1): e3341.

才能得到广泛应用[1]。

这并非简单的信息传递问题，其症结在于研究者与实践者两个社群在文化、语言、激励机制和知识观上的深层断裂。研究者追求理论的普适性和学术的严谨性，以发表高质量论文为主要目标，使用专业化的学术语言；实践者则更关注具体情境下的实际问题解决，需要易于理解和操作的方案，并且必须考虑现实中的各种约束条件。这种结构性的脱节，使得教育知识的转化异常困难，也是教育改革屡屡收效甚微的深层原因。

21 世纪初兴起的循证教育（Evidence-Based Education）运动试图通过强调科学证据来弥合这一鸿沟。然而，这一运动过度强调随机对照试验等特定研究方法，被批评为推行狭隘的后实证主义世界观，忽视了质性研究和批判性研究的价值[2]。这一争议，在某种程度上，可被视为引发了教育研究领域的第二次范式战争。为应对这一挑战，教育研究界发展出了新的方法论路径，如设计型研究（Design-Based Research, DBR）强调在真实情境中进行迭代设计，知识转化（Knowledge Translation, KT）和实施科学（Implementation Science）则系统研究如何将研究成果有效应用于实践。

这两个核心特征共同构成了教育研究的独特生态。任何新技术或新方法要在这一领域产生实质性影响，都必须直面哲学多元性的挑战，并为弥合研究与实践的鸿沟提供切实可行的解决方案。这也是评价 AI4S 在教育研究中应用价值的基本标准。

1.2 技术对于教育研究的适应性

探讨 AI4S 在教育研究中的应用，不仅需要理解教育研究的独特性，还需要审视技术本身与教育研究的契合程度。传统 AI 技术在教育研究领域的应用相对有限，这一现象可以通过莫拉维克悖论

[1] Bauer M S, Kirchner J. Implementation science: What is it and why should I care?[J]. VSI:Implementation Science, 2020, 283: 112376.

[2] Ponce O A, Pagán-Maldonado N, Gómez Galán J. Philosophy of educational research: new epistemological, methodological and historical approach[J]. International Journal of Educational Excellence, 2020, 6(2): 63-79.

（Moravec's Paradox）得到深刻解释。

莫拉维克悖论由机器人与人工智能专家汉斯·莫拉维克于 1988 年提出，其核心观察是：让计算机在智力测试或棋类游戏中达到成年人水平相对容易，而赋予它们一岁孩童的感知和行动能力却极其困难[1]。这一发现颠覆了早期计算机科学家的直觉认识。人类需要长期学习才能掌握的高级认知任务，如数学运算、逻辑推理和策略游戏，对计算机而言反而容易实现；而人类与生俱来的基础能力，如视觉感知、物体操控和自然语言理解，对计算机而言却构成了巨大挑战。

从进化论视角看，这一悖论有其深层原因。人类的基础感知和运动能力是数亿年进化的产物，这些能力已深深编码在大脑的感觉运动区域中。莫拉维克将人类有意识的推理过程形容为“人类思想最薄的一层表皮”，其之所以有效，完全依赖于更为古老、更为强大但通常是无意识的感觉运动知识的支撑。相比之下，抽象思维能力，如数学和逻辑，是演化史上非常晚近的“新技巧”，人类对其掌握程度有限，生物学实现也不够高效。因此，用 AI 来逆向工程一个演化了数亿年的技能，远比复制一个仅有近万年历史（甚至可能还要短得多）的技能要困难得多。

莫拉维克悖论为理解传统技术与教育研究的关系提供了重要视角。教育研究者的核心能力恰恰体现了莫拉维克悖论所描述的困难领域：理解学生话语的深层含义和情感色彩；将课堂观察置于学校文化、社区背景和政策环境的复杂语境中；把握学习过程、教学互动中的微妙动态；运用批判性思维解读教育现象的多重意义。这些能力都深度依赖于直觉判断和情境理解，而非简单的逻辑推理。

当教育研究者分析课堂对话记录时，其工作远不止于处理文字信息。研究者需要调动关于青少年心理、社会互动模式、文化规范、

[1] Moravec H. Mind children: the future of robot and human intelligence[M]. 4. print. Cambridge: Harvard Univ. Press, 1995.

权力关系等方面的内隐知识，才能准确把握言外之意、识别情感变化、理解互动的深层含义。这种理解能力，源于人类在漫长进化中，为适应复杂社会环境和进行有效协作所塑造的深层机制。

传统 AI 基于符号逻辑设计，擅长处理形式化、结构化的知识，但在面对自然语言的模糊性、多义性和情境依赖性时显得力不从心。这解释了为何传统 AI 在教育研究中的应用主要集中在莫拉维克悖论的“容易部分”。例如，智能辅导系统（ITS）将知识分解为原子化的知识点进行路径规划，教育数据挖掘（EDM）则从学生的点击流等行为数据中寻找可预测的模式。这些应用无疑具有价值，但在很大程度上回避了教育研究的核心挑战：理解和诠释充满意义的人类经验。

大语言模型的出现标志着 AI 在应对莫拉维克悖论方面取得了重大突破。通过在海量文本数据上的预训练，大语言模型获得了对人类语言、文化和社会常识的统计学理解能力。尽管这种理解并非真正的意识或情感，但它使 AI 首次能够有效处理承载复杂意义的自然语言。大语言模型能够识别访谈记录中的情感细微差别[1]，理解课堂对话中的隐含关系，生成符合特定教育情境的恰当反馈。

这种技术突破的意义在于，AI 不再局限于处理结构化数据的工具角色，而是能够参与到教育研究的核心任务中，即理解和诠释人类的学习经验与教育互动。大语言模型使 AI 能够在保持分析规模优势的同时，深入到教育现象的意义层面，这正是传统技术长期无法企及的领域。正是这种“跨越悖论”的能力，使得生成式 AI 对教育研究的冲击是根本性的、范式级别的。这一根本性的技术突破，正是本文所探讨的新的范式的起点。

1.3 定义教育研究中的 AI4S

在理解了教育研究的独特性质以及技术适应性的挑战后，我们

[1] 陈向东，陈鹏，张蕾. 基于大语言模型的教育质性研究：理论与实践.上海：华东师范大学出版社，2026。

可以更准确地界定“教育研究中的 AI4S”这一新兴概念。本节将正式提出 AI4S-Ed（AI for Science in Education Research）的定义，并通过与教育中的人工智能（AIED）和学习分析（LA）等相关概念的比较，明确其独特定位。这种概念辨析的重要性在于，它揭示了 AI4S-Ed 的本质特征：从运用 AI 改进教学实践转向运用 AI 变革教育研究过程本身。

教育研究中的 AI4S 可以界定为：一个新兴的研究范式，系统性地运用人工智能技术，特别是以大语言模型、生成式人工智能和自主智能体系统为代表的技术，来增强、自动化和加速关于学习、教学及教育系统的科学探究的整个过程。这一范式涵盖从海量非结构化数据中生成研究假设，设计复杂因果模型，自动化文献综述与理论构建，以及模拟教育政策干预效果等多个关键环节。

为准确把握 AI4S-Ed 的独特性，有必要将其与教育领域已有的相关概念进行系统比较：

教育中的人工智能（AIED）作为一个成熟的交叉学科领域，其核心目标是设计、开发和评估支持教学的 AI 驱动系统。AIED 的典型成果包括智能辅导系统、自适应学习平台、教育机器人和个性化推荐系统[1]。AIED 的根本目标是教育干预，它关心的问题是“如何利用 AI 让学生学得更好？”。而 AI4S-Ed 的目标是科学发现，它关心的问题是“关于学习和教学，我们能发现什么新的、可泛化的知识？”。

学习分析（LA）根据学习分析研究学会（SoLAR）的定义[2]，是“关于学习者及其学习过程的数据的收集、分析、解释和交流，旨在提供理论上相关且可操作的洞见，以增强学习与教学”。LA 虽然运用数据科学和 AI 技术，但主要聚焦于优化特定教育系统内的学习

[1] Wang S, Wang F, Zhu Z, et al. Artificial intelligence in education: a systematic literature review[J]. Expert Systems with Applications, 2024, 252: 124167.

[2] SoLAR. What is learning analytics[EB/OL]. 2025[2025-10-08]. https://www.solaresearch.org/about/what-is-learning-analytics/.

过程。其典型产出包括数据仪表盘、风险预警模型和个性化学习路径推荐。LA 通常在具体情境中解决实践问题，而 AI4S-Ed 则致力于生成和检验具有普遍意义的教育理论。简言之，LA 关注特定系统中的个体风险识别，AI4S-Ed 则探究跨系统的普遍性因果机制。

从思想渊源看，AI4S-Ed 与计算社会科学（CSS）有着密切联系[1]。CSS 利用计算工具、大规模数据集和模拟方法研究复杂社会现象，为传统社会科学方法难以触及的问题提供新的研究路径。AI4S-Ed 可视为 CSS 在人工智能时代的延伸和深化，不仅运用计算工具分析数据，更将 AI 作为知识发现的合作者，实现从数据分析到理论生成的跨越。

表 1-1 展示了这些相关概念在主要维度上的差异：

表 1-1 教育领域 AI4S 相关概念比较

特征	自然科学中的 AI4S	学习分析 (LA)	教育中的人工智能 (AIED)	教育研究中的科学智能 (AI4S-Ed)
主要目标	发现基本规律，生成新颖的科学理论。	优化学习过程，为利益相关者提供可操作的反馈。	开发智能工具以支持和传递教学。	生成和检验关于学习与教育的新颖科学理论。
探究对象	自然现象、物理定律、化学反应、生物过程以及宇宙演化等自然世界中的客观规律和机制。	特定数字学习环境（如 LMS、MOOC 平台）中学习者的行为数据、互动数据和表现数据。	人工智能技术在教育领域中的应用、设计与实施，包括智能教学系统、自适应学习系统、AI 赋能的评估工具、教育机器人、AI 辅助管理等。	教育领域中普遍性的学习规律、教学机制、教育政策的因果关系，以及教育系统的深层运行原理。
本体论立场	实在论：假设存在一个客观的、可被发现的实在。	实用主义/后实证主义：关注可测量的结果和概率性真理。	实用主义为主，但其底层理论基础广泛，关注 AI 工具和系统在教育实践中能否有效达成教育	认识论上备受争议：在 AI 的实证主义倾向与教育的建构主义传统之间的紧张地带运作。

[1] Lazer D, Pentland A, Adamic L, et al. Computational social science[J]. Science, 2009, 323(5915): 721-723.

特征	自然科学中的 AI4S	学习分析 (LA)	教育中的人工智能 (AIED)	教育研究中的科学智能 (AI4S-Ed)
			目标。	
核心产出	新知识、可证伪的假设、经同行评审的发现。	数据仪表盘、预测、干预措施、推荐。	各类 AI 赋能的教育技术产品与系统，例如智能辅导系统、自适应平台、教育机器人。	新研究问题、因果模型、理论洞见、经同行评审的发现。
典型问题	“这个蛋白质的三维结构是什么？”	“哪些学生在这门课程中有不及格的风险？”	“这个聊天机器人如何能最好地教授分数？”	“教学策略 X 与不同学生群体元认知能力 Y 的发展之间存在怎样的因果关系？”

通过这一比较框架可以看出，AI4S-Ed 代表着一种全新的研究议程。它试图将 AI 在自然科学中展现的知识发现能力，应用于充满复杂性、情境性和价值判断的教育研究领域。这种应用并非简单的技术迁移，而是需要在 AI 的实证主义倾向与教育研究的多元传统之间寻找平衡。

AI4S-Ed 作为新的知识生产范式，其独特性体现在三个关键维度。在认识论层面，它实现了从寻找客观规律到理解主观意义的转向。生成式 AI 的意义理解能力使教育研究能够深入探究学习体验、意义建构和价值追求等核心问题。这种能力并非要取代人类理解，而是增强和扩展研究者的认知能力，使我们能够在更大规模上把握教育的复杂性。

在方法论层面，AI4S-Ed 超越了单纯的预测和优化，发展出理解和生成并重的研究路径。通过生成式智能体、多模态分析等创新方法，AI 成为研究过程的积极参与者。这种参与旨在激发和放大研究者的创造潜能，开拓以往难以触及的研究领域。

在价值层面，AI4S-Ed 坚持教育的多元价值追求，强调价值敏感的技术应用。这要求在运用 AI 时保持批判性反思，确保技术服务于教育的根本目的，即促进人的全面发展和社会公平正义。

1.4 报告的结构

基于前述分析，本报告将系统探究 AI4S 应用于教育研究的全面影响。报告的核心论点是：AI4S-Ed 作为一场深刻变革，在重塑教育知识生产全流程的同时，也对教育研究的方法论基础、质量标准、伦理规范和研究者素养提出了根本性挑战。这场变革的发展方向，取决于我们能否在充分利用其潜力的同时，建立相应的批判性反思框架和审慎的实践规范。

为系统地展开这一核心论述，报告将围绕以下几个相互关联的核心议题，构建一个从宏观到微观、从理论到实践、从赋能到风险的完整分析框架。

第二章将探讨 AI4S-Ed 带来的范式变革。从科学哲学视角审视范式演进规律，分析大语言模型时代的独特性。重点论述 AI4S-Ed 如何通过拓展研究视域、创新研究方法和重构研究过程，从根本上改变教育研究的运作模式。

第三、四、九、十章聚焦方法论层面的变革。将分别探讨 AI 如何赋能质性研究和量化研究，分析其在课堂互动分析、政策评估、数据提取、合成数据生成等方面的应用，以及相应的风险挑战。特别关注 AI 合成数据这一颇具争议的新领域，深入讨论其应用价值和方法论争议。第十章将探讨 AI 在教育理论构建中的潜力，尤其是通过支持思想实验来拓展理论疆域的可能性。

第五、八章考察研究生态系统的重构。第五章探讨 AI 对研究质量标准的影响，包括如何建立新的可重复性、透明度标准，以及传统学术规范如何适应新环境。第八章从元研究视角分析 AI 如何改变知识生产体系，包括对研究者素养的新要求、人机协同模式的发展，以及新型研究组织形式的涌现。

第六章专门探讨 AI 在弥合研究与实践鸿沟方面的作用。将从理论层面分析 AI 如何促进知识转化的机制，探讨其如何实现从信息处

理到意义生成的跃升，并展望人机协同的知识转化新生态。

第七章系统审视 AI 带来的伦理挑战。除各章节中的具体讨论外，本章将从更宏观层面分析 AI 对研究议程的影响、研究伦理的新维度，以及构建实用伦理指南的必要性。

通过这一结构安排，本报告旨在为教育研究者、政策制定者和技术开发者提供一个全面的分析框架。需要明确的是，教育的核心问题，即人的全面发展、社会公平正义、知识与智慧的本质，具有深刻的哲学和伦理维度，并非纯粹的技术问题。AI 可以为探究这些问题提供新的可能性，如通过大规模数据分析揭示教育不平等的系统性特征，或通过模拟实验探索学习机制。然而，迄今为止作者仍然坚信，技术本身无法回答“我们应该追求什么样的教育？”以及“为了什么目的而教育？”这类根本性的价值问题。

因此，AI4S-Ed 的发展需要在技术创新与人文关怀之间保持平衡。这要求培养新一代研究者，他们既要掌握数据科学和 AI 技术，更要具备深厚的人文素养和批判性思维能力。他们需要理解研究范式的历史演进，认识知识的多样性本质；需要尊重实践的复杂性和教师的专业智慧；更需要对 AI 的能力边界和潜在偏见保持清醒认识。

正如前文所述，历史上的许多重大技术变革都曾深刻改变了人类的认知方式和社会结构。今天，大语言模型和生成式 AI 在教育研究中的应用，可能标志着另一次深刻转型。当 AI 开始参与意义理解和知识创造，我们正在见证的可能是人类认知方式和知识体系的一次根本性重构。

本研究报告将坚持一个基本立场：对教育研究 AI4S 的探讨保持开放性和反思性，避免陷入技术决定论或技术悲观主义的任何一端。我们既要认真对待 AI 带来的变革潜力，也要清醒认识其复杂性和不确定性。伦理考量应贯穿于所有讨论之中：在探讨方法创新时考虑公平性和包容性，在分析效率提升时反思价值预设，在展望未来时

评估社会影响。

本报告的目的不是提供确定的发展蓝图，而是全面呈现 AI4S-Ed 这一新兴范式的多个维度，识别机遇与风险，促进研究共同体的深入思考。在一个技术变革与全球地缘政治动荡交织、充满不确定性的时代，保持批判性的开放态度，在实践探索中不断反思和调整，可能是我们能够持守的最负责任的研究立场。只有通过持续的实践推进、理论探讨和伦理反思，教育研究共同体才能逐步形成对 AI4S 的成熟理解，并在技术可能性与教育价值之间寻找动态平衡。这将是一个需要不同背景研究者共同参与、在对话中推进集体理解的长期过程。

第 2 章 AI 重塑教育研究范式

人工智能（AI）正以一种颠覆性的力量重塑人类认知的疆域。以“人工智能驱动的科学研究”（AI4S）为标志，人工智能已经超越了智能工具的范畴，上升为一种全新的科学研究范式[1][2]。其核心理念是利用人工智能技术重塑科学研究的过程与方法，通过智能化手段拓展人类探索未知的边界[3]。这一范式已在多个领域取得突破，如蛋白质结构计算、分子动力学模拟、智能驱动材料设计等[4][5][6]，昭示着人工智能正在成为科学研究不可或缺的助推器。

教育研究领域同样面临 AI4S 浪潮的冲击。随着教育数字化转型的深入，教与学过程产生了海量的数字化痕迹，为深入理解复杂的教育现象提供了前所未有的数据基础[7]，但也对传统研究范式提出了严峻挑战。以大语言模型为代表的新一代人工智能技术的出现，为教育领域的 AI4S 开辟了关键路径。通过在海量数据上的预训练，这类 AI 技术习得了广泛、通用的世界知识表征，并展现出对复杂世界全面、深入的理解和建模能力[8][9]。

科学研究范式是研究共同体所认同的问题视域、方法规范、话语体系的有机统一[10]。以大语言模型为代表的 AI 技术的出现能否以及如何撼动传统范式结构，并最终引发教育研究的革命性变革，是一个亟待厘清的重大议题。近期研究明确指出，AI 智能技术正在为

[1] 王飞跃, 缪青海. 人工智能驱动的科学研究新范式：从 AI4S 到智能科学[J]. 中国科学院院刊, 2023, 38(4): 536-540.
[2] 李国杰. 智能化科研（AI4R）：第五科研范式[J]. 中国科学院院刊, 2024, 39(1): 1-9.
[3] Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence[J]. Nature, 2023, 620(7972): 47-60.
[4] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.
[5] Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2[J]. Nature Communications, 2022, 13(1): 1265.
[6] Leng C, Tang Z, Zhou Y G, et al. Fifth Paradigm in Science: A Case Study of an Intelligence-Driven Material Design[J]. Engineering, 2023, 24: 126-137.
[7] Hakimi L, Eynon R, Murphy V A. The Ethics of Using Digital Trace Data in Education: A Thematic Review of the Research Landscape[J]. Review of Educational Research, 2021, 91(5): 671-717.
[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[C]//Guyon I, Luxburg U V, Bengio S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.
[9] Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models[EB/OL]. (2022-07-12)[2024-04-14]. http://arxiv.org/abs/2108.07258.
[10] 托马斯·库恩, 伊安·哈金. 科学革命的结构[M]. 金吾伦, 胡新和, 译. 北京大学出版社, 2012.

社会科学研究带来“范式层面的变革”[1]。作为社会科学的重要分支，教育研究因其独特的人文关怀内核而具有更加复杂的学理基础，AI4S 对其影响也因此更加深远和复杂。

这种重塑远非简单的技术叠加，而是一场深刻的生态系统变革。它催生了一个以人类智慧为战略核心、以 AI 智能体为协作伙伴、以开放共享的社会-技术基础设施为支撑的全新研究范式。为了在开篇直观地呈现这一变革的全貌，我们构建了“人机协同研究生态系统”机制图（如图 2-1 所示）。

该图描绘了人类研究者如何扮演战略制定与价值引领的角色，AI 智能体如何从工具演进为协作者乃至科学家，以及所有参与者如何通过一个共享的基础设施进行交互，从而实现网络化、累积性、并行化的知识生产。在接下来的内容中，我们将首先奠定分析的理论基石，随后遵循“问题-方法-过程”的整合性框架，系统剖析这张图景背后的具体变革。

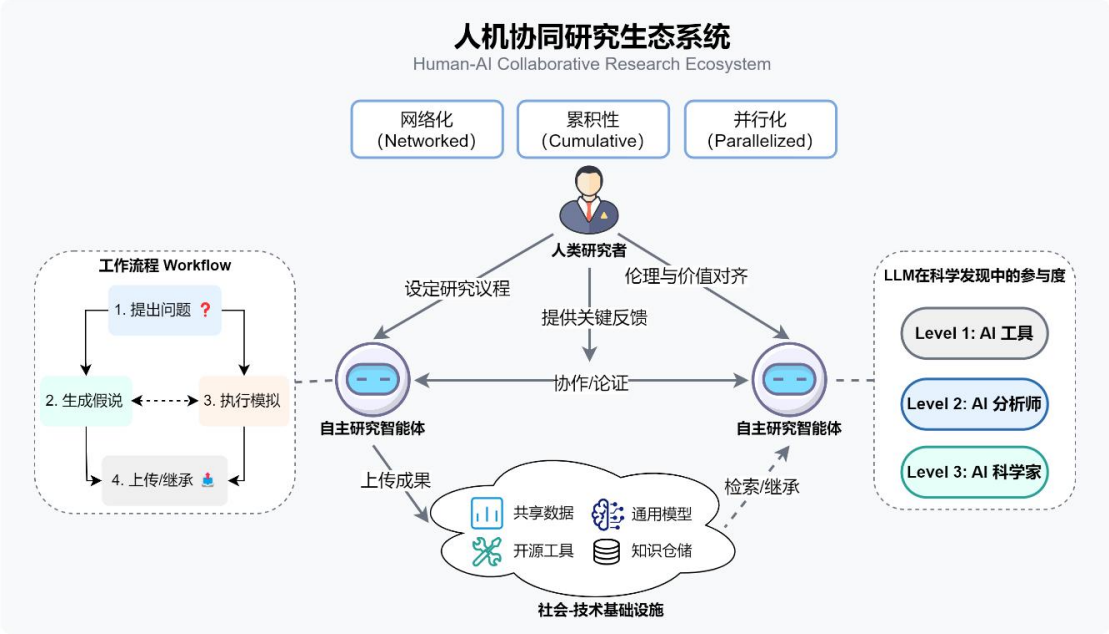

图 2-1 人机协同研究生态系统机制图

[1] Grossmann I, Feinberg M, Parker D C, et al. AI and the transformation of social science research[J]. Science, 2023, 380(6650): 1108-1109.

2.1 教育研究范式变革的理论基石

2.1.1 科学研究的范式

托马斯·库恩（Thomas Kuhn）在《科学革命的结构》中，将“范式”定义为特定科学共同体在特定时期从事研究的理论基础和实践规范。一个范式不仅包含抽象的理论和法则，更具体地规定了哪些是值得研究的“谜题”，以及解决这些“谜题”的合法工具和方法[1]。在范式主导下的“常规科学”时期，研究者的主要工作是在范式框架内进行“解谜”，而非挑战范式本身。

然而，当“常规科学”不断遭遇既有范式无法解释的“异常现象”时，危机便会产生。持续的危机会动摇科学家对原有范式的信心，最终可能引发“科学革命”，即旧范式被一个全新的、更具解释力的范式所取代的非累积性发展过程[2]。

库恩的理论因其强调革命的“非理性”和“整体转换”而引发争议。作为回应和发展，伊姆雷·拉卡托斯（Imre Lakatos）提出了“研究纲领”的概念。他认为科学发展并非剧烈的断裂，而是一系列相互竞争的研究纲领的演替[3]。每个研究纲领都拥有一个不容置疑的“硬核”，以及一个由辅助假说构成的、可灵活修改的“保护带”。一个进步的纲领能够不断预测新事实，而一个退化的纲领则只能在事后勉强修补。

尽管视角不同，他们都为我们提供了审视科学发展的分析工具。库恩的“范式-异常-革命”模型为我们识别根本性变革提供了判据，而拉卡托斯的“研究纲领”则让我们关注变革过程中的延续与竞争。本节以“问题-方法-过程”为整合性框架，系统考察以大语言模型为代表的 AI 技术引发的教育研究范式之变。

[1] 托马斯·库恩, 伊安·哈金. 科学革命的结构[M]. 金吾伦, 胡新和, 译. 北京大学出版社, 2012.
[2] 托马斯·库恩, 伊安·哈金. 科学革命的结构[M]. 金吾伦, 胡新和, 译. 北京大学出版社, 2012.
[3] Lakatos I. Falsification and the Methodology of Scientific Research Programmes[M]//Harding S G. Can Theories be Refuted? Essays on the Duhem-Quine Thesis. Dordrecht: Springer Netherlands, 1976: 205-259.

2.1.2 教育研究范式演进

将科学哲学的理论应用于教育研究的历史，可以看到一条范式演进的脉络。吉姆·格雷[1]提出的科学研究范式四个阶段，即经验科学、理论科学、计算科学和数据密集科学的演进逻辑，也大致可以套用于教育研究：早期的经验科学范式通过访谈、观察等质性方法积累事实[2]；随后的理论科学范式致力于构建稳定的解释框架以回答“为何”[3]；20 世纪 90 年代，计算科学范式借助建模与模拟推动了理论与实证的融合[4]；进入 21 世纪，数据密集科学范式（或称第四范式）则利用数据挖掘和学习分析等方法，从海量数据中发现隐藏的关联与模式[5]。数据驱动范式的代表性成果有基于学习痕迹数据对比不同学习风格下的认知建构差异[6]，依托学习管理系统日志动态监测学习参与度的演化模式[7]等。

第四范式极大地提升了教育研究的广度和效率，但随着教育数字化转型的深入，其局限性也日益凸显：

（1）海量非结构化数据带来的深度理解难题：教育过程产生了海量、多模态、非结构化的数据（如课堂录像、讨论文本、操作日志等）[8]。第四范式擅长处理结构化数据和发现相关性，但在深度理解这些富含情境意义的非结构化数据、揭示其背后的因果机制方面却存在不足。数据量的激增与教育意义挖掘深度的矛盾日益突出。

（2）面对系统复杂性的方法论局限：教育是一个典型的复杂自适应系统，充满了动态交互、非线性效应和价值涉入。传统的统计

[1] Gray J, Szalay A. eScience-A transformed scientific method[J]. presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, 2007.
[2] 黄荣怀, 王欢欢, 张慕华, 等. 面向智能时代的教育社会实验研究[J]. 电化教育研究, 2020, 41(10): 5-14.
[3] Swaminathan V, Lamberton C, Sridhar S, et al. Paradigms for Progress: An Anomaly-First Framework for Paradigm Development[J]. Journal of Marketing, 2023, 87(6): 816-825.
[4] 郑永和, 严晓梅, 王晶莹, 等. 计算教育学论纲：立场、范式与体系[J]. 华东师范大学学报(教育科学版), 2020, 38(6): 1-19.
[5] 米加宁, 章昌平, 李大宇, 等. 第四研究范式:大数据驱动的社会科学研究转型[J]. 学海, 2018(2): 11-27.
[6] van den Beemt A, Buijs J, van der Aalst W. Analysing Structured Learning Behaviour in Massive Open Online Courses (MOOCs): An Approach Based on Process Mining and Clustering[J]. The International Review of Research in Open and Distributed Learning, 2018, 19(5).
[7] Henrie C R, Bodily R, Larsen R, et al. Exploring the potential of LMS log data as a proxy measure of student engagement[J]. Journal of Computing in Higher Education, 2018, 30(2): 344-362.
[8] Hakimi L, Eynon R, Murphy V A. The Ethics of Using Digital Trace Data in Education: A Thematic Review of the Research Landscape[J]. Review of Educational Research, 2021, 91(5): 671-717.

模型和数据挖掘算法往往基于简化的假设，难以有效刻画和模拟这种系统层面的复杂性。例如，在评估一项教育政策的长期、间接和非预期的影响时，第四范式的方法工具便显现出局限性。

这些现象的不断累积，使得数据密集范式局限也日益凸显，并倒逼研究范式的革新突破。在此背景下，以大语言模型等新一代 AI 技术为核心驱动力的 AI4S 应运而生。

2.1.3 大语言模型时代的科学研究

如果仅仅将当前变革视为第四范式的简单延续，就会严重低估其革命性。以大语言模型为代表的 AI 技术引领的范式变革，在延续历史发展逻辑的同时，更在多个维度上实现了质的突破。

其一，研究主体性的拓展：从计算工具到智能协作者

从计算范式到数据密集范式，AI 在研究中主要扮演的是一个高效的计算工具或分析助手。而新一代 AI 技术引领的范式，则标志着 AI 的角色正从一个被动的工具转变为一个主动的协作者，甚至是具备一定主体性的智能体。近期研究将这类 AI 技术在科学发现中的参与度划分为从工具、分析师到科学家的三个递进层次，清晰地展示了 AI 正沿着一条从辅助执行到主动探索的路径演化[1]。当一个智能体系统能够自主提出假说、规划和执行实验并得出结论时[2]，它已经不再是一个简单的工具，而是在一定程度上参与了研究的主体性建构。这意味着，教育研究的智力探索，正从传统的研究者单中心主导，走向“人-机双主体协同”的崭新图景，这是对研究主体边界的一次根本性拓展。

其二，知识来源的革命：从被动发现到主动生成

数据密集范式使我们能够从数据中“发现”前人未见的关联和模式。而大语言模型为代表的 AI 技术，其“生成式”的特质，则开

[1] Zheng T, Deng Z, Tsang H T, et al. From automation to autonomy: a survey on large language models in scientific discovery[EB/OL]. (2025-05-19)[2025-07-30]. http://arxiv.org/abs/2505.13259.

[2] Gao C, Lan X, Li N, et al. Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives[EB/OL]. (2023-12-19)[2024-02-07]. http://arxiv.org/abs/2312.11970.

启了“生成”全新知识的可能性。这不仅体现在它可以生成新颖的科学假说、逼真的模拟情境[1]，甚至可以生成完整的科学论文[2]。这种生成能力，正在挑战传统的知识观。最新研究表明，大语言模型在假设生成方面展现出了前所未有的能力，能够通过迭代优化过程不断改进假设质量[3]。在医学研究中，GPT4 已经成功地生成了经过实验室验证的新颖药物组合假设，展现了 AI 从被动工具向主动假设生成者的转变[4]。

最新的哲学探讨为此提供了一些更为深刻的理论视角。例如最近有研究者提出了“适应性认识论”（Adaptive Epistemology）这一观点，以应对生成式 AI 为社会科学研究带来的认识论变革[5]。该理论认为，在算法与平台深度嵌入研究过程的后数字时代，知识不再是一个从外部现实中被动提取的稳定产物，而是由人类研究者与非人类智能体（如 AI）在持续的互动中动态地、共同建构和生成的。在此框架下，智能体的角色从现实的被动观察者，转变为能够主动建构本体论的参与者。这种转变标志着机器从规则跟随者转变为假设生成器和测试者，能够在没有人类干预的情况下进行自动化探索，从被动的知识库变成主动的假设探索者[6]。传统研究的重点是寻找答案，而适应性认识论则指出，在与生成式 AI 的互动中，最关键的研究活动转变为“建构问题”——即如何提出在计算上最优、在策略上有效且符合伦理反思的“提示”（prompts）。因此，AI 驱动的范式变革远不止于方法论层面。它标志着教育知识的来源，正在从对外部教育现象的“二次提炼”，转变为人机系统内部知识的“共

[1] Zhang Z, Zhang-Li D, Yu J, et al. Simulating classroom education with LLM-empowered agents[EB/OL]. (2024-11-27)[2025-07-30]. http://arxiv.org/abs/2406.19226.

[2] Zheng T, Deng Z, Tsang H T, et al. From automation to autonomy: a survey on large language models in scientific discovery[EB/OL]. (2025-05-19)[2025-07-30]. http://arxiv.org/abs/2505.13259.

[3] Zhou Y, Liu H, Srivastava T, et al. Hypothesis generation with large language models[C]//Peled-Cohen L, Calderon N, Lissak S, et al. Proceedings of the 1st Workshop on NLP for Science (nlp4science). Miami, FL, USA: Association for Computational Linguistics, 2024: 117-139.

[4] Abdel-Rehim A, Zenil H, Orhobor O, et al. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment[J]. Journal of the Royal Society, Interface, 2025, 22(227): 20240674.

[5] Punziano G. Adaptive epistemology: embracing generative AI as a paradigm shift in social science[J]. Societies, 2025, 15(7): 205.

[6] Dhar V. The paradigm shifts in artificial intelligence[J]. Communications of the ACM, 2024, 67(11): 50-59.

同生产”，这无疑是对学科认识论基础的一次深刻撼动。

其三，学科边界的消融：从交叉研究到整合创新

学科交叉一直是教育研究发展的内在要求，但传统的交叉研究常常受限于不同学科间的话语壁垒和方法论隔阂。AI4S 所倡导的“AI+X”跨界整合范式，为突破这一困境提供了新的可能。AI 擅长整合跨领域的数据和知识，打破学科壁垒，实现深度的跨学科整合，以应对基础性挑战。像大语言模型这类 AI 技术本身就是一个跨学科知识的巨大熔炉，其“世界知识”（World Knowledge）内在地包含了来自心理学、社会学、经济学、计算机科学等多个领域的概念和模式[1]。这使得它天然成为一个跨学科对话的“通用语”和“中介平台”。例如像 AgentRxiv 这样为自主研究智能体设计的协作生态系统，通过建立统一的、可共享的知识库，使得来自不同学科背景的研究者（或 AI 智能体）能够在一个共同的平台上，基于彼此的发现进行累积性创新[2]。这种模式不再是简单的学科并置，而是实现了深度的“知识整合”，正在催生一个更加开放、动态、无边界的科研新生态，推动教育研究从“多学科合作”迈向真正的“超学科融合”。

2.2 研究视域的拓展

科学研究的起点是提出有价值的问题。传统上，问题的发现高度依赖研究者的个人学识、理论敏感性与偶然的灵感迸发。然而，大语言模型等新一代 AI 技术的出现正在改变这一局面。其在海量数据预训练过程中形成的广博“世界知识”[3]，为系统性地发现与生成科学问题提供了前所未有的能力。这种转变使得问题发现的过程，从依赖个人灵感的非系统性活动，转变为一种可被系统化、规模化、协作化的科学流程。

[1] Bommasani R, Hudson D A, Adeli E, et al. On the Opportunities and Risks of Foundation Models[EB/OL]. (2022-07-12)[2024-04-14]. http://arxiv.org/abs/2108.07258.

[2] Schmidgall S, Moor M. AgentRxiv: towards collaborative autonomous research[EB/OL]. (2025-03-23)[2025-07-30]. http://arxiv.org/abs/2503.18102.

[3] Ha D, Schmidhuber J. World Models[EB/OL]. (2018-03-28)[2024-05-05]. http://arxiv.org/abs/1803.10122.

2.2.1 拓展问题视野

大语言模型为代表的生成式AI技术凭借其跨领域、多层次、动态演化的世界知识体系，能够从广度、深度、关联性和颠覆性等多个维度拓展教育研究的问题视野。

广度上，AI能够横跨不同学科，从海量文献数据中自动捕捉研究前沿，识别出尚待探究的理论分歧和实证空白，提出具有前瞻性的研究问题，这大大拓宽了问题视野的广度。例如，最新研究显示，AI在教育领域的应用已经从简单的内容生成扩展到复杂的学习者建模和个性化教学设计，能够整合心理学、认知科学、教育技术等多个领域的知识，识别跨学科研究机会[1]。

深度上，借助多模态融合能力，AI技术可将课堂教学的文本、语音、视频等异构数据纳入统一的分析框架，深度挖掘影响教学效果的关键因素，提出更加立体和本质的问题。研究表明，AI在教育中的应用正朝着全学习者支持的方向发展，不仅关注学术能力，还包括社会、情感、动机、文化和语言特征等非认知技能[2]。

关联性上，基于其世界知识，AI技术能够发现不同学科概念和原理间隐藏的关联，实现跨学科知识的创造性融合，引领研究者提出具有开拓性的交叉问题。例如，在智能教育系统中，AI能够整合学习分析、认知心理学和计算机科学的知识，为个性化学习提供可重用和可扩展的技术架构[3]。

颠覆性上，AI技术所具有的类比推理、迁移泛化等能力，使其能跳出既有理论框架，在世界知识中发现隐藏的关联，提出颠覆性的理论假设，为教育理论创新提供新的生长点。“科学假说论证”（Scientific Hypothesis Evidencing, SHE）任务的提出，便利用大语

[1] Peláez-Sánchez I C, Velarde-Camaqui D, Glasserman-Morales L D. The impact of large language models on higher education: exploring the connection between AI and education 4.0[J]. Frontiers in Education, 2024, 9: 1392091.

[2] Mannekote A, Davies A, Pinto J D, et al. Large language models for whole-learner support: opportunities and challenges[J]. Frontiers in Artificial Intelligence, 2024, 7: 1460364.

[3] Xing W, Nixon N, Crossley S, et al. The use of large language models in education[J]. International Journal of Artificial Intelligence in Education, 2025, 35(2): 439-443.

言模型根据科学文献摘要自动判断其是否支持或反驳给定的研究假说，实验表明这类 AI 技术已展现出一定的科学假说论证能力[1]。借助 AI 的强大认知能力，教育研究者能够跳出固有的知识结构和思维定势，发现传统路径难以触及的新问题域。

这些维度相互交织，共同勾勒出 AI 技术拓展教育研究问题视野的丰富图景。借助 AI 的强大认知能力，教育研究者能够跳出固有的知识结构和思维定势，发现传统路径难以触及的新问题域。

2.2.2 自动化假说生成

最新的研究进展不仅证实了上述判断，还进一步揭示了将这种潜力转化为现实的系统性方法。当前，研究界已不再满足于将 AI 作为简单的“灵感激发器”，而是致力于构建能够系统性地创造、筛选和验证科学假说的“认知引擎”[2][3][4]。

这一转变的核心在于将科学假说生成（Scientific Hypothesis Generation, SHG）明确地构建为一个自然语言生成（NLG）任务。为此，研究者开发了专门的结构化数据集，其中 HypoGen 数据集是一个具有代表性的例子[5]。该数据集包含了从顶级计算机科学会议中提取的约 5500 个结构化的问题-假说对，其核心是一个包含“Bit-Flip-Spark”与“推理链”（Chain-of-Reasoning）的格式，为 AI4S 创新学习科学的内在逻辑提供了清晰的蓝图：

- Bit（常规假设）：明确陈述研究领域中一个普遍存在但有局限性的传统假设或方法。
- Flip（创新论点）：清晰阐述论文提出的、旨在推动领域进步的新方法或反驳性论点。

[1] Koneru S, Wu J, Rajtmajer S. Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences[EB/OL]. (2024-03-25)[2024-04-16]. http://arxiv.org/abs/2309.06578.
[2] Alkan A K, Sourav S, Jablonska M, et al. A survey on hypothesis generation for scientific discovery in the era of large language models[EB/OL]. (2025-04-07)[2025-07-30]. http://arxiv.org/abs/2504.05496.
[3] O’Neill C, Ghosal T, Răileanu R, et al. Sparks of science: hypothesis generation using structured paper data[EB/OL]. (2025-04-17)[2025-07-30]. http://arxiv.org/abs/2504.12976.
[4] Xiong G, Xie E, Shariatmadari A H, et al. Improving scientific hypothesis generation with knowledge grounded large language models[EB/OL]. (2024-11-04)[2025-07-30]. http://arxiv.org/abs/2411.02382.
[5] O’Neill C, Ghosal T, Răileanu R, et al. Sparks of science: hypothesis generation using structured paper data[EB/OL]. (2025-04-17)[2025-07-30]. http://arxiv.org/abs/2504.12976.

- Spark（核心洞见）：用简短的词语捕捉从“Bit”到“Flip”的核心思想飞跃。
- Chain-of-Reasoning（推理链）：提供一个详尽的叙事，记录科学家从识别问题到形成解决方案的完整心路历程，包括提出探究性问题、进行方法论反思、报告实验观察与结果等。

通过在这种结构化数据上进行微调，像大语言模型这类生成式AI被训练为以一种“条件性语言建模”的方式来生成假说。在推理时，研究者仅需向模型提供“Bit”（即问题陈述），模型便能自动生成相应的“Spark”和详细的“Chain-of-Reasoning”，最终形成一个完整、新颖且可行的假说。这种方法将抽象的问题发现潜力，转化为一个具体的、可重复的生成过程，为教育研究者提供了一个强大的工具，能够系统性地审视本领域的“常识”，并从中发现可能被忽略的创新点。

在实际应用中，谷歌的AI共同科学家系统（AI Co-Scientist）已经展现了这种框架的强大潜力。该系统基于Gemini 2.0构建，作为虚拟科学合作者帮助科学家生成新颖假说和研究提案，在药物重新定位和靶点发现等领域取得了显著成果[1]。这种方法将抽象的问题发现潜力，转化为一个具体的、可重复的生成过程，为教育研究者提供了一个强大的工具，能够系统性地审视本领域的“常识”，并从中发现可能被忽略的创新点。

2.2.3 整合多元知识源

单纯依赖AI模型内部的“世界知识”进行假说生成存在局限，可能导致模型重复其训练数据中已有的观点，缺乏真正的创新性。因此，前沿研究强调，必须将AI模型嵌入一个更广阔的知识生态系统中，整合多元化的信息来源，以生成真正有价值的假说（Brand et al., 2024）。

[1] Gottweis J, Natarajan V. Accelerating scientific breakthroughs with an AI co-scientist[EB/OL]. (2025-02-19)[2025-08-04]. https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/.

知识图谱增强的假说生成是一种有效的方法。最新研究提出了KG-CoI（Knowledge Grounded Chain of Ideas）系统，通过将外部结构化知识图谱整合到大语言模型中，显著提高了科学假说生成的准确性，并减少了推理链中的幻觉现象[1]。该系统包含三个关键模块：知识图谱引导的上下文检索、知识图谱增强的思维链生成，以及知识图谱支持的幻觉检测。

多智能体协作框架提供了另一种整合路径。HILMA（Human-in-the-loop Multi-agent Framework）框架基于结构化智能理论中的发散思维和收敛思维原理，为可靠的科学假说生成提供了创新性的人机协作方案[2]。该框架融合了实时、系统的知识检索增强机制，动态整合最新研究进展构建引用网络子图，为AI模型提供全面且最新的科学知识调研。此外，该框架通过模拟科学同行评议过程的多智能体论证方法增强假说生成，同时利用人类专家的直觉和专业知识进一步完善和多样化生成的假说。

科学知识图谱驱动的假说生成是第三种重要方法。最新研究开发了基于科学知识图谱的研究假说生成系统，该系统能够创建精确、可验证且逻辑有效的陈述，为教育研究中的假说形成提供了强大工具[3]。

文献与数据双驱动整合是另一种被验证的有效路径。这种方法认为，高质量的科学假说应同时具备理论自洽性（植根于现有知识体系）和经验有效性（从经验数据中发现新模式）。其具体流程包括：首先，利用AI技术对海量学术文献进行分析和综述，提取关键概念、识别研究空白，从而生成一系列基于现有知识的文献驱动假说。其次，让AI模型直接从大规模观测数据中发现潜在的模式和关

[1] Xiong G, Xie E, Shariatmadari A H, et al. Improving scientific hypothesis generation with knowledge grounded large language models[EB/OL]. (2024-11-04)[2025-07-30]. http://arxiv.org/abs/2411.02382.

[2] 陈子阳, 赵翔, 赵润豪, 等. 基于人机协作的多智能体科学假设生成[J]. 计算机研究与发展, 2025, 62(7): 1639-1652.

[3] Borrego A, Dessì D, Ayala D, et al. Research hypothesis generation over scientific knowledge graphs[J]. Knowledge-Based Systems, 2025, 315: 113280.

联，生成数据驱动假说。最后，将两种假说进行整合与精炼，例如以文献假说为理论指导，对数据驱动发现的模式进行解释，反之亦然[1]。

对于教育研究而言，这意味着未来的问题发现过程将是一个动态的三角验证过程：研究者提出初步问题，AI 从海量文献中寻找理论支撑和研究空白，同时从教育大数据（如学习管理系统日志、课堂互动数据）中挖掘经验证据，三者相互印证、相互启发，共同催生出既有理论深度又具现实关照的重大研究课题。

2.2.4 研究者角色的重塑

尽管自动化假说生成前景广阔，但它也带来了一个核心挑战，即如何平衡假说的新颖性（Novelty）与可行性（Feasibility）。研究发现，AI 在科学假说生成中面临着准确性和创新性的权衡问题，需要通过微调和人类专家的深度参与来实现最佳效果[2]。

这一挑战重新定义了研究者在知识发现初始阶段的角色。研究者的核心任务不再仅仅是作为思想的唯一“源头”，而是转变为一个多重角色的扮演者：

（1）议程设定者（Agenda Setter）：研究者负责定义研究的问题假设，即识别出领域内最值得被挑战的核心问题和传统假设。这是整个创新过程的起点，需要深厚的领域知识和批判性思维。

（2）假说策展人（Hypothesis Curator）：面对 AI 生成的海量假说，研究者需要运用其专业判断力进行筛选和策展，评估哪些假说具有真正的理论价值和实践可行性。

（3）创造性诠释者（Creative Interpreter）：一个看似“错误”或“幻觉”的输出，在富有经验的研究者眼中，可能蕴含着颠覆性

[1] Zhou Y, Liu H, Srivastava T, et al. Hypothesis generation with large language models[C]//Peled-Cohen L, Calderon N, Lissak S, et al. Proceedings of the 1st Workshop on NLP for Science (nlp4science). Miami, FL, USA: Association for Computational Linguistics, 2024: 117-139.

[2] Liu H, Zhou Y, Li M, et al. Literature meets data: a synergistic approach to hypothesis generation[C]//Che W, Nabende J, Shutova E, et al. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers). Vienna, Austria: Association for Computational Linguistics, 2025: 245-281.

的洞见。最新研究表明，AI 的“幻觉”现象实际上是其创造力的体现，对科学发现具有重要价值[1]。

（4）协作伙伴（Collaborative Partner）：在人机协作的新模式下，研究者与 AI 系统形成真正的合作关系。MIT 的 FutureHouse 平台展示了这种协作的可能性，其 AI 代理能够自动化科学进步路径上的许多关键步骤，包括文献检索、数据分析和假说生成，研究者则专注于更高层次的创新思考和价值判断[2]。

从这个角度看，问题发现范式的真正变革，并非 AI 取代了人类的创造力，而是研究的起点从一种依赖个人、偶发的顿悟式创新，演变为一种人机协同的、系统化的协作式构想。我们正在构建的，是一种能够大规模探索假说空间的认知生态系统，而人类研究者则是这个生态系统的设计师、管理者和最终的价值判断者。

2.3 研究过程的重构

以大语言模型为代表的 AI 技术对教育研究范式的影响，不仅体现在问题发现和方法创新上，更深刻地体现于对整个研究过程的系统性重构。在这一重构过程中，“端到端”（End-to-End）和“人在回路”（Human-in-the-Loop）理念的融合，正在开启人机协同的新范式。

“端到端”概念常见于机器学习领域，旨在构建一个统一的模型架构，将原本流程中各个独立的步骤整合为一个完整的系统，以自动学习数据中的特征表示和映射关系[3]。而“人在回路”则强调将人的知识、经验和反馈引入到人工智能系统的训练迭代过程中，以动态调整和优化系统，形成人机协同[4]。尽管两者背景不同，但在 AI4S 的语境下却有着天然的融合点。AI4S 的核心理念之一便是

[1] Cai W, Gao M. Beyond hallucination: generative AI as a catalyst for human creativity and cognitive evolution[J]. IECE Transactions on Emerging Topics in Artificial Intelligence, 2025, 2(1): 36-42.
[2] Winn Z. Accelerating scientific discovery with AI[EB/OL]. (2025-06-30)[2025-08-04]. https://news.mit.edu/2025/futurehouse-accelerates-scientific-discovery-with-ai-0630.
[3] Glasmachers T. Limits of End-to-End Learning[C]//Zhang M L, Noh Y K. Proceedings of the Ninth Asian Conference on Machine Learning: Vol. 77. Yonsei University, Seoul, Republic of Korea: PMLR, 2017: 17-32.
[4] Wu X, Xiao L, Sun Y, et al. A survey of human-in-the-loop for machine learning[J]. Future Generation Computer Systems, 2022, 135: 364-381.

利用 AI 技术对科学研究的全流程进行“端到端”的赋能[1]。然而，教育研究涉及复杂的情境理解和价值判断，难以完全交由 AI 系统自动处理，因此必须嵌入人的参与，这使得“人在回路”理念的引入显得尤为必要。

新一代 AI 技术的出现，为这种理念的融合提供了理想的技术载体。一方面，这类 AI 技术能够全面参与并优化研究过程的始端（问题凝练）、中端（数据分析）、末端（理论构建）以及各环节之间的系统串联。另一方面，这类技术强大的自然语言处理能力，为“人在回路”的实施提供了基础。研究者可以通过自然语言交互，随时介入研究流程，输入指令引导 AI 模型执行任务，并基于反馈动态调整研究路径。值得注意的是，“人在回路”理念在新一代 AI 模型的训练范式中有着特殊的体现，即通过人类反馈实现大语言模型行为与人类意图和价值观的“对齐”（Alignment），其中最具代表性的是基于人类反馈的强化学习（RLHF）[2]。

2.3.1 AI 角色演化

要理解研究过程的重构，首先需要清晰地认识到 AI 在其中的角色正在发生演变。近期的一系列综述研究为我们提供了一个极具洞察力的分类框架，将 AI 在科学发现中的参与度划分为三个递进的层次[3]，这个框架为“端到端”和“人在回路”的融合提供了更具体、更动态的演化路径。

表 2-1 人工智能在科学发现中参与度的三层次分类

层次（Level）	角色定义（Role Definition）	任务范围（Task Scope）	自主性与工作流（Autonomy & Workflow）
Level 1: AI 作	AI 作为在人类直	执行科学方法中单个	自主性有限，完全基于

[1] 李国杰. 智能化科研（AI4R）：第五科研范式[J]. 中国科学院院刊, 2024, 39(1): 1-9.

[2] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C]//Koyejo S, Mohamed S, Agarwal A, et al. Advances in Neural Information Processing Systems: Vol. 35. Curran Associates, Inc., 2022: 27730-27744.

[3] Zheng T, Deng Z, Tsang H T, et al. From automation to autonomy: a survey on large language models in scientific discovery[EB/OL]. (2025-05-19)[2025-07-30]. http://arxiv.org/abs/2505.13259.

层次（Level）	角色定义（Role Definition）	任务范围（Task Scope）	自主性与工作流（Autonomy & Workflow）
为工具（AI as Tool）	接监督下的专用工具，增强研究者的能力。	阶段的、明确定义的离散任务，如文献摘要、手稿初稿撰写、数据处理代码片段生成等。	人类的明确提示和指令操作。其输出通常需要人类验证，并由人类整合到更广泛的研究流程中。
Level 2: AI 作为分析师（AI as Analyst）	AI 作为被动智能体，具备更复杂的信息处理、数据建模和分析推理能力。	能够独立管理一系列任务序列，如分析实验数据集以识别趋势、解释复杂模拟的输出、对模型进行迭代优化等。	自主性增强，在人类设定的总体目标框架内运作，减少了对中间步骤的人工干预。人类定义分析目标、提供数据并评估最终洞见。
Level 3: AI 作为科学家（AI as Scientist）	AI 作为主动智能体，能够以相当高的独立性，策划和导航科学发现过程的多个阶段。	能够主动提出假说、规划和执行实验、分析结果数据、得出初步结论，甚至提出后续研究问题或探索途径。	自主性最高，能够以最少的人类干预驱动研究周期的主要部分，展现出战略性和迭代性的工作流程，迈向开放式的科学探索。

这个三层分类法清晰地揭示了 AI 在研究过程中的角色演进轨迹。它不再是一个静态的工具，而是一个动态的协作者，其自主性和责任边界正在不断拓展。在教育研究中，这意味着：

- **在 Level 1**，研究者使用 AI 工具来润色论文、总结文献或生成课堂活动方案。
- **在 Level 2**，研究者将一个学期的学生学习数据交给一个 AI 分析师，指令其“识别出与学生辍学风险最相关的行为模式”，AI 自主完成数据清洗、特征提取、模型训练和结果可视化等一系列步骤，并提交一份分析报告。
- **在 Level 3**，研究者设定一个宏观目标，如“探索项目式学习对学生批判性思维的影响”，一个 AI 科学家智能体则可能自主完成文献综述、提出具体研究假说、设计线上实验方案、招募（模

拟的）被试、执行实验、分析数据，并撰写出一份包含初步结论的研究报告初稿。

2.3.2 自主研究智能体的涌现

“AI 作为科学家”的图景并非遥远的科幻。在化学、生物学和材料科学等领域，能够自主执行复杂研究任务的智能体系统已经涌现，为教育研究的未来指明了方向。例如，ChemCrow 系统通过集成 18 个专家设计的化学工具，能够自主完成有机合成、药物发现和材料设计等复杂任务[1]。ProtAgents 则可以自主地生成、测试和优化蛋白质序列，以满足特定的生化属性要求[2]。这些系统展示了“端到端”研究过程的真正潜力：AI 不仅能处理流程中的某个环节，而是有能力将从问题定义到解决方案生成的多个环节串联起来，形成一个闭环。

在教育领域，虽然尚未出现如此成熟的自主研究智能体，但其雏形已经显现。最新研究表明，AI 模型在神经科学预测任务中已经超越了人类专家的表现[3]，这预示着 AI 在复杂认知任务上的巨大潜力。结合前述的自动化假说生成框架和生成式智能体模拟框架，可以预见未来的教育研究智能体，能够基于具体的研究场景首先通过分析文献和数据生成关于“最佳教学干预策略”的假说，然后在一个构建的虚拟课堂中设计并执行一项 A/B 测试来验证该假说，最后根据模拟实验的结果，自动撰写分析报告并提出对理论的修正建议。这种高度整合的自主研究流程，将极大地加速教育知识的生产和迭代速度。

2.3.3 生成式智能体模拟

基于智能体的建模与仿真（Agent-based Modeling and Simulation,

[1] M. Bran A, Cox S, Schilter O, et al. Augmenting large language models with chemistry tools[J]. Nature Machine Intelligence, 2024, 6(5): 525-535.

[2] Ghafarollahi A, Buehler M J. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning[J]. Digital Discovery, 2024, 3(7): 1389-1409.

[3] Luo X, Rechardt A, Sun G, et al. Large language models surpass human experts in predicting neuroscience results[J]. Nature Human Behaviour, 2024, 9(2): 305-315.

ABMS）是通过构建个体计算模型来研究其在环境中行为和交互，以及在复杂系统中涌现行为的仿真模拟技术，一直是理解复杂教育系统动态演化的重要工具[1]。然而，传统 ABMS 构建的教育智能体多为被动型，缺乏对环境的主动感知和复杂决策能力，难以准确模拟教与学过程中智能体的适应性行为，长期以来制约了其在教育研究中的应用深度。

（1）生成式智能体的技术基石

新一代人工智能的出现，特别是生成式智能体（Generative Agents）的崛起，正引发 ABMS 领域的范式革命。生成式智能体的核心突破在于，它们不再依赖于预先编写的、僵化的行为规则，而是利用 AI 模型强大的语言理解、推理和生成能力，在复杂的社会环境中进行动态决策[2]。近期研究系统性地勾勒出了一个先进的 AI 智能体的标准架构，它通常由画像、感知、记忆、规划和行动等核心模块构成，这些模块的协同工作赋予了智能体前所未有的“类人性”特征[3][4]。表 2-2 对这一通用架构进行了梳理：

表 2-2 基于 AI 的生成式智能体核心模块架构

模块（Module）	功能（Function）	在教育模拟中的应用价值
画像模块（Profile Module）	为智能体设定并动态调整其身份，包括静态的人口统计学特征（如年龄、性格、知识背景）和动态的社会属性（如目标、情绪、价值观、人际关系）。	模拟具有不同学习风格、认知水平、动机和文化背景的学生或教师，使其行为更具多样性和真实性。例如，可以构建一个“好奇心强但基础薄弱”的学生智能体，观察其在不同教学策略下的反应[5]。

[1] Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems[J]. Proceedings of the National Academy of Sciences, 2002, 99(suppl_3): 7280-7287.

[2] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.

[3] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[4] Gao C, Lan X, Li N, et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1259.

[5] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. Frontiers of

模块（Module）	功能（Function）	在教育模拟中的应用价值
感知模块（Perception Module）	使智能体能够观察和理解不断变化的社会环境，包括物理环境（如教室布局、学习资源）、其他智能体的状态和行为，以及潜在的社会规范。	智能体能够“感知”到课堂氛围的变化、教师的指令、同学的讨论，并据此调整自身行为。例如，一个学生智能体可以感知到小组讨论陷入僵局，并主动提出新的观点来推动讨论[1]。
记忆模块（Memory Module）	赋予智能体持续学习的能力，使其能够积累、总结和反思历史经验，以克服大语言模型固有的上下文窗口限制。这通常通过外部知识库和检索增强机制来实现[2]。	模拟学习的累积过程。学生智能体可以“记住”先前课程的内容，并在新学习中应用旧知识，这与教育中的知识追踪理念相契合。它还可以根据过去的成功或失败经验，反思并调整自己的学习策略。
规划模块（Planning Module）	使智能体能够根据长期目标和当前情境制定并分解行为策略，动态调整计划，并在多智能体环境中进行协作或竞争[3]。	模拟学生的学习规划与自我调节能力。智能体可以为自己制定一个学习计划（例如，先复习前置知识，再学习新概念，最后做练习），并在遇到困难时动态调整计划[4]。
行动模块（Action Module）	赋予智能体在环境中执行决策和采取行动的能力。行动可以是预定义的（从一个固定集合中选择），也可以是自由生成的，或是两者的混合。	将智能体的内部决策（如“我应该向老师提问”）转化为具体的、可观察的行为（如在聊天框中输入问题）。这使得研究者可以追踪和分析智能体的完整行为链条，进行精细的过程性分析[5]。

这个五模块架构清晰地表明，生成式智能体已远非简单的聊天机器人。它们是复杂的、具备一定自主性的计算实体，其设计初衷就是为了在模拟世界中再现人类复杂的社会行为和认知过程。

Computer Science, 2024, 18(6): 186345.
[1] Gao C, Lan X, Li N, et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1259.
[2] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.
[3] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.
[4] Gao C, Lan X, Li N, et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1259.
[5] Gao C, Lan X, Li N, et al. Large language models empowered agent-based modeling and simulation: a survey and perspectives[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1259.

（2）生成式智能体在教育研究中的具体应用

基于这一强大的技术底座，生成式智能体正在个体、群体和系统三个层面，为教育模拟研究提供坚实的实证支持[1]：

- **个体层面：模拟认知多样性。**传统模拟难以展现不同认知水平学生的真实行为，包括常见的错误和认知偏差。最新研究提出了基于“课堂模拟”的创新方法，通过构建上下文化的学生生成式智能体来解决这一挑战[2]。这项研究通过 6 周的教育工作坊收集了 60 名学生的细粒度数据，开发了转移性迭代反思（TIR）模块，显著提升了 AI 模型在学习行为模拟方面的准确性。通过定制的在线教育系统，研究者记录了学生随时间与讲座材料交互的学习行为，实现了对个体学习差异的高保真模拟。
- **群体层面：模拟课堂互动生态。**教育的本质是社会性的。例如，著名的“斯坦福小镇”利用大语言模型构建了一个交互式环境，其中的虚拟人物能够自主展现个体行为，也能彼此互动，形成复杂的社会行为[3]。在教育领域，SimClass 框架通过构建一个由多个 AI 智能体（扮演教师、助教以及具有不同性格的同学）组成的虚拟教室，生动地展示了模拟复杂课堂互动的巨大潜力[4]。该框架识别了代表性的课堂角色，并引入了新颖的课堂控制机制来实现自动化课堂教学。实验表明，由多智能体构成的丰富社交环境能够激发用户进行更深入的思考和更积极的互动，从而获得更好的学习效果。最新的 Agent4EDU 框架进一步推进了这一领域的发展，通过智能工作流将 AI 代理应用于教育，实现了更高效

[1] Mou X, Ding X, He Q, et al. From individual to society: a survey on social simulation driven by large language model-based agents[EB/OL]. (2024-12-04)[2025-07-30]. http://arxiv.org/abs/2412.03563.

[2] Xu S, Wen H N, Pan H, et al. Classroom simulacra: building contextual student generative agents in online education for learning behavioral simulation[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Yokohama Japan: ACM, 2025: 1-26.

[3] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.

[4] Zhang Z, Zhang-Li D, Yu J, et al. Simulating classroom education with LLM-empowered agents[EB/OL]. (2024-11-27)[2025-07-30]. http://arxiv.org/abs/2406.19226.

的教学和学习过程[1]。

- **系统层面：模拟教育政策影响。**生成式智能体的应用已超越微观课堂，开始涉足宏观的政策模拟领域。例如，VACSIM 框架利用 100 个基于人口普查数据初始化的生成式智能体，在一个模拟的社会网络中互动，以评估不同公共卫生干预措施（如疫苗推广策略）对群体态度和行为的影响 [2]。尽管该研究聚焦于健康领域，但其方法论对教育政策研究具有直接的借鉴意义。研究者可以构建一个由学生、家长、教师、校长等利益相关者组成的智能体社会，模拟他们对某项教育政策（如招生制度改革、课程标准变化）的反应和博弈过程，从而为基于仿真论证的政策评估和优化机制提供了强大的技术工具[3]。

2.3.4 超越“人在回路”

然而，将研究过程的重构仅仅理解为“更强大的 AI”和“更聪明的回路”是片面的。最具革命性的变化在于，这些自主的智能体和人类研究者正在被连接成一个前所未有的、分布式的协作研究生态系统。

例如 AgentRxiv 框架的提出，这是一个专为自主研究智能体设计的、模仿 arXiv 的开放式预印本服务器[4]。在这个平台上，任何一个“智能体实验室”完成的研究（例如，一篇由 AI 自主生成的论文），都可以被上传和归档。随后，其他正在进行研究的智能体实验室可以异步地检索和引用这些 AI 生成的文献，将其作为自己新研究的起点。该平台的实证研究显示，拥有协作研究访问权限的智能体在 MATH-500 基准测试中实现了 13.7%的相对性能提升。

[1] Dai L, Jiang Y H, Chen Y, et al. Agent4EDU: advancing AI for education with agentic workflows[C]//Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education. Xiamen China: ACM, 2024: 180-185.

[2] Hou A B, Du H, Wang Y, et al. Can a society of generative agents simulate human behavior and inform public health policy? A case study on vaccine hesitancy[EB/OL]. (2025-07-13)[2025-07-30]. http://arxiv.org/abs/2503.09639.

[3] Large Language Model-Empowered Agents for Simulating Macroeconomic Activities[EB/OL]. (2023-10-16)[2024-02-05]. http://arxiv.org/abs/2310.10436.

[4] Schmidgall S, Moor M. AgentRxiv: towards collaborative autonomous research[EB/OL]. (2025-03-23)[2025-07-30]. http://arxiv.org/abs/2503.18102.

AgentRxiv 将研究过程从传统的、孤立的、线性的模式，转变为一种网络化、累积性和并行化的全新模式。这种设想与更广泛的“开放科学”（Open Science）和“科学自动化”（Automation of Science）的理念不谋而合[1]。它所代表的不仅仅是一个技术平台，更是一种全新的科研组织范式，其核心特征包括：

- **网络化（Networked）**：研究不再局限于单个研究者或团队的内部循环，而是变成了一个由众多人类和 AI 节点组成的、相互连接的知识网络。
- **累积性（Cumulative）**：知识的增长是累积性的。一个智能体的发现可以被另一个智能体继承和发展，实现了跨越时空的知识传递和叠加，这正是科学进步的本质。
- **并行化（Parallelized）**：多个智能体实验室可以同时对同一个或不同的研究问题展开探索，极大地提升了研究的广度和效率，实现了真正意义上的大规模并行科研。

这个生态系统彻底超越了简单的“人在回路”模型。在“人在回路”中，人是操作者和监督者，与一个 AI 进行点对点的互动。而在 AgentRxiv 所代表的生态系统中，人是整个生态系统的架构师、管理者和价值引领者。

2.3.5 人类研究者的转型

在这个新兴的研究生态系统中，人类研究者的价值非但没有被削弱，其核心地位反而得到了重塑与凸显。繁琐的、劳动密集型的研究任务被大规模地自动化，使得研究者能够将精力聚焦于那些机器无法替代的核心智力活动上。最新的教育研究表明，人机协同的混合智能学习环境正在重新定义教师和学生的角色[2]，这种协同模式在研究领域同样适用。

[1] Nielsen M A. Reinventing discovery: the new era of networked science[M]. First paperback printing. Princeton: Princeton University Press, 2014.

[2] Nguyen A, Hong Y, Dang B, et al. Human-AI collaboration patterns in AI-assisted academic writing[J]. Studies in Higher Education, 2024, 49(5): 847-864.

学术洞见与创新能力根植于研究者深厚的理论根基和敏锐的问题意识，人工智能技术则作为一种认知增强工具，用以放大和拓展这种智慧的边界。在此人机交互过程中，研究者必须调用其学科专业知识与实践智慧，对 AI 生成的多元可能性进行批判性甄别与审思，并做出最终的研究抉择。具体而言，研究者的主体性和专业判断力将更加聚焦于以下几个方面：

- **战略方向设定**：确定研究的宏大愿景和关键问题，为整个 AI 研究生态系统设定探索的方向。
- **批判性验证与反馈**：对 AI 生成的假说、实验设计、数据分析和结论进行严格的批判性审查，提供高质量的反馈，引导 AI 系统的迭代和优化。这正是早期研究中强调的“人在回路”的精髓所在，即通过人类反馈实现模型行为与人类意图和价值观的对齐。
- **跨领域知识整合**：利用人类独特的创造性联想能力，将不同 AI 实验室在不同领域取得的突破进行整合，形成更宏大、更具原创性的理论框架。
- **伦理与价值守护**：确保整个研究生态系统的运行符合学术伦理和人类社会的共同价值观，防止 AI 产生和放大偏见，并对其社会影响负责。
- **人文关怀与情境理解**：在教育研究这一具有深厚人文内涵的领域，人类研究者的情感智能、伦理判断和价值关怀仍然是不可替代的[1]。

总而言之，研究过程的重构，其终极图景并非一个由 AI 主导的全自动“知识工厂”，而是一个充满活力的、人机共生的社会-技术基础设施。在这个基础设施中，AI 的计算能力和人的批判性智慧深度融合，研究过程本身从一系列孤立的步骤，转变为一个持续演化、自我完善的生态系统。这或许是 AI4S 为教育研究带来的最深刻、最

[1] Atchley P, Pannell H, Wofford K, et al. Human and AI collaboration in the higher education environment: opportunities and concerns[J]. Cognitive Research: Principles and Implications, 2024, 9(1): 20.

激动人心的变革。

第 3 章 AI 辅助的教育质性研究

质性研究在揭示教育实践的复杂机制、理解学习者经验与教育情境中的意义建构方面具有不可替代的作用，然而，传统教育质性研究长期面临着效率低下、主观性强、一致性难以保证等挑战[1]，特别是在面对日益增长的教育大数据时，传统的手动编码和解释方法显得力不从心[2]。

长期以来，受限于早期 AI 技术在语言理解、情境感知和意义阐释等方面的能力不足，人工智能始终难以真正介入这一高度依赖人类认知的研究领域。然而，以 GPT、Claude、Gemini 等为代表的大语言模型在语言理解和生成能力上实现了质的飞跃[3][4]，使其具备了深度参与质性研究的技术条件[5]。大语言模型在自动编码、主题识别、多模态信息整合等方面的卓越表现，为提升研究效率、增强分析一致性、减少主观偏差提供了新的技术路径[6]。

本章旨在深入剖析大语言模型的技术特性及其在质性研究各流程中的适用性，结合课堂互动分析、学习体验评估、教育政策研究等典型应用场景的具体案例，力图为教育研究者提供关于大语言模型辅助质性研究的系统性理论框架和实践指导。

3.1 基于 AI 的教育质性研究流程

教育质性研究正面临着数据规模增长与分析复杂度提升的挑战。随着在线教育平台、数字化学习环境的普及，研究者需要处理的文本数据呈指数级增长，传统的人工编码分析方法已难以应对这一现

[1] Anuradha I, Mitkov R, Nahar V. Evaluating of Large Language Models in Relationship Extraction from Unstructured Data: Empirical Study from Holocaust Testimonies[C]//Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. 2023: 117-123.

[2] 陈鹏,张靖沅,陈向东.大模型如何融入教育的质性研究:理论潜力与案例实践[J].现代教育技术,2024,34(10):32-41.

[3] Thirunavukarasu A J, Ting D S J, Elangovan K, et al. Large language models in medicine[J]. Nature medicine, 2023, 29(8): 1930-1940.

[4] Liu Y, Han T, Ma S, et al. Summary of chatgpt-related research and perspective towards the future of large language models[J]. Meta-Radiology, 2023,(2):100017.

[5] Roberts J, Baker M, Andrew J. Artificial intelligence and qualitative research: The promise and perils of large language model (LLM)‘assistance’[J]. Critical Perspectives on Accounting, 2024, 99:102722 .

[6] Parker J L, Richard V, Becker K. Flexibility & iteration: Exploring the potential of large language models in developing and refining interview protocols[J]. The qualitative report, 2023, 28(9): 2772-2790.

实。与此同时，教育场景中的多元化表达、隐性知识传递、情感互动等现象要求研究工具具备更强的语义理解和推理能力。

大语言模型的出现为解决这些挑战提供了新的可能性。相关研究表明，大语言模型在主题识别、情感分析、意图理解等方面展现出显著优势[1][2][3]，特别是在处理教育语境中的复杂表达和多层次意义方面表现突出。例如，GPT-4 在分析日本医疗领域的半结构化访谈数据时，能够识别出与人类研究者相似的主题，并处理具有文化敏感性的概念[4]。这表明：研究者理论上可以应用模型的这种能力处理质性材料，在计算能力的加持下，快速、准确地分析材料中反映的价值取向、情感态度与意义建构倾向，从而减少研究者很多重复劳动的麻烦。

然而，大语言模型的技术特性与质性研究的认识论基础之间的契合程度仍需深入考察。正如 Gillen[5]与 Christou[6]所指出的，AI 在质性研究中的应用应主要限定于前期的数据清洗、辅助编码和文献管理等技术性环节，其在理论阐释、意义建构以及伦理敏感型研究中的能力边界尚不明确。基本共识在于，质性研究强调的主观性、情境性和意义建构等核心理念，与 AI 模型基于概率计算的运作机制之间存在着本质差异。因此，尽管大语言模型为质性研究带来了方法论创新的机遇，也出现了关于研究有效性、可信度以及伦理规范的新争议。

基于上述考虑，本节将深入探讨大语言模型技术特性与教育质

[1] De Paoli S. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach[J]. Social Science Computer Review, 2024, 42(4): 997-1019.

[2] Koch M A. Turning chaos into meaning: A chat GPT-Assisted exploration of COVID-19 narratives[D]. Enschede: University of Twente, 2023.

[3] Törnberg P. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages with zero-shot learning[EB/OL]. (2023-04-12)[2025-10-21]. https://arxiv.org/abs/2304.06588

[4] Sakaguchi K, Sakama R, Watari T. Evaluating ChatGPT in Qualitative Thematic Analysis With Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study[J]. Journal of Medical Internet Research, 2025, 27: e71521.

[5] Gillen A L. Can we trust AI in qualitative research[EB/OL]. (2024-05-16)[2025-10-21]. https://www.insidehighered.com/opinion/views/2024/10/09/can-we-trust-ai-qualitative-research-opinion

[6] Christou P A. Reliability and validity in qualitative research revisited and the role of AI[J]. The Qualitative Report, 2025, 30(3): 3306-3314.

性研究的理论兼容性，并系统阐述其在研究实践各环节中的应用机制，为构建科学合理的 AI 辅助质性研究方法论提供理论支撑。

3.1.1 大语言模型对于教育质性研究的适用性

当前关于大语言模型在质性研究中应用的相关研究[1][2]，主要体现在利用世界知识理解复杂语境、借助心理理论深化材料理解、通过非结构化数据分析提供深层洞见、多模态数据融合分析、数据增强丰富研究视角等五个核心维度，如图 3-1 所示。

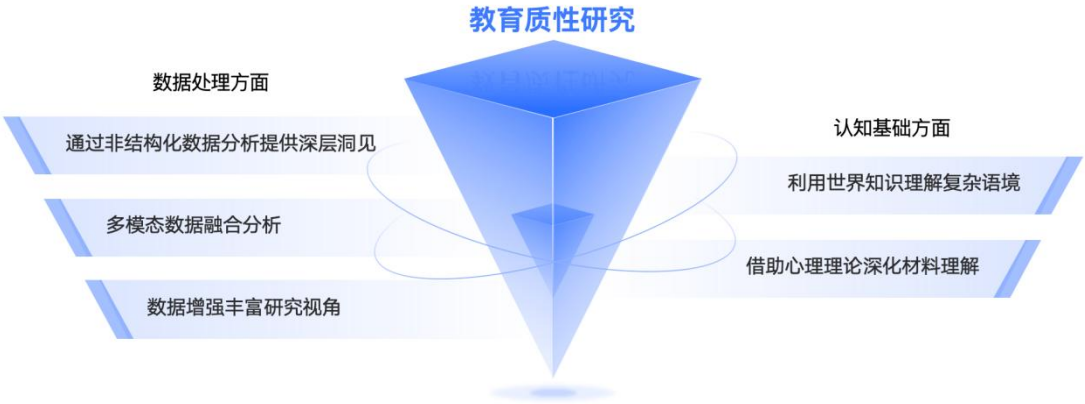

图 3-1 大语言模型对于教育质性研究的适用性

（1）利用世界知识理解复杂语境

质性研究的核心在于深入理解参与者的主观经验和行为意义，而这种理解的有效性很大程度上建立在研究者对其所处社会文化语境的准确把握之上。然而在实践中，研究者主要依靠长期的田野工作来浸入、观察并逐步理解与推断这些复杂的语境信息，但这种传统方式往往面临困境：一方面，深度的田野调查耗时较长，数据积累效率有限；另一方面，语境的识别和诠释容易受到研究者个人知识结构、学科背景和经验视野的制约，难以实现对多元语境要素的全面捕捉。

大语言模型通过大规模预训练获得了丰富的世界知识（World Knowledge），可以将其作为辅助性的语境理解工具，提供了新的解

[1] 陈鹏,张靖沅,陈向东.大模型如何融入教育的质性研究:理论潜力与案例实践[J].现代教育技术,2024,34(10):32-41.

[2] 马颂歌,左靖文.生成式人工智能何以融入质性研究——以“生成田野”为起点的功能评价与伦理审思[J].远程教育杂志,2024,42(06):83-91.

决路径。世界知识，是指关于客观世界中实体、事件、关系及其运作规律的结构化认知，涵盖社会常识、文化规范、专业领域知识、历史背景等多个维度。从理论上看，世界知识能够为语境理解提供多层次的参照框架：一方面，它包含了不同文化群体的符号系统和意义体系，有助于研究者识别特定话语背后的文化逻辑。质性研究强调文化敏感性和伦理考量，要求研究者充分尊重参与者的文化背景和价值观念[1]。当研究者分析访谈文本或观察记录时，模型可以识别其中的文化符号、专业术语和隐含概念，帮助研究者快速理解参与者话语的深层含义[2]。例如，在分析东亚文化背景下的课堂沉默现象时，模型能够提供关于集体主义文化、师生等级关系等相关知识，帮助研究者从文化内部视角理解这一行为的真实意义。

另一方面，它整合了跨学科的概念网络，能够弥补单一研究者知识结构的盲点。正如达令·哈蒙德（Darling-Hammond）等人[3]所指出，随着关于人类发展和学习方式的知识迅速增长，教育研究需要利用涵盖神经科学、心理学、社会学、学习科学等多领域的知识引入教育现象的分析中。大语言模型这种理论整合能力有助于研究者构建更丰富的分析框架，深化对复杂教育现象的理解，符合质性研究追求整体性和深度性的方法论要求[4]。

综上所述，大语言模型凭借其丰富的世界知识，帮助研究者突破个人知识与文化视角的局限，实现对教育现象更全面、深入的诠释。

（2）借助心理理论深化材料理解

质性研究的本质在于探索个体的主观体验、感受和意义建构过程，这要求研究者能够深入参与者的内心世界，理解其行为背后的

[1] 陈向东,卢淑怡,易乐湘.文化冲突:大语言模型教育应用中的张力与调适[J].远程教育杂志,2025,43(03):3-15+43.

[2] Kalla D, Smith N, Kuraku S, et al. Study and analysis of Chat GPT and its impact on different fields of study[J]. International Journal of Innovative Science and Research Technology, 2023,(3):827-833.

[3] Darling-Hammond L, Flook L, Cook-Harvey C, et al. Implications for educational practice of the science of learning and development[J]. Applied developmental science, 2020, 24(2): 97-140.

[4] Rossman G B, Rallis S F. Learning in the field: An introduction to qualitative research(third edition)[M]. Thousand Oaks, CA: SAGE, 2012.

信念、动机和情感状态。心理理论（Theory of Mind，ToM）作为认知科学的重要概念，为研究者提供了理解他人心理状态的理论框架。在教育质性研究中，心理理论的应用帮助研究者超越表面行为，深入探究学习者、教师以及其他教育参与者的内在认知过程和情感体验。

大语言模型在心理理论测试中展现出的卓越表现为质性分析提供了技术支撑。例如，GPT-4 在 93%的心理理论任务中表现出色，能够有效推理他人在特定场景下的心理状态，包括错误信念推理和意图识别等复杂认知任务[1]。基于这种能力，大语言模型在处理教育质性材料时展现出独特的优势：

首先，推理话语中的隐含心理状态。例如，当一位教师说“我觉得学生们应该能理解这个概念”时，大语言模型不仅能识别出教师对学生能力的信念，还能推断其中可能包含的教学期待、对课程难度的评估，以及潜在的焦虑或自信情绪。已有研究借鉴这方面的技术优势，设计了群体感知工具，为支持协作学习（CSCL）环境下的情感识别与调节[2]。该类工具借助大语言模型分析学习者互动中的情绪变化，生成情感报告，以支持师生及时感知并调节群体情绪与协作状态。这种多层次的心理推理为研究者提供了更深入的理解视角，帮助其把握参与者真实的认知框架和情感状态。

其次，解析象征性语言的深层意义。在教育实践情境中，教师与学生在表达复杂的学习体验和教学感受时，常常借助隐喻、类比等象征性语言。这些修辞手段反映了不同的主体对教育现象的认知结构和意义赋予方式。Prystawski 等人的研究证实了 GPT-3 在隐喻理解方面的有效性，这意味着模型能够识别隐喻背后的概念映射关

[1] Kosinski M. Theory of Mind May Have Spontaneously Emerged in Large Language Models[OL].<https://www.pasqualeborriello.com/wp-content/uploads/2023/02/Theory-of-Mind-2302.02083.pdf>

[2] 陈佳雯,褚乐阳,潘香霖,等.共享调节中的群体情感感知工具开发与应用——基于大语言模型技术框架[J].远程教育杂志,2024,42(03):79-92.

系，解析其承载的情感色彩和价值判断[1]。当学生将学习比作“攀登高峰”“航海探险”时，模型能够理解这些隐喻所暗示的挑战感、成就动机或不确定性体验。这种解析能力帮助研究者捕捉到参与者难以用直白语言表达的微妙心理状态，丰富了对教育体验的理解维度。

第三，揭示互动关系中的心理动力机制。教育现象往往涉及多个参与者之间的复杂互动，每个人都基于自己对他人心理状态的理解来调整行为策略，形成动态的心理博弈过程。大语言模型能够模拟这种多层次的心理推理——不仅理解“A 如何想”，还能分析“A 认为 B 如何想”以及“A 如何基于对 B 的理解来行动”。在实践层面，有研究者利用大语言模型模拟校长、家长、学生等不同角色，深入探讨各方对教育议题的不同观点和利益诉求[2]，这种模拟分析揭示了不同角色之间的权力关系、合作模式和心理互动机制。

实际上，质性研究中的意义建构过程本身就体现了心理理论的深度应用。参与者在叙述自己的经历时，不仅在描述事实，更在解释自己和他人的行为动机，表达对情境的理解和评价。这些叙述实质上是参与者运用心理理论对经验进行的意义编码。大语言模型能够识别这些叙述中的因果推理链条、归因模式以及隐含的价值判断，帮助研究者理解参与者如何构建自己对教育经历的意义理解。这种深度的心理分析为质性研究提供了更丰富的理论建构基础，使研究者能够更全面地把握教育现象的复杂性和深层机制。

（3）通过非结构化数据分析提供深层洞见

教育领域产生的大量文本数据——访谈记录、学习日志、开放式问卷回答、课堂讨论记录等构成了丰富的非结构化质性材料。这些自由表达的数据承载着参与者的真实语言和观点，蕴含着复杂的认

[1] Prystawski B, Thibodeau P, Potts C, et al. Psychologically informed chain-of-thought prompts for metaphor understanding in large language models[EB/OL]. (2022-09-16)[2025-10-21]. https://arxiv.org/abs/2209.08141

[2] 陈向东,靳旭莹.大模型教育应用展望：基于技术预见的方法[J].苏州大学学报(教育科学版),2025,13(01):13-24.

知、情感和社会互动模式。传统的人工细读和编码方式虽然能够深入挖掘文本内涵，但面对海量数据时往往效率低下且难以维持编码一致性。

大语言模型为非结构化数据分析开辟了新路径。BERT、GPT 等模型具备自动学习文本高维语义特征的能力，能够捕捉词语顺序关系和上下文语义，这种深度语义理解为识别歧义编码、进行主题挖掘和概念提取等分析任务提供了技术基础[1][2]。具体而言，大语言模型在处理教育质性数据时展现出三个层面的分析优势：

第一，高效处理海量文本并保持分析一致性。质性研究项目常常积累数十乃至上百份访谈记录、数千条学习日志，人工逐一细读和编码不仅耗时巨大，还面临“编码漂移”问题，即研究者在不同时间对相同内容的理解可能发生微妙变化，导致编码标准前后不一致。大语言模型能够同时处理大量文本，在统一的语义框架下进行分析，有效避免了人工编码中的时间效应和疲劳效应。更重要的是，模型的分析逻辑是可追溯和可复现的。研究者可以通过相同的提示词和参数设置对数据进行反复分析，确保分析标准的稳定性。例如，思维链提示能引导模型“逐步说明分析逻辑，先识别关键词，再生成编码理由，最后给出标签”这一结构，使分析过程透明且复现。与此同时，模型还能够发现人工编码可能遗漏的潜在模式——例如，某些词汇共现关系、跨文本的主题呼应、或是微妙的语气变化趋势。这种“弱信号分析”能力为研究者提供了全局性的数据图景，帮助其在深入细读前建立整体认知框架。

第二，消解语义模糊，准确把握复杂表达意图。教育情境中的参与者往往使用非正式语言、方言表达、口语化句式或隐喻性描述来表达复杂的学习体验，这些表达方式增加了理解难度。例如，一

[1] Ma Z, Wu W, Zheng Z, et al. Leveraging speech PTM, text LLM, and emotional TTS for speech emotion recognition[A]. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)[C]. Seoul, Korea: IEEE, 2024:11146-11150.

[2] Loureiro D, Rezaee K, Pilehvar M T, et al. Analysis and evaluation of language models for word sense disambiguation[J]. Computational Linguistics, 2021,47(2):387-443.

位学生可能说“那个知识点就是不进脑子”，这种口语化表述包含了认知困难、挫败感和对学习过程的隐喻性理解。大语言模型能够基于广泛的语言知识和上下文信息，识别出这类表述背后的多重含义[1]。因此，大语言模型尤为擅长处理表述不完整、逻辑跳跃或前后矛盾的文本片段。在口语化的访谈记录中，参与者常常出现话语中断、重新组织语言或补充说明的情况，这些不流畅的表达蕴含着丰富的心理活动信息。模型通过上下文推理能够重构完整的语义，减少因表述欠清晰而产生的理解偏差，帮助研究者更准确地把握参与者的真实意图和感受，避免因字面理解造成的意义遗失。

第三，挖掘深层模式，揭示隐含意义结构。超越表层语义的分析是质性研究达到理论深度的关键。大语言模型在识别文本中隐含模式和深层主题方面的能力为教育研究提供了新的分析维度。以文学分析领域为例，RoBERTa 模型能够分析诗歌中的象征性手法，如隐喻和暗喻，其分析结果与专业文学评论家相近[2]。这种能力在教育文本分析中同样适用：模型能够识别学习者表达中的隐喻（如“学习是一场战斗”）、象征（如“黑暗隧道”代表困难时期）和情感色彩（挫败、兴奋、焦虑等），揭示其深层的学习体验和认知状态。

通过将模型的计算效率与研究者的情境判断相结合，非结构化数据的分析能够在保持传统质性研究深度解释的优势之上，获得了应对大规模非结构化材料的处理能力。随着大语言模型在对话交互和语用知识学习方面的持续改进，其理解复杂文本语义的能力将不断提升，为深入探索教育现象的多层次含义提供更强大的技术支撑。

（4）多模态数据融合分析

传统的质性研究方法对多模态的质性材料通常采用分离处理的

[1] Jansen B J, Jung S, Salminen J. Employing large language models in survey research[J]. Natural Language Processing Journal, 2023,4:100020.

[2] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2025-10-21]. https://arxiv.org/abs/1907.11692

思路：视频记录交给行为分析、话语转录交给话语分析、教学材料交给内容分析，这些工作在早期的分析阶段通常是独立进行的。这种分析方式相对割裂，难以整体上捕捉不同模态之间的动态关联和相互影响，可能遗失掉发生在模态交汇处的关键意义。

多模态大模型的跨模态理解能力为解决这一挑战提供了技术基础[1]。模型能够同时处理文本、图像、音频等不同类型的数据，并建立它们之间的语义关联。因此，在教育质性研究中，多模态大模型的分析优势并非局限于信息处理效率，而体现在语义、结构与模式三个层面的融合性理解上。

在语义层面，模型能够建立跨模态的语义关联，揭示语言、手势、表情等多种符号系统的协同机制。教育情境中的意义传达往往不是单一模态的产物，而是多种模态协同作用的结果。当教师在课堂上使用手势配合语言说明时，手势不仅是语言的“装饰”，而是承载着独立或互补的语义信息[2]。例如，教师在讲解“函数”时，可能用上扬的手势配合“上升”这一词语，也可能在说“保持不变”时用水平手势强化语义。多模态大模型能够分析手势动作与语言内容之间的语义一致性、互补性或矛盾性，理解这种多模态表达如何共同构建教学意义。除此之外，模型能够识别跨模态的语义错配现象。当教师的语言表达与面部表情、语调或肢体语言出现不一致时（如口头鼓励但面露失望），这种错配往往传递着复杂的社会性信息或情感状态。模型对这类跨模态关系的敏感性帮助研究者捕捉到单一模态分析中容易忽视的微妙互动，为理解师生交往的真实性和复杂性提供了更立体的分析视角。

在结构层面，模型能整合分散的信息片段，构建反映参与者认知与情感投入的整体图景。基于跨模态关联能力，模型进一步展现

[1] Abdüsselam M S. Qualitative data analysis in the age of artificial general intelligence[J]. International Journal of Advanced Natural Sciences and Engineering Researches, 2023,(4):1-5.

[2] Deng X, Wu X. Exploring the role of multimodal metaphor through gestures in middle school English education[J]. Journal of Contemporary Language Research, 2024, 3(1): 1-9.

出数据整合处理的优势。原本分散的信息片段能够被融合为统一的语义表示，使研究者能够从整体角度理解参与者的表达和行为。以小组协作学习场景为例，学生在讨论中的言语表达、肢体动作、面部表情和使用的学习材料不再是孤立的数据点，而是形成了反映其完整参与模式和认知投入程度的有机整体。这种整合能力使研究者能够重构参与者的学习过程。例如，当学生用手比划一个圆形的同时说“这个概念很完整”，模型能够识别出手势与语言共同指向“完整性”这一抽象概念的具体表达。通过追踪这些多模态表达在时间序列上的演变，可为理解学习的发生机制提供过程性证据。

在模式层面，模型可识别跨模态表达的深层规律，揭示学习与互动的隐含机制。在语义理解的深层层面，模型能够把握视觉元素、听觉信息和文本内容在特定语境中的综合含义，识别它们之间的隐含关系和深层结构。这种理解能够深入到意义的建构过程中，帮助研究者理解参与者如何通过多种渠道同时传达复杂的想法和感受。例如，Whitehead 等人[1]利用多模态大语言模型对课堂视频进行姿态与行为分析，发现模型能够在统一语义框架下识别教师与学习者的非言语即时性模式，从而揭示协作学习过程中的潜在互动规律。总体而言，这些跨模态模式的识别可以为教育理论的构建提供实证基础，也为教学实践提供可观察的行为指标。

通过这些技术能力的综合运用，研究者不再需要将复杂的教育现象人为分解为单一模态的数据片段，而是可以保持其整体性和动态性。这种转变使质性研究能够更真实地反映教育现象的复杂性，呼应了具身认知理论和情境学习理论对学习整体性的强调。

（5）数据增强丰富研究视角

教育质性研究经常面临数据获取的现实限制。某些敏感话题（如学业压力、师生冲突等）的研究受到严格的伦理审查限制，研

[1] Whitehead R, Nguyen A, Järvelä S. Utilizing multimodal large language models for video analysis of posture in studying collaborative learning: A case study[J]. Journal of Learning Analytics, 2025, 12(1): 186-200.

究者难以获得充足的第一手资料。这些限制不仅影响研究的代表性，也可能导致理论建构过程中出现视角限制，某些边缘化群体的声音、非主流的教育体验、或隐蔽的教育机制可能因数据缺失而无法进入分析视野。目前，一些质性研究通常通过多样化现有数据来扩展研究视野，其在质性研究中展现出三个层面的增强优势：

首先，语义保持下的表达多样化，发现隐含概念结构。在文本数据处理中，大语言模型能够通过同义词替换、句子重构、语言风格调整等方式，生成语义相近但表达形式不同的文本变体。这种技术手段的价值不仅在于创造出更多数据，更在于帮助研究者发现隐藏在不同表达方式背后的共同主题和概念，减少因表达习惯差异而产生的理解偏差。具体而言，研究者可以使用少量高质量的访谈记录或观察笔记作为种子数据，引导模型学习其中的表达模式和语义特征，进而生成语义一致但形式略有差异的新文本[1]。例如，当一位学生说“老师讲得太快，我跟不上”，模型可能生成变体如“授课节奏超出了我的理解速度”“感觉课堂进度和我的接受能力不匹配”等。这些变体虽然表述不同，但都指向同一个核心概念——学习节奏失配。通过分析这些变体，研究者能够识别出概念的核心语义结构（教学节奏与学习能力的关系），而不被特定的表达方式所局限。更重要的是，这种生成过程不是简单的复制或模仿，而是在保持原始语义内核的基础上，探索不同的表达可能性。当研究者的种子数据主要来自某一特定群体（如城市学生）时，模型生成的变体可能呈现出不同社会阶层、年龄段或表达习惯的话语特征，提醒研究者注意到原始数据中可能存在的表达单一性，从而在后续数据收集中有意识地补充多元视角。

第二，跨语言跨文化视角扩展，揭示意义的文化境遇性。基于DeBERTa等模型的增强方法进一步展现了跨语言和跨文化的应用潜

[1] Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: A survey[J]. ACM Computing Surveys, 2023,56(2):1-40.

力。通过回译技术（将文本翻译成另一种语言再翻译回来）和重述增强，模型能够生成多语言版本的研究材料，帮助研究者理解同一教育现象在不同文化背景下的表达差异[1]。这种跨语言的数据增强为比较教育研究和多元文化教育研究提供了新的分析工具。研究者可以将中国学生关于“学习压力”的访谈文本进行多语言增强，对比其在英语、日语、芬兰语等文化语境中的表达变化，从而识别出哪些压力来源是跨文化普遍的（如考试焦虑），哪些是文化特定的（如对家庭期待的回应）。这种分析使研究者能够超越单一语言和文化的局限性，构建更具跨文化适用性的理论框架。

第三，基于多任务学习的质量深化，提升理论建构能力。多任务学习框架的引入使数据增强具备了更强的理论建构能力。通过将相关的分析任务整合到同一个模型中进行联合训练——例如同时进行情感分析、主题提取、因果关系识别和修辞策略识别——模型能够从不同任务中学习更广泛和深入的特征表示[2]。这种学习方式的优势在于，增强后的数据不仅在数量上得到扩充，在质量和理论深度上也得到提升。传统的单任务增强可能只是生成更多“相似”的文本，而多任务学习框架下的增强能够生成多维度丰富的文本。更深层次地，这种多任务增强帮助研究者发现变量之间的潜在关联。当模型在生成过程中自然地将“高学业压力”与“低自我效能感”“回避型应对策略”关联在一起时，这种关联模式本身就构成了值得深入探究的理论假设。研究者可以基于这些生成文本中的关联模式，返回原始数据进行验证性编码，或在后续研究中有针对性地收集数据，形成“增强—发现—验证”的迭代研究循环。这为质性研究的理论发现提供了更坚实的数据基础和更清晰的概念网络。

[1] d’Sa A G, Illina I, Fohr D, et al. Exploring conditional language model based data augmentation approaches for hate speech classification[A]. Text, Speech, and Dialogue[C]. Cham: Springer International Publishing, 2021:135-146.

[2] Modzelewski A, Sosnowski W, Wilczynska M, et al. Dshacker at semeval-2023 task 3: Genres and persuasion techniques detection with multilingual data augmentation through machine translation and text generation[A]. Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)[C]. 2023:1582-1591.

数据增强技术的应用改变了质性研究中数据稀缺问题的传统解决思路。研究者不再完全依赖于扩大样本规模或延长研究周期，而是可以通过技术手段深度挖掘现有数据的潜在价值。这种方法论转变使质性研究能够更灵活地应对各种实际约束，同时保持研究的深度和理论价值，为处理复杂教育现象提供了更强的分析能力。

综合以上分析，大语言模型对教育质性研究的适用性主要体现在以上五个维度。这些维度从理解、分析到建构，为质性研究提供了系统性的技术支撑，既保持了质性研究对语境敏感性、意义建构和深度理解的传统优势，又为应对当代教育现象的复杂性和研究方法的创新需求提供了新的可能性，推动教育质性研究朝着更加精细化和理论化的方向发展。

3.1.2 大语言模型在教育质性研究中的应用

教育质性研究强调情境理解、意义建构和理论生成，其过程具有迭代性、涌现性和非线性特征。通过梳理扎根理论、现象学研究、叙事研究等主流质性研究方法论，可将大语言模型辅助的教育质性研究过程概括为四个相互关联、循环递进的核心环节：研究设计与理论准备、数据收集与参与者交互、深度数据分析与模式识别，以及研究成果的展示、验证与交互。已有研究表明，大语言模型凭借其强大的语言理解、知识整合和模式识别能力，能够在各环节提供智能化支持，在保持质性研究开放性和灵活性的前提下，实现研究效率与深度的双重提升，并可能催生新的方法论创新（见图 3-2）[1]。

[1] 陈鹏,张靖沅,陈向东.大模型如何融入教育的质性研究:理论潜力与案例实践[J].现代教育技术,2024,34(10):32-41.

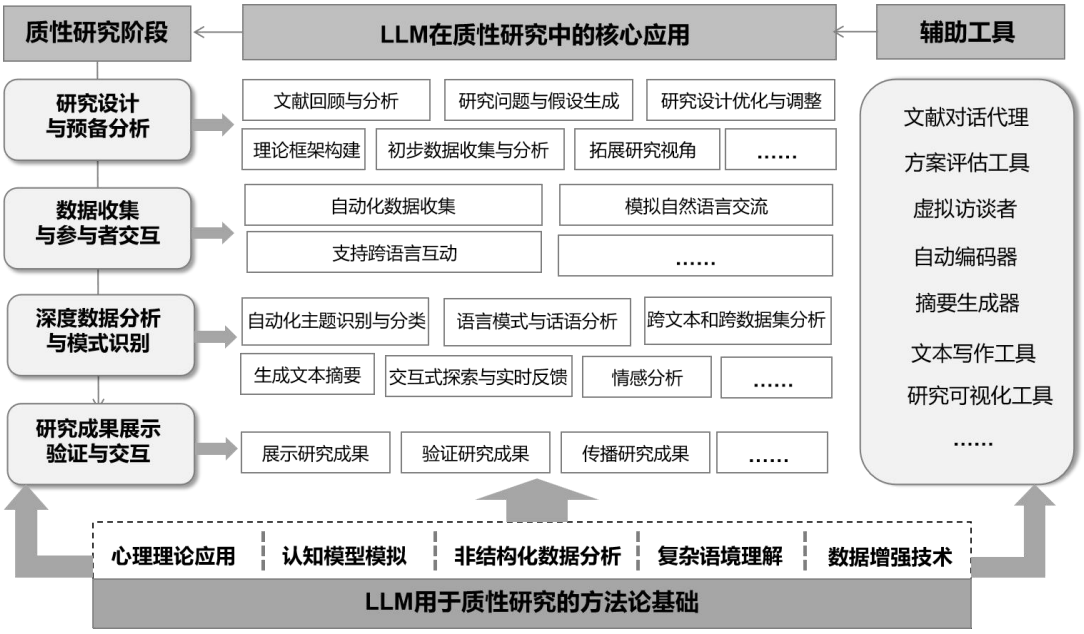

图 3-2 大语言模型在质性研究不同阶段的应用

同时，大语言模型在教育研究中可承担多重角色：既可作为研究辅助工具协助文献分析与问题生成，也可作为研究对象或数据来源参与理论建构过程。这种多元化的角色定位为教育研究提供了灵活的技术选择与应用策略。以下将详细阐述大语言模型在这四个关键环节中的作用机制与应用策略。

（1）研究设计与理论准备

研究设计与理论准备构成质性教育研究的前置阶段，其质量直接影响后续数据收集、分析与阐释的科学性与有效性。在传统质性研究中，研究者需要通过大量文献阅读来构建理论框架、识别研究空白、明确研究问题，这一过程不仅耗时费力，而且容易受到研究者知识结构与认知偏见的局限。特别是在教育领域，跨学科文献的复杂性与研究主题的多元化，使得全面、客观的文献综述与精准的问题定位成为重大挑战。大语言模型在文本挖掘、语义分析与知识整合方面的技术优势，为质性研究的前期准备提供了有力支撑。通过智能化的文献检索与主题归纳，大语言模型能够协助研究者快速识别相关理论脉络、梳理研究进展，并在此基础上生成具有针对性的研究问题与理论假设。这种技术辅助不仅提升了理论准备的系统

性与全面性，也为研究者提供了多元化的理论视角与创新性的问题思路，进一步增强了对教育现象的理论敏感性与概念抽象能力。以下将从文献回顾与主题归纳、研究问题与假设生成两个维度，具体探讨大语言模型在质性研究设计阶段的应用策略与方法路径。

1）文献回顾与主题归纳

传统文献回顾通常依赖关键词搜索与引用追踪等方法，通过知网、Google Scholar、Web of Science 等数据库进行文献检索。然而，这种基于关键词匹配的检索方式存在明显局限性，检索精度与准确度高度依赖特定关键词的存在与否，使得文献回顾过程如同“大海捞针”，难以全面覆盖语义相关但表述不同的文献资源。此外，面对海量文献数据的快速增长，研究者在信息筛选、主题归纳与趋势识别方面面临前所未有的挑战。

近期的研究探索了多种运用大语言模型来解决传统文献回顾的局限性，分别从语义检索系统构建、自动化综述流程设计以及全文分析技术优化等方面进行研究。首先，在语义检索系统构建方面，An 等人[1]开发的 vitaLITy 2 系统代表了这一技术路径的典型应用。该系统构建了包含 66,692 篇论文(1970-2023)的大规模语料库，通过三种语言模型创建文本嵌入，实现了在文本嵌入空间中识别语义相关文献的功能。vitaLITy 2 的技术架构包含一个新颖的检索增强生成(RAG)框架，支持通过增强提示与大语言模型进行交互，能够对论文集合进行自动化总结。系统提供的聊天界面允许用户执行复杂查询而无需学习新的编程语言，充分利用了大语言模型在大规模训练语料库中获得的知识。该系统的创新在于将传统的基于关键词的检索转换为基于语义理解的智能检索，显著提升了文献发现的准确性与全面性。

其次，在自动化综述流程设计方面，Han 等人[2]从系统文献综述

[1] An H, Narechania A, Wall E, et al. Vitality 2: Reviewing academic literature using large language models[EB/OL]. (2024-08-25)[2025-10-21]. https://arxiv.org/abs/2408.13450

[2] Han B, Susnjak T, Mathrani A. Automating systematic literature reviews with retrieval-augmented generation: a

自动化的宏观框架角度出发，深入分析了检索增强生成技术的三个关键过程，即检索、增强和生成，并提出了一个涵盖文献搜索、文献筛选、数据提取和信息综合四个阶段的完整自动化框架。该研究强调 RAG 技术通过整合大语言模型的生成能力与实时信息检索的精确性，能够有效缓解大语言模型对静态预训练知识依赖所导致的不准确性和幻觉问题。而 Agarwal 等人[1]则从具体实施策略的角度，通过零样本能力评估探索了任务分解与优化方法。他们将文献综述任务分解为检索和生成两个核心组件，针对检索环节引入了两步搜索策略和基于提示的重新排序机制，使标准化召回率提升至朴素搜索方法的两倍。在生成阶段，研究提出的基于规划的两步方法能够将生成综述中的虚假引用减少 18-26%，实现更高质量的综述生成。

最后，在全文分析技术优化方面，Brett 等人[2]的研究针对科学文献中关键信息往往蕴含在全文内容而非仅限于摘要的特点，构建了专门用于全文检索与信息提取的大语言模型系统。该系统通过生物学相关问题的基准测试，验证了稀疏检索方法能够在无需密集检索及其相关基础设施和复杂性开销的情况下，展现出接近最先进水平的结果。研究强调文档内语句往往需要更广泛的文章上下文才能被完全理解，因此系统设计注重处理完整文档的上下文信息，并演示了如何提高文献综述生成的相关文档覆盖率。

上述研究从不同层面解决了传统文献回顾面临的效率、准确性和全面性挑战。通过语义理解、智能归纳和上下文分析，这些技术为研究者提供了更系统、更客观的文献分析支持，从而为后续的研究问题生成奠定坚实的理论基础。

2）假设生成能力与质量评估

研究问题生成的系统性与创新性直接决定质性教育研究的学术

comprehensive overview[J]. Applied Sciences, 2024, 14(19): 9103.

[1] Agarwal S, Sahu G, Puri A, et al. LitLLMs, LLMs for literature review: Are we there yet?[EB/OL]. (2024-12-22)[2025-10-21]. https://arxiv.org/abs/2412.15249

[2] Brett D, Myatt A. Patience is all you need! An agentic system for performing scientific literature review[EB/OL]. (2025-04-18)[2025-10-21]. https://arxiv.org/abs/2504.08752

贡献。传统假设生成高度依赖研究者的知识储备与理论敏感性，但面临知识整合能力有限、认知偏见影响、创新思路受限等挑战。随着文献的指数增长，研究者难以全面掌握相关领域最新进展，可能导致重要研究关联的遗漏。大语言模型在假设生成领域的应用研究呈现三个递进层次：能力验证、预测效果和局限反思。

针对假设生成的能力验证，Banker 和 Chatterjee 等人[1]的研究采用了两种方法论路径来评估大语言模型的假设生成质量。第一种方法通过在过去 55 年超过 50 个社会心理学期刊发表的数千篇摘要以及预印本存储库(PsyArXiv)上对 GPT-3 进行微调，社会心理学专家对模型生成和人类生成的假设在清晰度、原创性和影响力维度上给出了相似的评价。第二种方法在未经微调的情况下使用 GPT-4 生成假设，结果显示社会心理学专家认为这些生成的假设在清晰度、原创性、影响力、合理性和相关性等维度上的质量均高于人类生成的假设。Yang 等人[2]进一步推进了这一研究领域，提出了首个用于社会科学学术假设发现的 NLP 数据集，包含 50 篇近期顶级社会科学出版物以及包含足够信息的原始网络语料库。该研究开发了一个多模块框架，并设计了三种不同的反馈机制，实证结果显示该框架在 GPT-4 评估和专家评估中均表现出优越性能，表明大语言模型能够生成文献中不存在的新颖且反映现实的有效科学假设的研究。Park 等人[3]的探索性研究同样证实了 ChatGPT 等大语言模型具备生成科学假设的能力，尽管存在较高的错误率，但生成式 AI 似乎能够有效结构化大量科学知识并提供有趣且可测试的假设。

在预测准确性方面，研究结果显示了大语言模型在预测研究结果方面的显著潜力。Rosenbusch 等人[4]通过测试 GPT-3 预测社会科

[1] Banker S, Chatterjee P, Mishra H, et al. Machine-assisted social psychology hypothesis generation[J]. American Psychologist, 2024, 79(6): 789.

[2] Yang Z, Du X, Li J, et al. Large language models for automated open-domain scientific hypotheses discovery[EB/OL]. (2023-09-05)[2025-10-21]. https://arxiv.org/abs/2309.02726

[3] Park Y J, Kaplan D, Ren Z, et al. Can ChatGPT be used to generate scientific hypotheses?[J]. Journal of Materiomics, 2024, 10(3): 578-584.

[4] Rosenbusch H, Stevenson C E, van der Maas H L J. How accurate are GPT-3's hypotheses about social science phenomena?[J]. Digital Society, 2023, 2(2): 26.

学简单研究结果的能力，获取了600名美国成年公民关于政治态度的真实调查数据作为基准（ground truth）。结果显示，机器生成的假设在零样本、五样本链式提示以及广泛微调条件下的准确率分别达到78%、94%和97%。该研究鼓励了在更具挑战性情境下开发假设引擎的发展，同时强调了解决假设自动化所带来的伦理和哲学挑战的重要性。Lippert等人[1]的研究同样验证了大语言模型在复杂行为科学实验预测中的表现。研究测试了GPT-3.5和GPT-4在预测大规模情感、性别和社会感知实验研究结果方面的能力，发现GPT-4的表现与119名人类专家的表现相匹配，聚合预测与实际效应量之间的相关性分别为0.89(GPT-4)、0.07(GPT-3.5)和0.87(人类专家)。此外，为大学受试者提供查询GPT-4聊天机器人的机会显著提高了他们预测的准确性。

然而，大语言模型在假设生成应用中也存在显著的局限性与挑战。Park等人[2]的研究发现了一个被称为“正确答案”效应的现象，即GPT-3.5在回答涉及政治倾向、经济偏好、判断和道德哲学的细致问题时，不同运行之间的回应变化为零或接近零，总是给出所谓的正确答案。在道德基础理论调查结果中，研究发现GPT-3.5在99.6%的情况下认同为政治保守派，在逆序条件下99.3%的情况下认同为自由派，但无论自报为“GPT保守派”还是“GPT自由派”，都表现出右倾的道德基础。这些结果质疑了将大语言模型作为社会科学中人类参与者一般替代品的有效性，并引发了对人工智能主导的未来可能面临思维多样性减少的担忧。

这些研究成果表明，大语言模型在研究问题与假设生成方面展现出了显著的技术潜力，能够在一定程度上辅助研究者进行创新性思考和理论建构。

[1] Lippert S, Dreber A, Johannesson M, et al. Can large language models help predict results from a complex behavioural science study?[J]. Royal Society Open Science, 2024, 11(9): 240682.
[2] Park P S, Schoenegger P, Zhu C. Diminished diversity-of-thought in a standard large language model[J]. Behavior Research Methods, 2024, 56(6): 5754-5770.

（2）数据收集与参与者交互

质性研究的数据具有来源多样、结构复杂、格式异构等特征，涵盖课堂观察记录、访谈转录、学习日志、反思笔记等多种语料类型。传统人工处理方式在面对大规模、多模态教育数据时存在效率瓶颈和一致性挑战，而大语言模型凭借其自然语言理解和文本结构化能力，为数据收集与处理提供了新的技术路径。下面将从自动化收集、智能扩充、智能分析三个层面探讨大语言模型在数据环节的应用，并揭示各层面的技术优势与存在的挑战。

首先，大语言模型通过 AI 访谈系统实现了数据收集的自动化。Chopra 等人[1]开发的 AI 访谈系统代表了这一技术方向的重要突破。该系统成功对 381 名受试者进行了关于股票市场参与动机的定性访谈，能够揭示影响非参与行为的丰富证据，特别是发现了积极投资心理模型的突出作用。研究显示，AI 访谈员识别的因素在初始即时反应和后续深入回应之间存在系统性差异，心理模型往往在访谈后期显现，且访谈数据在收集 8 个月后仍能预测经济行为，缓解了访谈中“廉价谈话”的担忧，证明了 AI 主导的访谈能够以远低于人工访谈的成本生成丰富、高质量的数据。然而，AI 驱动的数据收集系统在实际应用中也面临着数据质量与丰富性的挑战。Villalba 和 Brown 等人[2]通过大规模用户研究(n=399)评估了三种不同聊天机器人的数据收集效果，其中两种基于大语言模型，一种采用硬编码问题作为基线。评估结果显示，虽然聊天机器人能够根据既定评估标准引出高质量回应，但在捕捉参与者具体动机或个性化例子方面表现不佳，在“丰富性”评估中得分较低。研究还发现大语言模型与人类在质量和丰富性指标评估上的一致性较低，为使用大语言模型扩展和评估定性研究提供了重要警示。

[1] Chopra F, Haaland I. Conducting qualitative interviews with AI[EB/OL]. (2023-11-01)[2025-10-21]. CESifo Working Paper No. 10666. https://ssrn.com/abstract=4583756

[2] Cuevas A, Scurrell J V, Brown E M, et al. Collecting Qualitative Data at Scale with Large Language Models: A Case Study[J]. Proceedings of the ACM on Human-Computer Interaction, 2025, 9(2): 1-27.、

然而，当真实数据收集面临质量和规模限制时，合成数据生成提供了补充路径。Arora 等人[1]提出的人类-LLM 混合营销研究方法论证了大语言模型在质性研究中协助数据生成和分析的有效性。该研究显示大语言模型能够有效创建样本特征、生成合成受试者并进行深度访谈调节，AI-人类混合方法生成的信息丰富、连贯的数据在深度和洞察力方面超越了纯人工数据，在生成主题和摘要的数据分析任务中与人类表现相匹配。专家评判证据表明人类和大语言模型具有互补技能，人类-LLM 混合方法的表现优于纯人工或纯大语言模型方法。

在数据收集之后，大语言模型进一步应用于主题识别与内容分析。Bhaduri 等人[2]提出了基于检索增强生成(RAG)的大语言模型方法来分析访谈转录，将大语言模型定位为初级定性研究助手。该研究将方法扩展到半结构化访谈数据的主题建模，展示了这些模型在传统信息检索和搜索用途之外的多功能性。研究发现该方法能够成功提取感兴趣的主题，与同一数据集的人工生成主题相比具有显著覆盖率，建立了将大语言模型用作初级定性研究助手的可行性。尽管大语言模型在数据处理方面展现出技术优势，但其与人类分析师在推理方式和结果一致性方面仍存在显著差异。Bano 等人[3]通过对比人类和大语言模型的理解能力，使用小样本 Alexa 应用评论进行了实验性研究。研究显示人类与 ChatGPT 3.5 分类之间有三分之一的显著一致性，与 GPT-4 的一致性略低，约为四分之一。在推理比较中，人类分析师更多依赖个人经验，而大语言模型基于应用评论中的具体词汇选择和应用功能组件进行推理。另外，Dengel 等人[4]

[1] Arora N, Chakraborty I, Nishimura Y. AI–human hybrids for marketing research: Leveraging large language models (LLMs) as collaborators[J]. Journal of Marketing, 2025, 89(2): 43-70.

[2] Bhaduri S, Kapoor S, Gil A, et al. Reconciling methodological paradigms: Employing large language models as novice qualitative research assistants in talent management research[EB/OL]. (2024-08-21)[2025-10-21]. https://arxiv.org/abs/2408.11043

[3] Bano M, Zowghi D, Whittle J. Exploring qualitative research using LLMs[EB/OL]. (2023-06-23)[2025-10-21]. https://arxiv.org/abs/2306.13298.

[4] Dengel A, Gehrlein R, Fernes D, et al. Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on computer science education[C]//Informatics. MDPI, 2023, 10(4): 78.

的研究通过对 ChatGPT 和 BARD 进行半结构化访谈，研究发现这些模型的答案强烈依赖于提供的上下文，同一模型对相同问题可能产生截然不同的结果。

综上，大语言模型在质性数据收集与处理的全流程中提供了从自动化收集、合成生成到智能分析的技术支持，显著提升了数据处理效率。

（3）深度数据分析与模式识别

数据编码过程是质性研究中较为耗时且要求严格的分析环节，传统的人工编码不仅需要研究者投入大量时间精力，还面临编码一致性难以保证、主观性强以及跨编码员可靠性不稳定等挑战。大语言模型技术在自然语言理解与文本分类方面的能力为质性研究的编码过程提供了新的技术支撑。相关研究沿着两个层面展开：基于预设框架的演绎编码应用，以及归纳与演绎结合的混合编码探索，逐步揭示了 AI 辅助编码的技术路径与应用效果。

演绎编码作为基于预设理论框架的分析方法，是大语言模型最早介入的编码领域。Tai 等人[1]提出了使用大语言模型支持传统演绎编码的方法论，通过将样本文本和编码簿输入大语言模型，要求模型判断代码是否存在于提供的样本文本中并请求支持编码的证据。研究将样本文本输入 160 次以记录大语言模型响应迭代之间的变化，每次迭代类似于新编码员使用编码簿信息对文本进行演绎分析，结果显示大语言模型分析能够提供代码识别的系统平台并避免分析偏差。Balt 等人[2]则在心理社会解剖学研究中采用 few-shot learning 方法测试 LLAMA3 模型对访谈数据的演绎编码能力，通过 38 个半结构化访谈数据在三项任务中的表现测试，结果显示大语言模型在二元分类任务中与研究者达到了实质性一致(准确率：0.84)，在滑动窗

[1] Tai R H, Bentley L R, Xia X, et al. An examination of the use of large language models to aid analysis of textual data[J]. International Journal of Qualitative Methods, 2024, 23: 16094069241231168.

[2] Balt E, Salmi S, Bhulai S, et al. Deductively coding psychosocial autopsy interview data using a few-shot learning large language model[J]. Frontiers in Public Health, 2025, 13: 1512537.

口任务中的准确率为 0.67。

在演绎编码基础上，研究者进一步探索归纳与演绎结合的混合编码方法，构建更为灵活的人机协作框架。Zhang 等人[1]开发的 QualiGPT 工具专门解决使用 ChatGPT 进行定性分析时面临的挑战，通过结合归纳和演绎编码方法，在模拟和真实数据集上的比较分析显示该工具显著改善了定性分析过程，在效率、透明度和可访问性方面表现出色。使用编码间可靠性(IRR)措施评估显示，在各种编码场景中人类编码员和 QualiGPT 之间存在实质性一致。Than 等人[2]测试了生成式大语言模型复制和增强传统定性编码的能力，通过在四个封闭和开源大语言模型上实验多种提示结构，提出了使用大语言模型进行定性编码的工作流程。研究发现大语言模型在准确匹配手工编码输出方面的表现几乎与先前的监督机器学习模型一样好，使用大语言模型作为自然语言对话者能够密切复制传统质性方法。此外，Meng 等人[3]提出了 CHALET 方法，该方法涉及大语言模型支持的数据收集、执行人类和大语言模型演绎编码以识别分歧，并对这些分歧案例执行协作归纳编码以获得新的概念性见解。通过在心理疾病污名归因模型中的应用，该方法揭示了认知、情感和行为维度上的隐性污名化主题，展示了人机协作在发现新理论洞察方面的潜力。

从演绎编码的自动化实现到混合编码的灵活应用，这些研究为理解 AI 技术在提升编码效率、保证编码一致性方面的作用提供了重要参考，也为研究者在实际应用中选择合适的技术路径和协作模式提供了科学依据。

（4）研究成果的展示、验证与交互

[1] Zhang H, Wu C, Xie J, et al. When qualitative research meets large language model: Exploring the potential of QualiGPT as a tool for qualitative coding[EB/OL].(2024-07-25)[2025-10-21]. https://arxiv.org/abs/2407.14925

[2] Than N, Fan L, Law T, et al. Updating “The Future of Coding”: Qualitative Coding with Generative Large Language Models[J]. Sociological Methods & Research, 2025, 54(3): 849-888.

[3] Meng H, Yang Y, Li Y, et al. Exploring the potential of human-llm synergy in advancing qualitative analysis: A case study on mental-illness stigma[EB/OL]. (2024-05-09)[2025-10-21]. https://arxiv.org/abs/2405.05758.

研究成果的展示、验证与交互直接影响研究发现的传播效果和学术影响力。传统的成果呈现方式往往局限于文字叙述，难以充分展现质性数据的丰富性和研究发现的多维性，同时在成果验证和读者互动方面也存在诸多限制。大语言模型技术在自然语言生成、多模态处理和交互式对话方面的能力，为质性研究成果的展示、验证与传播提供了新的可能性。下面从可视化生成与写作辅助两个方面，探讨大语言模型如何改进研究成果的呈现。

首先，大语言模型通过代码生成能力实现了从文本到可视化的直接转化。利用 Codex 和 GPT-3 等模型，研究者能够自动生成创建数据可视化的代码，将复杂的质性分析结果转化为直观的图表形式[1]。这种技术不仅能处理结构化的编码结果，还能从非结构化文本中提取信息并生成相应的可视化表达。例如，使用 GPT-3.5-turbo 创建的知识洞察图能够有效展示材料属性和组成关系等复杂概念间的联系[2]。大语言模型在多模态数据的知识表达方面也展现出独特优势，能够整合文本、图像等不同形式的研究数据，生成更为丰富和立体的研究成果展示[3]。这种可视化能力使得复杂的研究发现更容易被学术界和公众理解，显著提升了研究成果的可及性和传播效果。

与此同时，大语言模型在写作辅助方面已从简单的文本生成工具演变为能够深度参与知识转化过程的写作伙伴。在信息综合与主题提炼层面，Lancaster 的研究[4]展示了大语言模型在处理海量学术文献并识别主题方面的能力，能够快速完成传统文献综述需要耗费大量时间的工作。Parker 等人[5]则将这种能力应用于实证数据分析，从大量用户反馈中高效提取洞察，为研究发现的系统化整理提供了

[1] Maddigan P, Susnjak T. Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models[J]. IEEE Access, 2023,11:45181-45193.

[2] Jablonka K M, Ai Q, Al-Feghali A, et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon[J]. Digital Discovery, 2023, (5): 1233-1250.

[3] 罗江华,张玉柳.多模态大模型驱动的学科知识图谱进化及教育应用[J].现代教育技术. 2023, (12):76-88.

[4] Lancaster T. A large language model-supported synthesis of contemporary academic integrity research trends[EB/OL]. (2024-01-08)[2025-10-21]. https://arxiv.org/abs/2401.03481

[5] Parker M J, Anderson C, Stone C, et al. A large language model approach to educational survey feedback analysis[J]. International journal of artificial intelligence in education, 2024: 1-38.

有力支持。在结构化报告生成方面，De Paoli 等人[1]的研究展示了大语言模型将抽象分析结果转化为具体成果的能力，例如在完成主题分析后，大语言模型能够将提炼出的用户特征自动转化为格式规范、内容生动的“用户画像”报告，这种转化不仅保持了学术严谨性，还增强了研究成果的可读性和实用性。在写作的精细化环节，Lee 等人[2]的研究表明大语言模型能够协助研究者筛选并优化手稿中的引文，确保每个论点都有充分的证据支撑。这些应用表明，大语言模型已经从简单的文本生成工具演变为能够深度参与知识转化过程的写作伙伴，帮助研究者将复杂的分析结果高效转化为规范的学术文本。

从可视化表达到文本写作，大语言模型正在改变质性研究成果的呈现方式，使研究发现的传播变得更加高效和多元。这些技术应用为研究者提供了提升成果展示质量和传播效率的新工具。

综上所述，大语言模型在质性研究中的应用已经从早期的单点尝试发展为覆盖研究全流程的系统性探索。从研究设计阶段的文献分析和理论框架构建，到数据收集中的自动化访谈和智能转录，再到编码分析中的主题识别和模式发现，以及最终成果的可视化呈现和交互验证，技术介入的广度和深度都在不断拓展。现有文献主要通过实验对比、案例分析、工具开发等方式探索大语言模型在具体任务中的效能，形成了丰富的实证基础和多样的技术路径。这些探索为理解大语言模型与质性研究的结合提供了重要的经验积累，也为构建更加系统的应用框架奠定了基础。

3.2 典型应用场景

教育质性研究涵盖了从微观课堂互动到宏观政策制度的广阔领

[1] De Paoli S. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach[J]. Social Science Computer Review, 2024, 42(4): 997-1019.

[2] Lee J, Lee J, Yoo J J. The role of large language models in the peer-review process: opportunities and challenges for medical journal reviewers and editors[J]. Journal of Educational Evaluation for Health Professions, 2025, 22.

域，不同研究场景对 AI 辅助工具的需求特征和技术要求存在显著差异。本节选择了三个具有代表性的典型应用场景，每个场景都具有独特的数据特征、分析需求和方法论挑战，为全面评估大语言模型辅助教育质性研究的可行性与有效性提供了丰富的实证基础。

3.2.1 课堂互动与学习过程分析

课堂互动与学习过程蕴含着复杂的语言交流、意义协商和认知建构过程，传统质性研究虽能深入挖掘其内在机制，但面临着分析规模和效率的制约。大语言模型技术的引入为这一困境提供了解决方案，通过保持质性研究对意义理解和情境诠释的重视，同时扩展分析的广度和深度。当前研究从话语意义分析、互动过程理解和认知发展追踪三个方面，展现了大语言模型在课堂研究中的应用价值。

首先，在话语意义分析方面，大语言模型的引入使得大规模话语语料的深度分析成为可能，同时保持了对语言细微差别和深层含义的敏感性。Long 等人[1]通过 GPT-4 分析中学数学和语文课堂对话，将 15 个类别的编码方案向量化处理，不仅将编码时间从 150 分钟缩短至 5 分钟，更重要的是在详细邀请和详细阐述类别中达到 0.973-0.995 的编码一致性，证明了大语言模型能够理解话语背后的教学意图和交际功能。Wang 等人[2]比较 BERT 微调和 Llama3 提示工程在话语动作识别中的表现，发现 BERT 在理解复杂教育语境方面的优势，为教师专业发展提供了基于话语分析的反馈机制。Hayes 的对话式质性数据分析方法强调了大语言模型在处理隐喻、代码转换等复杂话语现象时的敏感性，体现了质性研究对语言细微差别的关注得以在技术支持下放大。这些研究表明，大语言模型能够在保持质性研究对话语深层含义关注的前提下，实现大规模的意义识别和分类。

[1] Long Y, Luo H, Zhang Y. Evaluating large language models in analysing classroom dialogue[J]. npj Science of Learning, 2024, 9(1): 60.

[2] Wang D, Chen G. Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy[J]. British Journal of Educational Technology, 2025.

其次，大语言模型技术为模拟和分析复杂互动提供了新的可能性，能够识别互动模式中的细微变化和潜在机制。尽管课堂中的社会互动过程历来受到质性研究者的重视，但复杂的多方互动往往难以全面捕捉和系统分析。Zhang 等人[1]构建的 SimClass 多智能体仿真系统通过 GPT-4 驱动不同角色智能体，在弗兰德斯互动分析框架下重现了教师话语 82.4%、学生话语 16.2%的课堂互动模式，并观察到协作教学、情感支持等自然涌现的教育现象。这种仿真不仅验证了理论模型，更为理解真实课堂中的复杂互动提供了新的观察视角。

最后，大语言模型在理解语言背后认知状态方面展现出独特优势，为质性研究提供了新的认知过程分析工具。认知发展过程的追踪长期以来面临着内在思维难以直接观察的挑战，质性研究虽能通过话语分析间接推断认知状态，但缺乏大规模系统性分析的手段。基于实用探究模型的认知临场感分析采用七步 LACA 方法，在触发、探索、整合、解决四个认知阶段中，大语言模型在整合阶段表现最佳，能够有效识别学生的概念构建和意义建构过程[2]。Tai 等人[3]通过 160 次重复实验验证大语言模型在演绎编码中的可靠性，对自主性、坚持性、自我认知等概念代码的分析展现了质性研究对个体内在状态的关注。混合式工作流程的反思性内容分析通过思维链提示技术，在 GPT-4 的 9 个编码任务中有 8 个达到 0.6 以上的实质一致性，同时保持了质性研究的反思性和解释性特质[4]。这些研究显示了大语言模型在支持质性研究追踪认知过程、理解学习发展方面的价值。

3.2.2 制度评估与政策分析

教育制度和政策层面的质性分析涉及大量文档资料和访谈数据

[1] Zhang Z, Zhang-Li D, Yu J, et al. Simulating classroom education with LLM-empowered agents[EB/OL]. (2024-06-27)[2025-10-21]. https://arxiv.org/abs/2406.19226.

[2] Wang Y. A study on the efficacy of ChatGPT-4 in enhancing students' English communication skills[J]. Sage Open, 2025, 15(1): 21582440241310644.

[3] Tai R H, Bentley L R, Xia X, et al. An examination of the use of large language models to aid analysis of textual data[J]. International Journal of Qualitative Methods, 2024, 23: 16094069241231168.

[4] Dunivin Z O. Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive content analysis[J]. EPJ Data Science, 2025, 14(1): 28.

的处理，这一领域的复杂性要求研究者具备深厚的情境知识和专业判断能力。由于质性研究强调对现象的深度理解和意义诠释，研究者不仅需要识别显性的主题模式，更要挖掘隐藏在数据背后的深层逻辑和文化内涵，这使得大语言模型在该领域的应用既充满潜力又面临挑战。已有研究从不同角度探索了大语言模型在制度评估与政策分析中的应用，分别聚焦于能力评估、协作模式和工具开发等方面。

从能力评估角度来看，研究者测试了大语言模型处理教育政策数据的表现与局限。一项研究测试了 Google Bard 和 ChatGPT 在分析教师年度评估报告中质性数据的有效性[1]。研究数据包括关于多样性、公平性和包容性举措的图表和 30 份开放式回应，任务是进行主题分析。研究设计了三轮分析流程，系统地比较了独立的人类分析、协作的人类分析以及人机协作分析。研究发现 AI 工具虽然分析速度快，3 到 15 秒即可完成，但非常容易出错、过度简化，并且缺乏人类研究者所具备的校园特定情境知识。例如，Bard 在解读一个简单的条形图时就遇到了困难，而 ChatGPT 最初只是提供了内容的摘要，而非真正的主题分析。

从协作模式角度来看，研究者探索了如何将大语言模型与人类专业知识有效结合。一项研究探索了如何将 GPT-4 与人类专业知识相结合，以增强对 K-12 教育政策利益相关者访谈的文本分析[2]。研究数据为访谈转录稿，任务是同时进行主题分析和情感分析。研究比较了由人类专家、GPT-4 以及一个经过微调的 BERT 模型所生成的主题之间的一致性。研究发现，人机交互的分析方法能够提升分析的效率、效度和可解释性。GPT-4 能够提供新的视角和更高的一致性，但研究者强调，人类的领域知识在指导分析方向和解释结果

[1] Slotnick R C, Boeing J Z. Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale[J]. New Directions for Teaching and Learning, 2024.

[2] Liu A, Sun M. From voices to validity: Leveraging large language models for textual analysis of policy stakeholder interviews[EB/OL]. (2023-12-03)[2025-10-21]. https://arxiv.org/abs/2312.01202

方面扮演着互补性的、不可或缺的角色。

从工具开发角度来说，研究者构建了面向教育政策分析的辅助系统。一项研究以中国人工智能学会《大型语言模型的教育应用》研究报告来评估大语言模型在未来教育中的挑战与机遇[1]。项目在数字化转型背景下展开多方论证，开发了轻量级辅助研究工具 GPT-EDU4SIGHT。如图 3-3 所示，该工具支持访谈设计、智能访谈、自动编码与深度分析等环节。该工具利用大语言模型的认知建模与语境理解能力，帮助优化研究设计、识别趋势与弱信号，并揭示不同群体在教育改革议题上的共识与分歧，为实践探索和政策制定提供参考。

图 3-3 GPT-EDU4SIGHT 工具

在教育质性研究的制度评估与政策分析中，这些应用实践既展现出大语言模型提升处理效率和发现新分析视角的优势，也暴露出在情境理解、深度分析和准确性方面的局限性。在这一应用场景中，人类专业知识在维护研究深度和完整性方面都发挥着关键作用，AI 更多地承担着辅助和增强的功能。

[1] 陈鹏,张靖沅,陈向东.大模型如何融入教育的质性研究:理论潜力与案例实践[J].现代教育技术,2024,34(10):32-41.

3.2.3 医学教育应用研究

当前，大语言模型在教育质性研究中的应用不断拓展，已被广泛用于教师教育、心理教育、护理教育及医学教育等领域的学习过程与教学实践分析[1][2][3]，其中医学教育的应用尤为突出。

在医学教育领域，质性研究是理解学习者专业认同建构、教学互动机制和临床学习体验的重要途径。研究者已开始利用模型对访谈、学习反思和教学文本进行主题提炼与模式识别，以提升分析效率并拓展研究视角。现有应用主要集中在两个方面：一是分析医学生学习体验与教学互动，以识别学习规律与课堂交流模式；二是探索人机协同的质性研究方法，以优化编码与主题分析流程。

在学习体验与教学互动分析方面，大语言模型已被应用于医学生反思日志、访谈资料和课程反馈的分析之中，用以识别学习过程中的心理变化与行为模式。Haider 等人[4]在对医学生反思性写作的研究中，比较了 ChatGPT 生成的反馈与人工评阅结果，发现模型能够识别出常见主题并生成结构化建议，为教学者提供了高效的文本分析与形成性反馈支持。Zhang 等人[5]则在混合方法研究中发现，医学生使用 ChatGPT 进行学习记录整理和反思撰写时，模型输出的文本具有较高一致性，可为教师和研究者提供可量化的质性材料。类似研究表明，大语言模型 在医学教育的学习体验研究中能快速捕捉反思性文本中的主题模式，提升分析效率与资料利用率。

在研究方法层面，大语言模型的引入正在推动医学教育质性研究迈向人机协同的分析模式。Kondo 等人[6]开展了一项比较性研究，

[1] Park Y, Hong Y. Sentiment analysis of preservice teachers' reflections using a large language model[C]//2024 6th International Workshop on Artificial Intelligence and Education (WAIE). IEEE, 2024: 61-65.

[2] Smirnov E. Enhancing qualitative research in psychology with large language models: a methodological exploration and examples of simulations[J]. Qualitative Research in Psychology, 2025, 22(2): 482-512.

[3] Xu Y, Xie H, Zeng Y, et al. Using Large Language Models to analyze factors influencing academic supervision relationships in qualitative interviews with postgraduate nursing students[J]. Nurse educator, 2025, 50(4): E201-E206.

[4] Haider N, Morjaria L, Sheth U, et al. MD Student Perceptions of ChatGPT for Reflective Writing Feedback in Undergraduate Medical Education[J]. International Medical Education, 2025, 4(3): 27.

[5] Zhang J S, Yoon C, Williams D K A, et al. Exploring the usage of ChatGPT among medical students in the United States[J]. Journal of Medical Education and Curricular Development, 2024, 11: 23821205241264695.

[6] Kondo T, Miyachi J, Jönsson A, et al. A mixed-methods study comparing human-led and ChatGPT-driven qualitative analysis in medical education research[J]. Nagoya journal of medical science, 2024, 86(4): 620.

将 ChatGPT 辅助的主题分析与研究者主导的人工分析相对照，结果显示模型在初步聚类与模式识别上具有较高一致性与效率，虽然在情感细节和语境理解方面仍受限，但其分析框架可为医学教育研究提供可借鉴的自动化分析路径。Levit 等人[1]的研究结果亦表明，大语言模型能在重复模式识别上展现优势，其“人工+模型混合分析”的思路可迁移至医学教育的访谈与课堂文本处理中。通过此类协同分析，研究者可利用模型执行初步编码与主题归纳，再由人工补充教育语义与情境解释，以平衡效率与解释深度。

在医学教育具体应用背景下，Mondal 等人[2]进行了一项关于医学生对大语言模型教育应用感知的质性调研，虽然该研究本身使用传统质性方法，但为理解大语言模型在医学教育中的接受度提供了重要视角。研究通过对来自 8 个印度州的 25 名医学生进行深度访谈，访谈录音平均时长 55.28±18.04 分钟，经转录后使用 QDA Miner Lite v.2.0.8 进行主题分析。研究识别出三个主要主题：使用场景、增强学习和大语言模型局限性，为理解医学教育中大语言模型应用的复杂性提供了质性研究视角。综合现有文献可以看出，对于医学教育研究者而言，掌握大语言模型辅助质性研究的技能将成为提升研究能力的重要途径，但同时也需要保持对技术局限性的清醒认识，确保研究质量和学术诚信不因技术应用而受到影响。

3.3 质性研究的独特伦理风险

大语言模型在教育质性研究中的应用在带来技术便利的同时，也引发了一系列亟待解决的风险和伦理问题。与其他研究领域相比，教育质性研究的伦理问题具有其独特的复杂性，主要体现在三个方面：质性研究方法的深度性和亲密性特征使得数据隐私保护面临更大挑战；质性研究对复杂教育现象的解释性要求使得结果准确性和

[1] Shanwetter Levit N, Saban M. When investigator meets large language models: a qualitative analysis of cancer patient decision-making journeys[J]. npj Digital Medicine, 2025, 8(1): 336.

[2] Mondal H, Karri J K K, Ramasubramanian S, et al. A qualitative survey on perception of medical students on the use of large language models for educational purposes[J]. Advances in Physiology Education, 2025, 49(1): 27-36.

可靠性问题更加关键；质性研究强调的文化敏感性和多元价值观理解与大语言模型的技术局限性之间存在潜在张力。

此外，教育质性研究承载着重要的社会责任和影响力，其结果往往直接关联教育实践改进和政策制定，这种广泛的社会影响力要求研究者在使用大语言模型时承担更大的伦理责任。质性研究方法论强调的反思性、批判性思考和研究者—参与者关系的建构性，在大语言模型技术介入后都面临着新的伦理考验。

3.3.1 数据隐私保护风险

教育质性研究数据具有高度敏感性，往往包含未成年学生的个人信息、学习表现、家庭背景，以及教师的教学反思和职业评价。这些数据一旦泄露，不仅会侵犯个人隐私，还可能对学生的教育轨迹和教师的职业发展产生长期影响。在教育质性研究中使用大语言模型时，隐私保护面临多种挑战：学生与教师敏感信息的泄露风险、商业平台处理教育数据的合规冲突、以及课堂观察与访谈数据关联导致的隐私累积风险。

首先，教育场景涉及学生、教师、家长等多方参与者，其敏感信息泄露与知情同意机制的复杂性构成了首要挑战。研究表明，研究者必须特别注意避免分享个人身份信息或损害弱势群体的数据，这些风险在使用专有大语言模型时尤为突出[1]。在教育质性研究中，这一风险表现得更为复杂。学生访谈中可能包含学习困难、同伴关系、家庭问题等私密信息，教师访谈可能涉及对同事的评价、对学校管理的看法等职业敏感内容。这些信息一旦通过大语言模型泄露，可能对当事人造成持续性伤害。有研究发现，尽管研究者对隐私保护存在广泛担忧，但大语言模型仍在缺乏最佳实践教育的情况下被大量采用，这表明存在结构性的研究者培训失败[2]。

[1] Schroeder H, Aubin Le Quéré M, Randazzo C, et al. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025: 1-17.

[2] Schroeder H, Aubin Le Quéré M, Randazzo C, et al. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions[EB/OL]. (2024-10-15)[2025-10-21]. https://arxiv.org/abs/2410.07362

同时，教育研究者在使用大语言模型时往往缺乏充分的风险意识，可能没有意识到删除姓名和学校名称并不足以保护参与者身份。大语言模型强大的模式识别能力可能通过课程设置、教学活动、地理位置等信息推断出具体的学校和个人。教育环境中的多重利益相关者使隐私保护变得更加复杂。虽然研究指出需要更新知情同意书以准确反映大语言模型的预期用途，包括详细的隐私保护措施说明和潜在风险教育[1]，但在教育质性研究中，需要获得学生、家长、教师、学校管理层等多方同意。这些不同群体对 AI 技术的理解程度差异很大，年龄较小的学生可能无法充分理解数据处理的复杂性，家长可能对技术细节缺乏了解，而教师则可能担心职业评价受到影响。

其次，商业大语言模型平台的数据处理机制与教育领域的法律保护要求之间存在深层冲突。这些平台通常将用户输入的数据用于模型训练和改进，意味着教育访谈记录可能被永久保存并用于其他目的。对于教育领域而言，这种做法相当危险，因为教育数据涉及未成年人群体，受到更严格的法律保护。学生的课堂表现、学习进展、行为问题等信息具有长期性影响，不当使用可能影响其升学机会或同伴关系。而跨境数据传输在教育领域面临更严格的监管要求。主流商业大语言模型服务多由海外公司提供，教育数据的跨境传输可能违反当地关于未成年人数据保护的法规。不同司法管辖区对教育数据的保护标准、存储要求、访问权限等方面存在差异，研究者往往难以确保完全合规。当涉及国际学校、交换项目等跨国教育活动的研究时，合规要求变得更加复杂。

最后，课堂观察、师生访谈、学习档案等多维度教育数据的关联分析进一步放大了隐私风险。一个学生的多次访谈记录、不同教师对同一事件的描述、跨时间段的观察数据等，这些看似分散的信

[1] Schroeder H, Aubin Le Quéré M, Randazzo C, et al. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025: 1-17.

息在大语言模型的分析下可能被关联起来，形成更完整的个人画像。这种数据关联能力虽然有助于深度分析，但也大大增加了隐私泄露的风险和影响范围。

3.3.2 结果准确性与可靠性问题

在大语言模型辅助的教育质性研究中，结果准确性与可靠性问题直接影响研究的科学价值和实践应用效果。从理论层面来看，幻觉现象在大语言模型中具有不可避免的特征[1]，这一固有属性在处理教育质性研究的复杂语境时表现得尤为明显。当研究者依赖大语言模型进行访谈分析、文本编码或主题识别时，模型输出中的不确定因素可能系统性地影响研究结论，从而削弱教育政策制定和教学实践改进的科学基础。这些问题主要体现在两个方面：内容生成中的幻觉现象以及编码过程中的可靠性不足。

一方面，幻觉（hallucination）现象在教育文本分析中表现为多种形式的偏差，即生成看似合理但实际上不准确、不一致或不相关的内容。大语言模型基于统计模式预测词汇序列而非基于真正的理解能力[2]，使其在面对教育领域特有的情境复杂性时容易产生误判。在处理教育文本时，模型可能生成不准确的历史信息、科学事实或教育理论表述，这种事实性偏差在分析教育政策文件或学术文献时构成了严重风险。部分原因在于，它往往以高度流畅、专业化的语言呈现，研究者如果缺乏相关领域的深厚知识积累，很难在第一时间识别出错误，可能将这些虚假信息纳入分析框架，导致研究结论建立在错误的知识基础之上。

同时，幻觉的出现也经常会导致大语言模型在同一研究任务中产生相互矛盾的分析结果，破坏了质性研究所要求的内在逻辑一致性。例如，对同一概念的重复编码使用不同的类别标签；在分析因

[1] Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: An innate limitation of large language models[EB/OL]. (2024-01-24)[2025-10-21]. https://arxiv.org/abs/2401.11817

[2] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

果关系时，前后颠倒原因与结果；对参与者态度的判断在不同文本片段中出现对立性评价。尤其是当模型的上下文理解能力受限或对研究问题把握不准确时，会产生这些幻觉问题。当研究者利用大语言模型进行大规模文本编码时，幻觉会严重损害编码的信度，使研究发现缺乏内部效度支撑，无法形成连贯的理论叙事。

另一方面，编码可靠性问题则源于大语言模型对教育现象理解的表层性质。虽然研究显示大语言模型在识别预设编码类别方面具备一定能力，但在深度解释和细微差异识别方面存在明显不足[1]。当大语言模型分析课堂互动数据时，它可能准确识别“提问”这一行为类别，却无法精确判断该提问属于探索性、评估性还是引导性，因为这需要对教育情境的深度把握和教学意图的准确理解。这种表层识别与深层理解之间的差距，使得大语言模型在需要细致判断和情境化解释的质性分析任务中面临可靠性挑战。

当前的研究实践表明，大语言模型在教育质性研究中的应用需要审慎对待[2]。传统质性研究中的信度和效度标准依然适用于大语言模型辅助研究，但需要根据算法特性进行相应调整[3]。这要求研究者不仅建立技术层面的验证机制，更需要深入理解大语言模型工作原理与质性研究方法论之间的张力。

3.3.3 冲突和价值观影响

大语言模型辅助教育质性研究中的文化冲突和价值观影响问题源于技术发展的文化偏向性与教育研究多元化需求之间的根本矛盾。研究表明，当缺乏明确文化身份提示时，大语言模型表现出与西方文化价值观的高度一致性，这种倾向在不同版本模型中呈现出惊人的稳定性[4]。分析发现，大语言模型训练数据的地理分布、文化多

[1] Tai R H, Bentley L R, Xia X, et al. An examination of the use of large language models to aid analysis of textual data[J]. International Journal of Qualitative Methods, 2024, 23: 16094069241231168.

[2] Amirova A, Fteropoulli T, Ahmed N, et al. Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity[J]. Plos one, 2024, 19(3): e0300024.

[3] Morse J M, Barrett M, Mayan M, et al. Verification strategies for establishing reliability and validity in qualitative research[J]. International journal of qualitative methods, 2002, 1(2): 13-22.

[4] Tao Y, Viberg O, Baker R S, et al. Cultural bias and cultural alignment of large language models[J]. PNAS nexus, 2024, 3(9): pgae346.

样性和语言覆盖度存在显著失衡，这种数据层面的不平等直接塑造了模型的文化认知框架[1][2]。当研究者依赖这样的模型分析来自不同文化背景的教育现象时，必然面临偏向来源、具体表现、应对策略三个层面的挑战。

文化偏向在教育质性研究的具体实践中产生了多层次的影响。个性化教育场景评估揭示，大语言模型在为不同人群生成教育内容时表现出系统性偏见，这种偏见不仅体现在显性特征上，更隐蔽地体现在对教育价值观和学习方式的判断中[3]。当分析亚洲教育中的集体学习模式时，大语言模型可能将其误读为缺乏个体创新性，而非理解其在培养合作精神方面的文化价值。文化偏见的隐蔽性使其格外危险，因为它往往隐藏在看似中性的语言表达中，并通过特定文化传统的反复强化而根深蒂固[4]。这种偏向在编码和主题识别过程中可能导致对非西方教育实践的系统性低估或误解。

尽管具备多语言处理能力，大语言模型虽然在社会文化规范判断上表现出与人类相似的能力，但明显偏向西方文化价值体系，多语言训练只能部分缓解而无法根本解决这一问题[5]。当德语“Bildung”或中文“师父”等具有深厚文化内涵的教育概念需要在研究中被准确理解时，简单的词汇转换无法传达其完整的文化意义，从而影响整个研究的理论建构和结论形成。

面对这一困境，研究者需要重新审视大语言模型辅助研究与质性研究核心价值之间的兼容性问题。质性研究强调理解研究参与者的主观体验和意义建构，追求对教育现象的深度阐释和情境化理解，而大语言模型的文化偏向性可能系统性地扭曲这种理解过程。研究

[1] 陈向东,卢淑怡,易乐湘.文化冲突:大语言模型教育应用中的张力与调适[J].远程教育杂志,2025,43(03):3-15+43.

[2] Guo Y, Guo M, Su J, et al. Bias in large language models: Origin, evaluation, and mitigation[EB/OL]. (2024-11-19)[2025-10-21]. https://arxiv.org/abs/2411.10915.

[3] Weissburg I, Anand S, Levy S, et al. LLMs are biased teachers: Evaluating LLM bias in personalized education[EB/OL]. (2024-10-25)[2025-10-21]. https://arxiv.org/abs/2410.14012

[4] Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion[J]. ACM Journal of Data and Information Quality, 2023, 15(2): 1-21.

[5] Kim M, Baek S. Exploring large language models on cross-cultural values in connection with training methodology[EB/OL]. (2024-12-20)[2025-10-21]. https://arxiv.org/abs/2412.08846

者必须在数据收集、编码分析和理论建构的各个环节建立文化反思机制，确保大语言模型的参与不会削弱质性研究对多元文化视角的敏感性和包容性。只有在深入理解并有效控制文化冲突的前提下，同时保持对质性研究方法论原则的坚持，大语言模型才能真正成为增强而非削弱教育质性研究文化敏感性的有效工具。

本章通过系统梳理和分析现有文献，全面审视了大语言模型在教育质性研究中的应用现状与发展态势。综合发现，当前研究已从早期的技术可行性探索发展为覆盖研究全流程的系统性应用，在提升分析效率、扩展研究规模和深化理解层次方面展现出显著潜力，但同时暴露出数据隐私保护、结果准确性控制和文化价值观冲突等亟待解决的问题。现有研究为理解这一领域的发展轨迹和核心挑战提供了重要基础，但在理论框架整合、应用标准制定等方面仍存在不足，需要未来研究进一步深化和完善。

第 4 章 AI 驱动的教育量化研究

在教育研究的发展历程中，技术手段的演进始终与量化方法的完善紧密相连。无论是早期的计算器、统计软件，还是后来的数据库系统和可视化工具，技术工具贯穿于从数据采集、处理到统计分析、结果呈现的全流程，帮助研究者将教育现象转化为可度量的量化指标，并通过统计方法揭示变量间的关联。教育量化研究是指运用数学与统计学方法，对教育现象进行量化测量、分析与建模，从而揭示教育规律、验证教育理论并为教育决策提供依据的一种研究范式。

随着人工智能的快速发展，大语言模型已在社会学、教育学和医学等领域广泛应用于数据分析、统计建模等任务，显著提升了研究效率和精度[1]。在教育领域，大语言模型使研究者能够自动化处理海量数据，将课堂互动、学习动机、教育公平等难以量化的教育现象转化为可测量的指标，并通过算法模型挖掘数据中的潜在模式与关联。由此，这一技术驱动的研究模式进一步拓展了研究的深度与广度，推动教育量化研究从传统的描述性分析向预测性与决策支持系统转变，为教育理论发展、政策制定与教学实践提供了更科学的依据。

4.1 面向人工智能的教育量化研究

教育量化研究长期依赖问卷、测验与实验数据，以统计方法为主要分析工具，但在当下教育实践中逐渐显现出不足。一方面，在线学习平台与智能教育环境的普及带来了海量数据，包括学习日志、课堂互动文本与多模态数据，超出传统方法的处理能力。另一方面，研究者越来越关注学习行为模式、个性化发展轨迹以及教育资源配置等复杂问题，传统的变量构建与回归分析在解释力和预测力上受到局限。

[1] Ashikhmin E G, Levchenko V V, Gyuzel'I S. Experience in applying large language models to analyse quantitative sociological data[J]. Вестник университета, 2024: 206.

大语言模型等前沿人工智能技术的兴起为突破这些瓶颈提供了新的路径。已有研究表明，大语言模型在非结构化教育数据的量化处理、合成数据生成以及跨模态信息整合方面展现出较强潜力[1][2][3]。研究者利用其自然语言处理和生成能力，不仅能够自动化提取教学与学习过程中的关键变量，还能在建模预测与结果呈现上提升效率与精度。这些进展使教育量化研究逐渐由“单一统计驱动”转向“模型与数据驱动”并行的格局。

4.1.1 AI 对于教育量化研究的影响

大语言模型等技术的应用正在拓展教育量化研究的实践边界。近年来，越来越多的研究团队开始探索将大语言模型整合到量化分析流程中的可能性。这一探索的深层动因在于教育数据生态的根本性变化。数字化学习环境产生的大量非结构化数据长期处于研究视野的边缘，不是因为其价值不被认可，而在于缺乏有效的处理工具。大语言模型的出现打破了这一技术瓶颈，使研究者能够将文本、音频、图像等多模态教育数据纳入量化分析框架。

这种技术能力的突破在研究实践中产生了实质影响。传统量化研究受制于变量操作化的复杂性，往往将注意力集中在易于测量的显性指标上，而对学习过程中的隐性认知变化、社会情感因素、个体差异表现等复杂现象缺乏有效的量化手段。大语言模型的语义理解和模式识别能力为这些复杂现象的量化分析提供了新的可能性。研究者不再需要将复杂的教育现象简化为有限的量表条目，而是可以直接从学习者的自然表达中提取有意义的测量指标。这种转变不仅提升了测量的生态效度，也为教育量化研究拓展了新的问题域。

基于当前文献的系统梳理和实证证据的深入分析，大语言模型

[1] Remadi A, El Hage K, Hobeika Y, et al. To prompt or not to prompt: Navigating the use of large language models for integrating and modeling heterogeneous data[J]. Data & Knowledge Engineering, 2024, 152: 102313.

[2] Kang A, Chen J Y, Lee-Youngzie Z, et al. Synthetic data generation with LLM for improved depression prediction[EB/OL]. (2024-11-15)[2025-10-21]. http://arxiv.org/abs/2411.17672

[3] Palakonda V, Kumar A V S M A, Wesly D A, et al. A Study on Large Language Models in Multi Modal Learning[C]//2024 4th International Conference on Sustainable Expert Systems (ICSES). IEEE, 2024: 902-905.

对教育量化研究的影响主要体现在教育数据来源的多样化、研究议题的拓展与深化两个层面。

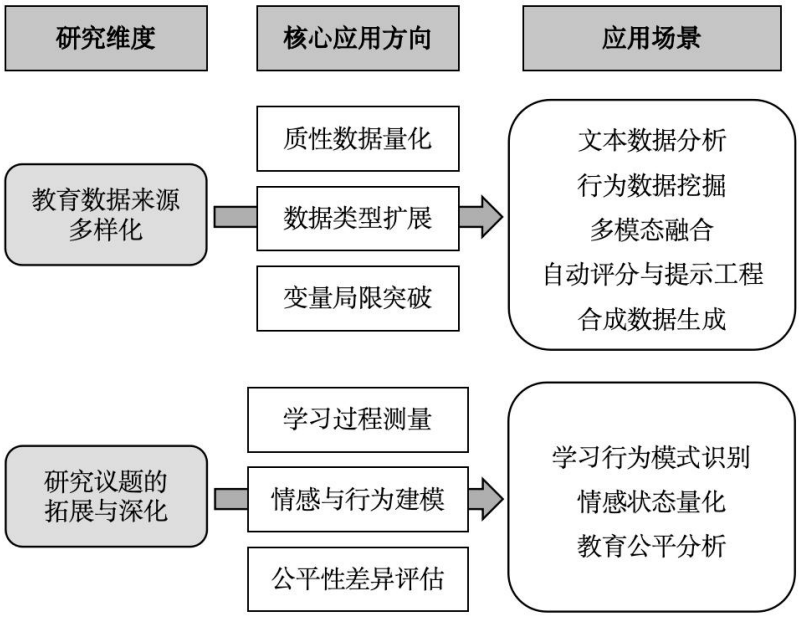

图 4-1 大语言模型对教育量化研究的拓展

（1）教育数据来源的多样化

教育量化研究正在经历数据来源的多样化扩展。这种扩展从传统的单一结构化数据源向多元化、多模态数据转变。这种转变的核心在于大语言模型技术使原本无法量化的教育过程数据变得可测量、可分析，打破了量化研究长期受限于标准化工具的桎梏，为更全面地理解教育现象提供了数据基础。下文将从三个层面展开论述：首先，分析传统量化研究在数据来源上的结构性局限；其次，阐述大语言模型技术如何为突破这些局限提供技术支撑；最后，具体说明扩展后的数据类型及其量化转化机制。

传统教育量化研究的数据局限主要体现在数据覆盖面的不足，利用标准化测验、李克特量表问卷和控制实验等收集方式进行数据收集。根据 Peláez-Sánchez 等人[1]对高等教育中大语言模型应用的系统性分析，当前教育研究中量化方法的使用比例为 27.71%，其中样

[1] Peláez-Sánchez I C, Velarde-Camaqui D, Glasserman-Morales L D. The impact of large language models on higher education: exploring the connection between AI and Education 4.0[C]//Frontiers in Education. Frontiers Media SA, 2024, 9: 1392091.

本量超过250人的大规模研究占38.89%，这一统计数据反映了量化研究在教育领域的重要地位，同时也揭示了一个关键问题：即使是这些大样本研究，仍主要局限于预设变量的数据收集模式。传统量化工具受制于问卷条目的有限性和测验内容的标准化要求，可能存在将复杂多维的教育现象压缩为离散的数值评分，从而使学习过程中的情境依赖性变化、认知策略的动态调整、情感状态的微妙波动、社会互动的复杂模式等关键信息被排除在分析框架之外。

大语言模型技术的成熟为克服这些传统限制提供了技术基础。以短答案自动评分为例，大语言模型与教师评分达到中等一致性，适合作为评分辅助以提升效率[1]。研究表明，非结构化的答案文本可按预设评分维度映射为可计量指标，从而减少人工逐条评阅的工作量，并缓解传统评阅流程的人力瓶颈。Xing等人[2]的研究提到，通过采用链式思维等高级提示策略，大语言模型在处理教育调查反馈分析任务时能够达到人类专家水平的性能，并能够简化教育量化反馈的分析。

基于这种技术能力，教育量化研究的数据来源扩展到三个重要类别。第一类是文本类数据，包括学习者的反思日志、在线学习社区的讨论帖子、课堂教学的对话转录、学习支持的咨询记录等自然语言材料，这些数据现在可以通过情感分析、主题建模、语义相似度计算等算法转化为量化指标。第二类是行为轨迹数据，主要指学习管理系统自动记录的学习者点击序列、学习路径选择偏好、教育资源访问模式、任务完成时间分布等时间序列信息，这些数据可以通过序列模式挖掘、聚类分析等算法转化为学习策略类型、认知负荷水平等量化变量。第三类是多模态数据，涵盖教学录像中的师生面部表情变化、语音交流的韵律特征、学习材料的视觉复杂度等异

[1] Grévisse C. LLM-based automatic short answer grading in undergraduate medical education[J]. BMC Medical Education, 2024, 24(1): 1060.

[2] Xing W, Nixon N, Crossley S, et al. The Use of Large Language Models in Education[J]. International Journal of Artificial Intelligence in Education, 2025, 35(2): 439-443.

质信息，这些数据能够通过跨模态深度学习算法整合为综合性的教学质量评估指标。Tang 等人[1]的实证研究证明，利用大语言模型生成的合成数据进行模型训练后，命名实体识别任务的 F1 分数从 23.37%提升至 63.99%，关系抽取任务的 F1 分数从 75.86%跃升至 83.59%。

数据来源的这种多样化拓展对教育量化研究产生了深远影响。它不仅大幅增加了可用于分析的数据类型和数量，更重要的是改变了研究者理解和测量教育现象的方式，使量化研究能够更接近教育实践的真实复杂性，为构建更加精准和全面的教育理论提供了数据支撑。

（2）研究议题的拓展与深化

大语言模型技术的应用正在重新定义教育量化研究的议题边界，使研究者能够测量和分析以往难以有效处理的复杂教育现象。传统教育量化研究主要受限于数据分析方式与处理能力，对于多模态、非结构化的教育数据，如学习过程、情感变化与交互行为，往往面临高昂的时间成本和复杂的编码难度。因此，研究多集中于易于量化的显性指标，如标准化测试成绩、问卷量表评分、参与率统计等，这些变量虽具备良好的统计特性，却难以全面反映教育活动的动态性与复杂性。

大语言模型技术为突破这些局限提供了新的可能。通过文本分析、情感识别和模式挖掘等算法，大语言模型将大量原本“不可测量”的教育现象转化为可进行统计分析的量化变量。例如，学习者的反思日志可以转化为可量化的认知策略指标，课堂对话记录可以量化为师生互动模式，在线学习行为轨迹可以生成时间序列数据。这种量化能力的扩展使量化研究的议题范围得以拓展，本部分将重点阐述三个新兴量化研究议题：学习行为模式识别、情感状态量化

[1] Tang R, Han X, Jiang X, et al. Does synthetic data generation of LLMs help clinical text mining?[EB/OL]. (2023-03-08)[2025-10-21]. http://arxiv.org/abs/2303.04360

分析和教育公平差异评估。

1）学习行为模式识别

传统量化研究难以捕捉学习过程中的策略选择、认知参与和学习持续性等动态行为特征，而大语言模型技术使这些隐性行为得以显性化测量。研究者利用大语言模型分析学习者的文本表达（如讨论帖子、反思日志、问答记录），将其转化为学习策略偏好、认知负荷水平、自主学习能力等可量化指标，并通过聚类分析等统计方法识别不同学习者群体的行为模式差异。例如，研究者可以量化分析深度学习策略与浅层学习策略在不同学习阶段的分布特征，检验学习策略类型与学习成果之间的关系。现有研究表明，大语言模型通过提示工程能够在情感和行为分析任务中提供非侵入式和可扩展的解决方案[1]。这一议题使量化研究从测量学习结果拓展到测量学习过程，为理解学习的动态性提供了量化证据。

2）情感状态量化分析

情感状态量化分析为量化研究开辟了新的议题空间。传统量化研究主要依赖李克特量表测量学习者的情感状态，但这种方法只能获得离散的、回溯性的情感数据，无法捕捉学习过程中情感的动态变化及其与认知活动的实时交互。大语言模型技术使研究者能够对学习者的自然语言表达进行连续性情感分析，将学习动机、自我效能感、学习焦虑等心理变量转化为时间序列的量化指标。研究者可以构建情感状态与学习效果之间的回归模型，检验情感波动对认知表现的影响，或通过多层次模型分析个体内部情感变化轨迹与个体间差异的交互效应。Ruwe 和 Mayweg-Paus 的实证研究发现[2]，大语言模型在反馈提供任务中的表现接近人类专家水平，能够有效识别和量化学习者的情感状态变化。此外，陈佳雯等人[3]基于大语言模

[1] Tanaka K, Tan B, Wong B. Leveraging language models for emotion and behavior analysis in education[EB/OL]. (2024-08-13)[2025-10-21]. http://arxiv.org/abs/2408.06874

[2] Ruwe T, Mayweg-Paus E. Embracing LLM Feedback: the role of feedback providers and provider information for feedback effectiveness[C]//Frontiers in Education. Frontiers Media SA, 2024, 9: 1461362.

[3] 陈佳雯,褚乐阳,潘香霖,等.共享调节中的群体情感感知工具开发与应用——基于大语言模型技术框架[J].

型开发的 SenGAware 群体情感感知工具，进一步验证了大语言模型在教育情感研究中的可行性。研究通过对协作文本进行情感识别与趋势建模，实现了群体情绪变化的连续量化分析。上述研究表明大语言模型的引入使教育研究者得以从静态测量转向动态追踪，从个体感受转向群体交互，推动了情感状态从质性描述走向量化建模，拓展了教育量化研究的边界。

3）教育公平差异评估

传统量化研究主要通过比较不同群体的平均成绩或入学率来评估教育公平性,但这种粗粒度的测量难以揭示教育不公平的深层机制和隐蔽模式。大语言模型技术使研究者能够对教育过程中的微观互动进行量化分析,从而识别更细致的群体差异。研究者可以量化分析不同社会经济背景、性别、种族学生在教育资源获取、课堂互动参与、教师反馈质量等方面的差异模式,并通过方差分析、回归分析等方法检验这些差异的统计显著性及其对学习成果的影响。例如,通过分析教师对不同群体学生的反馈文本,研究者可以量化教师反馈的鼓励性、具体性和建设性程度,检验是否存在系统性的群体差异。Yan 等人[1]的系统性综述指出，当前大语言模型教育应用研究中存在公平性评估不足的问题，只有九项研究公布了不同样本群体的描述性数据，研究发现通过抽样策略平衡人口统计学分布可以改善模型公平性和准确性[2]。

这些新兴议题的拓展，改变了教育量化研究的测量边界和分析能力。从测量范围看，量化研究从传统的少数几个预设变量扩展到从非结构化数据中提取的多维特征变量；从分析深度看，统计建模从简单的线性关系检验发展到动态过程建模和多因素交互效应分析。随着大语言模型技术的普及，这些新兴议题有望推动量化研究在教

远程教育杂志,2024,42(03):79-92.

[1] Yan L, Sha L, Zhao L, et al. Practical and ethical challenges of large language models in education: A systematic scoping review[J]. British Journal of Educational Technology, 2024, 55(1): 90-112.

[2] Yan L, Sha L, Zhao L, et al. Practical and ethical challenges of large language models in education: A systematic scoping review[J]. British Journal of Educational Technology, 2024, 55(1): 90-112.

育领域发挥更大作用。

综合而言，大语言模型技术正在从数据处理能力、研究议题范围等维度重新定义教育量化研究的理论基础，这种变革为理解其在具体研究环节中的操作化应用奠定了基础。

4.1.2 大语言模型的主要应用方式

大语言模型技术的引入进一步拓展了量化研究的边界。如图 4-2 所示，典型的应用主要体现在四个核心环节：从文本中提取结构化定量数据、生成合成定量数据、辅助数据分析、结果解读与呈现。

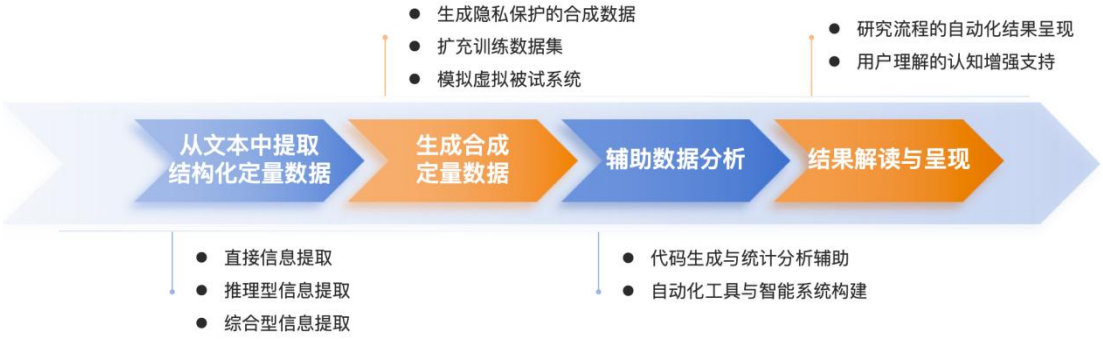

图 4-2 大语言模型在教育量化研究中的主要应用

（1）从文本中提取结构化定量数据

大语言模型的核心优势之一就是能够大规模、自动化地将这些文本转化为可用于计量分析的定量数据。例如，Vijayan 提出的使用对话式大语言模型进行结构化数据提取的提示工程方法，能够将自然语言书写的非结构化文本转换为可存储在数据库中并使用 SQL 等数据库访问语言查询的结构化信息[1]。这种转化能力在处理速度、准确性和成本控制方面展现出明显优势，为量化研究开辟了新的数据获取途径。从技术实现的复杂程度来看，大语言模型的文本数据提取能力可以分为直接信息提取、推理型信息提取和综合型信息提取三个层次。

首先是直接信息提取，即从文本中识别和提取明确记录的数值

[1] Vijayan A. A prompt engineering approach for structured data extraction from unstructured text using conversational llms[C]//Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence. 2023: 183-189.

或分类信息，主要处理文本中显性呈现的结构化要素。例如，Mackay 等人[1]使用基于共识的大语言模型集成方法从术中超声心动图报告的非结构化文本中提取结构化数据，实现了高共识准确率和低错误率。Peng 等人[2]研究了使用领域无关的通用预训练大语言模型从农业文档中提取结构化数据，通过基于嵌入的检索和大语言模型问答相结合的方法，在最小人工干预下实现了比现有方法更高的准确性。在教育研究中，这一层次的提取可以应用于从学生档案、课程记录等文本中自动提取学生基本信息、成绩数据、课程属性等明确记录的量化指标。

其次是推理型信息提取，要求大语言模型不仅识别文本表面信息，还要基于上下文进行逻辑推断。在科学文献分析中，Rettenberger 等人[3]提出了一种新颖的迭代方法，使用大语言模型自动化从冗长的科学文献中提取结构化信息，该方法首先压缩科学文献并保留密集格式的重要信息，然后检索预定义的属性。Schilling-Wilhelmi 的综述研究表明[4]，大语言模型能够高效地从非结构化文本中提取结构化、可操作的数据，为非专家用户提供了强大的工具。Dagdelen 等人[5]展示了使用 GPT-3 和 Llama-2 等大语言模型从科学文本中提取结构化科学知识的简单且可获得的方法，在材料科学任务中表现出优异性能。在教育研究中，推理型信息提取可以用于从学习日志和作业反馈中推断学生的学习策略类型、从开放式评教问卷中识别教学质量的评价维度、从教学观察记录中提取师生互动模式等隐性指标。

[1] MacKay E J, Goldfinger S, Chan T J, et al. Automated structured data extraction from intraoperative echocardiography reports using large language models[J]. British Journal of Anaesthesia, 2025, 134(5): 1308-1317.

[2] Peng R, Liu K, Yang P, et al. Embedding-based retrieval with LLM for effective agriculture information extracting from unstructured data[EB/OL]. (2023-08-07)[2025-10-21]. http://arxiv.org/abs/2308.03107

[3] Rettenberger L, Münker M F, Schutera M, et al. Using Large Language Models for Extracting Structured Information From Scientific Texts[C]//Current Directions in Biomedical Engineering. De Gruyter, 2024, 10(4): 526-529.

[4] Schilling-Wilhelmi M, Ríos-García M, Shabih S, et al. From text to insight: large language models for chemical data extraction[J]. Chemical Society Reviews, 2025.

[5] Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models[J]. Nature communications, 2024, 15(1): 1418.

最后，综合型信息提取是最复杂的层次，需要大语言模型整合多处信息源并进行复杂的关联分析。要求模型跨越文本的不同部分，建立信息之间的关联并进行综合判断。在证据综合研究中，Gartlehner 等人[1]对 Claude 2 的概念验证研究显示，相较人工数据提取，其在 160 个数据元素上的总体准确率为 96.3%，且测试、重测可靠性较高。Konet 等人[2]比较了两个广泛可用的大语言模型（Claude 2 和 GPT-4）在证据综合数据提取中的性能，发现当提供 PDF 中选定的文本时，Claude 2 和 GPT-4 分别准确提取了 98.7%和 100%的数据元素。在教育研究中，这一层次可应用于从多个教学报告中综合提取教学效果的多维指标，从政策文档和实施报告中整合提取政策目标与实施效果的关联数据，从跨学期的学习记录中提取学生成长轨迹的量化特征。

通过上述三个层次的提取能力，大语言模型为教育量化研究提供了强大的数据基础，使研究者能够将传统上难以大规模量化的教育文本数据转化为可供统计分析的结构化数据集，拓展了量化研究的数据来源和分析范围。

（2）生成合成定量数据

大语言模型还可以用于生成合成定量数据，以缓解教育量化研究中因隐私保护、伦理约束或样本稀缺导致的数据获取困难。通过学习已有数据的统计特征，大语言模型能够生成在分布上与真实数据相似、但不包含任何个体信息的合成数据集，从而为研究提供安全、可扩展的数据来源。该方法在量化研究中主要有三类应用场景：一是生成隐私保护的合成数据，用于替代真实敏感数据，降低数据泄露风险；二是扩充训练数据集，在样本有限的情况下提升模型训练与验证的稳健性；三是构建虚拟被试系统，模拟学习者行为与教

[1] Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: A proof-of-concept study[J]. Research synthesis methods, 2024, 15(4): 576-589.

[2] Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis[J]. Research synthesis methods, 2024, 15(5): 818-824.

育干预效果，为定量实验提供可控、可重复的研究条件。

第一类应用是在遵循严格隐私法规的前提下，生成可用于研究的合成数据。例如，Hastings 等人的研究[1]演示了如何利用大语言模型生成包含结构化字段（如诊断代码）与非结构化文本（如临床笔记）的合成电子健康记录（EHR）。这类合成数据的可用性取决于其质量，因此评估框架至关重要。另外，Hämäläinen 等人[2]的研究通过对比真实与合成数据在统计分布、变量相关性等方面的差异，来检验合成的表格化健康记录是否为真实数据的有效统计代理。在教育研究中，这一方法可应用于生成符合隐私保护要求的学生学习数据、成绩记录和行为轨迹数据，使研究者能够在不触及真实学生信息的情况下进行数据分析和模型测试，为涉及敏感教育数据的量化研究提供了可行路径。

第二类应用是数据扩充，即生成数据以扩大机器学习模型的训练集。在心理健康研究中，Kang 等人[3]便利用大语言模型生成了大量模拟抑郁症患者语言风格的社交媒体帖子，在将这些合成数据加入训练集后，其抑郁症分类模型的预测准确率得到了提升。这一技术具有很高的通用性，Li 等人[4]将其应用于生成更多样化的仇恨言论样本，以增强内容审核模型的鲁棒性；而 Modi 等人[5]则通过自动生成大量专业的“问题-答案”数据对，来精炼和优化医疗问答系统。这一方法可用于扩充有限的教育数据集，如生成多样化的学生作业样本、课堂对话记录或学习反馈文本，从而提升教育数据分析模型的泛化能力和预测精度，特别适用于样本量较小或数据分布不均衡

[1] Hastings J D, Weitl-Harms S, Doty J, et al. Utilizing large language models to synthesize product desirability datasets[C]//2024 IEEE International Conference on Big Data (BigData). IEEE, 2024: 5352-5360.

[2] Hämäläinen P, Tavast M, Kunnari A. Evaluating large language models in generating synthetic hci research data: a case study[C]//Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023: 1-19.

[3] Kang A, Chen J Y, Lee-Youngzie Z, et al. Synthetic data generation with LLM for improved depression prediction[EB/OL]. (2024-11-15)[2025-10-21]. http://arxiv.org/abs/2411.17672

[4] Li Y, Bonatti R, Abdali S, et al. Data generation using large language models for text classification: An empirical case study[EB/OL]. (2024-07-22)[2025-10-21]. http://arxiv.org/abs/2407.12813

[5] Modi S, Bokkena B, Avhad P, et al. Automated synthetic data generation pipeline using large language models for enhanced model robustness and fairness in deep learning systems[C]//2024 3rd International Conference for Advancement in Technology (ICONAT). IEEE, 2024: 1-6.

的教育研究场景。

第三类应用是将大语言模型用作虚拟被试模拟器，为量化研究提供计算机模拟路径。这类应用的实现依赖于解决一个核心技术问题：如何让原生于处理文本的大语言模型生成结构化的表格数据。Aygün 等人[1]与 Nacu 等人[2]的研究均探讨了相关技术路径，其核心思路是将表格的每一行序列化为一种统一的文本格式（如 CSV 字符串或键值对），大语言模型通过学习这种格式，便能生成保留了原始数据字段间复杂关系的全新表格行。基于这一技术基础，Singh 等人[3]提出将大语言模型用作“大规模人类被试模拟器”的可能性，为社会科学的假设检验提供了一种全新的计算机模拟（in silico）路径。在此路径下，当获取真实的调查数据成本高昂或不切实际时，研究者可以设定详细的虚拟受访者画像，让大语言模型扮演具有特定人口统计学特征（如年龄、性别、收入）和社会背景的角色，并对模拟的调查问卷做出回答[4]。通过这种方式，研究者可以快速生成大规模的、结构化的数据矩阵，用于探索性因子分析、聚类分析或测试初步的理论模型，为后续的真人数据收集提供参考。

通过上述三类应用场景，大语言模型为量化研究提供了突破隐私限制和样本稀缺的技术路径，使研究者能够在数据受限的情况下开展量化分析和模型构建。

（3）辅助数据分析

大语言模型还具备辅助数据分析的能力，为教育量化研究提供了新的技术路径。通过自然语言交互，研究者可以利用大语言模型实现代码生成、统计分析与结果解释等任务，从而显著降低数据分

[1] Aygün İ, Mehmet K. Use of large language models for medical synthetic data generation in mental illness[C]//7th IET Smart Cities Symposium (SCS 2023). IET, 2023, 2023: 652-656.
[2] Nadas M, Diosan L, Tomescu A. Synthetic data generation using large language models: Advances in text and code[EB/OL]. (2025-03-24)[2025-10-21]. http://arxiv.org/abs/2503.14023
[3] Singh M, Challagundla J, Karnal P, et al. LLMs as Master Forgers: Generating Synthetic Time Series Data for Manufacturing[C]//2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2024: 2053-2059.
[4] Anthis J R, Liu R, Richardson S M, et al. LLM social simulations are a promising research method[EB/OL]. (2025-04-03)[2025-10-21]. http://arxiv.org/abs/2504.02234

析的技术门槛，提升研究效率与可操作性。大语言模型在辅助数据分析中的应用主要体现在两个层面：代码生成与统计分析辅助、自动化工具与智能系统构建。

在代码生成与统计分析辅助层面，Prandner 等人的研究[1]以 ChatGPT-3.5 为对象，尝试让其复现已发表的量化分析过程，并生成用于 SPSS 的语法代码。研究表明，ChatGPT 能够提出合适的分析方案并产出代码，但其输出往往停留在一般性层面，部分语法存在错误，需要研究者进行校正。Selby 等人[2]则探讨了大语言模型在统计推断中的新角色。他提出利用大语言模型辅助贝叶斯建模中的先验分布设定，以及缺失数据的智能填补。通过在医疗和环境数据集上的实验，研究发现大语言模型能够生成具有“专家经验特征”的先验分布，并提供可行的缺失值补全方案，从而在一定程度上减轻了研究者的数据准备负担。但研究也提醒，若缺乏严格的验证机制，大语言模型生成的先验可能带来隐性偏差，对模型推断的可靠性构成风险。

在自动化分析工具和智能系统构建方面，Jansen 等人[3]推出了基于大语言模型的 R 包 mergen 工具。该工具允许研究者通过自然语言直接描述数据分析需求，系统便会调用大语言模型生成相应的 R 代码并执行分析流程。为提升可执行性，mergen 结合了提示工程与错误反馈机制，使得大语言模型在复杂数据分析任务中能够持续改进代码质量。实验结果显示，该工具在常规任务上表现良好，但在高复杂度、多步骤分析中仍存在执行失败的风险。这一成果凸显了“人机交互优化”在教育和科研数据分析工具开发中的关键地位。Ma 等人[4]提出的 InsightPilot 系统则从数据探索角度切入。该系统利

[1] Prandner D, Wetzelhütter D, Hese S. ChatGPT as a data analyst: an exploratory study on AI-supported quantitative data analysis in empirical research[C]//Frontiers in Education. Frontiers Media SA, 2025, 9: 1417900.

[2] Selby D, Iwashita Y, Spriestersbach K, et al. Had enough of experts? quantitative knowledge retrieval from large language models[J]. Stat, 2025, 14(2): e70054.

[3] Jansen J A, Manukyan A, Al Khoury N, et al. Leveraging large language models for data analysis automation[J]. PloS one, 2025, 20(2): e0317084.

[4] Ma P, Ding R, Wang S, et al. Demonstration of InsightPilot: An LLM-empowered automated data exploration

用大语言模型自动识别研究者的分析意图，并将其转化为具体的查询步骤（IQueries），通过与现有的数据探索引擎对接，实现了由自然语言驱动的数据探索链条。案例研究表明，InsightPilot 能够帮助非专业用户快速获取数据模式和趋势，降低了探索性分析的难度，但同时也需要用户具备一定的任务管理与语境判断能力，以避免分析结果的片面化。通过代码生成辅助和智能系统构建，大语言模型为量化研究降低了技术门槛，提升了研究效率。

综上所述，现有研究表明，大语言模型已能够在量化研究中承担从方法设计、统计建模到代码生成、数据探索等多层次的辅助作用。这些探索不仅涵盖了研究流程的关键环节，如质性资料的量化转化、贝叶斯推断中的参数设定，以及自动化生成可执行的分析语法，还延伸至跨学科领域的数据处理与工具开发。与此同时，不同学科的实证研究进一步验证了大语言模型的可操作性，显示其在提升研究效率、降低技术门槛和丰富结果解释方面具有显著价值，为量化研究的开展提供了新的资源与方法支持。

（4）结果解读与呈现

大语言模型能够支持量化研究结果的解读与呈现，以其自然语言处理与生成能力，为研究成果的理解与传播提供了新的可能。通过自然语言处理能力，大语言模型能够将量化结果转换为易于理解的叙述，并支持交互式探索。基于现有研究，大语言模型在结果解读与呈现中的应用主要体现在两个方向：研究流程的自动化结果呈现和用户理解的认知增强支持。

首先，在自动化结果呈现方面，大语言模型能够将复杂统计结果整合为连贯文本，并支持面向不同受众的格式转换。例如，Khan 等人[1]开发的 MetaLLMReporter 系统展示了元分析报告的全流程自动化能力。该系统不仅执行标准的元分析程序，能够将异质性评估、

system[EB/OL]. (2023-04-01)[2025-10-21]. http://arxiv.org/abs/2304.00477

[1] khan, khan M, Rzayev N, et al. MetaLLMReporter: An R Shiny App Integrating Meta-Analysis Execution with LLM-Assisted Reporting[J]. F1000Research, 2025, 14: 724.

敏感性分析、发表偏倚检验等复杂统计结果自动整合为连贯的文本报告，并支持 Cochrane、NEJM、Lancet 等多种期刊格式以及面向公众的通俗语言版本。与此类似，Reason 等人[1]在网络元分析领域的研究显示，大语言模型在数据提取方面达到超过 99%的准确率，能够自动生成完整可执行的 R 代码，并对分析结果进行统计学和临床意义的准确解释。在数据探索层面，Ma 等人[2]的 InsightPilot 系统实现了基于自然语言交互的自动化分析，系统能够智能选择分析策略，并以对话形式实时呈现数据洞察，使非专业用户能够通过简单的语言询问获得深入的数据理解。

其次，在认知增强支持方面，大语言模型通过降低理解门槛、创新呈现方式和交互式解释等手段，帮助不同背景用户深入理解量化结果。Choe 等人[3]的研究发现，大语言模型能够通过文本和视觉双重交互模式为数据素养较低的用户提供图表理解支持，有效回答具体的图表相关问题并引导深入探索，从而降低数据可视化的理解门槛。在呈现方式创新上，Strömel 等人[4]的研究展示了数据叙事化的价值，将量化的健身指标转换为定性的故事描述，显著增强了用户的反思性参与和关注度，使数据呈现更具人文色彩并促进行为改变。Weitl-Harms 等人[5]则探索了隐性信息的显性化呈现，通过分析用户体验数据中的细微表达，大语言模型能够提取并量化隐性的情感倾向，提供包含置信度的情感评分，使原本难以量化的主观体验得以科学化呈现。从理论层面看，Singh 等人[6]提出大语言模型重新

[1] Reason T, Benbow E, Langham J, et al. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models[J]. PharmacoEconomics-Open, 2024, 8(2): 205-220.

[2] Ma P, Ding R, Wang S, et al. Demonstration of InsightPilot: An LLM-empowered automated data exploration system[EB/OL]. (2023-04-01)[2025-10-21]. http://arxiv.org/abs/2304.00477

[3] Choe K, Lee C, Lee S, et al. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation[J]. IEEE Transactions on Visualization and Computer Graphics, 2024.

[4] Strömel K R, Henry S, Johansson T, et al. Narrating fitness: Leveraging large language models for reflective fitness tracker data interpretation[C]//Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 2024: 1-16.

[5] Weitl-Harms S, Hastings J D, Lum J. Using LLMs to establish implicit user sentiment of software desirability[C]//2024 International Conference on Machine Learning and Applications (ICMLA). IEEE, 2024: 1645-1650.

[6] Singh C, Inala J P, Galley M, et al. Rethinking interpretability in the era of large language models[EB/OL]. (2024-02-03)[2025-10-21]. http://arxiv.org/abs/2402.01761

定义了可解释性的范畴和方法。传统的机器学习解释主要局限于特征重要性和模型行为，而大语言模型的自然语言解释能力使得可解释性扩展到更大规模和更高复杂度的模式识别，同时支持交互式解释，用户可以根据自己的理解水平和关注点主动询问，获得个性化的深度解释。

总体而言，大语言模型正在重塑量化研究的知识传播方式，从静态呈现转向动态交互。然而，过度依赖可能削弱独立思考能力，解释的准确性仍需关注。未来需在技术便利性与批判性思维培养之间寻求平衡，使大语言模型成为促进理解而非替代思考的工具。

4.2 典型应用场景

在大语言模型的辅助下，教育量化研究正从传统的宏观统计分析，转向对“教”与“学”全过程的精细化、智能化洞察。这一转变的核心在于利用大语言模型的技术优势，围绕教育活动中的关键环节构建起数据驱动的应用范式。这些应用不仅致力于科学衡量教学工具本身的效能，也深入探究学习者的动态过程，并最终旨在通过数据预测来优化学习成果。下文将分别对这三大典型应用场景进行介绍，分别是（1）教学能力评估与测量，（2）多模态学习分析与评估，以及（3）教育数据挖掘与学生表现预测。

4.2.1 教学能力评估与测量

随着大语言模型在教育领域的广泛应用，其教学能力的科学评估已成为确保 AI 教育质量的关键环节。传统评估体系依赖主观判断和定性分析，难以满足大规模 AI 教学应用的客观化、标准化需求。针对大语言模型教学能力评估，本节从四个方面展开，分别是：多智能体评估架构的构建，标准化评估工具的开发，教学计划的量化评估和反馈效果的实证验证。

首先，在评估架构方面，多智能体系统为突破传统单一评估模

式提供了新路径。Maurya 等人[1]构建的 EducationQ 框架通过三元智能体系统重构了传统师生评估模式，该架构包含教师智能体、学生智能体和评估者智能体，在 GPQA 和 MMLU-Pro 数据集上的评估一致性较单一大语言模型直接评估提升 23%。这一架构的创新之处在于将静态的能力测试转化为动态的交互评估，为教育量化研究提供了更精确的测量工具。基于学习科学理论，研究者[2]建立了涵盖错误识别、错误定位、答案控制、指导提供、行动导向、逻辑连贯、语调适宜和自然表达的八维度量化评估体系。该体系在 1,498 个跨 13 学科问题的大规模测试中显示，各维度的平均准确率分布在 45%-74%之间，其中错误识别维度的 60%准确率远低于专业教师的 85%水平，揭示了显著的能力差距。此外，非正式形成性评估（IFAs）的数字化实现解决了传统课堂评估的即时性难题，研究团队记录了每轮师生互动的时间戳、响应长度和策略变化，构建了包含 1,596 个评估样本的动态交互数据库，将传统 20 分钟的课堂观察压缩至 2 分钟的自动化评估[3]。

其次，在评估工具方面，标准化基准的建立使跨研究对比成为可能。MRBench 数据集作为专门针对 AI 导师能力的综合评估基准，通过整合 Bridge 和 MathDial 两个公开数据源，形成了包含 192 个对话场景的标准测试集。每个场景记录学生错误类型、教师干预策略和学习结果变化。在该基准上，GPT-4 的综合得分为 67.3%，Llama-3.1-405B 为 58.9%，Claude-3.5-Sonnet 为 61.2%，显示出明显的模型性能差异[4]。MRBench 采用期望注释匹配率（DAMR）在 8 个教学

[1] Shi Y, Liang R, Xu Y. EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework[EB/OL]. (2025-04-22)[2025-10-21]. http://arxiv.org/abs/2504.14928

[2] Maurya K K, Srivatsa K V, Petukhova K, et al. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors[EB/OL]. (2024-12-20)[2025-10-21]. http://arxiv.org/abs/2412.09416

[3] Hikal B, Basem M, Oshallah I, et al. MSA at BEA 2025 Shared Task: Disagreement-Aware Instruction Tuning for Multi-Dimensional Evaluation of LLMs as Math Tutors[EB/OL]. (2025-05-31)[2025-10-21]. http://arxiv.org/abs/2505.18549

[4] Maurya K K, Srivatsa K V, Petukhova K, et al. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors[EB/OL]. (2024-12-20)[2025-10-21]. http://arxiv.org/abs/2412.09416

维度上评估教师回应质量，并以人类新手、专家与多种大语言模型为对照给出分维度基准。同样，MathTutorBench 通过奖励模型的偏好胜率量化教师回应的脚手架质量，用以比较不同模型的教学能力[1]。introEduAI 平台集成了先进的自然语言处理技术，实现了编程作业的自动评分功能。对比分析表明，该平台与人工评分整体一致性较高，且在代码类题目上的一致性普遍优于开放性问题[2]。

再次，在教学计划评估方面，基于教学内容知识（PCK）理论的评估框架提供了多维度测量标准。基于教学内容知识（PCK）理论的量化评估框架显示，GPT-4 生成的高中数学教学计划在目标设定、内容组织和方法选择三个维度的得分分别为 8.2、7.9 和 7.4 分（10 分制），与人类教师的 8.7、8.3 和 8.6 分相比差距在可接受范围内[3]。教学仿真增强方法通过模拟课堂互动提升计划质量，另有研究对比了 240 个原始计划与经过仿真优化的计划，发现优化后的计划在教学逻辑、问题设计和师生互动三个维度分别提升 1.3、1.7 和 2.1 分，统计检验显示这些改进达到显著水平（$p<0.001$），证实了量化研究方法在评估教学改进效果方面的有效性[4]。LessonPlanner 系统的量化实验表明，使用该系统的教师准备时间减少 42%（从 3.2 小时降至 1.9 小时），而计划质量评分仅下降 5%[5]。医学教育领域的专业化量化评估研究显示，GPT-4 在医学知识问答上准确率达 78%，但在临床诊断推理任务中降至 54%，这种显著的量化差异为构建学科特异性评估框架提供了实证依据，强调了教育量化研究在不同专业领域中的适应性需求[6]。

[1] Macina J, Daheim N, Hakimi I, et al. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors[EB/OL]. (2025-02-28)[2025-10-21]. http://arxiv.org/abs/2502.18940
[2] Mendonça P C, Quintal F, Mendonça F. Evaluating LLMs for automated scoring in formative assessments[J]. Applied Sciences, 2025, 15(5): 2787.
[3] Hu B, Zheng L, Zhu J, et al. Teaching plan generation and evaluation with GPT-4: Unleashing the potential of LLM in instructional design[J]. IEEE Transactions on Learning Technologies, 2024, 17: 1445-1459.
[4] Hu B, Zhu J, Pei Y, et al. Exploring the potential of LLM to enhance teaching plans through teaching simulation[J]. npj Science of Learning, 2025, 10(1): 7.
[5] Fan H, Chen G, Wang X, et al. LessonPlanner: Assisting novice teachers to prepare pedagogy-driven lesson plans with large language models[C]//Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. 2024: 1-20.
[6] Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions[J]. JMIR medical education, 2023, 9(1): e48291.

最后，在反馈效果验证方面，多项实证研究证明了大语言模型生成反馈的有效性。Meyer 等人[1]基于 459 名德国高中英语学习者的随机对照实验，显示接受 GPT-3.5-turbo 生成个性化反馈的实验组在文本质量提升幅度上比标准反馈组高出 1.7 分（10 分制，$p<0.05$），为大语言模型反馈效果提供了客观的量化证据。德国职前教师的诊断推理量化研究同样验证了这一结论，269 名参与者的数据显示，ChatGPT 生成的自适应反馈在证实质量维度显著优于静态反馈，效应量为 d=0.34，接受大语言模型反馈的教师平均处理时间延长 18%，但后续写作长度增加 31%，表明更深入的任务参与和学习效果[2]。自动化评分系统的量化可靠性验证研究覆盖了计算机编程课程的多种题型，LLaMA 3.2 与人类评分的皮尔逊相关系数为 0.84，GPT-4o 达到 0.89，在代码调试题目上，两个模型的评分准确率分别为 76.3%和 82.7%，证明了自动评分的可靠性[3]。BART 模型在学生项目报告评估任务中的量化表现显示，使用 80%训练数据时 ROUGE-L 指标得分 0.73，BERT-Score 达 0.81，训练数据减至 20%时相应指标降至 0.61 和 0.72，这一结果强调了高质量标注数据的重要性[4]。

通过对大语言模型辅助教学能力评估体系的全面分析，教育量化研究已在架构设计、工具开发和效果验证三个层面形成了相对完整的研究框架，但仍面临评估可靠性、标准化程度和长期追踪等方面的挑战。例如，评估可靠性受限于人类标注质量，平均一致性仅为 82%，Prometheus2 等大语言模型评估器与人类判断的相关性仅 0.67，远低于教育测量的 0.85 标准要求[5]。标准化程度不足导致跨

[1] Meyer J, Jansen T, Schiller R, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions[J]. Computers and Education: Artificial Intelligence, 2024, 6: 100199.

[2] Kinder A, Briese F J, Jacobs M, et al. Effects of adaptive feedback generated by a large language model: A case study in teacher education[J]. Computers and Education: Artificial Intelligence, 2025, 8: 100349.

[3] Mendonça P C, Quintal F, Mendonça F. Evaluating LLMs for automated scoring in formative assessments[J]. Applied Sciences, 2025, 15(5): 2787.

[4] Du H, Jia Q, Gehringer E, et al. Harnessing large language models to auto-evaluate the student project reports[J]. Computers and Education: Artificial Intelligence, 2024, 7: 100268.

[5] Maurya K K, Srivatsa K V, Petukhova K, et al. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors[EB/OL]. (2024-12-20)[2025-10-21]. http://arxiv.org/abs/2412.09416

研究比较困难，长期教学效果追踪机制的缺失限制了因果关系的建立。这些挑战表明，大语言模型教学能力的量化评估仍需在理论深化、工具优化和验证完善等方面持续发展，以更好地服务于AI教育应用的科学化推进。

4.2.2 多模态学习分析与评估

多模态学习分析与评估(Multimodal Learning Analytics, MMLA)作为教育量化研究的重要应用，通过整合来自不同感知通道的数据源来全面理解学习过程[1]。随着高频传感器技术的成熟和大语言模型能力的突破，该领域正经历着从传统单一数据源分析向智能化多模态数据融合的重要转变[2]。大语言模型在多模态教育数据处理方面展现出独特优势，能够将眼动追踪数据、学习成果、评估内容和教学标准等多样化数据源转化为教师可理解的分析报告。这种技术融合解决了传统多模态数据分析中语义解释薄弱、难以直接解释学习过程的关键问题。当前研究表明，大语言模型可以有效充当教育分析师角色，将复杂的多维数据转换为教师友好的洞察，这些洞察受到教育工作者的广泛认可[3]。多模态学习分析的核心价值在于突破传统基于结果的评估局限，通过捕获学习者行为、认知和情感等多维度信息，提供对学习过程更准确和全面的理解。研究显示，多模态数据的分析虽然提供了学习的整体图景，但其固有复杂性使得理解和解释变得困难，而大语言模型技术的引入为解决这一挑战提供了新的技术路径[4]。

在具体应用方面，大语言模型辅助多模态学习分析与评估的应用主要集中在学习行为深度分析、教学过程量化评估和学习成果智

[1] Ochoa X, Worsley M. Editorial: Augmenting learning analytics with multimodal sensory data[J]. Journal of Learning Analytics, 2016, 3(2): 213-219.
[2] HARVARD UNIVERSITY. LIT LAB. Multimodal Learning Analytics[EB/OL]. [2025-09-06]. https://lit.gse.harvard.edu/multimodal-learning-analytics.
[3] Davalos E, Zhang Y, Srivastava N, et al. LLMs as Educational Analysts: Transforming Multimodal Data Traces into Actionable Reading Assessment Reports[C]//International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2025: 191-204.
[4] Mu S, Cui M, Huang X. Multimodal data fusion in learning analytics: A systematic review[J]. Sensors, 2020, 20(23): 6856.

能预测三个核心领域。在学习行为分析方面，现有研究通过整合多种生理和行为数据源，实现了对学习过程的精细量化。Davalos 等人[1]开发了基于注视点聚类方法结合大语言模型驱动的教育分析师系统，该系统整合了眼动追踪数据、学习成果、评估内容和教学标准，通过无监督学习技术识别不同的阅读行为模式，然后由大语言模型将衍生信息综合成针对教育工作者的可操作报告。该研究利用无监督聚类技术识别有意义的学生群体，同时大语言模型将原始分析数据转换为可解释和可操作的洞察。协作学习模式的多模态分析同样验证了该技术的有效性。Chen 等人[2]的研究应用多模态学习分析收集了 19 对本科生的多模态过程和产品数据，收集了学生在结对编程过程中的语言音频、计算机屏幕录制和面部表情录制等多模态数据。研究利用 K-means 聚类进行协作聚类检测，并使用定量内容分析、点击流分析和视频分析来分析学生的言语交流、操作行为和面部表情维度，最终揭示了四种协作模式：共识达成模式、论证驱动模式、个体导向模式和试错模式。

在教学过程量化评估领域，大语言模型技术展现出对教学能力进行多维度自动化评估的潜力。Yang 等人[3]开发的 EducationQ 框架通过多智能体对话框架高效评估教学能力，测试了来自主要 AI 组织的 14 个大语言模型在 1,498 个问题上的表现，涵盖 13 个学科和 10 个难度级别。该研究的重要发现是教学效果与模型规模或一般推理能力不呈线性相关，一些较小的开源模型在教学环境中表现优于较大的商业模型。研究采用混合方法评估，结合定量指标与定性分析和专家案例研究，识别出顶级表现的模型采用的不同教学策略，如复杂的提问策略和适应性反馈机制。教育内容理解的多模态评估也

[1] Davalos E, Zhang Y, Srivastava N, et al. LLMs as Educational Analysts: Transforming Multimodal Data Traces into Actionable Reading Assessment Reports[C]//International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2025: 191-204.

[2] Xu W, Wu Y, Ouyang F. Multimodal learning analytics of collaborative patterns during pair programming in higher education[J]. International Journal of Educational Technology in Higher Education, 2023, 20(1): 8.

[3] Shi Y, Liang R, Xu Y. EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework[EB/OL]. (2025-04-22)[2025-10-21]. http://arxiv.org/abs/2504.14928

扩展了应用边界。Jain 等人[1]首次评估了最先进的多模态大语言模型在 CK12-QA 数据集上的表现，该数据集包含超过 26,000 个问题，来源于 CK-12 基金会的开放教育资源。研究采用多模态检索增强生成策略，将每个问题与两个上下文输入配对：从 CK12-QA 语料库检索的图表和课程文本段落，以评估现代 MLLM 回答基于包含文本和图表的教育内容的问题的能力。另外，学习成果预测作为该领域的重要应用方向，通过多模态数据融合实现了对学习表现的精准预测。Tsai 等人[2]的研究采用创新的基于问题的学习教学法，结合学习分析来识别数字分心、量化同伴学习参与的质量并预测学习表现。该研究涉及 51 名台湾研究生在基于问题学习教学法下的混合统计课程，多模态学习分析模型包含来自 Facebook 群组学习者话语和问卷的数据，包括学习者特征、感知的数字分心、主观同伴学习取向以及由机器学习模型认可的客观同伴学习参与。研究结果显示，报告更多数字分心问题的学生获得了较低的期末课程成绩，而报告更强同伴学习取向的学生获得了较高的期末课程成绩，更重要的是，由机器学习模型客观识别的同伴学习参与在学术表现预测方面比自我感知的同伴学习取向具有更好的预测有效性。

总而言之，大语言模型辅助多模态学习分析与评估呈现出技术融合性强、应用场景多元化、验证方法科学化的显著特征。在技术层面，该领域实现了从传统的手动数据处理向智能化自动分析的重要转变，大语言模型能够有效综合多模态学习分析数据，生成结构化的教师友好报告，教育工作者对此反应积极，特别是对学生聚类洞察和内容分析的认可。随着智能技术的进步，自动机器标注相比手动标注和自我报告标注受到越来越多的关注，因为后两者对于大规模自动分析来说过于耗时和主观。在应用方面，该领域实现了从

[1] Alawwad H A, Zafar A, Alhothali A, et al. Evaluating Multimodal Large Language Models on Educational Textbook Question Answering[EB/OL]. (2025-06-25)[2025-10-21]. http://arxiv.org/abs/2506.21596.

[2] Liao C H, Wu J Y. Deploying multimodal learning analytics models to explore the impact of digital distraction and peer learning on student performance[J]. Computers & Education, 2022, 190: 104599.

单一维度评估向多维度综合分析的转变，应用范围从课堂环境扩展到专门的训练场景，作为评估工具和实时教育干预的机制[1]。该领域的核心价值体现在显著提高了教育评估的精度和教学决策支持的有效性，已经能够处理包括眼动追踪、语音分析、视频识别、文本挖掘等多种数据类型的融合分析，为教育研究提供了前所未有的数据洞察深度。

4.2.3 教育数据挖掘与学生表现预测

教育数据挖掘与学生表现预测专注于从大规模教育数据中发现潜在模式，并构建数学模型对学生的学习成果进行量化预测。该领域的核心在于将学生的学习行为、答题记录、学习轨迹等数据转化为可量化的特征变量，通过统计建模和机器学习方法建立预测函数，实现对学生成绩、知识掌握水平、学习风险等指标的数值化预测。传统的教育数据挖掘通常依赖结构化数据与手工特征工程，而大语言模型的引入拓展了量化研究的建模范式。大语言模型能够从非结构化教育文本中自动提取量化特征，将复杂语义信息转化为高维向量，从而构建更精准的预测模型。当前大语言模型辅助教育数据挖掘与学生表现预测的量化研究可分为四个主要类别，分别是量化知识状态追踪与建模、多维学习表现的综合量化预测、自动化量化评分与成绩预测、特定领域的量化建模。

（1）量化知识状态追踪与建模

该方向旨在将学生对知识点的掌握程度转化为连续数值变量，并利用数学模型刻画其动态变化过程。研究者通常通过分析学生的答题序列，估计其对不同知识概念的掌握概率，从而预测未来表现。Neshaei 等人[2]提出将学生的答题历史表示为自然语言序列，利用大语言模型计算每个答案的条件概率，构建基于概率分布的量化知识

[1] Cohn C, Davalos E, Vatral C, et al. Multimodal methods for analyzing learning and training environments: A systematic literature review[EB/OL]. (2024-08-28)[2025-10-21]. http://arxiv.org/abs/2408.14491

[2] Neshaei S P, Davis R L, Hazimeh A, et al. Towards modeling learner performance with large language models[EB/OL]. (2024-03-25)[2025-10-21]. http://arxiv.org/abs/2403.14661

状态模型。实验在三个基准数据集上验证了该方法在AUC、准确率、F1分数等量化指标上的有效性，微调后的BERT模型在知识追踪任务中达到了0.75的AUC值。Wang等人[1]针对编程学习提出DPKT模型，利用大语言模型量化编程问题的文本理解难度分数（0-1标准化）和知识概念难度分数，将这些连续数值作为特征输入到神经网络中，实现对编程学习状态的精确量化追踪。这些研究表明，基于大语言模型的知识追踪模型能够更准确地量化学生学习动态，为个性化教学提供支持。

（2）多维学习表现的综合量化预测

多维学习表现的综合量化预测专注于整合学生的多源数据（成绩记录、行为日志、课程内容等），构建回归或分类模型对学习成果进行数值化预测。该研究的关键在于如何将异构数据统一量化并建立有效的预测函数。Oh等人[2]提出LMgMF框架，通过大语言模型将学生的文本信息转化为512维的语义向量，然后利用矩阵分解算法构建学生-课程交互的量化预测模型，在i-ScreamEdu数据集（包含1.6M交互记录）上实现了85.3%的预测准确率。Zhang等人[3]在成人识字研究中开发了基于GPT-4的量化预测方法，将学习行为数据编码为数值特征向量，建立回归模型预测学习表现分数，与传统方法相比在RMSE指标上提升了12.4%。

（3）自动化量化评分与成绩预测

该方向聚焦利用模型将学生的开放式回答自动转化为标准化分数，从而实现客观、可扩展的量化评估。研究通常通过回归或分类算法建立“文本特征-分数”的映射关系。Yilmaz等人[4]通过微调

[1] Yang L, Sun X, Li H, et al. Difficulty aware programming knowledge tracing via large language models[J]. Scientific Reports, 2025, 15(1): 11475.

[2] Oh C, Park M, Lim S, et al. Language model-guided student performance prediction with multimodal auxiliary information[J]. Expert Systems with Applications, 2024, 250: 123960.

[3] Zhang L, Lin J, Borchers C, et al. Predicting learning performance with large language models: a study in adult literacy[C]//International Conference on Human-Computer Interaction. Cham: Springer Nature Switzerland, 2024: 333-353.

[4] Latif E, Zhai X. Fine-tuning ChatGPT for automatic scoring[J]. Computers and Education: Artificial Intelligence, 2024, 6: 100210.

GPT-3.5-turbo 建立自动评分的量化模型，将学生的科学问答转化为 0-4 分的离散分数或 0-100 分的连续分数，在多个评分任务中实现了 0.82 的皮尔逊相关系数，显著优于传统 BERT 模型的 0.71 相关系数。Flodén 等人[1]进行的大规模量化分析通过 463 个样本的统计检验，发现 ChatGPT 评分与人工评分的相关系数为 0.76，标准差为 8.3 分，为大语言模型自动评分的量化可靠性提供了统计学证据。

（4）特定领域的量化建模

该方向针对不同学科的特殊性质，开发领域适配的量化指标和预测模型。不同学科在能力测量、难度评估、学习路径等方面需要专门的量化方法。在编程教育中，研究者利用大语言模型量化代码复杂度、算法时间复杂度、编程概念难度等学科特定指标，构建编程能力的多维量化评估模型[2]。在数学教育中，大语言模型被用于量化数学问题的认知负荷指数、推理步骤复杂度、概念关联度等指标，为数学学习进度的精确测量提供量化工具。

总而言之，大语言模型辅助的教育数据挖掘研究展现出突出的量化特征与实证优势。技术层面上，大语言模型实现了从非结构化文本到结构化特征的自动转换，突破了传统量化研究的特征提取瓶颈；性能层面上，其预测精度在 AUC、RMSE 等指标上普遍优于传统模型；应用层面上，量化研究已从成绩预测扩展至知识追踪、学习路径规划与风险预警等场景，展现出广泛的教育适应性与推广潜力。

4.3 混合研究方法的应用

现如今，许多研究已不再局限于单纯的量化研究或质性研究，而是越来越多地采用混合研究方法，通过整合定量和定性数据收集

[1] Flodén J. Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT[J]. British educational research journal, 2025, 51(1): 201-224.

[2] Bernik A, Radošević D, Čep A. A Comparative Study of Large Language Models in Programming Education: Accuracy, Efficiency, and Feedback in Student Assignment Grading[J]. Applied Sciences, 2025, 15(18): 10055.

与分析技术，以获得对研究问题更全面、更深入的洞察。混合方法研究结合定性和定量方法，提供规模和深度，即可测量的模式以及丰富的背景信息[1]。然而，传统混合研究在实践中面临着数据整合深度不足、分析流程相对固化、验证机制主观性较强等挑战。而大语言模型辅助的混合研究可以系统性地整合大语言模型的数据处理、分析和整合能力，实现质性与量化数据在语义层面深度融合，并通过人机协作优化研究流程和提升验证质量。这种方法论创新不仅保持了传统混合研究的核心优势，更通过技术赋能实现了研究深度和效率的双重提升。这种技术赋能不仅仅是工具层面的简单应用，而是深刻地重塑了混合研究的理论基础和实践模式。通过对现有理论探讨的系统分析，大语言模型辅助的混合研究主要呈现出三大特征：整合深度化、流程动态化与验证智能化。

首先，在整合深度方面，大语言模型推动了质性与量化数据的语义融合。传统混合研究主要在解释和报告阶段进行数据整合，整合在研究设计层次、方法层次、解释和报告层次发生，质性和量化数据往往分别分析后再进行比较和综合[2]。大语言模型辅助混合研究则将整合前移至数据处理和分析的早期阶段，实现了真正意义上的方法融合。大语言模型强大的自然语言理解能力使其能够识别和连接不同数据类型之间的语义关系，自动建立质性文本中的概念与量化变量之间的对应关系。例如，研究者可以要求大语言模型确定样本文本中是否存在特定编码，并要求提供支持该编码的证据，这种能力使得质性描述向量化指标的转换，以及统计模式向叙述性解释的转化成为可能[3]。在不同的研究实践中，大语言模型的整合潜力已得到充分体现。以周春红[4]关于教师教研的混合研究为例，研

[1] Bryman A. Mixed methods research; combining qualitative and quantitative research[J]. Social Research Methods, 2012: 627-651.

[2] Fetters M D, Curry L A, Creswell J W. Achieving integration in mixed methods designs—principles and practices[J]. Health services research, 2013, 48(6pt2): 2134-2156.

[3] Tai R H, Bentley L R, Xia X, et al. An examination of the use of large language models to aid analysis of textual data[J]. International Journal of Qualitative Methods, 2024, 23: 16094069241231168.

[4] 周春红. 教师教研中的共享任务理解研究[D].华东师范大学,2025.

究团队引入大语言模型，对教师协作过程中的访谈与反思文本进行自动编码与证据提取，构建了质性文本概念与量化变量之间的语义映射。该研究建立了“量化测量-质性验证-智能辅助”的分析体系，不仅提升了非结构化文本的处理效率，也为量化路径模型提供了可解释的语义支撑，展现出数据整合前移与方法协同融合的实践路径。大语言模型还可以跨越方法边界识别共同的主题、模式和概念，为后续分析提供统一的概念框架。这种深度整合推动混合研究从简单的方法叠加转向真正的协同效应。

其次，在研究流程层面，大语言模型使混合研究具备动态性与自适应性。传统混合研究通常采用预设固定的研究设计，研究者在研究开始时就确定具体的设计类型，整个执行过程相对刚性。相比之下，大语言模型辅助混合研究支持响应式的动态调整，能够根据数据发现实时优化研究策略。量化和质性分析应该在生成式 AI 中分别进行，然后再请求聊天机器人帮助处理混合方法结果，这种分离后整合的方式使得研究设计能够基于初步发现进行自适应调整[1]。研究者可以根据一个阶段的数据发现动态决定后续采用何种研究策略和数据收集重点。研究工作流程保留了传统的编码手册开发和完整循环，但增加了一个迭代步骤来调整机器理解的定义，通过人机协作实现研究过程的持续优化[2]。大语言模型还能基于数据模式分析为研究者提供下一步行动的智能建议，使混合研究变得更加灵活和响应性。

最后，在研究验证环节，大语言模型推动了混合研究的智能化与客观化。混合研究的验证一直是方法论难点，传统方式主要依靠研究者的主观判断进行三角验证，存在人为偏见和一致性检验不充分的问题。大语言模型辅助混合研究发展出了算法辅助的智能验证

[1] Combrinck C. A tutorial for integrating generative AI in mixed methods data analysis[J]. Discover Education, 2024, 3(1): 116.

[2] Dunivin Z O. Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive content analysis[J]. EPJ Data Science, 2025, 14(1): 28.

机制，显著提升了验证的客观性和系统性。例如，研究者可以将样本文本输入160次以记录大语言模型响应迭代之间的变化，通过重复运行来检测分析结果的稳定性和可靠性[1]。大语言模型能够系统性地比较不同方法获得的发现，识别一致性、互补性和矛盾性，为研究者提供客观的验证报告。更进一步，大语言模型可以基于一种方法的发现预测另一种方法可能的结果，通过预测准确性来评估数据整合的质量。这种验证智能化不仅提高了混合研究结果的可靠性和客观性，还为建立更加严谨的混合研究质量标准提供了技术支撑。

总体来看，大语言模型辅助的混合研究在理论与实践两个层面都具有重要价值。在理论层面，整合深度化、流程动态化与验证智能化共同推动混合研究由机械组合走向有机融合，拓展了其认识论与方法论基础；在实践层面，大语言模型显著提升了研究处理复杂社会现象的能力，使研究者能够在更大规模的数据基础上获得更深入的洞察。然而，这一新范式仍面临幻觉与偏见问题、领域知识局限，以及可能削弱质性研究“情境化理解”的风险[2]。未来研究需要在保持学术严谨性的同时，建立适用于大语言模型辅助混合研究的质量评估标准，并关注数据隐私与研究伦理等新议题。只有在技术与方法论的平衡中，大语言模型才能真正推动混合研究走向更加成熟和高效的发展阶段。

[1] Tai R H, Bentley L R, Xia X, et al. An examination of the use of large language models to aid analysis of textual data[J]. International Journal of Qualitative Methods, 2024, 23: 16094069241231168.

[2] Schroeder H, Aubin Le Quéré M, Randazzo C, et al. Large language models in qualitative research: Can we do the data justice?[EB/OL]. (2024-10-11)[2025-10-21]. http://arxiv.org/abs/2410.07362

第 5 章 研究质量与标准的重构

AI4S 对既有的科研质量评估、学术诚信及知识产权归属等核心规范构成了深刻挑战。一方面，AI 在“知识发现”中的贡献正获得学术共同体的认可。例如，谷歌 DeepMind 的科学家因其在蛋白质结构预测领域的 AI 驱动性突破而荣获 2024 年诺贝尔化学奖。另一方面，在“成果署名”环节，AI 的合法性与规范性却充满争议，众多权威学术期刊明确拒绝生成式人工智能作为论文作者。这种在贡献认定与署名资格之间的价值割裂，凸显了技术发展与现行学术评价与诚信体系之间的张力。

因此，建立人机协同研究的学术价值评估标准、清晰界定贡献归属，并据此更新学术诚信标准，已成为维护学术共同体严谨性的重要议题。传统的同行评议体系在面对此类新型成果的原创性与可靠性评估时，面临着前所未有的压力。

5.1 学术作品中 AIGC 的兴起

人工智能在研究中的应用已成为当代伦理探讨的重要议题。研究人员可将人工智能作为提升数据分析效率、拓展研究思路的辅助工具，但必须对所有 AI 生成内容的输出进行系统性、批判性的审查、多源验证与修订，并以第一责任人身份承担全部研究结果的学术责任与伦理义务。

5.1.1 人工智能辅助的学术写作

（1）人工智能在研究中的应用潜力

在核心科学发现层面，人工智能展现出超越人类认知局限的潜力。通过对海量文献与实验数据的深度分析，人工智能能够揭示出人类研究者可能忽略的潜在关联，从而生成全新的科学假说。例如，在生物学中，预测性 AI 工具实现了对蛋白质功能与细胞类型的自动注释[1]。在社会科学领域，生成式 AI 则被用于处理传统上需要巨大

[1] MESSERI L, CROCKETT M J. Artificial intelligence and illusions of understanding in scientific research[J]. Nature, 2024, 627(8002): 49-58.

劳动力的数据（如文本、图像）的标注与释义工作。在定性研究中，大语言模型可以分析教育研究中的非结构化文本，借助思维链的方式接近人类的表现并简化教育定性反馈的分析[1]。在提升科研生产力与学术交流效率方面，AI被广泛用作高效的辅助工具，其功能涵盖了学术写作中的语法校订、语言润色、文献格式规范、内容摘要及多语言翻译等。这对于非英语母语的研究者尤为重要，能够显著提升其学术论文的语言流畅度与专业性，使其更符合国际学术规范。Covington等人评估大语言模型在生成本科神经科学课程作业答案的准确性时，发现其表现出色[2]。Norberg等人测试大语言模型在提升数学应用题可读性上的能力，发现GPT-4可以帮助提高可读性指标，例如词频、句子复杂度和语义相似性[3]。

在教育与人才培养领域，AI正引发对传统教学法的深刻反思与革新。正如联合国教科文组织（UNESCO）所倡导的，AI可扮演“个性化导师”“协作学习伙伴”乃至“苏格拉底式提问者”等多元角色，旨在帮助教师实现过去难以达成的个性化与深度学习目标[4]。在高等教育中，大语言模型可以用于提高学生参与度、促进小组活动、创建交互式学习工具以及提供即时反馈和评估。例如，在大学个性化教育实践中，AI可以通过智能导学系统构建“测—学—练—评—辅”全流程闭环，结合学习者画像技术实现精准化学习支持与动态化路径调整，使个性化学习收益提升[5]。Abdelghani等人利用大语言模型自动生成激发好奇心的线索，以激励儿童提出更多、

[1] PARKER M J, ANDERSON C, STONE C, et al. A large language model approach to educational survey feedback analysis[J]. International Journal of Artificial Intelligence in Education, 2025, 35(2): 444-481.

[2] COVINGTON N V, VRUWINK O. ChatGPT in undergraduate education: Performance of GPT-3.5 and identification of AI-generated text in introductory neuroscience[J]. International Journal of Artificial Intelligence in Education, 2025, 35(2): 627-650.

[3] NORBERG K A, ALMOUBAYYED H, DE LEY L, et al. Rewriting content with GPT-4 to support emerging readers in adaptive mathematics software[J]. International Journal of Artificial Intelligence in Education, 2025, 35(2): 587-626.

[4] MCDONALD N, Aditya Johri, Areej Ali, et al. Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines[J]. Computers in Human Behavior: Artificial Humans, 2025, 3: 100121.

[5] 乔思辉, 眭依凡. 数智时代大学的个性化教育：价值理路、潜在挑战与变革策略[J]. 江苏高教, 2025, (06): 78-84.

更深入的问题[1]。Goslen 等人探索使用大语言模型在基于游戏的学习环境中自动生成计划，以支持学生的自主学习，发现利用大语言模型生成的计划有效且具多样化[2]。Dijkstra 等人使用大语言模型来生成测验和抽认卡等交互式教育材料，这些材料减轻了教师手动测验设计的负担，提高了学生的学习和参与度[3]。Bernius 等人使用机器学习为大型课程中学生的文本答案生成反馈，评分工作量减少 85%，准确率高，学生感知到的教学质量也得到提升[4]。

（2）人工智能幻觉现象的挑战

人工智能基于统计概率运行，本质上是文本生成器。最典型的“幻觉”现象会导致虚假文献的泛滥，当被要求提供文献来源时，人工智能并非检索学术数据库，而是生成看似合理的引文。模型可能将真实存在的学者、真实的出版社与虚构书名组合成一个看似可信但完全不存在的参考文献。

无论使用者意图如何，提交一篇包含“幻觉”文献的论文本身就构成了学术诚信违规，因为并未进行真实的研究。Haman 等人的研究显示，使用 ChatGPT 生成的参考文献中 66%为编造内容[5]。即使存在的参考文献也缺乏完整性和相关性，增加了研究者的验证负担[6]。这也使得不少学者认为，人工智能推动学术欺诈手段的升级，因其应用范围涵盖了生成完整论文、作业到编写代码、论坛帖子等研究涉及的方方面面。同时，为了规避传统的抄袭检测软件，人工智能能够执行复杂的释义和改写。这种能力使得高质量的伪造作品

[1] Abdelghani R，Wang Y H，Yuan X，et al. GPT-3-driven pedagogical agents for training children’s curious question-asking skills［J］. International Journal of Artificial Intelligence in Education，2024，34（2）：483-518.

[2] GOSLEN A, KIM Y J, ROWE J, et al. LLM-based student plan generation for adaptive scaffolding in game-based learning environments[J]. International Journal of Artificial Intelligence in Education, 2025, 35(2): 533-558.

[3] DIJKSTRA R, GENÇ Z, KAYAL S, et al. Reading comprehension quiz generation using generative pre-trained transformers[J]. 2022.

[4] Bernius J P，Krusche S，Bruegge B. Machine learning based feedback on textual student answers in large courses［J］. Computers and Education: Artificial Intelligence，2022，3：100081.

[5] Haman M, Školník M. Using ChatGPT to conduct a literature review[J]. Accountability in Research, 2024.

[6] Franzoni Velázquez A L, Huerta E, Jensen S. Retracting ChatGPT: Completeness and relevance of academic references[J]. Discover Education, 2024, 3(1): 226.

比过去更容易获得，远超传统的“论文工厂”模式。部分工具还通过模拟人类常见的错误等方式进一步混淆视听，例如添加拼写错误或改变写作风格，从而增加了检测难度。具体而言，AI 通过动态调整生成策略（如消除规律性特征、模拟人类写作风格变异），显著降低检测工具识别准确率，加剧学术欺诈的隐蔽性[1]。

研究表明，AI 生成文本还通过逻辑结构程式化与语义关联机械化（如过度使用连接词、时空错位信息拼接），使现有检测工具误判风险增加，同时对抗性技术（如同义词替换）可轻易破解文本水印，形成检测技术的动态博弈[2]。李启正等学者将这种持续对抗界定为 AI 文本检测中的“猫鼠游戏”，AI 通过版本迭代与文本伪装不断突破检测防线，而检测工具对升级后的规避策略响应滞后，同时人类自身对 AI 文本的识别能力局限进一步放大了检测困境[3]。

（3）人工智能领域的可重复性危机与科学可靠性挑战

《Science》期刊指出，人工智能领域面临严峻的“可重复性危机”，许多关键研究因缺乏公开的源代码而难以复现，这与过去十年困扰心理学、医学等领域的危机类似，从根本上动摇了科学知识的可靠性根基[4]。王阳等学者进一步指出，人工智能的“可重复性危机”具有独特的技术与认识论特征：模型迭代的动态性（如高版本模型输出差异）与算法黑箱的叠加效应，使基于特定版本的研究结果难以复现；数据私有性与生成式 AI 的情境依赖性，则突破了传统封闭系统假设，导致直接重复失效，亟需转向理论层面的概念重复验证[5]。

人工智能技术特别是大语言模型已经成为“论文工厂”的强大工具，使其能够以工业化规模生产伪造论文。目前 AI 伪造研究是一

[1] 沈锡宾, 王立磊. 人工智能生成学术期刊文本的检测研究[J]. 科技与出版, 2023, (08): 56-62.
[2] 王建磊, 解玲玲. AI 痕迹与数字灵韵：看待人工智能生成内容的另一种视角[J]. 福建师范大学学报(哲学社会科学版), 2025, (04): 98-107+171.
[3] 李启正, 胡崴琳, 祝成炎. 摘要 AI 文本检测中“猫鼠游戏”的行为界定和能力分析[J]. 情报杂志, 2024, 43 (11): 139-143+138.
[4] Hutson M. Artificial intelligence faces reproducibility crisis[J]. Science, 2018.
[5] 王阳, 何进彬. 人工智能时代疏解可重复性危机的认识论路径新探[J]. 自然辩证法研究, 2025, 41 (07): 82-89.

个突出的问题。以教育场景为例，教师在区分学生原创作品与AI生成的内容方面面临挑战：英国《时代》杂志在对剑桥大学400名学生的调查中显示，近一半的学生曾经使用ChatGPT等聊天机器人来帮助完成学位课程[1]。《福布斯》杂志称，三分之二的美国学生每周多次使用人工智能完成学业[2]。这些研究凸显了人工智能技术对学术诚信的影响，强调了在教育中使用ChatGPT时需要谨慎监管，生成式人工智能工具能够模仿学术写作风格，重写现有内容，甚至生成看似合理但完全虚假的文本、数据和图像。

值得警惕的是，AIGC技术还通过“洗稿”式重述规避查重系统，或全程替代学生完成论文创作，形成“技术性剽窃”与“AI代写”的新型学术不端[3]。更严峻的是，作者通过剔除AI特征词主动规避检测，而期刊同行评审系统未能有效识别，导致含AI生成内容的论文经多轮审核后仍得以发表，暴露出学术出版体系对技术伦理风险的应对滞后。

目前，学术撤稿数量因AI滥用急剧增加。2023年《Nature》研究指出，学术论文撤稿数量已超过1万篇，沙特阿拉伯、巴基斯坦、俄罗斯和中国在过去二十年中撤稿率最高，撤稿数量增长的速度远超科学论文数量的增长速度[4]。部分撤稿源于未声明AI辅助撰写，例如《Physica Scripta》杂志发表的一篇旨在揭示一个复杂数学方程新解的论文，在手稿中却出现ChatGPT指令，故而被撤稿[5]。这不仅表明了作者未按规定披露使用AI辅助学术论文，而且暴露了期刊对于论文的审核存在问题。《Springer Nature》在2025年撤回了一本关于机器学习相关的书籍，原因是书籍许多章节末尾的引文存在

[1] Hennessey L S | M. Almost half of cambridge students admit they have used ChatGPT[EB/OL]. (2023-04-21)[2025-08-25]. https://www.thetimes.com/business-money/technology/article/cambridge-university-students-chatgpt-ai-degree-2023-rnsv7mw7z.

[2] Carter S. Should kids use ChatGPT for school? Experts weigh in[EB/OL]. [2025-08-25]. https://www.forbes.com/sites/digital-assets/2025/08/06/should-kids-use-chatgpt-ai-for-school-parents-are-divided/.

[3] 王霞, 龚向和. AIGC介入学位论文之学术不端的定性与制度因应[J]. 中国高教研究, 2025, (06): 84-92.

[4] Van Noorden R. More than 10,000 research papers were retracted in 2023 — a new record[J]. Nature, 2023, 624(7992): 479-481.

[5] Conroy G. Scientific sleuths spot dishonest ChatGPT use in papers[J]. Nature, 2023.

严重错误，引用了并不存在的著作或书籍章节。这一事件表明，即便是知名出版商，其编辑和审查流程也可能存在严重漏洞，无法有效阻止此类低劣内容的出版[1]。《Neurosurgical Review》撤回了超百篇学术论文，原因是怀疑论文的真实性，其中很多论文来自于同一大学[2]。

单纯的人工智能原始输出，无论形式是精炼的文本、复杂的数据分析结果，还是新颖的科学假说，其本身都不能被视为完整的“研究成果”，因其始终缺失人类研究者所赋予的意图、批判性验证、科学解释和最终的责任承担。在这些情况下，许多研究者认为，人工智能的产出并未创造新的知识或思想，仅优化知识的表达形式，不具备原创性。因此，他们认为，开发并应用有效的人工智能内容检测工具，不仅是技术上的需求，更是维护学术诚信、防止未经审查的衍生内容泛滥、保障知识生产严谨性的必要举措。

5.1.2 人工智能检测工具的伦理风险

传统学术评估的前提是，学术成果要能够反映其背后的智力活动过程。然而，人工智能通过生成一个看似合理却脱离学生真实知识体系的产物，切断了这一联系。学生并未实际参与研究过程，教师评估的不再是学生的研究与批判性思维能力，这迫使教师重新设计评价体系，合理检测作业内容的生成。

在此背景下，人工智能内容检测工具曾被视为一种技术性的解决方案。然而，大量实证研究表明，主流 AI 抄袭检测工具（如 Turnitin、Originality.AI、GPTZero、DetectGPT 等）在技术层面存在根本性缺陷。到目前为止，这些工具的误报率（将人类写作标记为人工智能生成）和漏报率（未能识别出人工智能生成的内容）都非

[1] Aksenfeld R. Springer nature to retract machine learning book following retraction watch coverage[EB/OL]. (2025-07-16)[2025-08-23]. https://retractionwatch.com/2025/07/16/springer-nature-to-retract-machine-learning-book-following-retraction-watch-coverage/.

[2] Orrall A. As springer nature journal clears AI papers, one university’s retractions rise drastically[EB/OL]. (2025-02-10)[2025-08-23]. https://retractionwatch.com/2025/02/10/as-springer-nature-journal-clears-ai-papers-one-universitys-retractions-rise-drastically/.

常高。Turnitin 公开承认其系统存在显著误差，并调整了其准确率声明，承认为降低误报率而放过了一定比例的人工智能写作。Fergus 等人通过 Turnitin 学术不端检测软件对 ChatGPT 输出的化学答案进行检测，发现 ChatGPT 生成内容中伪造的参考文献，并没有被 Turnitin 检测出[1]。ChatGPT 可以通过重写问题的答案逃避 Turnitin 对生成答案的相似性检查：许多学生可以多次重新生成问题、提示或案例的答案，而不会被检测出抄袭[2]。在一项研究中，由 ChatGPT 生成的摘要被提交给了审稿人，而审稿人只发现其中 68% 的造假行为[3]。同样，如果期刊或出版商使用未经验证的人工智能文本检测器，并错误地指责作者使用人工智能生成文本，可能会给作者带来困扰和伤害，因为假阳性结果可能会损害作者的声誉，并对他们未来发表产生影响[4]。此外，现有的 AIGC 检测器在区分人工编写的代码和 AI 生成的代码方面也表现不佳[5]。

检测工具的伦理争议另一核心问题是其内嵌的算法偏见。斯坦福大学一项研究证实，人工智能检测工具对非英语母语者的写作存在偏见，会持续地将其作品错误归类为人工智能生成[6]。这种偏见的技术根源在于，检测器通常使用文本的“困惑度”，即词汇选择的可预测性，作为关键判断指标。非英语母语者的写作中词汇多样性较低、句法结构相对简单，导致其文本困惑度较低，这与人工智能生成文本的统计模式相似。有研究者进一步揭示，AI 系统可能通过生成过程自我强化算法歧视——训练数据中的历史偏见经模型迭

[1] Fergus S, Botha M, Ostovar M. Evaluating academic answers generated using ChatGPT[J]. Journal of Chemical Education, 2023, 100(4): 1672-1675.

[2] AlAfnan M A, Dishari S, Jovic M, et al. ChatGPT as an educational tool: Opportunities, challenges, and recommendations for communication, business writing, and composition courses[J]. Journal of Artificial Intelligence and Technology, 2023, 3(2): 60-68.

[3] Thorp H H. ChatGPT is fun, but not an author[J]. Science, 2023, 379(6630): 313-313.

[4] Bahammam A S, Trabelsi K, Pandi-Perumal S R, et al. Adapting to the impact of artificial intelligence in scientific writing: Balancing benefits and drawbacks while developing policies and regulations[J]. Journal of Nature and Science of Medicine, 2023, 6(3): 152.

[5] Pan W H, Chok M J, Jonathan Leong Shan Wong, et al. Assessing AI detectors in identifying AI-generated code: Implications for education | proceedings of the 46th international conference on software engineering: software engineering education and training[EB/OL]. 2024[2025-09-06]. https://dl.acm.org/doi/10.1145/3639474.3640068.

[6] Liang W，Yuksekgonul M，Mao Y，et al. GPT detectors are biased against non-native English writers［J］. Patterns，2023，4（7）.

代被放大，而现有审查标准缺乏对生成内容偏见的动态监测，导致检测工具本身成为歧视的载体[1]。当学生知道他们的论文将被一个算法审查时，他们可能会变得更加保守和公式化，从而抑制了批判性思维和创造性写作能力的发展。

AI 检测工具的失效，给人工智能辅助研究的质量判断建立了模糊地带，急需构建质量标准以对人工智能相关的研究进行评价和判断。这一转变要求我们的评估体系也随之进化。如果仅仅审视最终的论文文本，我们将无法区分一项研究是源于研究者深刻的洞察力，还是源于未经审慎思索的人工智能提示。因此，一个健全的评估框架必须能够深入考察其背后的人机协同，这不仅是对新研究范式的承认，更是维护学术严谨性的必然要求。

5.2 AI 辅助研究的质量标准

随着 AI 在科研领域的深度应用，传统的质量标准面临调整和扩展的必要性。主动探讨并建立 AI 辅助研究的质量标准，不仅是顺应技术发展的需要，更是维护学术研究价值的必然要求，对维护整个学术共同体的健康发展至关重要。

5.2.1 AI4S 中的透明度问题

AI 对于研究的介入，不只是引入了高效的辅助工具，而是从根本上引发了学术界对 AI4S 的稳健性与可靠性的深刻关切。这一关切主要源于以下几个相互关联的核心挑战。首先，问题源于现代 AI 模型其固有的“黑箱”特性与现代科学方法论的根本冲突。传统的科学研究方法要求研究过程的透明性和研究结果的可解释性，而 AI 模型的输入和输出都是在“黑箱”中进行操作的。这些模型的结构具有高度的复杂性，其内部包含数以亿计的参数，这些参数以高度非线性的方式进行交互与转换。与传统的基于显式规则的程序不同，这些模型主要依赖于从数据中习得的统计相关性，而非因果机制，

[1] 卜素. 人工智能中的“算法歧视”问题及其审查标准[J]. 山西大学学报(哲学社会科学版), 2019, 42 (04): 124-129.

这导致其推理路径具有天然的模糊性与不可解析性。这种模糊性因建档文件的缺失而进一步加剧。

以模型训练关键的超参数为例，超参数是训练前设置的外部配置变量（例如，学习率、神经网络的层数），它们对模型的性能和泛化能力有至关重要的影响。其选择与调优过程若记录不善，将直接导致研究的不可复现。一项文献对政治科学领域使用机器学习的文献进行审查，发现超过56%的使用机器学习的出版物没有报告其最终的超参数值，只有20%提供了关于值和调优过程的完整说明[1]，这种“调优透明度”匮乏不仅阻碍了研究人员对其学习到的潜在规律或决策依据的深入洞察，使得审稿人和读者无法评估研究结果的稳健性，从而削弱了研究结果的可信度，更对现代科学方法论的奠基性原则之一——卡尔·波普尔提出的可证伪性原则构成了实质性的挑战。

科学方法的精髓在于提出并检验能够被证伪的假设，以探究其背后的因果机制。然而，一个无法解释其决策逻辑的“黑箱”模型，其结论本质上是难以被系统性地证伪的。“黑箱”模型可能通过学习数据中存在的伪相关而非真实的因果联系，来达成极高的预测准确率，但这会导向在科学上毫无根据的结论。

例如，一个用于区分狼和哈士奇犬的图像分类模型，尽管在测试集上表现优异，但后续分析揭示，该模型并非学习到了动物的内在生物学特征，而是将“背景中是否存在雪”作为了关键分类依据，因为其训练数据中几乎所有的狼的图像都碰巧拍摄于雪景环境。因此，该模型实质上沦为了一个高效的“雪景探测器”，其结论对于动物学研究是无效的[2]。此案例深刻表明，AI模型可能通过学习数据中的伪相关而并非因果机制来达到高精度预测。这种“精确的谬

[1] ARNOLD C, BIEDEBACH L, KÜPFER A, et al. The role of hyperparameters in machine learning models and how to tune them[J]. Political Science Research and Methods, 2024, 12(4): 841-848.

[2] BUHRMESTER V, MÜNCH D, ARENS M. Analysis of explainers of black box deep neural networks for computer vision: A survey[J]. Machine Learning and Knowledge Extraction, 2021, 3(4): 966-989.

误”不仅无异于科学知识的准确产出，更因为其看似合理化的信息而迷惑大众，背离科学研究的初衷。

其次，AI模型存在路径依赖和数据污染的问题。AI模型的输出质量高度依赖于输入数据的质量。即便算法完全透明，若其所依赖的底层数据来源不清、记录不善，模型将会放大这些缺陷，产生系统性谬误，其结果同样不可信。2023年麻省理工（MIT）、Cohere for AI等11个机构共同发布了数据溯源倡议（The Data Provenance Initiative）[1]，旨在应对AI领域的数据透明危机。然而，数据溯源倡议对超过1800个数据集的审计发现了系统性问题，在流行的数据托管平台上，许可证遗漏率高达72%以上，错误率高达50%以上[2]。这种溯源信息的缺乏不仅带来了显著的法律和伦理风险，更使得理解和复现基于这些数据的研究变得异常困难。为了解决标准化文档缺失的问题，Gebru等人提出了“数据集的数据表”的概念。类似于电子元件的数据表，该框架要求为每个数据集都配备一份标准化的“说明书”，详细说明其动机、构成、收集过程、推荐用途和潜在的伦理问题[3]。其目的是促进数据创建者和使用者之间的沟通，提高数据生态的透明度和问责制，并减少数据滥用。

人工智能伦理领域就提高研究透明度问题发表了大量文献。人们呼吁增加透明度措施来维护可复制性，通过确保研究过程和人工制品的充分共享来维护研究的科学严谨性和完整性，以方便重新进行研究，验证研究结果的有效性。人工智能产品、参与者和开发过程接受外部审计和监管，将更容易预防、识别和减轻危害，并追究相关方的责任。

在认识到可复制性危机之后，许多领域都发现了原始研究结果与复制研究结果之间的差异。这些问题存在于科学领域以及机器学

[1] Data Provenance Initiative[EB/OL]. [2025-10-12]. https://www.dataprovenance.org/about.
[2] LONGPRE S, MAHARI R, CHEN A, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in AI[J].
[3] GEBRU T, MORGENSTERN J, VECCHIONE B, et al. Datasheets for datasets[A]. arXiv, 2021.

习领域。为了回应这些担忧，DARPA 和 ACM 等机构发起了“开放科学运动”，KDD、NeurIPS、ICLR 和 ICML 等顶级机器学习会议制定了提交和审查准则，旨在从制度层面提高研究成果和假设的透明度[1]。

增加透明度的方法之一是预注册（pre-registration），也就是说在进行实验之前，科学家将假设和数据分析计划在第三方平台进行存档，以防止日后只挑选具有统计显著性的结果进行汇报。这样做可以防止“择优汇报”问题。预注册又可以分为三种，分别是狭义上的预注册、注册报告和注册复制报告[2]。狭义上的预注册是指研究人员应尽可能详细地描述其研究计划，并将这些计划保存在带有时间戳且不可编辑的档案中。此记录可与审稿人、编辑和其他研究人员共享。注册报告是指研究人员在开展研究之前，会向期刊提交一份详细的研究方案。注册复制报告是注册报告的一种变体，侧重于直接重复一项或多项原始研究结果。

预注册原则作为实证社会科学的标准实践，正在迅速获得相当多的支持。最近心理学领域被引用最多的文章之一就是倡导使用预注册的文章[3]。在其他社会科学领域也同样倡导使用预注册。为了激励和奖励预先注册的研究，包括《心理科学》在内的各种期刊现在都以“开放科学徽章”的形式对涉及预先注册研究的已发表文章进行特别表彰。开放科学运动的支持者建议，期刊甚至应该考虑强制要求提交发表的研究进行预注册。

5.2.2 AI 应用于研究引发的偏见

AI 模型中存在的偏见是威胁科学有效性的核心因素之一。这些偏见可能在研究的每个阶段被引入，从而扭曲结果、损害科学的公

[1] Kou T. From model performance to claim: How a change of focus in machine learning replicability can help bridge the responsibility gap[C]//Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, 2024: 1002-1013.

[2] 赵加伟, 夏涛, 胡传鹏. 心理学研究中预注册的现状、挑战与建议[J]. 心理科学进展, 2024, 32(5): 715-727.

[3] Nosek B A, Ebersole C R, DeHaven A C, et al. The preregistration revolution[J]. Proceedings of the National Academy of Sciences, 2018, 115(11): 2600-2606.

正性。有研究深入探讨AI中的各种偏见，基于机器学习的各阶段，将偏见分为历史偏差、表征偏差、测量偏差、聚合偏差、评估偏差和部署偏差[1]。有研究者按照数据收集过程中产生的偏见将偏见分为选择偏差（某些个体更有可能被选中进行研究）、报告偏差（某类观察结果更有可能被报告，从而导致观察结果的某种选择偏差）以及检测偏差（某一现象更有可能被一组特定的研究对象观察到）[2]。根据偏见对现实世界的影响，可将偏见分为选择偏见、确认偏见、测量偏见、刻板印象偏见和外群体同质性偏见[3]。

（1）数据收集与标注阶段的偏见

在多模态数据分析中，很多研究的数据收集往往是有偏误的，指的是某些群体的代表性过高或过低，导致最后的错误率过高。例如，基于三个大型X光数据集利用人工智能分类技术对患者进行分类，发现对于黑人患者、西班牙裔患者的漏诊率更高，原因在于数据集的偏差[4]。在医疗保健领域的计算机辅助诊断系统时，绝大多数模型都是在欧洲血统比例过高的数据集上进行训练的，往往没有考虑到算法的公平性，在TCGA数据集中，来自33种癌症类型的8594份肿瘤样本中，82.0%的病例来自白人患者，10.1%来自黑人或非裔美国人[5]。

类似的问题在文本数据和图像数据也同样突出。一方面，许多研究揭示在公开的单词训练中存在大量与性别和种族攻击性相关的单词，即使是基于谷歌新闻文章训练的词向量，其也呈现出性别刻板印象，词向量的广泛使用往往会放大这些偏见[6]。另一方面，在

[1] Harini Suresh, John Guttag. A framework for understanding sources of harm throughout the machine learning life cycle | proceedings of the 1st ACM conference on equity and access in algorithms, mechanisms, and optimization[EB/OL]. 2021[2025-08-11]. https://dl.acm.org/doi/10.1145/3465416.3483305.

[2] Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems—an introductory survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020, 10(3): e1356.

[3] Bias in AI[EB/OL]. [2025-08-12]. https://www.chapman.edu/ai/bias-in-ai.aspx.

[4] Seyyed-Kalantari L, Zhang H, McDermott M B A, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations[J]. Nature Medicine, 2021, 27(12): 2176-2182.

[5] Chen R J, Wang J J, Williamson D F K, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare[J]. Nature Biomedical Engineering, 2023, 7(6): 719-742.

[6] Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker?

计算机视觉领域中，研究人员普遍依赖所谓的“基准图像数据集”进行模型训练，或直接使用基于这些数据开发的预训练模型作为特征提取器。这种做法的底层假设是：这些数据集是真实世界的无偏抽样。然而，这些图像数据集可能存在偏差。主流的面部分析数据集中，深肤色女性的样本严重不足，直接导致商业化模型在识别该群体时，表现出惊人的高错误率[1]。另外许多数据集的创建并不严格遵循统计原则，缺失值或缺失信息也可能产生偏见，导致数据集不能代表目标人群[2]。

例如，一个教育 App 通过自愿问卷收集用户数据来证明这款 App 能有效提高学生的高中数学成绩，但这导致了严重的“自选择偏见”，因为成绩提高的用户更愿意参与，这个充满缺失值和片面信息的数据集，仅代表一小部分积极用户，而无法代表全体学生。

在数据标注过程中，由于人类标注员的主观解释或不一致而可能会引入标签偏见。例如，在情感分析中，不同文化背景的标注员可能对同一段文本的情感有不同解读。在教育研究的场景中，根据学生的历史虚拟学习环境活动数据和其他数据，生成一个预测模型来预测学生的期末成绩[3]。这样的模型在获得更高精度的同时，也有可能存在偏见。更为隐蔽的是，即使模型在整体上达到较高准确率，也可能对特定人口群体表现出系统性的不公平——例如对某一性别或族裔的学生预测准确率显著偏低，或错误地将某些人口统计学特征（如性别、族裔）本身作为预测失败的依据，从而复制甚至放大社会中已存在的教育不平等。

在一项关于人口统计学在教育数据挖掘领域的作用的调查中，

Debiasing word embeddings[C]//Advances in Neural Information Processing Systems: 卷 29. Curran Associates, Inc., 2016.

[1] Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification[C]//Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR, 2018: 77-91.

[2] Dana Pessach, Erez Shmueli. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings[J]. Expert Systems with Applications, 2021, 185: 115667.

[3] Wolff A, View Profile, Zdrahal Z, et al. Improving retention[M]//Proceedings of the Third International Conference on Learning Analytics and Knowledge. 2013: 145-149.

Paquette 等人发现，大约有一半在分析中包含人口统计学的论文至少使用了一个人口统计学属性作为模型中的预测特征，以提升预测精度，而没有将人口统计学纳入模型测试或验证——即没有检验模型是否在不同性别、族裔或社会经济地位的学生群体中保持一致的预测性能。这意味着这些模型可能在不知情的情况下对某些群体产生歧视性结果，却因缺乏针对性的公平性验证而未被研究者察觉[1]。

（2）算法与模型阶段的偏见

算法偏见产生的原因是算法旨在最大程度地减少总体预测错误，因此使多数群体的结果比少数群体的结果更优。模型设计中的固有假设或开发者无意识中嵌入的偏见会加剧算法偏见。例如，一个过分强调收入或教育水平的模型可能会强化对边缘化群体的有害刻板印象。教育领域以外的研究已经聚焦于算法偏见可能会导致的一些真实的危害。这些危害可以分为分配性和代表性危害[2]。

分配性危害指的是人工智能系统不公正地分配、扣留或拒绝提供有形资源和机会，例如工作、贷款或医疗保健。亚马逊公司曾开发的 AI 招聘工具使用过去十年间收到的简历进行训练，发现系统对包含“女性”一词的简历进行惩罚性降分，并系统性地将女性候选人的排名置于末尾，直接剥夺了女性求职者的工作机会[3]。同样的，谷歌的广告工具被发现向女性提供的高薪工作广告明显少于男性[4]。

代表性危害指的是当算法系统以歪曲、刻板、贬低或抹杀的方式描绘特定社会群体时，所造成的文化和社会层面的损害。在刑事司法系统中被告的审前风险评估会将黑人更高预测为高风险人群[5]。Obermeyer 等人的研究揭示预测医疗资源需求的算法中存在种族偏

[1] Paquette L, Ocumpaugh J, Li Z, et al. Who’s learning? Using demographics in EDM research[J]. Journal of Educational Data Mining, 2020, 12(3): 1-30.

[2] Baker R S, Hawn A. Algorithmic bias in education[J]. International Journal of Artificial Intelligence in Education, 2022, 32(4): 1052-1092.

[3] Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women *[M]//Ethics of Data and Analytics. Auerbach Publications, 2022.

[4] Datta A, Tschantz M C, Datta A. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination[A]. arXiv, 2015.

[5] Angwin J, Larson J, Mattu S, 等. Machine bias *[M]//Ethics of Data and Analytics. Auerbach Publications, 2022.

见，该模型预测黑人患者比白人患者的医疗资源需求更低，因为过去黑人患者群体的医疗支出更低[1]。在谷歌的广告搜索中，黑人身份的名字被广告暗示有逮捕记录的可能性要高出 25%[2]。

在数据驱动的建模过程中，一个关键的公平性挑战源于代理特征（Proxy Features）所引发的隐性偏见。尽管在数据预处理阶段可以移除明确的敏感属性（如种族、性别），但模型仍有可能通过与这些属性存在强统计相关性的非敏感特征（例如，地理位置、经济行为指标）来间接推断它们。这些代理特征充当敏感信息的载体，将潜在的社会结构性偏见编码进模型中。

例如，Ocumpaugh 等人基于城市、郊区和农村学生的数据，对美国东北部中学数学进行分析，构建了四种与教育相关的情感状态模型，结果发现针对主要来自一个人口群体的人群训练的情感检测器无法推广到主要来自其他人口群体的人群，即使这些人群可能被视为同一国家或地区文化的一部分[3]。在另一项研究中，研究者从哥斯达黎加、菲律宾和美国三个国家收集学生数据集，生成寻求帮助行为与学习联系起来的模型，将每个模型在其他国家的数据上进行测试，研究发现，有效求助模型在不同国家之间不能相互转移[4]。也就是说，任何试图构建普适性人工智能教育系统的努力，都应充分考虑到学习行为背后深刻的文化和社会差异。

人工智能决策对某些人群的歧视性影响已经在众多案例中有所体现，可能会随时随地影响到每个人，为不同学科或日常生活中面临的问题提供解决方案，但同时也会带来风险，比如被剥夺工作或医疗机会。例如使用来自学术医疗机构的胸部 X 光数据集训练的卷积神经网络 (CNN) 被证明会对特定患者群体（包括女性、黑人、西

[1] Obermeyer Z, Powers B, Vogeli C, et al.. Dissecting racial bias in an algorithm used to manage the health of populations[J]. Science, 2019.
[2] Sweeney L. Discrimination in online ad delivery[J]. Communications of the ACM, 2013, 56(5): 44-54.
[3] Ocumpaugh J, Baker R, Gowda S, et al. Population validity for educational data mining models: A case study in affect detection[J]. British Journal of Educational Technology, 2014, 45(3): 487-501.
[4] Ogan A, Walker E, Baker R, et al. Towards understanding how to assess help-seeking behavior across cultures[J]. International Journal of Artificial Intelligence in Education, 2015, 25(2): 229-248.

班牙裔和社会经济地位低下的患者）的疾病检测不足[1]。

5.2.3 大语言模型的开源闭源之争

AI 模型的开源和闭源之争一直是人工智能时代科学研究范式面临的争议。在学术研究领域，这一争论聚焦于一个根本性矛盾：开放、透明、可复现的科学价值与数据隐私、技术壁垒和商业安全为主要追求的工业价值之间的冲突。尤其是伴随着 Deepseek、Qwen 等新型开源模型的问世，与闭源模型之间的对比愈发激烈。在开源模型中，又分为完全开源和部分开源，如 Stable Diffusion (Comp Vis 许可证）、BERT(Apache 2.0 许可证）的代码、训练数据和预训练权重均是开源的，而 Llama 2 和 3 (Meta 许可证）、Mistral7B (Apache 2.0 许可证）等模型的代码和预训练权重是开源，而数据是闭源的[2]。开源阵营认为大语言模型的开源可以迅速发现并修复漏洞，保障模型的准确性和机构的安全性，还可以推动技术共享与创新，为验证研究结论提供最直接的路径，快速迭代出更强大的 AI 系统[3]。

相比之下，在闭源模型中，研究者无法窥见其内部运作，这实际上是将研究过程置于“内部黑箱”中操作，使得同行评议和知识验证很难进行，直接动摇了科学研究的知识积累基础。闭源阵营的支持者则主张封闭环境下对企业和模型的保护以及长期发展。其中，知识产权保护是闭源策略的核心考量，开发和训练大语言模型需要巨大且长期的人力物力投资，企业通过闭源以获取回报利益，保障商业可持续化发展。同时，开发者也可以通过闭源实施工具监控、内容过滤和防止数据篡改，进而实现安全控制。

在学术研究领域中，相关的争论聚焦于科学研究的透明性和可重复性。可重复性指科研人员基于原作者提供的原始数据、算法代码及实验参数等核心要素，在相同操作条件下复现已发表研究成果

[1] Gaube S, Suresh H, Raue M, et al. Do as AI say: Susceptibility in deployment of clinical decision-aids[J]. npj Digital Medicine, 2021, 4(1): 31.

[2] 郑晓龙，李家彤．人工智能时代的开源与闭源技术模式探讨［J］．中国科学院院刊，2025，40（3）：459-464.

[3] 郑晓龙,李家彤.人工智能时代的开源与闭源技术模式探讨[J].中国科学院院刊,2025,40(3): 459-464.

的验证能力[1]。有研究者指出可重复性是指独立研究人员按照原始研究人员共享的文件从实验中得出相同结论的能力[2]。可重复性是科学研究的基石，在人工智能研究中，研究者采用多种方法实现可重复性。AI 模型的内在随机性，例如深度学习中常用的随机梯度下降算法，以及模型的高度复杂性和海量参数，使得传统意义上“相同代码、相同数据、产生完全相同的结果”的精确复现变得极具挑战性。

因此，以用于训练模型的实际计算机代码的形式的透明度对于研究可重复性至关重要。在没有代码的情况下，可重复性只能依靠文本描述中的复制方法[3]。然而，文本描述往往受研究者主观因素影响较大，难以涵盖研究的所有关键细节，从而导致复现结果与原结果之间的不匹配和偏差。

开源 AI 模型通过公开算法、代码和相关细节，提供了从研究伊始到研究结果的可验证性的核心路径和技术逻辑支持，帮助研究者深究模型机制，增加了研究的可信度。在教育研究领域，开源 AI 模型为研究者提供了宝贵的认识工具和信息获取样本，这种透明度有助于探究技术工具与学习过程的互动机制，探究 AI 模型与研究方法的相融共创，为教育研究方法和教育理论发展提供实证基础。

虽然通过开源代码和数据复制论文结果的能力很有价值，但也有研究者认为，共享人们实验中的所有东西并不是一件小事，这不仅需要论文作者做大量的额外工作，也需要审稿人做大量的额外工作[4]。也有一些论文作者反对将代码公开化，例如 McKinney 等人指出，模型的训练代码与内部工具、基础设施及硬件深度耦合，导致即使公开也难以复现。此外，公开训练参数也存在泄露数据属性的

[1] 于倩倩, 孟银涛, 钱力, 等. 计算领域的论文数据共享与可重复性问题分析[J]. 图书情报工作, 2024, 68(17): 3-15.

[2] Gundersen O E. The fundamental principles of reproducibility[J]. Philosophical Transactions of the Royal Society A, 2021.

[3] Haibe-Kains B, Adam G A, Hosny A, et al. Transparency and reproducibility in artificial intelligence[J]. Nature, 2020, 586(7829): E14-E16.

[4] Drummond C. Replicability is not reproducibility: Nor is it good science[J]. 2009.

风险，可能引发安全漏洞和攻击[1]。通过严格的产权保护和技术封锁，可以确保企业对技术实行全流程的控制，维护关键技术安全，不断优化模型性能，实现飞轮效应优势，在市场竞争中占据领先地位。

当代围绕大型人工智能模型开源与闭源路径的论争，正是该领域可重复性危机的体现。闭源是保护专有知识和巨额投资的必要手段，因为训练这些模型需要庞大的资金支持。通过 API 控制访问，开发者可以实施安全措施，防止模型被滥用，过滤有害内容，并阻止恶意行为。闭源模式有诸多好处，如性能稳定、提供专业支持、定期更新以及易于集成的用户友好界面，这些都是成熟商业产品所必须的。闭源倡导者提出的“安全”论点，实际上创造了一种与透明度之间的悖论关系。虽然他们主张控制对于安全是必要的，但正是这种控制导致了不透明，而不透明本身就是一个重大的风险。

在科学领域，开源理想意味着完全访问验证所需的所有组件：代码、数据和环境。然而，在 AI 领域，“开放”可能意味着多种情况：带有严格许可的开放权重、开放代码但数据专有等。不同层次的开放性提供了不同程度的可重复性。开放权重允许进行微调和推理分析，但无法复现原始的训练过程。开源生态系统创造了一种去中心化的风险与去中心化防御的动态平衡。与闭源模型的中心化控制不同，开源范式将滥用的权力和防御的能力分散到了一个全球社区中。同样的透明度，既让恶意行为者能够发现漏洞，也让成千上万的安全研究人员能够发现并修复它。

以商业公司为主导的闭源模型在方法论上对科学研究构成了挑战。斯坦福大学《基础模型透明度指数》白皮书（FMTI）通过系统性评估，从数据、算法、算力和人力等多个维度打分，量化闭源模型和开源模型在透明度上的差异，即使是得分最高的公司也未能达

[1] McKinney S M, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening[J]. Nature, 2020, 577(7788): 89-94.

到高水平的透明度，并且随着模型变得越来越强大和商业化，整体趋势是透明度下降[1]。Nature 的社论指出，闭源模型因其固有的不透明性，与科学探究的开放原则存在内在冲突，从而阻碍了知识的验证与积累[2]。Bender 等人将此类模型界定为认识论上的“黑箱”，其训练语料、模型架构与参数权重的高度保密性，使得独立验证和外部审计在实践中变得困难，这直接影响了实证研究的基础[3]。研究者将此问题从学术范畴延伸至社会伦理层面，警示在医疗等高风险领域部署不可复现系统所带来的潜在公共安全危害[4]。

开源大语言模型也为教育带来了变革。开源大语言模型以其代码的公开降低了知识和技术的获取和使用门槛，打破时空和地域的限制。对于个人来说，每个人不仅可以免费、无限次体验 AI 功能，还可以根据自身需要研究、理解、修改以适应自身学习需求的 AI 大语言模型，亲身感受 AI 的数据训练与技术迭代过程。对于教育机构来说，开源大语言模型使得有限的教育资源分配更为合理，偏远地区的学校有机会有渠道获取优质教育资源和丰富的研究实例，弥补资源差异带来的教育不公平影响，也为教育研究者提供了可复制的研究路径和研究模型。

但同时，在开放的环境中大语言模型的乱用、滥用也会妨碍学生数据隐私安全，增加数据泄露的伦理风险。闭源大语言模型的封闭环境与系统为研究者提供了一种低技术难度的交互选择，研究者不需要熟练掌握编程语言与技能，也可以随时直接使用最前沿的 AI 能力进行学科研究。然而，闭源大语言模型的科学价值受限于其工程属性。研究者无法直接了解到研究运行的内部机制、无法检验模型的实际作用过程，导致研究结果的无法复现，减低研究的可信度。

[1] Foundation model transparency index［EB/OL］．［2025-10-04］．https://crfm.stanford.edu/fmti/May-2024/index.html．

[2] ChatGPT is a black box: How AI research can break it open[J]. Nature, 2023, 619(7971): 671-672．

[3] BENDER E M, GEBRU T, MCMILLAN-MAJOR A, et al. On the dangers of stochastic parrots: Can language models be too big? [C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, 2021: 610-623．

[4] BALL P. Is AI leading to a reproducibility crisis in science?[J]. Nature, 2023, 624(7990): 22-25．

面对开源与闭源大语言模型，研究者不应该全盘接受或者否定，针对自身的研究目标、资源条件以及技术能力权衡、选择最为合适的模型，以保证研究的稳健性与可靠性。

5.3 学术评议与质量保障

随着 AI 在科研中应用的日益普及，国内外高校和学术出版商正积极制定和调整相关政策，以规范 AI 工具在稿件撰写、数据分析及同行评议过程中的使用。这些新标准和指南的核心目标在于平衡 AI 带来的效率提升与维护科研诚信、确保研究质量之间的关系，强调人类在决策中的中心作用，总体呈现出一种“人在回路中”的理念[1]。

作为维护学术质量的“守门人”，同行评议体系正面临着前所未有的双重挑战：一方面，审稿人需要评估日益增多的、包含 AI 生成内容的稿件；另一方面，审稿人自身使用 AI 工具也带来了新的伦理风险。因此，对同行评议机制进行深刻的革新，已成为保障学术生态系统健康发展的当务之急，其核心在于平衡技术效率与科研诚信，重申人类在知识创造中的主体性地位。

5.3.1 AI 辅助研究的规定

国外众多高校和学术出版商就 AI 在研究和出版中的观点以及使用政策虽然在细节上存在差异，但已形成以责任归属和透明度披露为核心的普遍共识，包括 AI 工具不能被列为作者；人类作者对包括 AI 生成内容在内的全部工作负责；必须透明地披露 AI 的使用情况等。

（1）作者责任与 AI 署名规定

不同期刊对于研究中使用人工智能技术的态度各不相同。一些研究人员将 AI 列为学术论文的共同作者，如在 Oncoscience（《肿

[1] 褚乐阳, 潘香霖, 陈向东. AI 大模型在教育应用中的伦理风险与应对[J]. 苏州大学学报(教育科学版), 2024, 12(1): 87-96.

瘤科学》）杂志上发表的一篇文章中将 ChatGPT 列为共同作者[1]。对此，伦敦 Taylor & Francis 出版社出版伦理与诚信编辑萨宾娜-阿拉姆（Sabina Alam）表示，作者应对其作品的有效性和完整性负责，并应在致谢部分注明使用了 AI[2]。Nature 和 Science 杂志也表达了他们的立场，即不能将大语言模型工具列为作者。根据 Nature 的规定，AI 不具备作者资格，因其无法承担作者身份所固有的学术责任。在研究中对 AI 的任何实质性应用，都必须在方法论部分予以声明。

然而，若仅为优化人类撰写文本的语言表达与格式（如提高可读性、修正语法错误）而使用 AI 工具，则无需声明。最终，人类作者须对稿件的全部内容负最终责任。世界医学编辑协会 (WAME)在其关于学术出版中聊天机器人和生成人工智能的建议中指出，“聊天机器人不能成为作者”。Sage 允许研究者使用生成式人工智能，但必须披露使用情况，也不能将人工智能列为作者[3]。而 Elsevier 允许作者在写作过程中使用生成式人工智能和人工智能辅助技术改进论文的语言和可读性，并需要进行适当披露[4]。

Science 主编 H.霍尔登-索普（H. Holden Thorp）指出作品中不能使用由 ChatGPT（或任何其他人工智能工具）生成的文本，数字，图像或图形也不能是此类工具的产物。人工智能程序不能成为作者，违反这些政策将构成科学不端行为，与篡改图像或剽窃现有作品无异[5]。根据出版伦理委员会 (COPE) 的指南，人工智能 (AI) 工具因无法承担作品责任、处理版权或许可、以及声明利益冲突，故不满足作者身份的要求。因此，在研究中利用 AI 工具进行数据采集分析、图文生成或稿件撰写时，作者必须在“材料与方法”部分明确披露

[1] ChatGPT Generative Pre-trained Transformer, Zhavoronkov A. Rapamycin in the context of pascal’s wager: Generative pre-trained transformer perspective[J]. Oncoscience, 2022, 9: 82-84.
[2] Stokel-Walker C. ChatGPT listed as author on research papers: Many scientists disapprove[J]. Nature, 2023, 613(7945): 620-621.
[3] Artificial intelligence policy[EB/OL]. [2025-08-26]. https://www.sagepub.com/journals/editorial-policies/artificial-intelligence-policy.
[4] Generative AI policies for journals[EB/OL]. [2025-08-26]. https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals.
[5] Thorp H H. ChatGPT is fun, but not an author[J]. Science, 2023, 379(6630): 313-313.

所用工具及具体使用方式。人类作者对稿件的全部内容（包括由 AI 生成的部分）承担完全责任，并对任何潜在的学术不端行为负责[1]。

随着人工智能产业的发展，学术界对人工智能和作者身份（以及发明人身份）进行了许多讨论。从现行版权法的角度来看，《中华人民共和国著作权法》中提到“创作作品的自然人是作者”，即作者需要是自然人的身份。AI 工具不能被列为论文作者，因为它们无法承担研究工作的知识责任和法律责任，也无法同意署名或处理版权事宜。

对于 AI 使用的图像，各出版机构的政策不尽相同但普遍持谨慎态度。Nature 不允许在发表中使用人工智能生成的图片和视频，除非是从有特定关系的机构获得的图像或者图像是研究的核心内容，并且必须在图注中清晰标注其由 AI 生成。但同时，Nature 表示政策会随领域的发展而调整[2]。ACS 允许在期刊封面使用 AI 生成的图像（前提是充分披露、获得商业使用许可且输出不归生成网站所有），如果在论文中使用，需要在致谢中需要披露生成图像的完整信息，但不允许在目录 (ToC) 图文摘要中使用，因为后者缺乏足够的空间进行透明解释[3]。

（2）AI 使用披露的强制性与具体性

在研究中披露使用人工智能对于维护学术研究写作的完整性和可信度至关重要。几乎所有机构都强制要求作者在稿件中（通常在“方法”部分或“致谢”部分）明确、详细地披露 AI 工具的使用情况。披露内容应包括所使用的具体 AI 工具名称、版本号、开发者/公司，以及 AI 在研究的哪个阶段（如文献回顾、数据分析、图像生成、稿件撰写或润色等）、以何种方式、在多大程度上被使用。

[1] Authorship and AI tools[EB/OL]. (2023-02-13)[2025-08-08]. https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools.
[2] Artificial intelligence (AI) | nature portfolio[EB/OL]. [2025-08-27]. https://www.nature.com/nature-portfolio/editorial-policies/ai.
[3] Banik G M, Baysinger G, Kamat P V, et al. The ACS guide to scholarly communication[M]. Washington, DC: American Chemical Society, 2020.

MAIEI 不允许作者使用人工智能生成框架和观点，要求披露生成内容的提示过程[1]，而 JMIR 则建议作者提交 AI 交互的完整记录（包括提示和回复）作为补充材料，供编辑和审稿人参考[2]。人类作者必须对提交稿件的全部内容，包括由 AI 生成的任何部分（文本、图像、代码、数据分析结果等）的准确性、原创性和完整性负最终责任。各种出版商的人工智能政策表明，由于该领域的不断发展，它们将不断更新和修订，其中的要求需要根据新的发展进行修改。

目前，大部分的期刊编辑保留最终裁量权，可以根据 AI 在稿件中使用的范围和方式，判断其是否恰当或过度。如果编辑认为 AI 的使用不当（例如，AI 生成了实质性的评论、核心论点或广泛的文献综述，而缺乏作者足够的原创贡献和批判性评估），可能会导致稿件被拒或要求作者进行重大修改以减少或移除 AI 生成的部分[3]。

5.3.2 AI 赋能同行评议

同行评议是学术出版的基石，但其过程漫长、负担沉重，且存在主观性和不一致性。人工智能技术被探索用于辅助同行评议流程这一核心环节，以期提高效率和质量。AI 目前可以协助进行一些初步的、程序性的检查，简化传统上由人类编辑和审稿人处理的众多任务。这些应用范围广泛，从自动语言和语法检查到抄袭检测、格式合规，甚至是研究意义和研究方法的初步评估。目前已有不同的人工智能工具来帮助出版商、编辑和同行评审在同行评审过程的不同阶段发挥作用，例如 RobotReviewer、SciScore、StatReviewer、AuthorOne，涉及到的功能包括研究设计评估、方法评估、统计分析评估、稿件格式评估等[4]。

使用人工智能系统对学术论文进行评价，很多人会对评价的准

[1] Our editorial stance on AI tools in submissions[EB/OL]. [2025-08-27]. https://montrealethics.ai/our-editorial-stance-on-ai-tools-in-submissions/.

[2] Leung T I, de Azevedo Cardoso T, Mavragani A, et al. Best practices for using AI tools as an author, peer reviewer, or editor[J]. Journal of Medical Internet Research, 2023, 25: e51584.

[3] Banik G M, Baysinger G, Kamat P V, et al. The ACS guide to scholarly communication[M]. Washington, DC: American Chemical Society, 2020.

[4] Shah F A, Jawaid S A. The inevitable future of peer review: Human and AI integrated peer review system[J]. Pakistan Journal of Medical Sciences, 2025, 41(4): 941-943.

确性、客观性、新颖性产生质疑。有研究者将 GPT-4 生成的反馈与 Nature 的 15 种期刊（共 3096 篇论文）和 ICLR 机器学习会议（1709 篇论文）上人类同行评审员的反馈进行了定量比较，发现 GPT-4 和人类审稿人提出的要点重叠率（Nature 期刊平均重叠率为 30.85%，ICLR 为 39.23%）与两名人类审稿人之间的重叠率（Nature 期刊平均重叠率为 28.58%，ICLR 为 35.25%）相当，说明大语言模型在同行评审上可以给予帮助[1]。20 位期刊编辑对人工智能生成的稿件审稿人建议进行评估，人工智能系统与编辑的选择有 42% 的重叠，并显著提高了时间效率[2]。

早在 2020 年，frontiers 就已使用人工智能审稿助手（AIRA）协助编辑、审稿人和作者进行稿件的自动质量检查和初步审阅工作，帮助审稿人更有效地做出编辑决策[3]。其产品开发总监 Daniel Petrariu 认为 AIRA 是机器与人类之间协作的系统，帮助我们实现高精度和效率。

多伦多论文匹配系统（TPMS）通过构建审稿人的个人资料库（见图 5-1），主要是存放提交的材料和元数据，还可以与其他论文管理框架交互，应用多个组合信息计算审稿人与每篇提交论文的适用性，依据全局最优原则将论文分配给审稿人[4]。

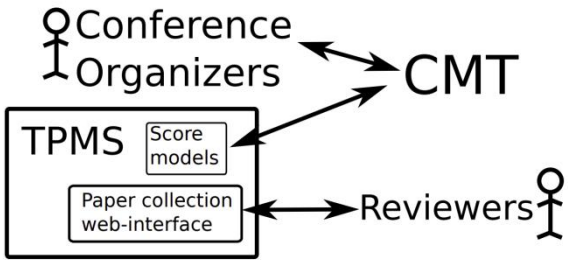

图 5-1 多伦多论文匹配系统

[1] Liang W, Zhang Y, Cao H, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis[J]. NEJM AI, 2024.

[2] Farber S. Enhancing peer review efficiency: A mixed-methods analysis of artificial intelligence-assisted reviewer selection across academic disciplines[J]. Learned Publishing, 2024, 37(4): e1638.

[3] Artificial Intelligence to help meet global demand for high-quality, objective peer-review in publishing[EB/OL]. [2025-10-10]. https://www.frontiersin.org/news/2020/07/01/artificial-intelligence-peer-review-assistant-aira/.

[4] CHARLIN L, ZEMEL R S. The Toronto Paper Matching System: An automated paper-reviewer assignment system[J].

荷兰阿姆斯特丹 World Brain Scholar (WBS) 公司研发的 Eliza 工具，拥有支持同行评审所需的关键自然语言处理（NLP）技术，可以帮助同行评审员进行内容分析、数目分析以及措辞改进，提供更为全面和有洞察力的反馈，同时工具也展现出跨语言处理智能特性：其不仅能够为人类评审推荐参考文献，还可将其他语言的评审内容实时翻译为英文。据相关信息显示，该技术已协助《欧洲物理评论》杂志处理了 23%的非英语评审内容。今年 Eliza 经过重新设计并推出，不仅具有以前工具的所有功能，还增添了许多新的模块，例如编辑支持、参考文献分析、范围检查和期刊建议等（见图 5-2），更为灵活和与使用者个人适配[1]。

图 5-2 Eliza 工具的功能

International Conference on Learning Representations（ICLR）引入了“评审反馈智能体”（Review Feedback Agent），让 AI 去识别审查中存在的潜在问题，并向审稿人反馈以进行改进。该系统不会

[1] World Brain Scholar[EB/OL]. [2025-10-11]. https://www.world-brain-scholar.eu/news/13.

取代审稿人，而是充当一个助手，使审稿对作者更具建设性和可作性，最终验收决定将由 AC、SAC 和审稿人做出[1]。系统有五个 LLM（参与者、聚合器、批评者和格式化器）组成，两个并行的参与者生成初始反馈，然后将其传递给聚合器、批评者，接着传递给格式化器，最后引入可靠性测试（Reliability Tests），对 AI 反馈的特定属性进行评估，只有通过所有的测试才会将反馈结果发送给审稿人，确保其质量（见图 5-3）。结果表明，收到反馈的审稿人有 27%更新了自己的评论并采纳了来自代理的 12000 条建议，极大提高了同行评审的质量和效率[2]。

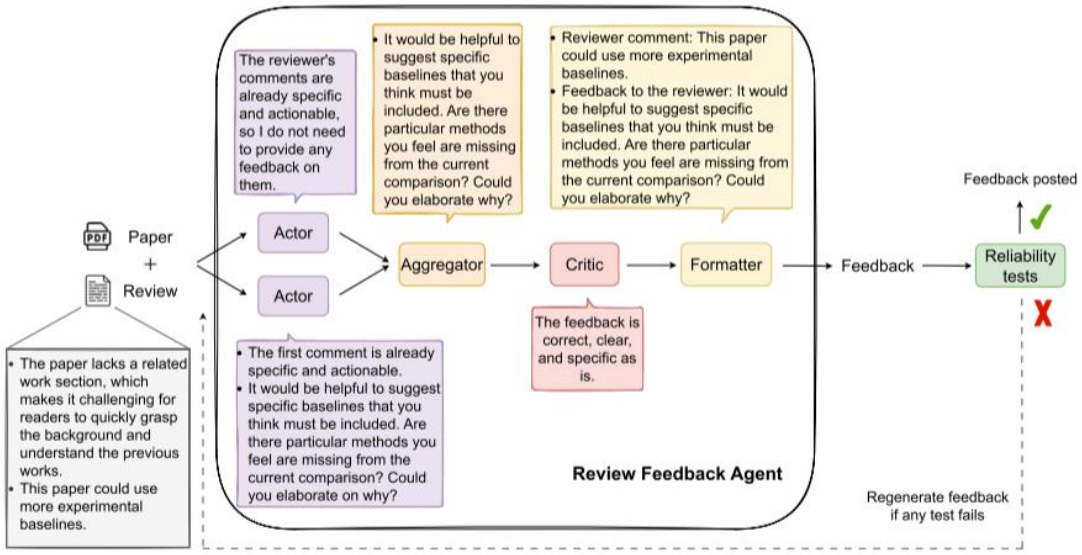

图 5-3 ICLR 系统运行程序

然而，学术界普遍认为，人工智能目前尚不能替代人类专家的核心评审职能，尤其是在评估研究的原创性、科学重要性、概念新颖性、方法论的深度和严谨性、以及研究结果的解释和潜在影响等方面。科学的同行评审需要深入了解专业课题、复杂方法和前沿发现，但人工智能会错过某些方法论缺陷或理论不一致之处，过度依赖人工智能可能会导致遗漏错误或对稿件关键要素的评估不足[3]。

[1] Assisting ICLR 2025 reviewers with feedback – ICLR Blog[EB/OL]. (2024-10-09)[2025-10-11]. https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/.

[2] THAKKAR N, YUKSEKGONUL M, SILBERG J, et al. Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025[EB/OL]. arXiv, 2025[2025-10-11]. http://arxiv.org/abs/2504.09737.

[3] Doskaliuk B, Zimba O, Yessirkepov M, et al. Artificial intelligence in peer review: Enhancing efficiency while preserving integrity[J]. Journal of Korean Medical Science, 2025, 40(7): e92.

作者的人为操控也是使用人工智能进行同行评审需要注意的要点，包括显性操纵和隐性操纵，显性操纵指使用极小的白色字体将操纵性评论内容嵌入稿件 PDF 中，使其在背景中几乎不可见，而隐形操控指作者主动披露稿件中无关紧要的局限性，这些操纵会影响大语言模型的评审结果[1]。

2023 年，美国国立卫生研究院（NIH）禁止使用 ChatGPT 等在线生成式人工智能工具分析和起草同行评审意见[2]。因此，人工智能在同行评议中的应用，正从最初被视为简单的“辅助性工具”，逐渐向更深度嵌入评审流程的“嵌入式辅助系统”发展，但其确切的角色定位、伦理边界和技术局限性仍在持续探索和界定中。

目前，针对审稿人如何使用人工智能工具，也出现了一些具体指南：

Nature 同行评审指南明确禁止审稿人将待审稿件上传到公开的生成式人工智能工具中，因为人工智能工具可能缺乏最新知识，并且可能生成无意义、有偏见或虚假的信息。稿件还可能包含敏感或专有信息，这些信息不应在同行评审流程之外共享。如果稿件中提出的任何声明的评估部分得到了人工智能工具的支持，要求同行评审员在同行评审报告中声明此类工具的使用[3]。

Science 同行评审指南中也明确不允许使用大语言模型和其他生成式人工智能工具生成审稿意见。同行评审员负责撰写自己的审稿意见，未经编辑许可，不能与同事分享稿件，且不得征求同事的意见。同行评审员严格遵循保密原则，审阅后，需要删除或丢弃稿件的所有副本[4]。

JMIR Publications 要求对同行评审中提交的内容负责，同行评审

[1] Ye R, Pang X, Chai J, et al. Are we there yet? Revealing the risks of utilizing large language models in scholarly peer review[A]. arXiv, 2024.

[2] Cheng K, Sun Z, Liu X, et al. Generative artificial intelligence is infiltrating peer review process[J]. Critical Care, 2024, 28(1): 149.

[3] Artificial intelligence (AI) | nature portfolio[EB/OL]. [2025-08-24]. https://www.nature.com/nature-portfolio/editorial-policies/ai.

[4] Peer review at science journals[EB/OL]. [2025-08-24]. https://www.science.org/content/page/peer-review-science-publications.

的质量和内容符合标准，对人工智能工具使用的风险保持谨慎，并明确披露所使用的 AI 工具及其具体用途，在使用过程遵循问责、透明和保密的原则。同时，禁止使用 ChatGPT 的免费版本这种可能存在数据泄露风险的人工智能工具协助生成同行评审意见[1]。

Sage 同行评审指南中指出虽然大语言模型可以模仿审稿报告，但无法体现人类专家的专业经验、对研究背景的深刻理解，以及对其社会影响的评估能力。使用生成式 AI 工具撰写审稿意见存在数据泄露和违反保密规定的风险。如果编辑怀疑审稿报告由 AI 生成，应立即向 Sage 出版社上报，以获得进一步的指导。

Elsevier 要求审稿人不应使用生成式人工智能或人工智能辅助技术来协助论文的科学评审，因为同行评审所需的批判性思维和原创性评估超出了该技术的范畴，并且有可能得出关于论文稿件的不正确、不完整或带有偏见的结论。审稿人应对评审报告的内容负责。

这些政策和指南的出现，在于将 AI 定位为受监管的辅助工具，而非决策主体，反映了学术界面对人工智能技术快速发展所带来的挑战，在努力寻求一种既能利用人工智能提高科研和出版效率，又能有效维护科研诚信和质量的平衡。

5.3.3 研究可信度建立的新方法

在 AI 深刻融入科学研究的背景下，传统的研究可信度建立方式正面临挑战，同时也催生了新的评估方法和保障机制。AI 既可能因其“黑箱”特性、潜在偏见或生成虚假内容的能力而削弱研究的可信度，也可能通过其强大的数据分析、模式识别和自动化验证能力，为建立更高标准的可信度提供新的工具和途径。AI 时代研究可信度的建立，正从主要依赖单一的、出版前的同行评议模式，演变为一个更加多元化、动态化、贯穿研究全生命周期的“可信度生态系统”。

[1] Leung T I, de Azevedo Cardoso T, Mavragani A, et al. Best practices for using AI tools as an author, peer reviewer, or editor[J]. Journal of Medical Internet Research, 2023, 25: e51584.

（1）建立 AI 使用的原则框架

针对 AI 研究及其成果，在教育研究中需要审查现有的人工智能整体管理政策，目前有诸多的伦理框架和指南。引起世界各国政府关注的一些问题包括人工智能的歧视和偏见、隐私泄露、侵犯人权以及人工智能的恶意使用。鉴于此，各国一直在制定国家政策和战略，为人工智能的使用提供更明确的指导，以最大限度地发挥其效益，同时减轻其带来的威胁，各个机构制定了相应的原则框架。例如 OECD 提出“倡导可信、创新的 AI，尊重人权、公平、透明、可解释、稳健、安全、可靠和问责”[1]，NIST 提出“提供了一个自愿性框架，用于治理、识别、衡量和管理 AI 风险，推广可信 AI 的特征（有效、可靠、安全、稳健、可问责、透明、可解释、隐私增强、公平）”[2]。开放科学的理念中的 FAIR 原则（可发现性 Findability, 可访问性 Accessibility, 互操作性 Interoperability, 可重用性 Reusability）[3]，对 AI 辅助的教育研究产生深远影响，使研究人员能够通过严格的测试和验证流程来证实研究结果并解决偏差问题，有助于最大限度地减少人工智能模型中的偏差，确保数据的高质量、代表性和可访问性。这些原则为 AI 教育研究注入强大动力，但也对其伦理规范、数据治理和研究者素养提出了新的、更高的要求。

联合国教科文组织在 2021 年制定《人工智能与教育》为政策制定者提供指南，指导政策制定者充分利用人工智能与教育深度融合带来的机遇以及应对随之而来的风险[4]。目前有一些学者建立使用人工智能的框架，如 Chan 采用定量和定性的研究方法，构建人工智能政策教育框架，框架涵盖教学法、治理和运营三个维度，教学维

[1] OECD. Recommendation of the Council on Artificial Intelligence[R]. Paris: OECD Publishing, 2019. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[2] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0)[R]. U.S. Department of Commerce, National Institute of Standards and Technology, 2023. https://doi.org/10.6028/NIST.AI.100-1.

[3] Hanna M G, Pantanowitz L, Jackson B, et al. Ethical and bias considerations in artificial intelligence/machine learning[J]. Modern Pathology, 2025, 38(3): 100686.

[4] 人工智能与教育: 政策制定者指南 - UNESCO digital library[EB/OL]. [2025-08-25]. https://unesdoc.unesco.org/ark:/48223/pf0000378648.

度集中于利用人工智能改善教学成果，治理维度集中于解决与隐私、安全和问责相关的问题，运营维度则涉及基础设施和培训等问题[1]。

所有关于将人工智能纳入高等教育的框架和指导方针的共同点包括：机构的愿景和目标；基础设施/资源和运营挑战；利益相关方（教师、学生和行政部门）的参与；结构化方法；为学生和教师提供指导方针；沟通渠道；以及对学生和教师进行人工智能培训（人工智能扫盲）[2]。然而，Moorhouse 等人调查了世界排名前 50 的大学有自己的指导方针，发现仅有不到一半的院校制定了公开的指导原则。指导原则主要包括三个方面：学术诚信、评估设计建议和与学生沟通[3]。McDonald 等人分析了 116 所美国大学发布的指导文件后发现，超过 60% 的大学鼓励使用生成式人工智能，约 40%的大学提供了课堂使用生成式人工智能的详细指导[4]。

（2）建立评估 AI 生成内容可信度的框架

针对 AI 可能直接参与科学假设生成或结论阐述的新情况，学术界和监管机构开始探索专门的评估框架。

FDA 的基于风险的可信度评估框架：该框架提出一个基于风险的 7 步流程，用于评估 AI 模型输出在特定“使用情境”下的可信度。它要求首先明确 AI 要解决的“问题”和 AI 模型的具体“使用情境”，然后评估模型在该情境下的风险（综合考量“模型影响力”和“决策后果”），并据此制定和执行详细的“可信度评估计划”（包括对模型本身、开发过程、训练数据、评估过程的全面描述和验证），最终判断模型对于该使用情境是否“充分可信”[5]。

[1] Chan C K Y. A comprehensive AI policy education framework for university teaching and learning[J]. International Journal of Educational Technology in Higher Education, 2023, 20(1): 38.

[2] Symeou L, Louca L, Kavadella A, et al. Development of Evidence-Based Guidelines for the Integration of Generative AI in University Education Through a Multidisciplinary, Consensus-Based Approach[J]. European Journal of Dental Education, 2025, 29(2): 285-303.

[3] Moorhouse B L, Yeo M A, Wan Y. Generative AI tools and assessment: Guidelines of the world’s top-ranking universities[J]. Computers and Education Open, 2023, 5: 100151.

[4] McDonald N, Aditya Johri, Areej Ali, et al. Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines[J]. Computers in Human Behavior: Artificial Humans, 2025, 3: 100121.

[5] Malek T W Roma Sharma, Linda. FDA proposes framework to assess AI model output credibility to support regulatory decision-making[EB/OL]. (2025-01-29)[2025-08-25]. https://www.cmhealthlaw.com/2025/01/fda-proposes-framework-to-assess-ai-model-output-credibility-to-support-regulatory-decision-making/.

FIU 图书馆的 FLACK 框架：这是一个指导用户批判性评估 AI 生成信息的实用框架，包括四个步骤：Fractionate the Information（将 AI 输出的叙述性内容拆解为具体、可查证的声明），Lateral Reading to Verify Information（通过横向阅读，查找多个独立来源来验证这些声明的准确性），Evaluate the Assumptions（评估用户提问时和 AI 生成回答时各自的隐含假设），以及 Add to Knowledge Base? (最终决定是否采信这些信息，并反思提问技巧)[1]。

Sourcely 等 AI 工具辅助评估与 CRAAP 框架：一些新兴的 AI 工具本身也被设计用来辅助评估其他 AI 生成内容或信息源的可信度，例如进行自动化事实核查、来源可信度分析、偏见检测等。在评估信息源时，可以借鉴经典的 CRAAP 框架（Currency 时效性，Relevance 相关性，Authority 权威性，Accuracy 准确性，Purpose 目的性）[2]。

MIT 的 SciAgents 框架：该框架通过构建一个由多个具有不同角色的 AI 智能体（如“本体论构建者”、“科学家 1”、“科学家 2”、“批评家”）组成的协作系统，来模拟科学发现的过程。SciAgents 利用知识图谱和图推理方法，不仅能自主生成研究假设，还能对其进行初步评估和批判（例如，“批评家”智能体会指出假设的优缺点并建议改进）[3]。

（3）增强研究者的伦理教育

对 AI 塑造研究可信度评议的革新，已成为保障学术生态系统健康发展的当务之急。然而，任何评议体系的改革，若想取得长效，其根基必须深植于教育之中。人工智能的复杂性要求在教育领域建立“一套全面适用的伦理原则”。因此，倡导将伦理纳入人工智能

[1] Jimenez C. FIU libraries: AI tools and the research process: assessing AI tools[EB/OL]. [2025-08-25]. https://library.fiu.edu/AI-Tools/assessing.
[2] Top 10 AI tools for ensuring content credibility and accuracy[EB/OL]. 2025[2025-08-25]. https://www.sourcely.net/resources/top-10-ai-tools-for-ensuring-content-credibility-and-accuracy.
[3] Need a research hypothesis? Ask AI.[EB/OL]. (2024-12-19)[2025-08-25]. https://news.mit.edu/2024/need-research-hypothesis-ask-ai-1219.

教育至关重要，以确保学生理解在学术研究中使用人工智能的伦理影响。

大学应将伦理道德融入人工智能和技术课程，涵盖数据隐私、算法偏见和社会影响等主题。伦理讨论应涵盖所有学科，以确保更广泛的理解。关于人工智能引发的抄袭、偏见性决策和知识产权问题的真实案例研究，有助于培养批判性思维。应制定清晰易懂的政策，定义学术工作中人工智能的伦理使用，以促进透明度和原创性。教师可以设计项目，探索人工智能的伦理困境，鼓励实践理解。围绕人工智能伦理影响的公开讨论、研讨会和辩论，邀请专家参与，将有助于学生反思负责任的人工智能使用方式。这种方法确保人工智能能够补充人类的能力，同时维护学术诚信和道德标准。

例如，康奈尔大学提出应将 AI 基础知识、核心方法论、伦理规范以及批判性思维能力的培养，全面融入科研人员和高等教育学生的培养体系中[1]。麻省理工学院媒体实验室团队为中学生和教师提供了一个关于人工智能和伦理的开放课程，通过一系列课程计划和实践活动，教师可以指导学生学习人工智能系统的技术术语以及人工智能的伦理和社会影响[2]。这不仅包括技术操作层面，更要强调对 AI 能力边界的认知和对其社会影响的学习。

当下的学生有可能成为未来审稿专家。教师需在其学术生涯的初始阶段，就灌输“人在回路”的核心理念，使其领悟到 AI 的价值在于增强而非取代人类专家的学术洞察力。只有这样，才能为正处于变革中的评议体系输送具备新素养、遵循新伦理的下一代守护者。

因此，人工智能与科学研究的深度融合，最终导向研究质量与标准的范式性转变。这一导向不仅依赖于构建互融共生的人机协同智能生产的伦理框架与强有力的透明度质量标准，更取决于我们能

[1] Generative AI in academic research[EB/OL]. [2025-08-25]. https://live-cu-research-innovation.pantheonsite.io/generative-ai-in-academic-research/.

[2] Akgun S, Greenhow C. Artificial intelligence in education: Addressing ethical challenges in K-12 settings[J]. AI and Ethics, 2022, 2(3): 431-440.

否借助教育体系，将这一框架内化为未来研究者的自觉意识和核心素养，弥合技术趋势下工具与价值之间的鸿沟。这表明，高等教育需要突破传统方法论传授的局限，系统地培育学生明确自身主体性驾驭 AI 工具以提升研究效率的能力。

第 6 章 AI 促进教育知识转化

教育研究与实践之间的“知行鸿沟”，是长期制约教育发展的核心瓶颈，也是一个困扰了数代教育者的难题。尽管在过去数十年间，教育领域积累了丰硕的理论创新与实证成果，但大量极具价值的发现却迟迟无法有效转化为一线教学实践，服务于真实的学与教[1]。海量的研究证据被束之高阁，理论的精巧建构与实践的迫切需求之间，始终存在着一道难以逾越的屏障。这种知识转化的结构性困境，不仅严重削弱了教育研究的社会价值，也深刻地延缓了教学质量提升与教育公平实现的进程。

在“人工智能驱动的科学研究”（AI4S）的范式浪潮席卷全球的背景下，利用 AI 驱动科学发现已成为新一轮科技革命的核心引擎。当这一范式应用于教育领域，“教育研究中的 AI4S”便为破解上述难题提供了全新的解题思路。教育知识转化之所以困难，根源在于教育实践的内在复杂性：它深深嵌入在具体情境之中，高度依赖于师生间的精微互动、独特的课堂文化与复杂的学校组织生态的动态适配[2]。

以大语言模型为代表的生成式人工智能技术，其展现出的卓越的知识理解、情境感知与动态推理能力，与教育知识转化对情境敏感、动态调整、策略再生的核心需求形成了惊人的契合。AI4S 的理念，正是要将这种强大的技术能力与教育科学的内在规律相结合，构建起全新的研究与实践通路。本章旨在系统性地探讨以大语言模型为代表的 AI 技术如何作为 AI4S 的关键赋能工具，促进教育知识转化的范式变革。本章将深入剖析当前知识转化面临的结构性障碍，阐明 AI 赋能的核心机制，并构建一套 AI 驱动的知识转化路径体系，为弥合“知行鸿沟”提供兼具前瞻性与实操性的理论指导与实践蓝

[1] Berliner D C. Comment: Educational Research:The Hardest Science of All[J]. Educational Researcher, 2002, 31(8): 18-20.

[2] Honig M I. complexity and policy implementation: challenges and opportunities for the field[M]//Honig M I. New Directions in Education Policy Implementation. SUNY Press, 2006: 1-23.

图。

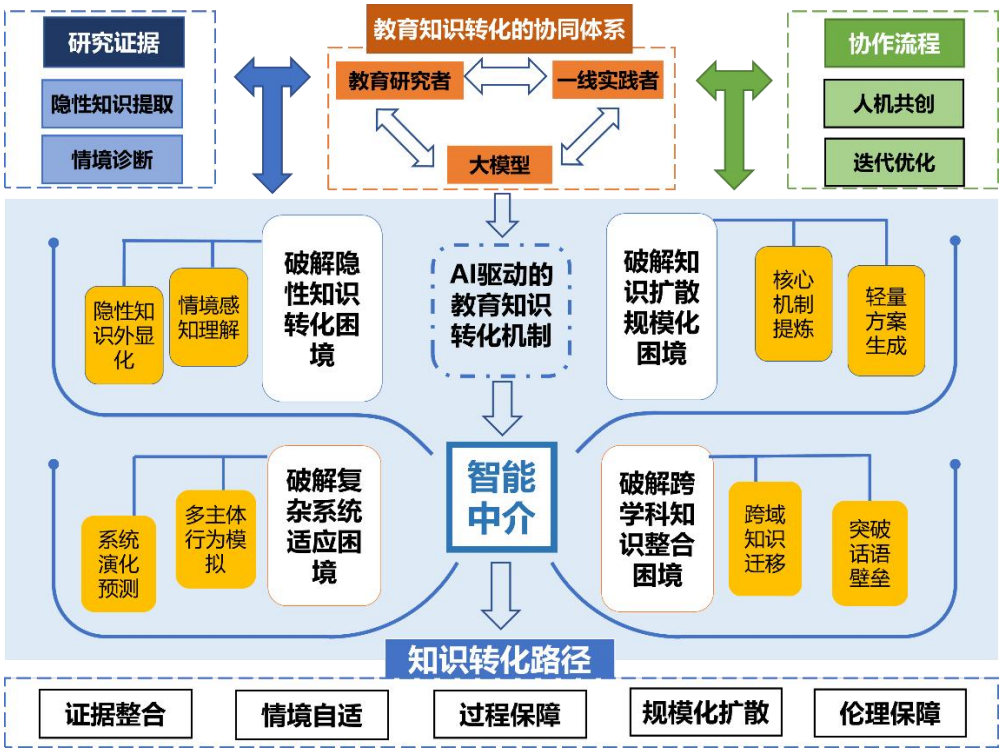

图 6-1 基于 AI4S 的教育知识转化机制模型

6.1 教育知识转化的现实困境

在自然科学领域，知识转化往往遵循相对清晰的“发现-验证-应用”路径。然而，教育领域的知识转化却呈现出截然不同的复杂图景。这种复杂性是由教育知识的独特本质和转化过程的系统性挑战共同塑造的。

6.1.1 教育知识的核心特征

教育知识之所以难以转化，其根本原因在于其独特的内在属性，这决定了任何简化的、线性的、试图“一刀切”的转化方案都难以奏效。与自然科学中那种可以被客观描述、普遍验证、并以命题形式存在的知识不同，教育知识更接近一种“实践智慧”，其本体论地位复杂而多维，深深植根于人类的实践活动之中。纵观历史，教育领域为弥合“知行鸿沟”所做的努力，本身就揭示了对知识转化复杂性认识的不断深化。这一过程并非简单的理论应用，而是围绕教育知识的几大核心特征展开的持续探索。

教育知识的首要特征，在于它是一种显性与隐性知识的共生体。哲学家迈克尔·波兰尼（Michael Polanyi）提出的“隐性知识”概念，即“我们知道的比我们能言说的多”，为理解这一特征提供了关键钥匙。教育知识中，能够被清晰编码、记录在册的课程标准、教学理论、学科内容等，仅仅是知识冰山浮于水面的显性部分。而在水面之下，更庞大、更关键的是教师大量的隐性知识，那种基于多年经验形成的专业直觉，在面对课堂突发状况时的即时判断与应对策略，对不同学生微表情的敏锐解读，以及营造特定课堂氛围的“教学艺术”。早期的知识转化探索，如借鉴野中郁次郎的 SECI 模型[1]，正是试图理解这种显性与隐性知识的转换机制。然而，研究者很快发现，任何只关注显性知识传递的转化方案，都注定会丢失掉那些最鲜活、最有效的实践精髓，因为教师的隐性知识难以被完全言说和形式化。

这种以隐性形态为主的实践智慧，必然表现出其第二个特征：深刻的情境嵌入性与实践根植性。知识并非存在于真空中的抽象实体，而是与孕育它的具体情境密不可分，这正是“情境认知”（Situated Cognition）理论的核心洞见[2]。一项教学策略的意义和有效性，高度依赖于其所在的具体情境，这包括了学生的年龄、认知水平、家庭背景，师生之间已经建立的关系模式，班级乃至学校所特有的文化氛围，以及更广泛的社区环境和政策导向。这意味着，教育知识无法被简单地“拿来”，而必须在新的土壤中经历一个“再创造”的过程。这一洞见推动了转化理论的重大突破。格斯基（Guskey）提出的教师变革模型指出[3]，教师的信念是在亲身尝试新实践并观察到学生学习改善后才会改变，这颠覆了“先理论后实践”的单向逻辑。在此思想指引下，一系列扎根实践的转化模式应

[1] Nonaka L, Takeuchi H, Umemoto K. A theory of organizational knowledge creation[J]. International Journal of Technology Management, 1996, 11(7-8): 833-845.
[2] Brown J S, Collins A, Duguid P. Situated Cognition and the Culture of Learning[J]. Educational Researcher, 1989, 18(1): 32-42.
[3] Guskey T R. Professional development and teacher change[J]. Teachers and Teaching, 2002, 8(3): 381-391.

运而生，如“专业学习共同体”（PLCs）和“课例研究”（Lesson Study）[1]。它们都将知识转化的主体定位为一线教师，将场域设定在真实教学情境中，通过协作探究、共同设计和反思，将抽象理论转化为具体的、情境化的教学决策。

当这种情境化的实践智慧在课堂、学校这样的社会单元中展开时，其第三个特征便显现出来：它是在一个复杂适应系统中的涌现性实践。当视线从单个教师放大到整个教育系统时，会发现教育场域并非机械系统，而是一个由众多拥有自主意识的主体（学生、教师、管理者、家长等）构成的、充满互动与反馈的生态系统。任何外部干预都可能会引发一系列非线性的、难以预料的“涌现”效应[2]。例如，爱尔兰“项目数学”改革因未能撼动既有的教学文化与评估体系的惯性，转化效果大打折扣[3]。而芬兰教育改革的成功，则得益于其系统性的策略：通过提升教师专业地位、建立研究型师范教育、营造信任文化，在宏观制度层面实现了研究与实践的良性互动[4]。这表明，有效的知识转化并非孤立的技术问题，而是一个牵一发而动全身的系统工程，其最终效果是在多方力量的持续博弈和动态演化中生成的。

最后，教育知识具有一个不容忽视的、也是其区别于许多自然科学知识的根本特征：内在的价值负载性。教育活动从根本上说，是一种价值负载的、具有明确目的性的实践活动。它不仅要回答“是什么”（事实）和“怎么办”（方法）的问题，更要深刻地回应“应该如何”（价值）的追问。我们选择教什么、如何教，背后都承载着我们对于“理想的人”和“理想的社会”的价值判断和伦理追求。因此，教育知识的转化，绝非一个纯粹的技术操作过程，

[1] DuFour R, Reeves D. The futility of PLC lite[J]. Phi Delta Kappan, 2016, 97(6): 69-71.
[2] Hawe P, Shiell A, Riley T. Theorising interventions as events in systems[J]. American Journal of Community Psychology, 2009, 43(3): 267-276.
[3] Prendergast M, O’Donoghue J. ‘students enjoyed and talked about the classes in the corridors’: pedagogical framework promoting interest in algebra[J]. International Journal of Mathematical Education in Science and Technology, 2014, 45(6): 795-812.
[4] Sahlberg P. Finnish schools and the global educational reform movement.[M]//Evers J, Kneyber R. Flip the system: Changing education from the ground up. Abingdon, Oxon New York, NY: Routledge, 2015: 162-174.

它必然触及深层的教育信念、文化传统和伦理规范，这使其转化过程充满了张力与审慎，需要持续的价值反思与伦理判断。

6.1.2 知识转化的结构性障碍

正是上述独特的本体论特征，共同催生了教育知识转化在现实中所面临的四个相互交织、根深蒂固的结构性障碍。它们系统性地解释了为何“知行鸿沟”如此难以逾越。

第一个，也是最基础的障碍，是隐性知识的转化困境。如前所述，教师高水平的实践智慧大量以隐性形态存在，它们个人化、情境化且难以言传。这直接导致了教育领域一个普遍的痛点：“名师”的经验难以被有效“解码”和复制，高水平的教学艺术难以通过标准化的培训大规模传承。当知识本身难以被清晰地表达、记录和共享时，其后续的传播和转化自然举步维艰。这构成了知识转化链条上最源头的瓶颈。

由此直接引发了第二个障碍，知识扩散的规模化困境。一项在特定实验环境或少数精英学校被验证有效的教育创新，一旦尝试大规模推广，其效果往往会急剧衰减，遭遇“水土不服”。这远非简单的技术复制问题。科伯恩[1]指出，真正的规模化不仅是数量上的扩张，更涉及在不同情境下保持其核心机制的深度（depth）、确保长期影响的可持续性（sustainability），以及实现理念和方法被新采纳者真正内化的所有权转移（shift of ownership）。然而，在推广过程中，由于不同区域、学校、教师之间的认知框架、资源条件、文化背景存在巨大差异，创新理念的核心机制常常在传递过程中被稀释、误解甚至扭曲[2]，最终导致“橘生淮南则为橘，生于淮北则为枳”的普遍现象。

第三个障碍，是复杂系统中的知识适应困境。即使能够提炼出

[1] Coburn C E. Rethinking scale: moving beyond numbers to deep and lasting change[J]. Educational Researcher, 2003, 32(6): 3-12.

[2] Spillane J P, Reiser B J, Reimer T. Policy implementation and cognition: reframing and refocusing implementation research[J]. Review of Educational Research, 2002, 72(3): 387-431.

创新方案的核心机制，并尝试进行推广，这些知识和策略在进入一个新的学校或课堂时，也必须面对一个动态演化的复杂适应系统。传统的“规划-控制”式线性实施思维在此完全失效。因为任何外部干预都会被系统视为一种“扰动”，它会与系统中已有的各种力量（如教师的既有习惯、学生的反应、同伴的压力、学校的行政指令）发生复杂的化学反应。这种反应充满了非线性反馈和延迟效应，使得实施结果难以被精确预测和控制[1]。实施者常常发现，精心设计的方案在现实面前会“走样”，其实质是知识未能成功适应复杂系统的动态生态。

最后，跨学科知识的整合困境为解决上述问题制造了最终的壁垒。教育是一个极其复杂的综合性人类活动，解决其面临的真实问题，通常需要整合来自教育学、心理学、社会学、神经科学、管理学等多个学科的知识。然而，在现实中，不同学科之间存在着显著的话语体系壁垒、认识论差异和方法论分歧。例如，对于同一个课堂互动现象，心理学家可能关注其认知负荷，社会学家可能分析其权力关系，而教育学家则可能着眼于其教学法意义。这种学科分割导致的知识碎片化，使得很难形成一个整合的、综合性的解决方案来应对复杂的教育挑战[2]。

这四大结构性障碍相互关联，层层递进，共同构成了教育知识转化领域难以撼动的系统性挑战。在知识生产与应用节奏日益加速的今天，以大语言模型为代表的AI技术的崛起，正为从根本上重构解题思路、突破这些传统障碍提供了全新的可能性。

6.2 AI赋能知识转化的理论机制

面对教育知识转化根深蒂固的结构性障碍，传统依赖人力与组织协同的路径已日益凸显成本高昂、规模化困难、情境适应能力有

[1] Chambers D A, Glasgow R E, Stange K C. The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change[J]. Implementation Science, 2013, 8(1): 117.

[2] Penuel W R, Allen ,Anna-Ruth, Coburn ,Cynthia E., et al. Conceptualizing research–practice partnerships as joint work at boundaries[J]. Journal of Education for Students Placed at Risk (JESPAR), 2015, 20(1-2): 182-197.

限等瓶颈。以大语言模型为代表的生成式 AI 的兴起，则为突破这些瓶颈提供了前所未有的技术可能性。与以往辅助性工具不同，这类 AI 技术作为新型的“智能中介”，其赋能价值源于多项核心技术能力的有机整合。世界知识是基础，为后续机制提供了丰富的知识储备；知识蒸馏是手段，将复杂知识转化为可规模化的形态；生成式仿真是工具，在虚拟空间中预演转化过程；泛化能力是保障，确保知识能够跨越学科边界实现融合。这四项核心机制并非孤立运作，而是相互支撑、协同发力的有机整体，共同构成了 AI 赋能知识转化的完整技术链条。

表 6-1 概括了 AI 赋能知识转化的四项核心机制及其针对的具体困境。这一对应关系为我们提供了分析框架，但每一项机制的理论内涵、技术实现和实践效果，仍需更细致的阐述。以下各小节将依次展开论述。

表 6-1 AI 赋能教育知识转化的核心概念与适用性

AI 核心能力	核心概念与出处	应对的转化困境	适用性阐释
世界知识（World Knowledge）	AI 通过海量数据预训练获得的广泛知识表征，以分布式参数形式编码了人类知识体系的丰富内容[1][2]。	隐性知识的转化困境	通过深层语义理解和情境感知能力，识别、提取并形式化教育实践中的隐性经验与专业智慧，在保持核心理念的同时实现情境化适应。
知识蒸馏（Knowledge Distillation）	将复杂模型的知识压缩并迁移到更精简模型的技术，保留关键能力的同时降低复杂度[3]。	知识扩散的规模化困境	从成功的教育创新中提炼核心机制，生成适应不同资源条件的简化方案，实现从资源密集型试点向大规模推广的有效转化。
生成式仿真（Generative Simulation）	基于概率推理和心理理论能力，模拟复杂系统中多主体的认知、情感	复杂系统中的知识适应困境	构建教育生态的生成式仿真，预测知识在不同层级、不同主体间的传播路径与适应过程，支持动态决策与风

[1] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners[C]//Larochelle H, Ranzato M, Hadsell R, et al. Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 1877-1901.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[C]//Guyon I, Luxburg U V, Bengio S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[3] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2025-09-27]. http://arxiv.org/abs/1503.02531.

AI 核心能力	核心概念与出处	应对的转化困境	适用性阐释
	与行为交互[1][2]。		险预判。
泛化能力（Generalization Ability）	AI 通过零样本或少样本学习，将已有知识迁移到新领域、新任务的能力[3][4]。	跨学科知识的整合困境	突破学科话语壁垒，在统一的语义空间中识别不同领域知识的深层联系，生成跨学科的整合性解决方案。

6.2.1 世界知识与隐性知识外显化

以大语言模型为代表的 AI 技术，以其广博的“世界知识”为基础，为破解隐性知识的转化困境提供了技术解法。通过对海量文本和数据的学习，这类 AI 技术不仅掌握了显性的教育理论，更在其参数网络中内隐地学习了蕴含在无数教育叙事、案例分析和实践反思中的经验模式。当一位教师尝试表达其教学智慧时，AI 能够凭借其强大的语言理解和生成能力，捕捉其话语中的深层含义，帮助其将模糊的、零散的思考变得结构化和清晰化，从而实现隐性知识的“外显化”。更重要的是，基于注意力机制[5]，这类 AI 技术具备了独特的情境感知能力。它不是僵化地存储知识，而是能够根据当前对话的上下文，动态地调用和重组最相关的知识，从而在守护一个教育理念核心原则的同时，生成高度适应特定情境的建议。

2024 年发表在《教育心理学评论》的一项研究表明[6]，大语言模型能够将概念性知识转化为经验性知识，并通过互动式对话帮助教师将其隐性的教学智慧外显化为可共享的教学策略。一个具体的案例是卡内基学习公司（Carnegie Learning）的 AI 驱动适应性学习

[1] Kosinski M. Evaluating large language models in theory of mind tasks[J]. Proceedings of the National Academy of Sciences, 2024, 121(45): e2405460121.

[2] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.

[3] Brown T, Mann B, Ryder N, et al. Language Models are Few-Shot Learners[C]//Larochelle H, Ranzato M, Hadsell R, et al. Advances in Neural Information Processing Systems: Vol. 33. Curran Associates, Inc., 2020: 1877-1901.

[4] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. (2023-01-10)[2025-09-26]. http://arxiv.org/abs/2201.11903.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need[C]//Guyon I, Luxburg U V, Bengio S, et al. Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.

[6] Huber S E, Kiili K, Nebel S, et al. Leveraging the potential of large language models in education through playful and game-based learning[J]. Educational Psychology Review, 2024, 36(1): 25.

平台[1]。该平台利用AI分析教师的教学反思日志，自动提取出隐含的教学模式和策略，并生成结构化的教学建议。研究显示，使用该系统的教师中有42%报告称AI帮助他们减少了行政任务时间，25%表示AI在个性化学习方面提供了显著帮助。

6.2.2 知识蒸馏与规模化扩散

通过“知识蒸馏”技术，以大语言模型为代表的AI技术为应对知识的规模化扩散困境提供了关键机制。知识蒸馏最初由辛顿（Hinton）等人[2]提出，其核心思想是将一个复杂、强大的“教师模型”的知识，压缩并迁移到一个更精简、高效的“学生模型”中。这一原理与教育创新的规模化推广需求高度契合。一项在资源丰富的学校被验证有效的复杂教学方案，往往难以直接在资源有限的学校实施。此时，AI可以扮演知识提炼与转化的角色：它能够深入分析复杂的原始方案，从中提炼出真正对效果起决定性作用的核心机制与关键要素，同时识别出哪些是可以根据情境调整的外围元素。近期研究表明，这一技术在教育领域的应用已经从简单的模型压缩扩展到了复杂教育方案的适应性迁移，能够将资源密集型的教育创新“蒸馏”为适合不同情境的轻量化版本[3]，从而实现从资源密集型试点向大规模、多样化情境的有效转化。

6.2.3 生成式仿真与系统复杂性应对

面对教育系统的复杂性，以大语言模型为代表的AI技术以其“生成式仿真”能力提供了前所未有的应对工具。近期多项研究表明[4][5]，如GPT-4等AI已经展现出相当程度的“心理理论”

[1] Slagg A. AI in education in 2024: educators express mixed feelings on the technology’s future[EB/OL]. [2025-08-28]. https://edtechmagazine.com/k12/article/2024/09/ai-education-2024-educators-express-mixed-feelings-technologys-future-perfcon.

[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[EB/OL]. (2015-03-09)[2025-09-27]. http://arxiv.org/abs/1503.02531.

[3] Cantini R, Orsino A, Talia D. Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices[J]. Journal of Big Data, 2024, 11(1): 63.

[4] Verma M, Bhambri S, Kambhampati S. Theory of mind abilities of large language models in human-robot interaction: an illusion?[C]//Companion of the 2024 ACM/IEEE International Conference on Human-robot Interaction. Boulder CO USA: ACM, 2024: 36-45.

[5] Kosinski M. Evaluating large language models in theory of mind tasks[J]. Proceedings of the National Academy of Sciences, 2024, 121(45): e2405460121.

（Theory of Mind）能力，在标准的错误信念任务中达到了较高水平，能够理解和预测人类的信念、意图和情感状态。

斯坦福大学 Park 等人的开创性工作尤其值得关注[1]。他们构建了包含 25 个生成式智能体（Generative Agents）的虚拟小镇，这些智能体能够模拟真实的人类社会行为，包括形成意见、相互交流、建立关系等。当研究者设定一个智能体想要举办情人节派对时，其他智能体会自主地传播邀请、建立新的社交关系、相约一起参加派对。将这一能力应用于教育领域，意味着可以构建出教育系统的“数字孪生”或“沙盘”。例如有研究者开发了 SimClassroom 系统[2]，能够模拟包含教师和多个学生智能体的完整课堂环境，用于测试新的教学方法和预测可能的实施效果。这种“事前预演”的能力，使得决策者和实践者能够超越传统的线性规划思维，更好地理解和驾驭复杂系统中的不确定性。

6.2.4 泛化能力与跨学科整合

大语言模型等 AI 技术的“泛化能力”（Generalization Ability）为突破跨学科知识整合的壁垒开辟了新路径。凭借其在高维语义空间中对不同领域知识的统一表征，这类 AI 技术能够发现不同学科术语、理论和方法论之间潜在的深层联系。例如，有研究展示了 AI 如何通过“知识迁移”将高资源语言（如 Python）的编程知识迁移到低资源语言（如 OCaml、Racket）[3]，这种跨域迁移的能力同样适用于教育领域的跨学科整合。通过这种跨领域的知识迁移和重组，AI 能够生成整合性的解决方案，有效回应长期存在的学科分割问题。特别是多模态 AI 模型的出现，使其能够整合文本、图像、数据等多

[1] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.

[2] Wen Q, Liang J, Sierra C, et al. AI for education (AI4EDU): advancing personalized education with LLM and adaptive learning[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona Spain: ACM, 2024: 6743-6744.

[3] Cassano F, Gouwar J, Lucchetti F, et al. Knowledge transfer from high-resource to low-resource programming languages for code LLMs[J]. Proceedings of the ACM on Programming Languages, 2024, 8(OOPSLA2): 677-708.

种来源的知识，进一步增强了其构建跨学科综合解决方案的能力。

综上所述，世界知识、知识蒸馏、生成式仿真与泛化能力相互协同，共同构成了一套 AI 赋能教育知识转化的整合性理论框架。它们系统性地回应了前述的隐性知识转化、规模化扩散、系统适应性与跨学科整合这四大结构性障碍，为在智能时代有效弥合“知行鸿-鸿沟”提供了兼具理论深度与实践前景的全新路径。

6.3 知识转化的技术路径

将 AI 的理论潜力转化为实践中的有效行动，需要超越零散的工具应用，设计一个贯穿知识转化全周期的整合性支持体系。借鉴实施科学研究中对转化过程的阶段性划分逻辑[1][2]，可以将 AI 赋能的知识转化解构为五个相互关联的核心环节（如图 6-2 所示）：证据的系统整合、策略的适应性转化、实施过程的保障、创新扩散与规模化，以及伦理与公平保障，为破解教育领域长期存在的“知行鸿沟”提供了系统性的操作蓝图。

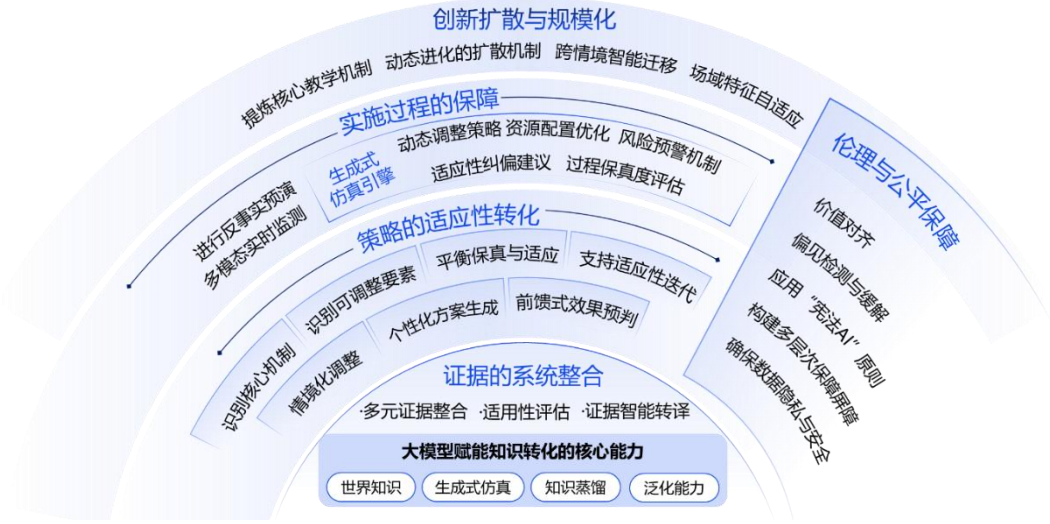

图 6-2 AI 赋能的教育知识转化路径框架

6.3.1 证据的系统整合

知识转化的逻辑起点，是对现有研究证据进行系统性综合与评

[1] Aarons G A, Hurlburt M, Horwitz S M. Advancing a conceptual model of evidence-based practice implementation in public service sectors[J]. Administration and Policy in Mental Health and Mental Health Services Research, 2011, 38(1): 4-23.

[2] Damschroder L J, Aron D C, Keith R E, et al. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science[J]. Implementation Science, 2009, 4(1): 50.

估，它旨在为后续的转化与实施提供坚实的循证基础。然而，这一环节在传统实践中面临着巨大挑战。首先是教育研究证据的高度异质性，它既包括基于随机对照试验的“硬”证据，也包括案例研究等“软”证据，如何有效整合并评估其综合价值是一大难题。其次是证据的情境依赖性，一项教育干预的效果高度依赖于实施情境，如何评估证据在新情境中的适用性至关重要。最后是从证据到行动的转译鸿沟，即如何将高度学术化的研究发现，转化为一线教师能够理解和操作的实践指南[1]。

AI 技术为此提供了构建证据综合智能化体系的可能，其核心价值在于从“信息检索”升级为“知识合成”与“智慧提炼”。以大语言模型为代表的 AI 技术凭借其世界知识与跨模态理解能力，能够高效处理不同来源的证据。最新工作表明，AI 结合检索增强生成（RAG）流水线能把传统的信息检索升级为“知识合成与因果链条重建”，既可自动抽取定量效应（如元分析中的效应量、样本特征），也能整合质性研究中有关实施机制与情境变量的描述，从而产出多维证据图谱[2][3]。

更进一步，该体系能够将证据的适用性评估转变为动态匹配过程，通过分析一项研究证据产生的原始情境特征（如学生年龄、社会经济背景、教师经验等），并与使用者当前所面临的目标情境进行智能比对，AI 能够预测该证据在新情境中可能的效果与风险，帮助实践者回答“什么方法，在哪里，对谁，以及在何种条件下最有效”这一循证实践的核心问题。最终，也是最具变革性的一点，该体系能够完成从“证据到行动”的智能转译。AI 的生成能力使其可以将充满学术术语的研究报告，转化为清晰、可操作的实践指南，

[1] Slavin R E. Perspectives on evidence-based research in education—what works? Issues in synthesizing educational program evaluations[J]. Educational Researcher, 2008, 37(1): 5-14.
[2] Luo X, Chen F, Zhu D, et al. Potential roles of large language models in the production of systematic reviews and meta-analyses[J]. Journal of Medical Internet Research, 2024, 26: e56780.
[3] Lieberum J L, Toews M, Metzendorf M I, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review[J]. Journal of Clinical Epidemiology, 2025, 181: 111746.

并根据不同受众（如新手教师、资深教师或学校管理者）的知识背景和需求，调整其语言风格和详略程度。如 GPT4EBP 平台所展示的，通过人机对话的交互形式，AI 可以引导教师将循证研究的原则，转化为一个真正适合其具体班级情况的教学方案[1]，从而从根本上降低循证实践的门槛。

6.3.2 策略的适应性转化

在完成对普适性证据的综合后，知识转化进入了针对特定情境进行“策略适应性转化”的核心环节。实施科学的一项核心洞察在于，任何干预措施都必须在“保真”（Fidelity）与“适应”（Adaptation）之间找到精妙平衡，既要保持干预的核心要素，又要根据本地条件进行必要调整。这正是解决“一刀切”方案在教育领域失灵的关键。

AI 的介入，为实现这种精妙平衡提供了前所未有的技术支持，能够将普适性的知识原则转化为高度情境化的实践策略。这一过程始于深度的情境诊断，AI 系统可通过与教育者的多轮结构化对话，全面诊断实践情境，这不仅包括学生的人口统计学特征、学业水平等“硬”数据，更涵盖课堂氛围、师生关系、教师自身的教学信念与风格、学校的资源与文化等难以量化的“软”信息，为后续的个性化定制奠定坚实基础。在此基础上，AI 并非简单推荐一个现成的方案，而是启动基于其核心原则的适应性生成机制。它将从证据中提炼出的核心教育原则作为“生成骨架”，将诊断出的本地情境特征作为“血肉”，动态生成一个或多个为当前情境量身定制的实践方案。例如，针对“文化相关教学法”（Culturally Relevant Pedagogy）这一原则，AI 可根据班级学生的文化背景，自动生成与学生生活经验高度相关的教学案例、调整教学语言风格、或推荐能够引发学生文化共鸣的补充阅读材料[2]。交互式教案生成系统（如

[1] 褚乐阳, 刘泽民, 王浩, 等. 大模型支持的教师循证实践：行动框架与案例应用[J]. 开放教育研究, 2024, 30(4): 91-103.

[2] Wang J, Xiao R, Hou X, et al. LLMs to support K-12 teachers in culturally relevant pedagogy: an AI literacy

LessonPlanner）可基于经典教学事件框架为不同经验水平教师生成分层教案草本，教师通过对生成内容的评判回馈与再编辑来形成最终方案[1]。

在技术层面，参数高效微调（PEFT）与低秩适配（LoRA / LoReFT 等）使得可以将通用的基础 AI 模型快速适配为面向具体教学任务或学科的专用 AI 模型，既节约算力又能保持教学机制的功能性保真，从而实现大规模可复制的个性化转译[2]。人机协同的迭代流程把 AI 作为“认知脚手架”，最终结果兼顾了证据内核与教师主体性。

6.3.3 实施过程的保障

即使拥有了完美的、量身定制的方案，知识转化的成功与否最终仍取决于实施过程的质量。大量研究表明，许多教育创新的失败，并非方案本身无效，而是“实施失败”，即干预方案在复杂多变的真实环境中未能被高质量地执行[3]。传统的实施支持（如定期的专家督导、实施日志等）往往存在反馈滞后、监测维度单一、难以捕捉复杂动态等局限。AI 的介入，则有望将实施支持从“事后复盘”的模式，升级为“事前预演”与“过程导航”的全新范式。

一方面，在正式实施一项新的教学干预之前，可以利用 AI 驱动的生成式智能体技术，构建一个模拟课堂或学校的“数字孪生”系统[4]。在这个虚拟沙盘中，可以输入新方案的各项参数，并设定不同的初始条件（如教师的熟练度、学生的配合度、不同家长的反应等），然后观察和预测该方案在系统中的演化轨迹。例如，近年出

example[EB/OL]. (2025-05-12)[2025-07-31]. http://arxiv.org/abs/2505.08083.

[1] Fan H, Chen G, Wang X, et al. LessonPlanner: assisting novice teachers to prepare pedagogy-driven lesson plans with large language models[C]//Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. Pittsburgh PA USA: ACM, 2024: 1-20.

[2] Balne C C S, Bhaduri S, Roy T, et al. Parameter efficient fine tuning: a comprehensive analysis across applications[EB/OL]. (2024-04-23)[2025-08-29]. http://arxiv.org/abs/2404.13506.

[3] Durlak J A, DuPre E P. Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation[J]. American Journal of Community Psychology, 2008, 41(3): 327-350.

[4] Park J S, O’Brien J, Cai C J, et al. Generative Agents: Interactive Simulacra of Human Behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco CA USA: ACM, 2023: 1-22.

现的 AI 驱动多智能体课堂仿真（SimClass/Simulating Classroom）为在上线前测试教学脚本、暴露流程薄弱点提供了低成本沙盘[1]。这种反事实的预演，能够帮助实施者提前识别潜在的风险点和关键的成功要素，从而在进入真实世界前就对方案进行优化，并制定更有针对性的应对策略。

另一方面，在实施过程中提供“过程导航”的实时动态支持。AI 系统可被授权分析课堂中的多模态数据流，如课堂对话的录音（分析提问类型、师生话轮转换频率）、学生的在线作业数据（分析常见错误点和学习进度），甚至在条件允许的情况下分析课堂录像（分析学生的参与度和专注度）。通过对这些数据的实时分析，AI 能够以一种远比传统观察量表更精细、更客观的方式，评估实施过程是否偏离了方案的核心原则。例如 Tutor CoPilot 的随机对照试验 AI 能在实地辅导情境中提升学生掌握率并修改教师即时策略[2]。这种即时的、非评判性的、高度情境化的反馈，如同一个经验丰富的“教练”在旁实时指导，能够极大地帮助教师在复杂的教学实践中不断调整和优化自己的行为，确保知识转化能够沿着正确的轨道前进。

6.3.4 创新扩散与规模化

知识转化的最终目标是从“局部有效”走向“系统采纳”，进入更具挑战性的创新扩散与规模化阶段。真正的规模化，不仅是数量上的扩展，更是在质量、持续性与实施主体主动性等多个维度上的同步推进[3]。然而，传统的教育创新推广方式往往陷于两种典型困境：一方面是过度强调“标准化”，忽视本土情境适配，导致方案在推广后“水土不服”；另一方面则是高度强调情境特异性，却

[1] Zhang Z, Zhang-Li D, Yu J, et al. Simulating classroom education with LLM-empowered agents[EB/OL]. (2024-11-27)[2025-08-28]. http://arxiv.org/abs/2406.19226.
[2] Wang R E, Ribeiro A T, Robinson C D, et al. Tutor CoPilot: a human-AI approach for scaling real-time expertise[EB/OL]. (2025-01-26)[2025-08-29]. http://arxiv.org/abs/2410.03017.
[3] Coburn C E. Rethinking scale: moving beyond numbers to deep and lasting change[J]. Educational Researcher, 2003, 32(6): 3-12.

无法在大范围内保持干预实施的保真度。这种张力迫切需要一种既具有结构稳定性又具备模块灵活性的实施设计逻辑。

以大语言模型为代表的 AI 技术为此提供了两大关键机制，有效回应了创新扩散与规模化的核心挑战。其一，通过知识蒸馏实现方案的轻量化迁移。该机制能够将一个在资源密集型环境下发展出的复杂干预方案，“浓缩”并提炼出其功能完整的核心教学机制，形成一个运行轻量的教学模型。这种蒸馏不仅是技术上的简化，更是策略结构的提纯，使得核心教学机制能够迁移至资源有限或结构差异显著的教育场景中。近年的一些研究已经展示了把复杂模型的推理能力“蒸馏”到轻量学生模型的可行方法，为在带宽/算力受限环境中部署教育 AI 大模型提供了实操路径[1][2]。

其二，借助泛化能力驱动策略的跨情境再生。通过学习多个成功案例，模型能够识别出跨越不同情境的共性逻辑、提炼迁移模式，并据此自动生成适配于新场域的推广策略。这种以模式识别为基础的策略重构，不仅提升了转化效率，更突破了人为归纳的瓶颈和机械复制的弊端，实现从“人找路径”到“模型推演”的革新。从更宏观的视角看，这类 AI 技术还为建立一个持续的创新生态系统提供了支持，通过不断收集和分析各地的实施数据，模型能持续优化推广策略，识别新的成功路径，并将这些经验反哺整个系统，形成创新的良性循环。

6.3.5 伦理与公平保障

当 AI 深度介入知识转化这一价值负载的实践活动，建立一个全面、稳健的伦理与公平保障框架就并非路径的最后一步，而是贯穿所有环节的价值前提与操作底线。

这首先要求进行严格的价值对齐（Value Alignment）。由于大

[1] Tian Y, Han Y, Chen X, et al. Beyond answers: transferring reasoning capabilities to smaller LLMs using multi-teacher knowledge distillation[EB/OL]. (2024-11-23)[2025-08-29]. http://arxiv.org/abs/2402.04616.

[2] Moslemi A, Briskina A, Dang Z, et al. A survey on knowledge distillation: recent advancements[J]. Machine Learning with Applications, 2024, 18: 100605.

语言模型在预训练阶段吸收了海量潜藏着各种社会偏见的人类数据，因此在应用于教育领域前，必须经过专门的微调过程，以确保其输出符合教育公平和多元包容的伦理标准。例如，以 Anthropic 公司提出的“宪法人工智能”（Constitutional AI）方法为代表的技术，并非依赖人工标注，而是通过让模型遵循一套明确的伦理原则宪法，进行自我监督和反复微调，使其行为与社会核心伦理预期保持一致[1]。

其次，必须建立持续的偏见检测与缓解机制。由于 AI 模型可能无意中习得甚至放大训练数据中固有的偏见，实施机构与开发人员必须建立持续的监控与校正流程。例如，Nadeem 等人[2]开发的 StereoSet 等方法已被成功用于量化评估模型中的社会刻板印象。在知识转化平台中引入类似工具，便可对 AI 生成的教学建议、评估反馈等内容进行实时扫描，一旦发现潜在的伦理风险，便能主动预警并启动纠偏程序。

此外，在利用多模态数据进行实施保障时，必须构建多层次的数据隐私与安全屏障，遵循相关法律法规与伦理原则，确保所有数据的采集、存储和使用都在合法、合规、透明的前提下进行，切实保护师生权益。协同运用以上机制，可为 AI 支持下的教育知识转化全过程，构建起一个从事前预防（价值对齐）、事中监控（偏见检测）到持续优化的多层次伦理保障体系。

知识转化是连接教育研究与实践的关键枢纽，其效能直接决定了教育研究能否真正服务于教学改进与学生发展。本章系统阐述了 AI 如何通过世界知识、知识蒸馏、生成式仿真与泛化能力这四项核心机制，破解教育知识转化的结构性障碍，并构建了从证据整合到伦理保障的技术路径体系。AI 赋能教育知识转化的意义不仅在于为

[1] Bai Y, Kadavath S, Kundu S, et al. Constitutional AI: harmlessness from AI feedback[EB/OL]. (2022-12-15)[2025-09-27]. http://arxiv.org/abs/2212.08073.

[2] Nadeem M, Bethke A, Reddy S. StereoSet: measuring stereotypical bias in pretrained language models[EB/OL]. (2020-04-20)[2025-09-27]. http://arxiv.org/abs/2004.09456.

弥合“知行鸿沟”提供了可操作的解决方案，更在于它深刻体现了AI4S-Ed的核心价值主张：运用AI变革教育研究过程本身。如果说传统教育研究受制于人力成本和认知局限，那么AI的介入则从根本上重构了知识转化的可能性边界，从线性的、延迟的、低效的转化模式，升级为动态的、实时的、智能化的知识流动生态。在AI4S-Ed的整体版图中，知识转化既是研究成果走向实践应用的“最后一公里”，也是实践智慧反哺理论创新的第一入口，构成了教育研究闭环中最具变革潜力的关键环节。

第 7 章 教育研究中新的伦理考量

随着人工智能技术在教育研究中的深度应用，研究伦理议题已超越传统框架。以往的伦理规范多聚焦于防范直接伤害、保障知情同意以及通过数据匿名化来保护个体隐私。然而，人工智能技术引发的风险更具系统性与延迟性，表现为数据跨境流动导致的隐私隐患[1]、算法偏差加剧教育不平等[2]，以及弱势群体代表性的缺失[3][4]。本章将对这些新兴伦理挑战进行梳理，揭示人工智能介入所引发的伦理范式的转变。同时，还将梳理政策制定者与研究者近年来提出或倡导的实用性伦理指南与治理框架，以回应教育研究在实践中的迫切需求。

7.1 数据安全与隐私治理

人工智能推动教育研究进入数据丰饶与风险并存的时代。细粒度、可支撑强大语言模型的数据，往往伴随着更高的隐私暴露与再识别风险。传统研究中，数据与模型的关系常被视作简单的、一次性的映射，但在现代 AI 模型中，数据一旦进入模型，便可能在持续训练与迁移学习中被反复调用。这意味着教育场景中的隐私治理，尤其涉及未成年人的研究，不能再依赖单次同意书与去标识化，而需转向以生命周期为单位、技术与制度并重的治理框架[5][6]。

教育数据的敏感性不仅来自身份标识，还源于语境耦合。学习日志、课堂视频、语音、作业文本、讨论互动、考试监控与行为检测等多模态数据，一旦拼接，便可能重构学生的学习轨迹、心理状态乃至家庭背景。即便在表面匿名的情况下，时间戳、地理线索、

[1] Frueh S. How AI is shaping scientific discovery[J]. National Academies, 2023.
[2] 沈苑,汪琼.人工智能教育应用的偏见风险分析与治理[J].电化教育研究,2021,42(08):12-18.
[3] Hartman S, Ong C S, Powles J, et al. Position: we need responsible, application-driven (RAD) AI research[EB/OL]. (2025-05-07)[2025-10-21]. http://arxiv.org/abs/2505.04104
[4] 刘革平,秦渝超.论人工智能作为教育类主体[J].教育研究,2025,46(05):30-42.
[5] United Nations Committee on the Rights of the Child. General comment No. 25 on children's rights in relation to the digital environment[R]. Geneva: Office of the United Nations High Commissioner for Human Rights, 2021.
[6] UNESCO. Recommendation on the ethics of artificial intelligence[R]. Paris: United Nations Educational, Scientific and Cultural Organization, 2021.

语言风格、设备指纹等仍可能实现再识别，其风险往往具有滞后性，并在后续调用与模型更新中逐渐显现[1]。近年的研究进一步强调了利益相关者之间的张力。教师、家长和开发者普遍担忧教育数据在被采集后会被不当使用，尤其是当未来用途不明确时。这种担忧不仅关乎技术漏洞，也涉及对教育主体自主权的侵蚀。新的研究提出隐私伦理协同的概念，认为不同群体对于隐私的期待差异较大，需要通过制度和协商机制予以回应。

在技术路径上，隐私增强技术逐渐成为共识，其中以联邦学习和差分隐私最具代表性。前者通过在数据源端本地训练，仅上传模型参数，由中心端聚合，从而避免原始数据跨域流动，适用于多校、多平台与多终端的教育生态，但也面临非独立同分布、模态异构与通信开销等挑战，需要配套鲁棒聚合策略[2]。差分隐私则通过在统计结果或模型参数中注入噪声，为个体的不可识别性提供形式化保障，但在教育数据规模较小、目标变量稀疏的情境下，噪声注入常显著降低模型性能，隐私预算的设定也需在可用性与保护度之间进行权衡[3]。

然而，仅依赖技术并不足以解决隐私风险。教育研究应将隐私治理工程化，即将原则落实到研究全生命周期：在前置环节遵循最小化原则，明确必需、可选与禁止采集的数据范围，并评估多模态拼接带来的再识别风险；在知情同意环节采用动态可撤回机制，在数据清洗、模型发布、跨机构共享等关键节点实现再告知与再授权；在使用与审计环节构建全链条追踪体系，对未成年人数据设置更短保留期与默认不跨境传输；最后，在成果发布与再利用环节建立条件开放机制，对超出原始研究目的的二次利用设立审查程序，并在

[1] UNESCO. Artificial intelligence and education: guidance for policy-makers[R]. Paris: United Nations Educational, Scientific and Cultural Organization, 2021.
[2] United Nations Children's Fund (UNICEF). Policy guidance on AI for children[R]. New York: UNICEF, 2021.
[3] Université de Montréal. Montréal declaration for a responsible development of artificial intelligence[EB/OL]. (2018-12-04)[2025-10-21]. https://www.montrealdeclaration-responsibleai.com/.

.

结果发布前开展隐私攻击演练，披露评估与缓解措施[1][2]。

值得注意的是，教育现场的隐私风险常伴随效率叙事而出现。例如，自动监考与课堂行为分析常被包装为提升效率与保障公平，但其代价可能是常态化的敏感数据采集与不成比例的监控扩张，从而越过必要性与相称性的边界[3]。跨境数据流动亦构成突出挑战。多校合作与国际比较研究往往涉及数据跨境传输，应优先采用联邦学习与安全聚合，使模型可移动而数据不流动。若必须传输汇总数据，则应进行再匿名化与风险评估，并在协议中明确用途限制与再转移禁止条款[4][5]。

由此，开放科学与隐私保护的均衡日益重要。数据共享不等于原始数据无条件开放，更合理的方式是分层开放：教学设计与代码尽量公开，数据则根据敏感度分为合成数据或强脱敏样本、受控访问与封闭三类。隐私治理应被视为研究质量的重要组成部分，明确的数据沿袭与可追踪的再利用机制不仅提升研究的可信度，也有助于赢得教育机构与公众的长期信任[6][7]。

综上，教育研究中的隐私与安全保护应被重构为技术与流程并行的复合体系。已有研究揭示了制度缺位带来的风险[8]，也强调了技术路径的可能与局限[9][10]。两者的结合提示我们：教育研究中的

[1] Organisation for Economic Co-operation and Development. Recommendation of the Council on Artificial Intelligence[R]. Paris: OECD, 2019.
[2] High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI[EB/OL]. (2019-04-08)[2025-10-21]. Brussels: European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[3] Selvakumar P, Sudheer P, Kannan N. Balancing innovation and privacy: understanding AI in the digital world[M]// Digital Citizenship and the Future of AI Engagement, Ethics, and Privacy. Hershey, PA: IGI Global, 2025: 279–304.
[4] Fjeld J, Achten N, Hilligoss H, et al. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI[R]. Cambridge, MA: Berkman Klein Center, Harvard University, 2020.
[5] Hagendorff T. The ethics of AI ethics: an evaluation of guidelines[J]. Minds and Machines, 2020, 30(1): 99–120.
[6] Mittelstadt B. Principles alone cannot guarantee ethical AI[J]. Nature Machine Intelligence, 2019, 1(11): 501–507.
[7] Whittlestone J, Nyrup R, Alexandrova A, et al. The role and limits of principles in AI ethics: towards a focus on tensions[C]// Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES). Honolulu, HI: ACM, 2019: 195–200.
[8] Ryan M, Stahl B C. Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications[J]. Journal of Information, Communication and Ethics in Society, 2021, 19(1): 61–86.
[9] Floridi L, Cowls J. A unified framework of five principles for AI in society[J]. Harvard Data Science Review, 2019, 1(1).
[10] Schiff D, Borenstein J, Laas K, et al. AI ethics in the public, private, and NGO sectors: a review of a global document collection[J]. IEEE Transactions on Technology and Society, 2021, 2(1): 31–42.

数据安全与隐私保护不应仅停留于合规要求，而应成为促进高质量研究与社会信任的积极力量。

7.2 算法公平与偏见管理

不同于传统研究中研究者对方法与变量的直接掌控，人工智能系统的决策不仅受设计影响，更深受数据质量与模型机制的制约。一旦训练数据存在偏差，算法往往在结果中再生产甚至放大不平等，从而对教育公平构成系统性威胁[1][2]。

在教育场景中，这一问题表现得尤为突出。学业预测模型通常依赖在线平台的大规模行为数据，但弱势群体或农村地区学生的数据往往稀缺，其学习特征在训练中被低估甚至忽视。结果是，模型更容易识别主流群体的模式，却难以准确反映边缘群体的真实情况。这种代表性不足直接导致预测误差，并可能在资源分配和教学干预中制造新的不公平[3][4]。进一步研究表明，当算法缺乏对社会文化与家庭背景等变量的考量时，弱势学生往往被过度识别为“高风险”，从而承受额外的标签化压力[5]。

算法的不公平不仅源于数据，还来自模型逻辑的黑箱性。许多机器学习算法在追求预测准确率时，往往牺牲了可解释性，使研究者和教育实践者难以识别其中潜在的偏差。例如，作文自动评分系统可能因训练语料主要来源于特定语言风格，而对具有方言特征或跨文化表达的学生表现出系统性不利。这种模型内隐偏差往往隐蔽而持久，对教育公平的侵蚀比数据偏差更为复杂[6]。

[1] Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting[C]// Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*)*. Atlanta, GA: ACM, 2019: 220–229.
[2] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389–399.
[3] Raji I D, Scheuerman M K, Amironesei R, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing[C]// Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency. Barcelona: ACM, 2020: 33–44.
[4] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets[J]. Communications of the ACM, 2021, 64(12): 86–92.
[5] Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification[C]// Proceedings of the 1st ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2018: 77–91.
[6] Corbett-Davies S, Pierson E, Feller A, et al. Algorithmic decision making and the cost of fairness[C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

应对这一问题需要多维度路径。在技术层面，公平感知算法、去偏技术与对抗性训练等方法能够在一定程度上缓解预测中的群体差异。例如，在训练过程中引入敏感属性控制变量，或通过重加权机制平衡不同群体的样本比例，从而降低结果偏差[1]。在制度层面，则需要建立透明性与问责机制。研究者在发表成果时，应披露数据来源、样本分布及可能的偏差来源，并对模型在不同群体中的表现进行公平性评估。这种可解释性与公平性审查的结合，有助于避免教育研究在不知不觉中固化不平等[2]。更进一步，算法公平性还涉及跨文化语境与价值取向的差异。在多元教育环境中，不同文化群体对公平的理解本身可能存在差异。因此，教育研究不仅要关注技术上的去偏，还需在议题设计与制度规范中引入多元价值观，确保弱势群体和少数文化群体在研究议程中拥有充分代表性[3]。

现有的研究已经表明，算法公平性与偏见管理应当成为教育研究中与隐私保护并行的重要伦理支柱。技术手段可以在一定程度上缓解偏差，但唯有与制度设计和价值导向相结合，才能避免人工智能在教育研究中成为放大不平等的工具，而真正发挥推动教育公平与社会正义的积极作用。

7.3 跨文化与弱势群体保护

人工智能在教育研究中的迅速扩展，使跨文化差异和弱势群体处境成为新的伦理焦点。传统研究强调对不同文化背景和社会群体的敏感性，而人工智能的引入却常在不经意间掩盖这些差异，导致研究议程和结果不可避免地呈现出主流偏向[4][5]。

(KDD). Halifax, NS: ACM, 2017: 797–806.
[1] Kizilcec R F, Lee H. Algorithmic fairness in education[M]// The Ethics of Artificial Intelligence in Education. London: Routledge, 2022: 174–202.
[2] Das S, Stanton R, Wallace N. Algorithmic fairness[J]. Annual Review of Financial Economics, 2023, 15(1): 565–593.
[3] Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning[C]//Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS). 2016: 3315–3323.
[4] Chouldechova A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments[J]. Big Data, 2017, 5(2): 153–163.
[5] Holmes W, Miao F. Guidance for generative AI in education and research[M]. Paris: UNESCO Publishing, 2023.

跨文化偏差首先体现于训练数据的单一性[1]。自动作文评分系统便是典型案例。如果训练数据主要来源于英语母语学生，非母语学生的语言风格与表达方式往往被系统性低估。这种隐性歧视不仅会影响成绩评估，还可能长期削弱学生的学习动机与学术自我效能感。类似风险同样出现在情感识别与课堂互动分析中。不同文化群体在表情、手势与语调上的差异显著，但若算法以单一文化标准为参照，就可能误判其他文化学生的情绪与参与度，使其在评价体系中长期处于不利位置[2][3]。

弱势群体的脆弱性在人工智能研究中表现得尤为突出。一方面，人工智能可能为资源不足群体提供支持，例如个性化学习平台能够为学习困难学生提供定制化路径。另一方面，这些平台通常伴随高强度数据采集与持续监控，而弱势群体由于缺乏话语权与法律保障，更容易成为数据剥削的对象。已有研究表明，机构在隐私保护方面的缺位，对弱势群体的影响尤为严重，因为他们缺乏足够能力对不当使用提出质疑或寻求救济[4]。这种风险在实践中已有体现。例如，高风险学生干预研究常利用学习日志预测辍学可能性。虽然有助于教育机构提前干预，但弱势群体的学习行为往往因家庭条件、网络不稳定或文化差异而呈现不同模式。这些差异极易被模型误判为风险信号，从而导致弱势学生被长期贴上问题标签。一旦固化，这种标签不仅限制其教育机会，还可能加深社会对其的负面认知。更为严重的是，人工智能系统常将弱势群体的经验噪声化，即在追求整体性能最优时忽略异质性数据点。由此，弱势群体在知识生产中被算法性抹除，教育研究逐渐呈现表面干净却内在失真的现实[5]。

[1] 陈向东,卢淑怡,易乐湘.文化冲突:大语言模型教育应用中的张力与调适[J].远程教育杂志,2025,43(03):15-43.

[2] Kusner M, Loftus J, Russell C, et al. Counterfactual fairness[C]// Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). Long Beach, CA: Curran Associates, Inc., 2017: 4066–4076.

[3] Barocas S, Selbst A D. Big data's disparate impact[J]. California Law Review, 2016, 104(3): 671–732.

[4] Friedler S A, Scheidegger C, Venkatasubramanian S, et al. A comparative study of fairness-enhancing interventions in machine learning[C]// Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*)*. Atlanta, GA: ACM, 2019: 329–338.

[5] Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification[C]// Proceedings of the 1st ACM Conference on Fairness, Accountability, and Transparency

为应对上述风险，研究者提出了伦理审查机制强化、数据多样性优化与参与式研究方法等不同的解决路径。其中，参与式研究方法的视角认为，跨文化与弱势群体不应仅作为研究对象，更应成为议题设定的共同参与者。在研究问题提出、数据使用与结果解释等关键环节，邀请其代表参与，不仅能提升研究的伦理合法性，还能增强结果的现实有效性。例如，在作文评分研究中，应让非母语学生与教师共同参与标准制定，确保训练语料涵盖多元文化表达；在学习行为预测研究中，应通过对弱势群体的访谈理解其特有的学习挑战，以避免算法误将差异解读为劣势[1][2]。

综上，跨文化与弱势群体的伦理考量并非教育研究的附属议题，而是人工智能介入教育研究后亟需直面的挑战。已有研究提醒我们代表性缺失的风险，也揭示弱势群体在隐私保护缺位时首当其冲。真正的解决路径在于技术优化与制度保障并行，更在于参与式研究与全球知识平等的共同推动。

7.4 伦理指南与治理框架

面对算法技术的广泛介入，传统以研究者自律和机构审查为核心的伦理体系已难以充分应对跨数据、跨算法与跨地域的复杂风险。研究治理因此需要从原则层面的宣示转向制度层面的落实，从静态规范走向动态监管。有效的伦理治理不再仅依赖抽象的价值共识，而是要求在价值导向与制度执行之间建立双重支撑，使伦理审查、风险评估与问责机制相互衔接、形成闭环，为 AI 时代的教育研究提供持续的规范保障。

7.4.1 教育研究伦理指南

传统的研究伦理主要依赖研究者自律与机构审查，关注研究过

(FAT*)*. New York: ACM, 2018: 77–91.

[1] Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting[C]// Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*)*. Atlanta, GA: ACM, 2019: 220–229.

[2] Raji I D, Buolamwini J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products[C]// Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES). Honolulu, HI: ACM, 2019: 429–435.

程中个体层面的风险与直接伤害。当人工智能逐渐成为研究对象、研究工具和知识生产机制的重要组成部分后，伦理风险的性质与边界随之发生深刻变化，呈现出跨算法、跨数据和跨地域的系统性特征。为应对这些新兴挑战，国际社会、各国政府及学术组织陆续制定了一系列人工智能伦理指南和政策框架，旨在为教育研究的技术应用提供价值导向和制度依据。这些指南共同推动伦理治理从原则倡议走向制度实践，成为教育研究在人工智能时代实现安全、可控与公平发展的重要支撑。

联合国教科文组织（UNESCO）发布的《人工智能伦理建议书》是首个全球性的 AI 伦理规范性文件[1]。该文件确立了“以人权为中心”的治理逻辑，提出了人类尊严、公平正义、多样性与包容性等核心原则，并明确要求成员国在教育与科研领域建立 AI 伦理治理机制。这份文件的重要意义在于，它首次将教育研究界纳入 AI 伦理治理体系之中，要求教育机构在数据收集、模型使用与科研传播中防止算法偏见和文化排他。UNESCO 随后发布的《生成式人工智能与教育指南》进一步提出[2]，应在高校和研究机构内普及 AI 伦理素养，建立数据透明度、模型问责与风险报告制度。这一系列文件标志着教育研究从“伦理被动防御”走向“伦理前置治理”的转向。

与 UNESCO 的价值导向框架相呼应，经济合作与发展组织（OECD）提出了《人工智能原则》，并在 2024 年对其进行了修订和扩展[3]。OECD 倡导“以人为本的可信 AI”，强调包容、公平、安全、透明与问责五项核心价值。2024 年的更新中引入“敏捷治理”理念，要求各成员国构建覆盖 AI 全生命周期的监管体系。这一框架

[1] United Nations Educational, Scientific and Cultural Organization (UNESCO). Recommendation on the ethics of artificial intelligence[R/OL]. Paris: UNESCO, (2021) [2025-10-21]. https://unesdoc.unesco.org/ark:/48223/pf0000381137

[2] United Nations Educational, Scientific and Cultural Organization (UNESCO). Guidance for generative AI in education and research[R/OL]. Paris: UNESCO, (2023) [2025-10-21]. https://unesdoc.unesco.org/ark:/48223/pf0000386829

.

[3] Organisation for Economic Co-operation and Development (OECD). Updated OECD principles on artificial intelligence (2024 revision)[R]. Paris: OECD, 2024a.

不仅强调规范制定，还提出“可验证的伦理审查机制”。OECD 下设的 AI 政策观察站收集并分享各成员国在科研与教育场景中落实伦理治理的实践案例，包括模型解释性测试、教育数据风险分级与算法公平性评估等，为教育研究提供了可操作的制度参考[1]。此外，联合国秘书长在 2024 年发布的《Governing AI for humanity: Final report》[2][3]提出了“预期治理”的概念，主张 AI 治理应提前识别潜在风险并在政策层面建立防范机制。该报告特别强调在科研和教育领域加强跨国合作与数据流动监管，防止算法殖民与知识生产的不平等。同年，七国集团（G7）通过《G7 Hiroshima Process international guiding principles for advanced AI systems》[4]，进一步细化了高风险 AI 系统的伦理标准，要求教育科研机构在使用生成式 AI 时履行透明披露与责任追踪义务。这些国际框架共同形成了从价值倡导到制度落实的演化路径，为教育研究伦理治理提供了系统的上位设计。

国家与区域层面的政策同样体现出 AI 治理的加速制度化。欧盟于 2024 年正式通过《人工智能法案》[5]，成为全球首部系统性 AI 立法文件。该法案采取风险分级管理模式，将教育与职业培训明确定义为高风险场景，要求相关机构在系统设计与使用过程中开展伦理与人类监督评估。法案还设立了算法透明度报告与伦理备案机制，规定教育研究中涉及 AI 系统的项目需提交风险评估文件并接受独立审查。这一立法标志着教育研究伦理进入可问责、可追踪的法律治理阶段。此外，中国于 2024 年发布《人工智能安全治理框架 1.0》

[1] Organisation for Economic Co-operation and Development (OECD). AI Policy Observatory[R/OL]. Paris: OECD, 2024b [2025-10-21]. https://oecd.ai/en/observatory.
[2] United Nations (UN). Governing AI for humanity: final report[R/OL]. New York: United Nations, 2024b [2025-03-25]. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf.
[3] United Nations (UN). UN News – global perspective human stories[EB/OL]. (2024-09-26)[2025-10-21]. https://news.un.org/zh/story/2024/09/1131551.
[4] Snoswell A. G7 Hiroshima Process international guiding principles for advanced AI systems[R]. Tokyo: G7 Cabinet Office, 2023.
[5] European Parliament & Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 on artificial intelligence (AI Act)*. Official Journal of the EU.

[1]，提出“包容审慎、安全可控、风险导向、敏捷治理、协同共治”的五项基本原则。该框架要求在科研与教育领域建立 AI 应用的全生命周期伦理风险评估制度，覆盖立项、研发、测试与发布等各个环节。文件特别强调数据安全、算法透明与科研诚信的统一治理，并提出由科研机构主导、教育部门协同、社会公众参与的多方治理结构。与欧盟模式相比，中国的框架更加注重制度协调与社会责任的结合，体现出对 AI 伦理的“系统工程化”倾向。

在英美等国，伦理治理体现出更强的灵活性与实践导向。美国 2023 年颁布的《安全、可信 AI 行政令》[2]要求所有联邦资助的 AI 研究项目在立项阶段提交“伦理影响声明”，并实施偏见检测与算法安全评估。这一制度将伦理从审查阶段前移到研究设计阶段，强调科研责任与可追溯性。英国政府同期发布《AI Regulation: A Pro-Innovation Approach》[3]，提出以“促进创新”为前提的柔性监管路径。该政策鼓励高校和科研机构设立 AI 伦理委员会与风险登记系统，在保持技术创新活力的同时强化研究者问责。这两种路径分别代表了监管驱动与自律驱动的伦理治理模式，为教育研究机构提供了不同的制度参照。除国家与国际政策外，行业与学术组织也在提供更具操作性的技术性伦理标准。其中，IEEE 发布的 7001、7002 与 2089 系列标准构成了人工智能伦理治理的核心技术规范体系[4]。IEEE 7001 强调系统透明度与可解释性，7002 提出隐私影响评估方法，2089 则聚焦适龄数字服务设计。这些标准在教育研究领域具有直接适用性，能够转化为研究项目的伦理审查清单，用以评估数据采集、模型训练和结果传播的伦理合规性。与此同时，美国教育研

[1] 国家标准化管理委员会. 人工智能安全治理框架 1.0[EB/OL]. (2024-10-09)[2025-10-10]. https://www.sac.gov.cn/sjzx/gzdt/202410/t20241009_568321.html.

[2] The White House. Executive order on safe, secure, and trustworthy artificial intelligence[R/OL]. Washington, D.C.: The White House, 2023 [2025-10-21]. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-safe-secure-and-trustworthy-ai/.

[3] UK Department for Science, Innovation and Technology (DSIT). AI regulation: a pro-innovation approach[R]. London: HM Government, 2023.

[4] IEEE Standards Association. IEEE 7000 series standards (7001, 7002, 2089)[R]. New York: IEEE, 2021–2023.

究协会发布的《AI 在教育研究中的伦理使用声明》[1]指出，研究者应明确披露 AI 工具的使用方式与影响，确保研究结果的可重复性与公平性。

从这些多层次举措中可以看出，全球 AI 治理的核心趋势是从价值倡导向制度实践转型。AI 教育研究的伦理指南正从价值宣示迈向可验证与可问责的制度体系：国际组织提供共识框架，国家与地区立法与政策将其法制化，行业标准把伦理目标转译为工程与流程要求，机构工具与清单进一步把要求落到研究与教学情境之中。这一多层结构使伦理不再是外部约束，而被内嵌为教育研究全流程中的常态化环节。

7.4.2 伦理审查清单与流程设计

人工智能的引入正在推动教育研究的伦理审查体系发生结构性转变。传统的审查机制通常在研究立项阶段进行一次性评估，侧重对潜在风险的事前识别与伦理批准。然而，人工智能研究具有高迭代性与强数据依赖性，其伦理风险并非静态存在，而是在算法优化、数据更新和结果应用的循环过程中不断生成与演化[2][3]。这使得伦理治理必须从静态合规转向动态监管，从单点审批转向持续评估。为回应这一变化，国际组织、专业协会以及高等教育机构陆续推出以“伦理清单与流程设计”为核心的 AI 审查工具，为教育研究领域提供了可借鉴的路径。

欧盟《人工智能法案》[4]首次规定高风险 AI 系统（包括教育与培训领域）必须在部署前完成“合规性评估”，并提交风险管理文件、数据治理说明与模型透明度报告。该法案形成了一套由风险识

[1] American Educational Research Association (AERA). Statement on ethical use of AI in educational research[R]. Washington, D.C.: AERA, 2023.

[2] Floridi L, Cowls J, Beltrametti M, Chatila R. AI ethics: its nature, importance, and practical challenges[J]. Philosophy & Technology, 2022, 35(4): 1–25. DOI:10.1007/s13347-022-00545-3.

[3] Leslie D. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector[R/OL]. London: The Alan Turing Institute, 2021 [2025-10-21]. https://doi.org/10.5281/zenodo.4404235.

[4] European Parliament and Council. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 March 2024 on artificial intelligence (AI Act)[R/OL]. Official Journal of the European Union, 2024 [2025-10-21]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

别、影响评估、合规备案和后续监测构成的闭环流程。与之配套的欧盟人工智能办公室提出了《高风险 AI 伦理与合规清单》，包括算法可解释性、训练数据代表性、用户知情权与申诉机制等条目。这种基于清单的监管结构强调可验证性与可问责性，逐渐成为国际通行的 AI 伦理治理模式[1]。

行业与专业组织层面，IEEE 发布的 IEEE 7000 系列标准为伦理设计提供了系统化模板。其中，IEEE 7001《系统透明性标准》提出在研究中建立算法决策的追踪日志；IEEE 7002《隐私影响评估》提供了从数据采集、建模到发布的逐项评估流程；IEEE 2089《适龄数字服务标准》则强调对未成年人数据的特殊保护[2]。同时，OECD 的 AI Policy Observatory 收录了多国教育科研机构的伦理审查工具包，如加拿大的“Algorithmic Impact Assessment（AIA）”、美国国立卫生研究院的“AI Research Ethics Checklist”、以及澳大利亚教育部开发的“Responsible AI in Education Self-Assessment Tool”[3]。这些清单均要求研究者在立项前明确 AI 介入的范围、目的、数据类型与风险等级，并在研究后期提交伦理合规报告。

此外，一些机构尝试引入“伦理沙盒”机制，即在受控环境中测试 AI 教育工具的伦理风险。英国教育部[4]与澳大利亚研究委员会[5]都已采用该方法，用以提前识别研究中可能出现的隐私、偏见与社会影响问题。此类机制强调实验性监管与学习型治理，为教育研究领域提供了低风险的伦理创新空间。

总体来看，伦理审查清单与流程的演化反映出治理理念的三重转向：从静态审批走向动态评估，从研究外部审查走向内部嵌入，

[1] European Union Artificial Intelligence Office (EU AI Office). Checklist for high-risk AI ethics and compliance[R]. Brussels: European Commission, 2024.
[2] IEEE Standards Association. IEEE 7000 series standards (IEEE 7001, IEEE 7002, IEEE 2089)[R/OL]. New York: IEEE, 2021–2023 [2025-10-21]. https://standards.ieee.org.
[3] Organisation for Economic Co-operation and Development (OECD). AI Policy Observatory[R/OL]. Paris: OECD, 2024b [2025-10-21]. https://oecd.ai/en/observatory.
[4] Department for Education (DfE). AI governance and ethical sandbox framework[R]. London: UK Government, 2024.
[5] Australian Research Council (ARC). Responsible AI in education self-assessment tool[R/OL]. Canberra: ARC, 2024 [2025-10-21]. https://www.arc.gov.au.

从被动合规走向反思性实践。其审查流程通常包括四个阶段：

- 预评估阶段：研究者需提交 AI 应用说明书，阐明算法功能、数据来源与潜在风险；
- 正式审查阶段：委员会依据伦理清单（涵盖隐私保护、公平性、透明度、未成年人权益等维度）进行分级评估；
- 动态监测阶段：研究者在模型更新或数据扩展时需提交补充报告；
- 结果披露阶段：要求在研究成果中公开 AI 工具使用范围与伦理影响声明。

教育研究的伦理治理不应仅依赖统一模板，而需建立适应性强、可扩展的审查体系，使伦理评估成为研究过程的组成部分而非外在负担。未来的 AI 教育研究伦理机制将更注重跨机构协同与持续问责，实现从“伦理备案”到“伦理生态”的结构性升级[1][2]。

7.4.3 自律机制与外部监管平衡

当人工智能深度嵌入教育研究后，单纯依赖研究者的自律已不足以应对复杂的伦理风险。从数据隐私泄露到算法偏见，从跨境流动到成果再利用，这些问题往往超越单一项目或团队的控制范围。因此，教育研究的治理核心在于在“自律机制”和“外部监管”之间寻求平衡。前者强调研究机构和研究者的内部责任与自我约束，后者依赖国家、国际组织和第三方的监督与规制。二者既存在张力，又相互补充，只有结合起来才能形成稳健的治理体系。

在自律层面，有效治理需要在研究机构内部建立清晰的责任链条。已有研究指出，项目负责人应成为研究设计与伦理适配的第一责任人，数据管理员需保障数据全生命周期的安全与合规，伦理专员则负责监督人工智能模块的透明性与偏差检测。学术委员会在其

[1] Floridi L, Leslie D, Cowls J. A governance framework for trustworthy AI[J]. AI & Society, 2023, 38(2): 389–405.

[2] United Nations Educational, Scientific and Cultural Organization (UNESCO). Guidance for generative AI in education and research[R/OL]. Paris: UNESCO, 2023 [2025-10-21]. https://unesdoc.unesco.org/ark:/48223/pf0000386829.

中扮演争议复核与裁决角色，从而形成多角色协作的内部问责机制。通过这种制度化分工，伦理治理能够嵌入日常研究流程，而不再仅依赖临时性的判断。美国教育研究协会发布的《AI 在教育研究中的伦理使用声明》强调[1]，教育研究者应在研究方案中主动披露 AI 工具的使用范围与影响，以确保学术共同体内部形成责任共识。英国高等教育技术联合会[2]开发的 AI 教学与研究伦理工具包为研究机构提供了内部审查模板和决策支持工具。这些实践说明，自律机制不仅是个人道德要求，更是一种制度化的集体责任。然而，自律机制也存在明显局限：容易沦为形式化合规文件的堆砌，缺乏对潜在风险的实质干预；研究团队兼任执行者与审查者时可能出现利益冲突；治理效果还高度依赖机构资源与能力。重点高校往往能建立较完善的伦理委员会，而中小机构可能仅能维持最低限度的文档合规[3]。

因此，外部监管成为必要补充。国家法律、政府政策和国际组织为教育研究提供制度化的纠偏机制。欧盟《通用数据保护条例》明确要求敏感教育数据本地存储，并对跨境传输设定严格限制，以保障学生隐私权[4]。美国教育部在《人工智能与教学的未来》报告中提出，应在联邦资助项目中设立 AI 伦理审查委员会，并要求研究机构提交 AI 影响评估报告[5]。中国发布的《人工智能安全治理框架 1.0》则提出协同共治与敏捷治理原则，推动教育科研机构与政府部门共同监管高风险 AI 应用[6]。此外，部分国家已将“公平性审计报告”作为科研资助与成果评估的前置条件，要求教育类研究在算法

[1] American Educational Research Association (AERA). Statement on ethical use of AI in educational research[R/OL]. Washington, D.C.: AERA, 2023 [2025-10-21]. https://www.aera.net.

[2] Joint Information Systems Committee (JISC). AI ethics toolkit for teaching and research[R/OL]. Bristol: JISC, 2023 [2025-10-21]. https://www.jisc.ac.uk.

[3] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines[J]. Nature Machine Intelligence, 2019, 1(9): 389–399.

[4] European Parliament and Council. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR)[R/OL]. Official Journal of the European Union, 2016 [2025-10-21]. https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[5] U.S. Department of Education. Artificial intelligence and the future of teaching and learning: insights and recommendations[R/OL]. Washington, D.C.: Office of Educational Technology, 2023 [2025-10-21]. https://tech.ed.gov.

[6] Australian Research Council (ARC). Responsible AI in education self-assessment tool[R/OL]. Canberra: ARC, 2024 [2025-10-21]. https://www.arc.gov.au.

使用阶段完成偏差检测与可解释性分析。国际组织也推动研究成果发布时附带“模型卡（model cards）”和“数据卡（data sheets）”，以便跨国互认与监管[1][2]。这些制度性创新体现出 AI 治理从自律到强制约束的演进逻辑。然而，外部监管也存在风险：过度规制可能抬高研究成本，抑制小规模团队创新；各国法律差异可能导致多重规制困境；政策制定的滞后性则可能在人工智能快速迭代的背景下造成治理真空[3]。

在此情境下，教育研究治理并不是自律与外部监管之间二选一，更多的是通过嵌入式治理和比例原则实现动态平衡。近年来，多国教育机构探索了“混合伦理委员会（Hybrid IRB）”模式，将传统人文伦理审查与 AI 技术审查相结合[4]。英国伦敦大学学院与澳大利亚教育研究委员会已建立“AI 伦理沙盒”，允许研究者在受控环境中测试算法的公平性与隐私影响，以提前识别潜在风险[5][6]。这种模式强调实验性监管与持续学习，使伦理治理在创新与风险防控之间取得平衡。实践中，许多机构并未重建独立的人工智能伦理审查体系，而是将风险管理嵌入既有 IRB 与科研管理流程。例如，在立项申请时要求提交人工智能使用说明，在成果发布阶段附加模型卡和偏差检测报告。这种嵌入式模式既降低了组织摩擦，又确保人工智能相关风险获得制度化关注。比例原则有助于实现分级治理：对涉及未成年人数据或弱势群体的高风险研究，应强化前置审查与外部监督；而对低风险研究，则可以采用更灵活的流程，以避免过度合

[1] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji I D, Gebru T. Model cards for model reporting[C]// Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*)*. Atlanta, GA: ACM, 2019: 220–229. DOI:10.1145/3287560.3287596.

[2] Gebru T, Morgenstern J, Vecchione B, Wortman Vaughan J, Wallach H, Daumé III H, Crawford K. Datasheets for datasets[J]. Communications of the ACM, 2021, 64(12): 86–92.

[3] Morley J, Floridi L, Kinsey L, et al. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices[J]. Science and Engineering Ethics, 2020, 26(4): 2141–2168.

[4] Leslie D. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector[R/OL]. London: The Alan Turing Institute, 2021 [2025-10-21]. https://doi.org/10.5281/zenodo.4404235..

[5] Australian Research Council (ARC). Responsible AI in education self-assessment tool and sandbox framework[R]. Canberra: ARC, 2024.

[6] Department for Education (DfE). AI governance and ethical sandbox framework[R]. London: UK Government, 2024.

规对学术创新造成阻碍[1]。

归根结底，自律与外部监管的平衡，本质上是权力、责任与能力的动态协调。研究者与机构在设定议程时拥有自主权，但这种权力必须以对学生、家长和社会的责任为前提，并建立在相应的技术与制度能力之上。只有当三者相互匹配时，人工智能治理才能避免失衡。UNESCO[2]提出的“ethics by design”理念与 OECD[3]倡导的“accountability across lifecycle”均强调，伦理治理不应停留于事后修正，而应前置至研究设计阶段。这种理念的落地意味着教育研究的治理正在从防御性合规转向建设性保障，从被动约束走向主动塑造科研公信力的机制。治理也不应被视为单纯的制度成本，而应理解为获取公共信任与社会合法性的生产性投资。如果研究机构能够在自律机制中落实 UNESCO 强调的人权与公平原则，在外部监管中对接 OECD 倡导的制度化路径，并结合 Batool 等[4]提出的全生命周期问责逻辑，人工智能便有可能从伦理风险的放大器转变为提升研究质量与社会价值的助推器。关键在于推动伦理治理从“防御性合规”升级为“建设性保障”，从而实现研究创新与教育公平的兼容共生。

[1] Floridi L, Cowls J. A unified framework of five principles for AI in society[J]// Machine Learning and the City: Applications in Architecture and Urban Design. Cham: Springer, 2022: 535–545.

[2] United Nations Educational, Scientific and Cultural Organization (UNESCO). Guidance for generative AI in education and research[R/OL]. Paris: UNESCO, 2023 [2025-10-21]. https://unesdoc.unesco.org/ark:/48223/pf0000386829.

[3] Organisation for Economic Co-operation and Development (OECD). Updated OECD principles on artificial intelligence (2024 revision)[R/OL]. Paris: OECD, 2024a [2025-10-21]. https://oecd.ai/en/ai-principles.

[4] Batool R, Lee M, Khan S. Accountability across the AI lifecycle: a governance model for responsible innovation[J]. Journal of Responsible Technology, 2023, 15: 100–118.

第 8 章 元研究视角：AI 如何改变知识生产

正如前文所言，人工智能持续扩展着教育研究者的认知边界，既带来前所未有的机遇，也伴随着新的挑战。相比于古腾堡印刷术引发的知识爆炸，现代智能技术的影响更为深远——它不仅重塑了社会关系，改变了知识的形态，更在根本上重构了教育研究者对事实与想象界限的认知。这场变革迫使教育研究者必须重新审视知识的本质及其在现代社会中的角色。

8.1 研究者的 AI 素养

AI 素养概念最早在 2010 年代中期被提出，用于描述公众应对人工智能技术的能力结构。小西洋子提出，AI 素养不仅关涉对相关技术的掌握，更是一种面向未来技术变革的思考与应对能力，这为后续研究奠定了理论起点[1]。随后，Kandlhofer 等将 AI 素养具体化为理解产品与服务背后的基本技术与概念，并据此划分七个核心主题：自动装置、智能代理、图与数据结构、排序、基于搜索的问题求解、经典规划与机器学习，标志着该领域从概念倡议走向内容框架的初步建构[2]。

在概念演进上，AI 素养经历了由“高校—专业化”向“公众—通用化”的转向。随着图形化编程、虚拟仿真与自然语言处理等技术的普及，学习门槛显著降低，AI 素养从面向专业学习者的术语，逐步成为强调基础知识与综合能力的公民基本素养之一[3]。在此基础上，Long 与 Magerko 对 AI 素养给出被广泛采用的功能性定义：AI 素养是一组综合能力，涵盖对 AI 技术的批判性评估、与 AI 系统的有效沟通与协作，以及在多种情境中合理应用 AI 的能力[4]。该定

[1] Konishi Y. What is needed for AI Literacy?[EB/OL].(2016-01-13)[2025-06-20].https://www.rieti.go.jp/en/columns/s16_0014.html.
[2] Kandlhofer M, Steinbauer G, Hirschmuglgaisch S, et al. Artificial intelligence and computer science in education:From kindergarten to university[EB/OL].(2016-11-01)[2025-06-20].https://ieeexplore.ieee.org/document/7757570.
[3] 陈力行,张文兰.近十五年技术支持环境下的学习动机研究[J].中国教育信息化,2022,28(7):50-58.
[4] Long D, Magerko B. What is ai literacy?Competencies and design considerations[EB/OL].(2020-04-23)[2025-05-10].https://dl.acm.org/doi/pdf/10.1145/3313831.3376727.

义突出了技术、批判性思维与沟通协作等软技能的多层面整合。

纵观现有研究，对 AI 素养的关注主要聚焦于三个维度：技术、认知与批判性思维[1]。基于本研究报告的目的，我们将聚焦“技术维度”，并以教育研究者为对象展开讨论。人工智能（AI）素养作为现代教育研究者必备的关键能力，涵盖对 AI 技术的本质理解、工具应用能力及其社会影响的全面认知。其中，AI 工具的理解是构建这一素养的基础。它不仅涉及技术术语的掌握，更强调去神秘化的认知过程，使教育研究者能够科学、理性地认识和评价 AI 的能力与局限。

8.1.1 AI 工具的理解

AI 工具的理解首先要求了解核心技术概念，包括机器学习、神经网络、自然语言处理等基础原理。然而，这一理解超越了传统的技术培训，侧重于知晓 AI 系统的运作机制及其历史发展脉络，从而帮助教育研究者打破“黑箱”效应，构建科学的认知框架[2]。例如，芬兰开展的全民 AI 素养教育项目，通过通俗易懂的语言向公众解释机器学习和神经网络的基本原理，阐明 AI 系统“以数据为食”的本质，促使学习者理性认识 AI 的能力与限制[3][4]。此类教育注重培养潜在的持久技能，而非局限于单一工具的使用，同时强调结合理论知识与实践经验，提高对 AI 的有效驾驭能力[5]。

Kong 等[6]进一步地从实践维度提出了 AI 素养的概念构成，尤其强调“有目的”的接触 AI。他们的定义主要由三个核心要素构成：对人工智能基本概念的理解、评估概念应用能力及将其应用于现实

[1] Long D,Magerko B.What is AI literacy?Competencies and design considerations[A].New York:ACM,2020:1-16.

[2] Wang S, Sun Z. Roles of artificial intelligence experience, information redundancy, and familiarity in sha** active learning: Insights from intelligent personal assistants[J]. Education and Information Technologies, 2025, 30(2): 2525-2546.

[3] UNESCO. International Forum on Ai and Education:Ensuring AI as a Common Good to Transform Education[EB/OL].(2021-12-07)[2025-04-12]. https://en.unesco.org/sites/default/files/ai-ineducation-forum-2021-cn-en.pdf.

[4] 郑燕林. 芬兰推进全民人工智能素养教育的路径选择[J]. 外国教育研究,2022,49(10):103-115.

[5] Yi Y. Establishing the concept of AI literacy[J]. Jahr - European Journal of Bioethics, 2021, 12(2): 353-368.

[6] Long D, Magerko B. What is AI literacy? Competencies and design considerations[C]//Proceedings of the 2020 CHI conference on human factors in computing systems. 2020: 1-16.

问题的能力。具体而言，首先要熟悉人工智能技术与方法，以识别AI驱动的工具和平台；其次需深入了解AI功能，能够有目的地与之交互；最后则要求批判性评估AI工具对个体生活、学习成果和社会的影响。这一定义与最新研究提出的AI素养结构高度契合，围绕如何正确地使用AI工具构建了包含识别、理解、使用、评估、创造及道德导航六个维度的素养结构[1][2]。

对于教育研究者而言，由于其研究使命与职业角色的特殊性，还需具备对AI发展趋势及应用场景的宏观把握。理解AI不仅限于掌握单一技术或算法，更需洞察其在教育、社会与文化等多重层面的作用，涵盖伦理、法律和社会责任等重要议题[3]。这一维度强调避免对AI的盲目信任或无端恐惧，培养批判性思维，关注算法偏见、数据隐私和决策透明度，从研究层面推动智能教育的公平与可持续发展[4]。

值得注意的是，人工智能素养水平的差异会显著影响个体对AI生成信息的接受与判断。例如，当面对带有“幻觉”——即不准确或虚假信息的AI响应时，AI素养较低的个体往往倾向于无条件接受，而素养较高的个体则更可能通过质疑和追问来批判性地分析答案[5]。这一现象凸显了AI素养在防止信息误导、促进健康认知态度中的关键作用。此外，缺乏AI素养可能导致过度依赖人工智能，缺少必要的脑力劳动，进而削弱批判性学习心态，阻碍长远学习发展[6]。这对于教育研究者而言，AI素养不足可能在学术研究中误信AI生成的不实内容，影响研究质量。

[1] Almatrafi O, Johri A, Lee H. A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023)[J]. Computers and Education Open, 2024, 6: 100173.
[2] Ng D T K, Leung J K L, Chu K W S, et al. AI literacy: Definition, teaching, evaluation and ethical issues[J]. Proceedings of the association for information science and technology, 2021, 58(1): 504-509.
[3] Long D, Magerko B. What is AI literacy? Competencies and design considerations[C]//Proceedings of the 2020 CHI conference on human factors in computing systems. 2020: 1-16.
[4] Wang V. Ethics and Equity in AI-Driven Education[M]//AI Integration Into Andragogical Education. IGI Global Scientific Publishing, 2025: 231-254.
[5] Oelschlager R. Evaluating the impact of hallucinations on user trust and satisfaction in llm-based systems[J]. 2024.
[6] Zhai C, Wibowo S, Li L D. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review[J]. Smart Learning Environments, 2024, 11(1): 28.

与此同时，AI 工具的理解还强调“知情者参与”的理念，即具备理解力和判断力的积极参与者，而非单纯的技术接受者或旁观者[1]。教育研究者不能只是“会用”AI 工具，而要“懂”AI 工具，并能主动、有意识地参与到 AI 的应用决策中。这与 21 世纪教育对公民素养的要求高度一致，激励教育领域人员深入探索 AI 与教育融合的路径，确保技术发展既推动教育创新，也尊重伦理原则与社会价值。

综上所述，AI 工具的理解不仅是技术知识的积累，更是一种面向未来社会的认知建构过程。通过系统的知识传授与批判性思维培养，教育研究者能够有效驾驭 AI 技术，理性评估其应用潜力和挑战。

8.1.2 AI 工具的应用

人工智能正成为各学科研究者的关键工具，深刻改变研究的组织方式与方法论。总体上，研究者通过引入各类人工智能算法与技术，显著提升了数据的采集、分析与解释效率，并拓展了知识发现的边界[2]。在应用层面，AI 的应用正从作为外部工具的“赋能”属性走向流程层面的“重构”，呈现出跨领域的广泛影响。

首先，在医学与医学教育相关研究中，AI 通过对大规模临床数据的学习与建模，支持更为精细的诊断与个性化治疗方案的制定[3]。例如，东京大学医学院团队基于新算法与序列参数开发了一套 AI 诊断系统：当该系统与深度学习程序集成时，样本患者组的最高诊断准确率为 83.5%；进一步与深度学习与决策树系统联合后，准确率提高至 87.3%[4]。这类案例显示，AI 的集成化应用能够在真实临床任务上获得可量化的性能增益，为循证医疗与教育提供更为坚实的

[1] Pinski M, Benlian A. AI literacy-towards measuring human competency in artificial intelligence[J]. 2023.
[2]Eremeev A P, Paniavin N A, Marenkov M A. An object-oriented approach to ontology modelling in specialists education of methods and technologies of artificial intelligence[C]//2022 VI International Conference on Information Technologies in Engineering Education (Inforino). IEEE, 2022: 1-4.
[3] Lee D H, Yoon S N. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges[J]. International journal of environmental research and public health, 2021, 18(1): 271.
[4] Sato M, Morimoto K, Kajihara S, et al. Machine-learning approach for the development of a novel predictive model for the diagnosis of hepatocellular carcinoma[J]. Scientific reports, 2019, 9(1): 7704.

数据基础。这一应用实践提示研究者需要具备集成化工具应用与效果评估的双重能力。单一 AI 系统的引入往往难以满足复杂研究任务的需求，研究者应学会根据具体问题特征，组合多种算法或技术路径以获得性能增益，并通过可评估的指标优化应用效果，从而将 AI 从辅助工具提升为可靠的研究支撑系统。

其次，在复杂数据处理与模式识别方面，AI 辅助研究工具能够在高维度、异质性的资料中识别难以通过传统方法发现的结构性特征或潜在趋势，从而提升研究发现的敏感度与广度[1]。在此基础上，AI 算法还可整合多源数据（如文本、图像、传感器日志），为研究者生成候选假设或优化实验设计路径，促进问题界定与研究方案的迭代完善[2]。这一特性对研究者提出了从工具赋能到流程重构的思维转变要求。传统研究范式中，数据分析通常处于研究流程的后端；而 AI 的引入使其得以前置到假设生成与设计优化阶段，研究者需要培养跨模态数据整合的意识与能力——不仅要掌握如何让 AI 处理单一类型的数据，更要理解如何协调文本、图像、音频等异质性信息源，使其在研究设计的早期阶段发挥启发性作用，从而实现研究方法论层面的创新。

再次，在教育与创意实践领域，AI 通过人机协作的方式提升学习与创新的产出质量。以服装设计为例，教师将 ChatGPT 与 Midjourney 等生成式技术纳入教学活动中，辅助学生进行概念发散、风格探索与原型迭代，从而在提升流程效率的同时增强创造力与想象力，促进学习者形成更为系统的设计思维[3]。这类应用表明，AI 不仅是信息处理的工具，亦可作为认知支架和协作伙伴嵌入到研究型学习场景。这一转变要求研究者重新审视 AI 在研究过程中的角色

[1] Nieto M, Senderos O, Otaegui O. Boosting AI applications: Labeling format for complex datasets[J]. SoftwareX, 2021, 13: 100653.

[2] Razia B, Awwad B, Taqi N. The relationship between artificial intelligence (AI) and its aspects in higher education[J]. Development and learning in organizations: an international journal, 2023, 37(3): 21-23.

[3] Lee J, Suh S. AI technology integrated education model for empowering fashion design ideation[J]. Sustainability, 2024, 16(17): 7262.

定位与协作边界。有效的 AI 应用不应是简单的任务外包或全盘自动化，而需要研究者明确人机分工，避免陷入过度依赖或盲目排斥的两极。

此外，基于自然语言处理的智能对话系统为研究中的数据收集（如调查或半结构访谈）提供了新的技术路径，降低了人力成本，并在一定程度上缓解了部分人际互动可能带来的偏差风险[1]。在临床应用层面，实践案例亦显示 AI 的场景化集成价值：克利夫兰诊所自 2016 年起将 Microsoft Cortana 融入电子医院系统，用于 ICU 高危患者的预测性识别与夜间监护；匹兹堡大学医学中心的 AI 辅助系统能够聆听并学习医生与患者的病房对话，从而支持临床流程优化与知识抽取[2]。位于洛杉矶的雪松西奈医院采用亚马逊 Alexa 作为病房虚拟护理助手，承担药物提醒、预约提示、常见问题应答等高频重复任务，释放护理人员时间并提升患者服务的连续性[3]。这些场景化应用案例凸显了研究者需要培养情境匹配与工具选择的判断能力。不同研究场景对 AI 工具的需求差异显著。研究者应避免“技术万能”或“单一工具依赖”，而应基于研究目标、数据特征、伦理约束等多维度考量，选择或组合最适配的 AI 方案，并在应用过程中持续监测其适切性与有效性。

综合而言，AI 与教育研究的深度整合在多个层面上带来显著增益：一是提升数据驱动的证据质量与分析速度；二是促进跨模态与多源数据的整合，从而增强假设生成与实验设计的科学性；三是通过人机协作提升创造性产出与教学研究的流程效率。在肯定其潜力的同时，也需在后续研究与应用中关注数据质量、可解释性、伦理与隐私等问题，以确保技术红利在学术与社会层面得到稳健转化。

[1] Al-Ansi A M, Jaboob M, Garad A, et al. Analyzing augmented reality (AR) and virtual reality (VR) recent development in education[J]. Social Sciences & Humanities Open, 2023, 8(1): 100532.
[2] Forbes. The Hospital Will See Vou Now[EB/OL].(2019-02-11)[2025-09-27]. https://www.forbes.com/sites/insights-intelai/2019/02/11/the-hospital-will-see-you-now/#4c9b42ae408a.
[3] Uzialko, A. Artificial Intelligence Will Change Healthcare as We Know it[EB/OL].(2019-06-09)[2025-04-27]. https://www.businessnewsdaily.com/15096-artificial-intelligence-in-healthcare.html .

可以认为，教育研究者的AI应用素养呈现出系统性与批判性并重的特征要求。系统性体现在需要从场景识别、工具选择、集成配置到效果评估的全流程能力建设；批判性则要求研究者在追求技术效能的同时，始终保持对数据质量、算法透明度、伦理风险的审慎反思。这种素养不是一次性习得的技能清单，而是在持续的研究实践中，通过“应用—评估—反思—优化”的迭代循环逐步形成的专业判断力。

8.2 实践：人机协同

在人机协同决策的范畴内，人与人工智能的互动正从传统的工具使用关系向更为复杂和深层的协同合作转变。具体而言，这一过程不仅强调“人在回路”中的主动参与和动态反馈，更注重通过“共享心智”实现认知融合与目标协同。二者相辅相成，共同构建了智能时代下教育研究者在决策过程中实现高效、精准和创新的核心能力框架。

8.2.1 人机协同理论

人机协同的理论谱系强调的是，智能活动并非仅由个体人脑完成，而是在人与技术、符号表征与情境环境构成的复合系统中生成：分布式认知指出认知过程可外化并在工具与界面中延展；行动者网络理论将算法、数据与平台视作具能动性的行动者，关注它们与人类角色通过“翻译”与关联而共同塑造实践；社会技术系统理论则主张社会子系统与技术子系统的联合优化，提醒在引入AI时同步调整组织流程、角色分工与规范治理；批判算法研究进一步揭示算法的社会—政治嵌入性，要求以可解释性、审计与问责来应对偏见与不平等；而后人类主义视角打破人类中心主义，将AI视为共生的知识与行动伙伴，重估主体性与伦理关系。综合看，这些理论从认知、网络、组织与价值层面提供互补框架，指向一种将信息流、制度安排与权力分配整合考量的协同设计与治理路径。

在上述人机协同的认知、网络、组织与伦理视角的综合框架下，本节将进一步聚焦于更具操作性的两个关键机制——“人在回路”（Human-in-the-Loop）与“共享心智”（Shared Mental Models），以连接宏观理论与微观实践，阐明人机决策控制权配置与团队认知对齐在实际系统设计与治理中的落地路径。

（1）人在回路

在人机协同决策中的“人在回路”是AI时代教育研究者必要的素养之一。随着机器学习，尤其是基于大规模数据驱动的深度学习技术的快速发展，人机关系迈入了一个关键的转折阶段——循环适应。在这一阶段，教育研究者与AI系统之间形成了一个高速运转、持续演化、双向作用的反馈回路（Human-AI Feedback Loop），这不仅改变了决策的模式，也深刻影响了教育研究的设计与实践。

首先，教育研究者在AI驱动的研究平台中，其研究行为及决策过程被全面数字化。无论是设计实验、调整参数、还是对AI分析结果的解读和修正，均会产生大量结构化与半结构化的交互数据。这些细微的反馈数据，成为AI系统自我学习和优化的基础资源。教育研究者因此不仅是数据的提供者，更是AI学习过程中的“训练者”，通过自身专业判断和策略调整，持续影响AI模型的发展方向。具体而言，教育技术领域已有多种应用充分体现了这一机制。例如，Outwrite和Grammarly等写作辅助平台利用用户输入和反馈数据，实时提供语法纠正与写作建议；ShortlyAI等创意写作工具则依托人类交互帮助完成故事构思和情节发展；编程领域的GitHub Copilot通过学习开发者的编码习惯，利用GPT-4模型辅助代码生成与调试；客户服务中的智能聊天机器人通过用户对话不断调整响应策略以提升服务质量。这些实例展示了教育研究者及相关用户通过交互行为，参与并推动AI系统的持续优化，体现了人在回路中的核心作用。因此，作为AI时代的教育研究者，具备有效管理和引导AI学习过程

的能力，理解并利用数字化交互数据对AI模型进行训练和调整，是其必备的重要素养。这不仅提升了研究效率和精准度，也促使教育研究实践向更智能化、个性化和动态适应的方向发展。

其次，AI模型基于教育研究者的反馈及使用行为，进行自适应优化。通过持续采集和分析教育者的交互数据，AI系统运用协同过滤、深度神经网络及强化学习等技术，动态调整其算法与参数，以更精确地支持研究者的认知需求和决策偏好。教育研究者在与AI工具的互动中，其决策支持功能从固定规则向个性化、动态响应转变，从而提升了教育研究的效能与创新能力。在实际应用中，AI系统通过对大量实时数据的监测与分析，实现了对教育环境中复杂变量的精准把控。例如，配备AI算法的可穿戴设备能持续监测学生的生理指标，如体温、心率和呼吸模式。对这些生命体征数据的动态分析，不仅能够及时发现潜在健康问题，还能辅助医疗干预，降低校园内疾病传播的可能性[1]。此外，AI还可以通过分析学生的行为模式，识别心理健康问题的早期迹象。例如，活动水平的突然变化、社交互动减少或学习成绩波动，可能预示焦虑、抑郁等心理障碍。基于这些模式识别，AI及时向学校辅导员或心理健康专家发出预警，促进及时干预和有效支持[2]。

这些实践案例体现了教育研究者作为“人在回路”的关键角色：他们不仅利用AI提供的动态数据和分析结果来优化决策，还通过专业判断对AI输出进行验证和调整，进一步引导AI模型的优化方向。教育研究者与AI系统形成双向互动的反馈机制，实现了研究决策的动态进化和智能提升，有助于推动教育领域的科学研究和实践向更加精准化、个性化及人本化的方向发展。

[1] Truong T D, Bui Q H, Duong C N, et al. Direcformer: A directed attention in transformer approach to robust action recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 20030-20040.

[2] Cheng K, Zhang Y, Cao C, et al. Decoupling gcn with dropgraph module for skeleton-based action recognition[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 536-553.

第三，经过优化的AI工具又反过来积极影响教育研究者的认知结构与研究路径。AI不仅为研究者提供个性化的信息检索、数据分析建议和实验设计优化方案，还通过人机交互界面引导研究者关注不同的数据维度及理论视角，从而潜移默化地塑造研究者的决策框架与学术判断。举例来说，推荐算法可能会引导教育研究者聚焦某些热点研究主题，或对某类教育数据模式赋予更高权重，影响他们对“权威知识”或研究趋势的认知。

最后，这种通过AI影响而变动的决策行为，再次形成新一轮的数据反馈，推动AI系统进入下一阶段的学习和优化循环。教育研究者的人机交互不仅实现了实时动态协同，也使其在教育科学知识生产中成为共创主体之一。人与AI的边界在此过程中趋于模糊：教育研究者既是AI模型的学习“引导者”，也是被AI结果影响和再塑的“学习者”。例如，在参与“人在回路”的机器学习应用中，研究者的专业判断和标注行为会随着与AI的持续互动而不断调整，形成一种协同进化的关系。有研究表明，在基于文本的解剖学内容查询任务中，经过个性化定制的聊天机器人在表现上优于通用的商业生成式人工智能（GenAI）系统[1]。相关研究评估了ChatGPT-4o和Claude 3.5 Sonnet对未标注解剖图像的解释能力，发现这些AI工具能够有效辅助教育者理解复杂内容[2]。更进一步，ChatGPT o1-preview版本被用作AI评估者，利用评分量规对AI生成答案进行评价，并与人类评估者的评分进行对比，结果显示其在教育辅助中的潜力和可靠性[3]。

这些实证研究不仅验证了AI在教育研究支持中的有效性，也凸显了教育研究者作为AI反馈环节不可或缺的角色。他们通过专业知

[1] Arun G, Perumal V, Urias F P J B, et al. ChatGPT versus a customized AI chatbot (Anatbuddy) for anatomy education: A comparative pilot study[J]. Anatomical Sciences Education, 2024, 17(7): 1396-1405.
[2] Hyeamang L J, Sekhar T C, Rush E, et al. AI's ability to interpret unlabeled anatomy images and supplement educational research as an AI rater[J]. Anatomical Sciences Education, 2025.
[3] Latif E, Zhou Y, Guo S, et al. A Systematic Assessment of OpenAI o1-Preview for Higher Order Thinking in Education[EB/OL]. (2024-10-11)[2025-09-27]. http://arxiv.org/abs/2410.21287.

识指导 AI 优化，同时借助 AI 洞察提升自身认知与决策水平，共同推动教育研究质量及创新能力的提升。由此，教育研究者与 AI 协同构建起一个动态、开放且不断进化的知识生成生态，实现教育科学的深化发展和智能变革。

总之，AI 时代下教育研究者在“人在回路”的人机协同决策中，应具备系统化理解并有效参与 AI 循环适应机制的能力。他们不仅要善于利用 AI 技术优化研究设计和决策，还要具备批判性反思 AI 对自身认知和行为模式影响的能力。通过这种动态的双向反馈互动，教育研究者与 AI 共同推动教育科学理论与实践的深度融合与创新发展。

（2）共享心智

在人机协同决策中的共享心智（Shared Mental Model）代表着人类智能与人工智能关系的深化与融合，体现为认知的共生与目标的协同。作为协同演进的前沿阶段，共享心智超越了传统简单的反馈循环，强调教育研究者与 AI 系统在认知层面和决策目标上的高度整合，这对于推动教育研究的创新和提升决策效率具有重大意义。尽管目前仍处于初步探索阶段，但这一趋势已在多个领域显现出明确的萌芽迹象。

首先，教育研究者需掌握混合主动性交互（Mixed-Initiative Interaction）的能力。在此模式下，人与 AI 系统之间的控制权不再是单向的，而是基于具体研究情境和任务需求实现动态切换。AI 具备主动介入的能力，能够基于研究目标、教育情境和交互历史，提前提供前瞻性建议，解释其推理过程，甚至主动请求关键的研究背景信息。这种情境感知式的主动互动模式，使得教育研究者能够更有效地引导 AI 辅助系统，同时获取更适切的支持与反馈，促进人机协同达到最佳水平[1]。

[1] Horvitz E. Principles of mixed-initiative user interfaces[C]//Proceedings of the SIGCHI conference on Human Factors in Computing Systems. 1999: 159-166.

其次，教育研究者应关注并应用可解释人工智能（Explainable AI，XAI）技术，将其作为构建人与AI共享心智的桥梁。在共生整合阶段，AI的决策过程和推理逻辑必须清晰透明，便于研究者理解、信任并合理利用AI建议。同时，通过研究者对AI解释结果的反馈与纠正，可以帮助AI修正偏差，完善模型认知。这种双向互动不仅提升了系统的鲁棒性和可靠性，也深化了研究者对AI能力界限及潜在风险的认识，推进了人机协同决策的可控性。

第三，教育研究者需要具备在复杂教育研究中与AI实现共同目标动态协商和演化的能力。面对开放性、多维度的教育问题，教育研究者与AI系统不只是协同完成既定任务，更需在目标设定、方案选择和价值取向上持续沟通与协调。AI通过其强大的模拟和模式识别能力，能够发现人类尚未察觉的潜在变量或新兴趋势，启发研究者重新审视研究命题或优化教学设计方案。教育研究者则基于自身的专业判断和伦理考虑，与AI共同调整研究方向，实现策略与目标的动态升级。

通过对共享心智的培养，教育研究者将在认知层面实现与AI的深度耦合，形成超越传统工具依赖的混合智能系统。这种系统已经被发现能促进原创性假设的生成、跨学科理论的整合，以及研究方法的创新发展[1][2]。教育研究者的创造力、价值判断和情境理解，与AI的强大运算与推理能力相互补充，构建出“1+1>2”的认知协同效应。

综上，AI时代下教育研究者在“共享心智”的素养培养中应注重“共享心智”的人机协同趋势，提升混合主动性交互能力、理解和应用可解释AI技术，以及开展动态协商式的目标管理与决策。这在推动了教育研究方法和实践革新的进程中，为构建人机协同科研

[1] Cukurova M. The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence[J]. British Journal of Educational Technology, 2025, 56(2): 469-488.

[2] Liu Y, Fu Z. Hybrid intelligence: design for sustainable multiverse via integrative cognitive creation model through human–computer collaboration[J]. Applied Sciences, 2024, 14(11): 4662.

范式奠定了坚实基础，助力应对日益复杂与多变的教育挑战。

8.2.2 人机协同决策

AI 时代的教育研究在推动科学发现和实践创新的关键领域，涵盖了协作能力与复杂问题解决技能两个核心方面。协作不仅增强了研究者之间的信息共享与资源整合，促进多样化视角的融合，还提升了整体研究的效率与质量；而复杂问题解决则要求研究者面对教育领域的多变挑战，运用系统性和创新性的策略，有效解析并应对高度关联且动态变化的问题。二者相辅相成，共同推动研究设计与实施向更加开放、灵活和精细化方向发展，体现了现代教育科研对认知与实践能力的综合要求。

协作作为现代教育研究不可或缺的组成部分，涵盖团队协作与多学科融合两大关键层面。团队协作强调成员间的有效沟通、角色分工与协同共创，是应对复杂教育问题的重要保障；而多学科融合则促使不同学科的理论、方法和技术交叉融合，打破知识壁垒，拓展研究视野和创新空间。通过强化这两个层面的协同互动，教育研究不仅能够培养跨专业的综合能力，还能实现更具深度与广度的学术突破，为智能时代的教育发展提供坚实支撑。

来自团队讨论的想法挖掘包括模拟或促进多个代理之间的协作头脑风暴的方法，无论是纯粹的算法还是涉及人类参与者，利用迭代批评、背景知识检索和重组来生成比单代理管道更丰富、更多样化的想法组合。

（1）AI-AI 协同

AI-AI 协作作为推动科学研究创新的重要方式，主要通过细化假设、批评建议与整合外部知识（即多智能体合作）来改进科学思维[1]。当前的研究方法大致可分为两类：

第一类是反馈导向的议题挖掘（feedback-guided exploration），

[1] Swanson K, Wu W, Bulaong N L, et al. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation[J]. bioRxiv, 2024: 2024.11. 11.623004.

该方法通过在不同研究阶段主体之间交换评论，利用迭代反馈不断优化和精炼假设。一些研究在想法挖掘、实验设计及结果解释过程中引入反馈循环以提升系统性能[1]，另一些则通过先前输出的反馈来改进假设生成[2]。此类方法具体体现为同行评议机制[3]、对假设的直接批评[4]以及对实验结果的评估反馈[5]，通过多轮互动促进假设不断完善。

第二类是团队讨论导向挖掘，旨在通过组建多个具备不同角色的智能代理，模拟人类研究团队的动态协作，从而丰富和多样化研究创意[6]。Su 等人[7]提出的虚拟研究团队 VirSci，通过代理间反复提出和评论想法，借助不断增长的想法，实现了比单一代理更多的新概念生成。Yang 等人[8]基于大语言模型开发了多智能体框架 MOOSE-CHEM，专注于化学领域的科学假设发现，具备灵感提取与背景驱动的假设生成能力。Li 等人[9]设计的思想链（Chain of Ideas, COI）代理，通过将文献组织成反映主题历史进程的连续链条，实现了与小型研究团队质量相当的研究产出，且极大降低了成本。Lagzian 等人[10]则提出通过多视角头脑风暴机制，进一步增强了生成结果的多样性和新颖性。

这一协作范式的兴起对教育研究者提出了新的素养要求。首先，

[1] Zhou Y, Liu H, Srivastava T, et al. Hypothesis Generation with Large Language Models[EB/OL]. (2024-04-05)[2025-09-27]. http://arxiv.org/abs/2404.04326.
[2] Hu X, Fu H, Wang J, et al. Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas[EB/OL]. (2024-10-18)[2025-09-27]. http://arxiv.org/abs/2410.14255.
[3] Lu C, Lu C, Lange R T, et al. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery[EB/OL]. (2024-08-12)[2025-09-27]. http://arxiv.org/abs/2408.06292.
[4] Baek J, Jauhar S K, Cucerzan S, et al. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models[EB/OL]. (2024-04-11)[2025-09-27]. http://arxiv.org/abs/2404.07738.
[5] Ma P, Wang T H, Guo M, et al. LLM and Simulation as Bilevel Optimizers: A New Paradigm to Advance Physical Scientific Discovery[EB/OL]. (2024-05-16)[2025-09-27]. http://arxiv.org/abs/2405.09783.
[6] Pu Y, Lin T, Chen H. PiFlow: Principle-Aware Scientific Discovery with Multi-Agent Collaboration[EB/OL]. (2025-05-22)[2025-09-27]. http://arxiv.org/abs/2505.15047.
[7] Su H, Chen R, Tang S, et al. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation[J]. 2024.
[8] Yang Z, Liu W, Gao B, et al. Moose-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses[EB/OL]. (2024-10-09)[2025-09-27]. http://arxiv.org/abs/2410.07076.
[9] Li L, Xu W, Guo J, et al. Chain of Ideas: Revolutionizing Research via Novel Idea Development with LLM Agents[EB/OL]. (2024-10-17)[2025-09-27]. http://arxiv.org/abs/2410.13185.
[10] Lagzian A, Anumasa S, Liu D. Multi-Novelty: Improve the Diversity and Novelty of Contents Generated by Large Language Models via Inference-Time Multi-Views Brainstorming[EB/OL]. (2025-02-19)[2025-09-27]. http://arxiv.org/abs/2502.12700.

研究者需要理解多智能体系统的运作机理，包括反馈循环如何优化输出、不同角色代理（如批评者、生成者、评估者）如何分工协作，从而能够根据研究任务特征选择合适的协同架构。其次，研究者应具备配置与编排多代理系统的能力，例如在假设生成阶段引入同行评议机制、在实验设计阶段融入背景知识检索代理，使多智能体的协同效应得以充分发挥。更为关键的是，研究者需要培养对 AI-AI 协作产出的批判性评估能力——尽管多智能体可生成更丰富的创意组合，但其中可能存在逻辑矛盾、事实错误或重复冗余，研究者必须保持专业判断力，对生成结果进行筛选、整合与验证，避免将技术输出直接等同于可靠的研究结论。总之，有效驾驭 AI-AI 协同不仅需要技术操作能力，更需要对多智能体生态的系统理解与质量把控意识。

因此，AI-AI 协作通过反馈制导的反复优化和多角色代理的动态协作，有效促进了科学假设的生成与完善，展示了多智能体系统在辅助科研中的广阔应用前景。

（2）人-AI 协同

在人-AI 协作领域，研究重点集中在人类研究者如何通过选择和管理人工智能生成的中间产物，引导大语言模型开展创新探索的过程[1]。例如，Radensky 等人[2]提出的 Scideator 系统，允许研究者从已有文献中选取不同的组成部分，如问题陈述、研究方法和数据集，随后由大语言模型对这些元素进行重新组合，生成新的候选研究想法，有效提升了创新质量。同样，Garikparthi 等人[3]开发的交互式研究思维系统 IRIS 通过验证研究动机和基于研究者查询提供综合方法论建议，促进了人类与人工智能之间的协同工作。

[1]Blodgett S L, Daumé III H, Madaio M, et al. Proceedings of the Second Workshop on Bridging Human--Computer Interaction and Natural Language Processing[C]//Proceedings of the Second Workshop on Bridging Human--Computer Interaction and Natural Language Processing. 2022.

[2] Radensky M, Shahid S, Fok R, et al. Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination[EB/OL]. (2024-09-22)[2025-09-27]. http://arxiv.org/abs/2409.14634.

[3] Garikaparthi A, Patwardhan M, Vig L, et al. IRIS: Interactive Research Ideation System for Accelerating Scientific Discovery[EB/OL]. (2025-04-23)[2025-09-27]. http://arxiv.org/abs/2504.16728.

然而，相关研究亦指出，尽管大语言模型辅助在短期内可增强研究创造力，但在缺乏辅助的情况下，用户的独立创造性可能受到抑制[1]，这引发了对 AI 长期影响人类创造力和认知能力的关注与审视。

团队合作作为现代知识生产的重要组成部分，其专业构成对组织绩效和政策制定具有关键影响[2]。其中，技术的易获取性成为影响团队专业知识构成的核心因素。以“微软 Kinect 被破解”事件为例，该事件使运动传感技术成本大幅降低，从而取代了对领域专家的部分依赖，并推动了更多外部领域专家的参与合作。换言之，技术成本的降低改变了团队合作中知识专业化的格局，促使跨领域的知识创造更加广泛。这一发现对组织与政策制定者提出了启示，即通过战略性投资降低技术门槛，可以优化团队结构，进而促进更具多样性的知识创新。

关于团队知识生产的动态机制，研究表明，在现代工作环境中，团队成员通过信息技术和专业知识的协同运用，产生新的知识成果。技术不仅支持团队内的沟通、协作与知识共享，还重塑了团队的内部工作流程。不同类型专家的知识贡献和整合对提升团队解决问题的能力至关重要，而技术与专长之间的互动更是推动创新和提高效率的关键要素。基于此，研究提供了实践指导，强调在快速变化的技术环境下科学设计和管理团队的重要性，以促进知识生产和应用的最优化。

人-AI 协同模式对教育研究者的素养提出了更为复杂的要求。一方面，研究者需要掌握主动引导 AI 的能力，包括如何从文献中提取和重组研究要素（如问题陈述、方法论、数据集）、如何通过精确的查询和约束条件引导 AI 大语言模型生成符合研究意图的创意，以

[1] Kumar H, Vincentius J, Jordan E, et al. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025: 1-18.

[2] Teodoridis F. Understanding team knowledge production: The interrelated roles of technology and expertise[J]. Management Science, 2018, 64(8): 3625-3648.

及如何在迭代过程中对 AI 的中间产物进行选择性保留或修正。这种能力要求研究者既要熟悉 AI 系统的输入输出机制，也要保持对研究目标的清晰认知，避免被 AI 的生成逻辑“牵着走”。

另一方面，研究者必须警惕技术依赖对独立创造力的潜在抑制作用。如前文所述，长期依赖 AI 辅助可能削弱研究者在无辅助情境下的创新能力，因此需要在借助 AI 拓展思维与保持独立思考之间建立动态平衡——例如在早期发散阶段借助 AI 激发灵感,在关键决策阶段回归人类判断。此外，在团队层面，研究者还需理解技术可及性如何重塑知识生产的专业构成：AI 工具的普及可能降低某些专业知识的门槛，促进跨领域协作，但也可能导致过度依赖技术而忽视领域专长的深度积累。

因此，教育研究者需要在战略层面思考如何通过技术投资优化团队结构，在战术层面培养“与 AI 共舞”的实践智慧，既充分利用 AI 的增强效应,又保持人类在价值判断、伦理审视和创造性飞跃上的核心主导地位。

8.3 知识生产生态系统的重构

人工智能与协同演进的浪潮，正以前所未有的深度和广度重构着知识生产的整个生态系统。传统的、以个体或孤立实验室为单位、以期刊为主要发表渠道的科研模式，正被一个更加开放、网络化、平台化且急剧加速的新范式所取代。这场变革并非单一维度的技术应用，而是体现在研究机构的形态、科研协作的机制以及科学发现的内在动力等多个层面上的系统性转型。

8.3.1 新型研究机构与平台涌现

在人工智能迅猛发展的背景下，知识生产的组织形式和基础结构正经历一场深刻的结构性变革。这一变革不仅打破了传统学术、产业与公共机构之间的边界，也催生了多种跨界融合的新型研究机构与平台，逐步构建出一个以协同、开放、高效为特征的智能时代

科研新生态。

首先，使命驱动的混合型研究机构正崛起为前沿创新的策源地。这类机构以 OpenAI、DeepMind（现为 Google DeepMind）、Anthropic 等为代表，它们通常设定极具前瞻性和公共性的发展目标，例如推动通用人工智能（AGI）的实现，或确保人工智能发展的安全性、可控性与社会福祉。相较于传统大学实验室，这些组织在资源调配、跨领域协作和研究节奏上具备更大灵活性和规模优势，能够承担更长周期、更高风险的基础研究任务。与此同时，它们又不完全等同于商业公司，而是兼具非营利组织的公共性特征：关注研究透明度、强调成果共享，致力于通过开放论文、发布模型与工具包等方式回馈全球科研社区。例如，阿里巴巴的新一代通义千问模型 Qwen3（简称“千问 3”）是一款通用开源大语言模型，在代码生成和模型评测方面表现突出，因其开放性而受到企业与个人的广泛欢迎。其开放协作的组织形态与高效迭代机制，正成为推动人工智能变革性进步的关键引擎。

与此同时，部分传统学术机构也在积极重塑自身角色，适应 AI 时代的跨学科要求。麻省理工学院的媒体实验室（MIT Media Lab）便是其中的典范。该实验室以“反学科”（Antidisciplinary）为核心理念，打破学科疆界，鼓励艺术、技术、设计与社会科学的深度交融。在 AI 时代，这一理念被赋予新的生命力——研究者不仅关注算法性能本身，更重视技术与人类经验、社会结构、文化表达之间的复杂交互。在媒体实验室中，AI 被用于探索声音生成、艺术创作、人机情感交互等富有想象力的方向，展现出人工智能作为文化媒介与社会实践的多重角色。

除了机构形态的变革，一个以平台为中心的科研新范式也正在快速成型，并日益成为知识生产的核心基础设施。在这一模式下，科研的各个关键环节——从成果发布、同行评议到模型开发与复

用——正迅速向高度开放、数字化的平台迁移，重塑学术生态与协作方式。

以预印本平台 arXiv 为例，它突破了传统学术出版体系的时效限制，使得研究者可以在正式发表前即刻共享自己的研究成果，大幅加快了学术传播的速度。尤其在 AI 等前沿领域，arXiv 已成为新思想、新方法的首发阵地，形成一种“先发共享、后期优化”的研究文化。研究表明预印本通过提前公开与扩大可见度，促进早期传播与讨论，从而在正式发表前就提升读者覆盖与引用影响[1]。类似地，OpenReview 平台通过开放同行评议流程，增加了评审的透明度与互动性，促使学术讨论不再局限于编辑部与匿名审稿人之间的封闭对话，而转化为一种开放、协同的社区评议模式。这种机制不仅提高了评审的质量与可信度，也为年轻研究者提供了更多参与学术治理的机会。

代码托管平台 GitHub 在这一新生态中扮演着不可替代的角色。它不仅是开源社区的集结地，更逐渐演变为科学研究的重要评价维度之一。一个研究项目若能同步开放代码，其复现性、透明度与影响力通常会显著增强。许多顶级会议与期刊已将代码开放视为投稿标准之一，GitHub 因此成为衡量研究可信度与社会影响力的重要指标。尤其是在人工智能领域，代码与论文之间的界限日益模糊，模型架构、训练参数与实验细节往往通过 GitHub 进行共享，成为研究成果不可分割的一部分。

在此基础上，新兴平台 Hugging Face 正在构建更具协同性与可扩展性的模型共享生态。该平台不仅托管了海量的预训练模型、数据集与评测基准，还通过“Transformers”库等工具集成，极大地降低了 AI 开发的门槛。研究人员与开发者可以轻松加载、微调、复用已有模型，加速了研究迭代与应用落地的过程。更重要的是，

[1] Tsunoda H, Sun Y, Nishizawa M, et al. The Impact of Preprints on the Citations of Journal Articles Related to COVID-19[C]//International Conference on Asian Digital Libraries. Singapore: Springer Nature Singapore, 2024: 57-69.

Hugging Face 以社区驱动为导向，鼓励全球开发者协作完善模型文档、优化算法性能，形成一种“模型即公共物”的共享文化，重塑了 AI 知识传播与演化的方式。

此外，Weights & Biases、Papers with Code、Kaggle 等辅助平台也在形成强大的技术协同网络。例如，Papers with Code 将论文与实现代码一一对应并自动追踪性能指标，使得研究进展更加透明与可追溯；Weights & Biases 提供实验管理、模型调优与可视化工具，促进了模型开发过程的标准化与可复现；而 Kaggle 则通过竞赛与数据集开放，成为连接研究者、企业与开发者的桥梁，进一步推动理论研究与实际应用之间的良性循环。

这一系列平台的兴起，不仅提升了知识生产的效率与开放性，更深刻地改变了科研的范式与价值体系。从“封闭发表”走向“协作开发”，从“论文至上”转向“模型与数据并重”，从“中心化权威”迈向“社区驱动创新”，平台所构建的是一种去中心化、可组合、共享导向的科研生态。这种结构性变革，不仅拓展了知识创造的空间，也为全球科研社群注入了前所未有的活力。

总而言之，新型研究机构与平台的崛起，标志着人工智能时代知识生产方式的根本性重构。在这个不断演化的生态系统中，研究不再是象牙塔内的孤立行为，而是一种跨机构、跨学科、跨平台的协作共建。它既回应了 AI 时代技术发展的复杂需求，也体现出对开放、透明、可持续科研伦理的追求，为未来的科技治理与社会创新奠定了坚实基础。

8.3.2 开放科学与共享机制的推进

伴随着科研组织形态的深刻变革，科学研究的运行机制与文化范式也正发生结构性的重塑。在人工智能引领的知识生产新生态中，“开放”不再是理想化的倡议，而正在转化为具有约束力的制度性规范。开放科学（Open Science），作为这一变革的核心理念，正日

益成为 AI 研究领域不可逆转的发展方向。

首先，开放获取（Open Access）、开放数据（Open Data）与开源代码（Open Source Code）已逐渐从边缘化的实践转变为前沿科研的“新常态”。在 NeurIPS、ICLR、CVPR 等顶级 AI 会议中，主办方普遍要求或强烈鼓励作者同步提交模型代码和训练数据，并提供详细的实验复现说明，以提高研究的可验证性与可重复性。这种机制不仅提高了成果的透明度，也推动了整个学术共同体对科学可复现性（Scientific Reproducibility）的重视。

为了确保数据和代码的可用性与可持续性，FAIR 原则（可查找 Findable,可访问 Accessible,可互操作 Interoperable,可复用 Reusable）被广泛采纳。许多平台和研究团队在发布数据集和模型时，遵循 FAIR 原则进行标准化处理，从而提高科研资源的可检索性、互操作性和重用性。例如，Zenodo 和 Figshare 等科研数据托管平台提供了 DOI 赋码、元数据注释和版本管理功能，使得数据本身成为独立可引用的学术产出；Papers with Code 则将论文、代码、数据集和性能评估标准系统性整合，构建了一个透明的 AI 研究知识图谱。

同时，开放获取出版也在快速制度化。除了 PLOS ONE、eLife 等原生开放获取期刊，传统顶刊如 Nature、Science 等也陆续推出开放出版渠道（如 Nature Communications、Science Advances），推动学术成果向公众与非营利研究机构全面开放。这种变革，不仅降低了全球研究者之间的知识鸿沟，也增强了科学与社会之间的双向联系。

在开放文化的推动下，去中心化、社群驱动的科研协作机制正在崛起。研究者不再严格依赖于大学或研究机构的正式身份，而是通过 GitHub、Discord、Reddit、Hugging Face Spaces 等分布式平台，自发组成兴趣导向、任务导向的合作网络。以 GitHub 为例，通过 Pull Request、Issue Tracker 等机制，支持全球开发者围绕同一项目

协同开发与持续迭代。Hugging Face Spaces 则进一步推动了模型协作的可视化和应用化，研究者可以在平台上即时部署模型 Demo，与用户交互获取反馈，缩短了从实验到实际应用的转化周期。Discord 社区则常被用于实时讨论和小型项目孵化，形成灵活的“虚拟实验室”；Reddit 上的子论坛（如 r/MachineLearning）则为新手与专家提供了一个开放的知识问答和趋势讨论场域。

这种去中心化协作模式促生了一系列制度性创新实践。例如，“预注册研究（Pre-Registered Studies）”鼓励研究者在实验前公开假设与方法，以减少事后修正数据以迎合结果的发表偏倚；“开放同行评议（Open Peer Review）”则通过平台如 OpenReview、F1000Research 等，公开审稿意见与作者回应，增强评审透明度和知识交流的深度。这些机制有效挑战了传统期刊封闭、不可见的知识审查逻辑，推动了科研治理的民主化与透明化。

这一开放与协作生态的集中体现，是 AI4S 范式的兴起，它标志着 AI 不仅是科研工具，更正在成为主动参与知识生成过程的“共生智能体”。DeepMind 开发的 AlphaFold 即为这一趋势的典范。该系统凭借深度学习技术成功预测了上亿种蛋白质的三维结构，解决了困扰生物学界数十年的难题。更重要的是，DeepMind 将 AlphaFold 模型及其结构数据库向全球科学界免费开放，通过与 EMBL-EBI（欧洲生物信息研究所）的合作，共同建立了全球最大的蛋白质结构数据库，为药物研发、疾病机制研究等多个领域带来了指数级加速。这不仅展现了 AI 在知识发现中的能力，也体现了开放科学所能释放的协同潜力与社会价值。

类似地，AI 系统在材料发现、量子化学、数学猜想验证等领域也开始展现出与人类科学家协作的新模式。例如，Microsoft Research 的 AI4Science 项目、Meta 的 Galactica 模型，以及 Google 的 Minerva 数学助手，皆试图将 AI 系统整合入从假设生成、实验设

计到理论归纳的全过程。这一趋势预示着科研将从“人类主导+AI辅助”逐步演化为“人类与AI共建”模式，拓展了科学本身的边界。

综上，从研究机构的重组，到基础设施平台化的演进，再到开放与共享机制的制度化，AI正驱动知识生产生态系统发生一场深刻而系统的重构。这一新生态的核心特征是：开放性（Openness）、协作性（Collaboration）与加速性（Acceleration）。它不仅重塑了科研的日常实践，也在重新定义科学发现的逻辑、速度与规模，预示着一个更加透明、共享与高效的知识社会正在加速到来。

8.3.3 科研评审

第五章讨论AI介入科研评审主要从风险、标准与制度安排入手，强调治理框架与规范边界。本章则将在此基础上由“议题分类”转向“流程导向”，以评审全流程为主线，系统梳理AI在稿件分流与初筛、领域与专家匹配、利益冲突识别与可重复性核验等关键环节中的功能定位、数据流转与责任划分。

（1）评审语料库

在AI时代的教育科学研究中，科研评审作为保障学术质量和推动知识进步的关键环节，正逐步引入人工智能技术以提升评审的效率与公正性。当前，基于多种数据集的AI系统已经能够完成从文本检测到评审生成、质量评估以及决策支持等多维度任务[1]。为模拟真实的同行评审互动场景，研究者构建了一系列代表性数据集，如PeerRead[2]、Spot[3]、NLPeer[4]等，这些数据集涵盖了来自顶尖会议和期刊的大量论文及其评审记录，支持对大语言模型在多轮交互、长上下文处理及角色扮演式同行评审中的训练与综合评估。

[1] Zhou R, Chen L, Yu K. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks[C]//Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024). 2024: 9340-9351.

[2] Kang D, Ammar W, Dalvi B, et al. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications[EB/OL]. (2018-04-25)[2025-09-27]. http://arxiv.org/abs/1804.09635.

[3] Son G, Hong J, Fan H, et al. When AI Co-Scientists Fail: SPOT-A Benchmark for Automated Verification of Scientific Research[EB/OL]. (2025-05-18)[2025-09-27]. http://arxiv.org/abs/2505.11855.

[4] Dycke N, Kuznetsov I, Gurevych I. NLPeer: A Unified Resource for the Computational Study of Peer Review[EB/OL]. (2022-11-12)[2025-09-27]. http://arxiv.org/abs/2211.06651.

为确保AI系统在敏感的学术环境中表现出足够的可靠性，LLMart[1]提供了一个用于对抗测试与快速优化的工具包，专门评估大语言模型的鲁棒性。与此同时，Shin等人[2]与Couto等人[3]通过预定义的多个维度，对同行评审内容的质量进行了细致分析，揭示了大语言模型生成内容与人工评审重点之间的差异。在评审质量检测方面，相关研究还探讨了多种质量特征的评估方法[4]，并特别关注AI系统在检测“懒惰思维”或体现礼貌性的评审语言方面的表现，如PurkayDisA[5]等和PolitePEER[6]等研究所示。

此外，针对AI生成评审的检测，AI-Peer-Review-Detect-Benchmark[7]和Try[8]等数据集汇集了成千上万条由AI产生的同行评审文本，同时包含ICLR和NeurIPS等顶级会议的人工评审，为相关检测方法的开发和验证提供了标准化语料库。

总之，AI辅助的科研评审技术不仅提高了同行评议的自动化和智能化水平，也促进了教育科学研究评价机制的革新，为构建更加高效、公正和透明的学术生态系统奠定了基础。教育研究者在这一进程中，应积极理解并把握AI技术在科研评审中的应用与挑战，推动其与人类专家判断的有效协同，实现科学评价的智能升级。

（2）评审流程

AI辅助论文评审的流程可概括为“智能分流—质量审查—推荐审稿人—人机协同决策”的闭环：在投稿进入系统后，AI先进行主

[1] Liang W, Zhang Y, Cao H, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis[J]. NEJM AI, 2024, 1(8): AIoa2400196.
[2] Shin H, Tang J, Lee Y, et al. Mind the Blind Spots: A Focus-Level Evaluation Framework for LLM Reviews[EB/OL]. (2025-02-25)[2025-09-27]. http://arxiv.org/abs/2502.17086.
[3] Couto P H, Ho Q P, Kumari N, et al. RelevAI-Reviewer: A Benchmark on AI Reviewers for Survey Paper Relevance[EB/OL]. (2024-06-14)[2025-09-27]. http://arxiv.org/abs/2406.10294.
[4] Ghosal T, Kumar S, Bharti P K, et al. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews[J]. Plos one, 2022, 17(1): e0259238.
[5] Purkayastha S, Li Z, Lauscher A, et al. LazyReview: A Dataset for Uncovering Lazy Thinking in NLP Peer Reviews[EB/OL]. (2025-04-17)[2025-09-27]. http://arxiv.org/abs/2504.11042.
[6] Bharti P K, Navlakha M, Agarwal M, et al. PolitePEER: does peer review hurt? A dataset to gauge politeness intensity in the peer reviews[J]. Language Resources and Evaluation, 2024, 58(4): 1291-1313.
[7] Yu S, Luo M, Madasu A, et al. Is Your Paper Being Reviewed by an LLM? Investigating AI Text Detectability in Peer Review[EB/OL]. (2024-10-03)[2025-09-27]. http://arxiv.org/abs/2410.03019.
[8] Liu R, Shah N B. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing[EB/OL]. (2023-06-01)[2025-09-27]. http://arxiv.org/abs/2306.00622.

题匹配与合规性筛查，并完成相似性检测与初步方法论、数据与伦理风险审计；随后依据内容特征与冲突规则推荐合适评审人，并提供结构化评审提纲、证据抽取与可重复性核验辅助；在评审与作者回应往返中，AI对意见一致性、偏见风险与关键争议点进行标注与汇总，为编委会决策提供可追溯的证据链；全流程以人机协作、透明记录与隐私保护为原则，实现效率提升与公正性的平衡。具体步骤如下：

第一步，论文匹配。这是选择合适的期刊或会议是发表研究成果的关键一步。不当的目标选择可能导致优质文章被拒稿或影响传播范围。因此，利用期刊推荐工具来辅助选择目标期刊变得至关重要。目前，许多出版商推出了基于文本相似度与机器学习的期刊推荐工具，支持首次投稿及被拒后重投的场景。这些工具通常通过与已发表文献进行匹配，利用论文的标题、摘要、关键词或参考文献来推荐期刊。例如，Springer Nature的Journal Suggester[1]、Wiley的Journal Finder[2]、IEEE的Publication Recommender[3]都基于已发表论文的数据进行主题匹配；EndNote的Manuscript Matcher则结合了Web of Science的数据，利用标题、摘要与参考文献来推荐目标期刊[4]；而JANE（Journal/Author Name Estimator）则是通过PubMed的数据与文本相似度推荐生物医学领域的期刊[5]。此外，JournalGuide（Research Square）[6]通过提取论文标题与摘要中的信息来进行匹配。

近年来，许多推荐工具进一步结合了AI技术，进一步提高了期刊推荐的精准度。例如，Elsevier的JournalFinder利用智能搜索和学科专用词汇进行匹配[7]；Taylor & Francis的Journal Suggester[8]与

[1] Springer Nature. Journal Suggester[EB/OL]. [2025-09-27]. https://journalsuggester.springer.com.
[2] Wiley. Journal Finder[EB/OL]. [2025-09-27]. https://journalfinder.wiley.com/search?type=match.
[3] IEEE. Publication Recommender[EB/OL]. [2025-09-27]. https://publication-recommender.ieee.org/home.
[4] Clarivate. Manuscript Matcher[EB/OL]. [2025-09-27]. https://endnote.com/product-details/manuscript-matcher.
[5] Biosemantics Group. JANE: Journal/Author Name Estimator[EB/OL]. [2025-09-27]. https://jane.biosemantics.org.
[6] JournalGuide[EB/OL]. [2025-09-27]. https://www.journalguide.com.
[7] Elsevier. Journal Finder[EB/OL]. [2025-09-27]. https://journalfinder.elsevier.com.
[8] Taylor & Francis. Journal Suggester[EB/OL]. [2025-09-27]. https://authorservices.taylorandfrancis.com/publishing-your-research/choosing-a-journal/journal-suggester.

Sage Journal Selector[1]则声称应用了先进的 AI 算法来推荐相关期刊。研究表明，机器学习在期刊推荐方面已经表现出较高的准确性。以 XGBoost 模型为例，在基于 Web of Science 的计算机类期刊数据集的实验中，准确率达到了 84%[2]。深度学习系统 Pubmender 在 1130 种 PubMed 期刊的 88 万篇论文摘要数据上达到 87%的准确率，显著优于部分商业工具[3]。此外，GraphConfRec 利用论文文本、合著与引文网络来推荐相关的计算机科学会议[4]。这些技术的不断发展，使得学者们在选择期刊时能够更加高效和精准地定位目标期刊或会议。

第二步，在稿件进入同行评审前，编辑部通常会进行桌面审查（Desk Review），评估稿件是否符合期刊的范围、质量要求以及规范。AI 技术在此阶段的应用主要体现在稿件路由与分类、质量检查与规范审查等多个方面。

首先，AI 在稿件的路由与分类方面发挥着重要作用。系统如 Elsevier Evise[5]和 Editorial Manager（EM）[6]通过索引和术语提取，将稿件精确地路由到合适的学科领域和编辑团队。而 IEEE Manuscript Central（基于 ScholarOne）[7]则结合元数据、作者关键词和学术网络，利用审稿人检索工具实现更加精准的稿件分类与审稿人匹配。Springer SNAPP 和 Nature 的 AI 分类工具[8]则进一步提升了初审的效率和准确性。此外，AnnotateGPT[9]能够生成稿件注释，帮

[1] JournalGuide[EB/OL]. [2025-09-27]. https://www.journalguide.com.
[2] Zheng W H, Tao M, Yanni Y, et al. Recommendation method for academic journal submission based on doc2vec and XGBoost[J]. Scientometrics, 2022, 127(5): 2381-2394.
[3] Feng X, Zhang H, Ren Y, et al. The deep learning‐based recommender system “Pubmender” for choosing a biomedical publication venue: Development and validation study[J]. Journal of medical Internet research, 2019, 21(5): e12957.
[4] Iana A, Paulheim H. Graphconfrec: A graph neural network-based conference recommender system[C]//2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE, 2021: 90-99.
[5] Elsevier. Rolling out our new editorial system: EVISE[EB/OL]. (2015-10-22)[2025-09-27]. https://www.elsevier.com/connect/reviewers-update/rolling-out-our-new-editorial-system-evise.
[6] Adrian Tedford. Helping editors find reviewers[EB/OL].(2015-09-23)[2025-06-27].https://www.elsevier.com/connect/helping-editors-find-reviewers.
[7] IEEE Computer Society. How to make peer review recommendations and decisions[EB/OL].[2025-05-27]. https://www.computer.org/publications/making-peer-review-recommendations.
[8] Springer Nature. Snapp: Springer nature's next-generation peer review system[EB/OL].(2023-12-01)[2025-04-20].https://www.springernature.com/gp/snapp.
[9] Díaz O, Garmendia X, Pereira J. Streamlining the review process: AI-generated annotations in research

助编辑快速判断稿件的研究范围和质量。

在质量与规范检查方面，AI 也扮演着关键角色。抄袭检测方面，iThenticate[1]系统通过对比系统内外文献，检测稿件中的全文或段落重复，同时，短语异常检测技术[2]能够识别“机翻回译”式的学术不端行为。为了应对 AI 内容生成带来的挑战，ZeroGPT 被用于识别由如 ChatGPT 等语言模型生成的文本[3]。此外，方法与统计检查工具如 SciScore 提供了方法学评分与透明度指数，StatReviewer[4]和 StatCheck[5]则帮助检测统计报告中的错误。透明度和可重复性方面，Dimensions Research Integrity preCheck[6]可以检查数据可得性声明及软件版本信息的准确性。

在稿件的结构与引用匹配方面，Penelope.ai 用于检查稿件结构与投稿指南的一致性[7]，而 Recite 则能够自动匹配文内引用与参考文献列表，确保引用的准确性和一致性[8]。最后，像 Frontiers AI Review Assistant[9]和 UNSILO Manuscript Evaluation[10]等多功能质控平台综合检查了稿件的写作质量与格式规范，从而提高了初审阶段的整体效率和质量。

总的来说，AI 技术通过智能化的工具和系统在稿件审查和质量控制的各个环节中提供了高效支持，不仅提升了编辑部的工作效率，也保证了稿件在质量和规范上的高标准要求。

第三步，审稿人匹配与利益冲突检测。其目标是提升评审的质

manuscripts[J]. arxiv preprint arxiv:2412.00281, 2024.

[1] Turnitin. iThenticate plagiarism detection software[EB/OL]. [2025-09-27]. http://www.ithenticate.com.

[2] Cabanac G, Labbé C, Magazinov A. Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals[EB/OL]. [2025-09-27]. http://arxiv.org/abs/2107.06751.

[3] Khalil M, Er E. Will chatgpt get you caught? rethinking of plagiarism detection[C]//International conference on human-computer interaction. Cham: Springer Nature Switzerland, 2023: 475-487.

[4] StatReviewer. StatReviewer: Automated statistical review[EB/OL]. [2025-09-27]. http://www.statreviewer.com.

[5] Nuijten M B, Epskamp S. statcheck: Extract statistics from articles and recompute p-values[EB/OL]. (2024)[2025-09-27]. http://statcheck.io.

[6] Ripeta. Research integrity platform[EB/OL]. [2025-03-27]. https://ripeta.com.

[7] Penelope.ai. Automated manuscript checking tool[EB/OL]. [2025-03-27]. https://www.penelope.ai/faq.

[8] Recite. APA and harvard citations checked instantly[EB/OL]. [2025-09-27]. https://reciteworks.com.

[9] Frontiers. Artificial intelligence to help meet global demand for high-quality, objective peer-review in publishing[EB/OL]. (2020-07-01)[2025-09-27]. https://blog.frontiersin.org/2020/07/01/artificial-intelligence-peer-review-assistant-aira.

[10] UNSILO. UNSILO manuscript evaluation[EB/OL]. [2025-09-27]. https://unsilo.ai/unsilo-manuscript-evaluation.

量与公平性。为此，系统需要将稿件分配给在相关领域具有高度契合度且无利益冲突的专家。AI 在这一过程中扮演着重要角色，尤其通过自然语言处理（NLP）与图谱分析技术来进行专家检索、匹配以及冲突检测。

在算法和模型方面，Charlin 等[1]提出了一种基于亲和度分数的整数规划模型，进一步将论文与审稿人档案嵌入潜在主题空间，从而增强匹配的准确性。Wu 等[2]开发了半自动的 COI 检测系统，并结合监督排序模型优化了审稿人的分配。Pradhan 等[3]则利用贪婪算法，在满足 COI 约束的前提下，进一步优化了审稿人分配的效果。此外，Leyton-Brown 等[4]开发了大型会议匹配（LCM）算法，在成千上万的论文中，平衡了专业领域的覆盖和审稿人的工作负载。Fu 等[5]与 Aitymbetov 和 Zorbas[6]则提出了跨学科的审稿团队组建方法，解决了多学科投稿的审稿人匹配问题。

在商业工具方面，多个平台通过集成数据资源来推荐审稿人。例如，Clarivate Reviewer Locator（集成于 ScholarOne）利用 Web of Science 和 Publons 数据进行审稿人推荐[7]；Aries Systems Reviewer Discovery（集成于 Editorial Manager）则基于 ProQuest 作者资料进行专家匹配[8]；Elsevier Reviewer Finder 基于 Scopus 数据库，帮助识别潜在的审稿人[9]。而 Frontiers COVID-19 Reviewer Recommender

[1] Charlin L, Zemel R. The Toronto paper matching system: an automated paper-reviewer assignment system[J]. 2013.

[2] Wu S, U L H, Bhowmick S S, et al. Pistis: A conflict of interest declaration and detection system for peer review management[C]//Proceedings of the 2018 International Conference on Management of Data. 2018: 1713-1716.

[3] Pradhan D K, Chakraborty J, Choudhary P, et al. An automated conflict of interest based greedy approach for conference paper assignment system[J]. Journal of Informetrics, 2020, 14(2): 101022.

[4] Leyton-Brown K, Nandwani Y, Zarkoob H, et al. Matching papers and reviewers at large conferences[J]. Artificial Intelligence, 2024, 331: 104119.

[5] Fu Y, Luo J, Nan G, et al. Peer review expert group recommendation: A multi-subject coverage-based approach[J]. Expert Systems with Applications, 2025, 264: 125971.

[6] Aitymbetov N, Zorbas D. Autonomous Machine Learning-Based Peer Reviewer Selection System[C]//Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations. 2025: 199-207.

[7] Clarivate. Web of Science Reviewer Locator[EB/OL]. [2025-09-27]. https://clarivate.com/webofsciencegroup/solutions/reviewerlocator.

[8] Aries Systems Corporation. Reviewer Discovery - free trial available[EB/OL]. (2013-06)[2025-09-27]. https://www.ariessys.com/newsletter/june-2013/reviewer-discovery-free-trial-available.

[9] Elsevier. Rolling out our new editorial system: EVISE[EB/OL]. (2015-10-22)[2025-09-27]. https://www.elsevier.com/connect/reviewers-update/rolling-out-our-new-editorial-system-evise.

则结合知识图谱与信息抽取方法，实现了更为精确的专家匹配[1]。中国自然科学基金委员会（NSFC）[2]的 AI 审稿人推荐系统结合了自然语言处理与语义分析，达到了约 80%的准确率，并内置了 COI 检测规则，进一步提高了审稿人推荐的效率和准确性。

在会议审稿人分配方面，多伦多论文匹配系统[3]（NIPS 2010）采用了主题建模与出版记录来自动建议审稿人分配。对于 AI 会议，许多研究提出了针对性的自动分配算法[4][5]。此外，模拟平台的使用也为审稿流程提供了优化方案，研究表明，这些平台可以缩短审稿时间约 30%[6]。

第四步，评审。这一阶段 AI 技术可显著提升审稿过程的效率和一致性。首先，基于大语言模型的工具可以为审稿人提供初步的评审报告框架，并自动生成要点提示，帮助审稿人快速启动评审工作。其次，AI 可以结合统计分析和方法学检查结果，为审稿人提供量化参考，辅助其打分和做出评判。此外，AI 还能够整合多位审稿人的意见，生成元评审报告，为期刊编辑提供摘要和比较分析，进一步支持决策过程。

评审过程通常包括两大步骤：同行评审和元评审。同行评审阶段，领域专家独立对稿件进行评价；而在元评审阶段，编辑将综合不同审稿人的意见，做出最终的稿件处理决定。由于审稿人意见常常存在分歧，且其建议可能因规范参考不充分而产生不一致，自动化系统在这一阶段显得尤为重要，尤其是当系统能够提供透明、可检查的决策依据时。基于 PeerRead 科学同行评审数据集，已有研究

[1] Frontiers. COVID-19 research funding monitor[EB/OL]. [2025-09-27]. https://coronavirus.frontiersin.org/covid-19-research-funding-monitor.

[2] Liu X, Wang X, Zhu D. Reviewer recommendation method for scientific research proposals: a case for NSFC[J]. Scientometrics, 2022, 127(6): 3343-3366.

[3] Charlin L, Zemel R. The Toronto paper matching system: an automated paper-reviewer assignment system[J]. 2013.

[4] Al Mahmud T, Hossain B M M, Ara D. Automatic reviewers assignment to a research paper based on allied references and publications weight[C]//2018 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018: 1-5.

[5] Kalmukov Y. An algorithm for automatic assignment of reviewers to papers[J]. Scientometrics, 2020, 124(3): 1811-1850.

[6] Mrowinski M J, Fronczak P, Fronczak A, et al. Artificial intelligence in peer review: How can evolutionary computation support journal editors?[J]. PloS one, 2017, 12(9): e0184711.

使用深度学习网络预测同行评审报告中的稿件接受或拒绝决策，并生成最终的荟萃评审。研究发现，系统的推荐决策与原始决策之间具有较高的一致性，且在二元决策（接受/拒绝）的预测中，准确率可达到 74%-86%，优于传统的机器学习算法和定制化的同行评审判断算法[1]。另外，对审稿报告情绪的分析也被证实能够预测会议论文的最终决策（接受或拒绝）以及资助计划的评分[2]。例如，PeerJudge 系统通过情绪分析技术来估计审稿报告中的赞扬与批评强度，从而为编辑的管理决策提供依据。PeerJudge 在预测 F1000Research 审稿人的决定时，表现出中等程度的准确性[3]。

总体而言，AI 在科研评审流程中的嵌入显著简化了环节、缩短了周期并提升了资源配置效率。与传统科研评审流程相比，力在不降低评审质量前提下，AI 有助于实现审稿时长的可观压缩、工作量从高认知任务向高价值判断的重分配，以及流程数据的端到端可追踪与审计化管理。综合来看，AI 为评审流程提供了标准化、可扩展与数据驱动的支撑框架，为期刊与资助机构在高负载情境下维持效率与质量的一致性提供了可验证的技术路径。

在大语言模型驱动下，教育研究的知识生产呈现出平台化、开放化与评价多元化的总体趋势。第一，研究结构与平台方面，生产范式由“论文—个人—课题组”逐步转向“模型—数据—算力—评测”的耦合架构，预印本与开放评审平台、模型/代码/数据托管平台以及基准与排行榜共同构成从产出到验证的基础设施链条，RAG、合成数据、人机协同标注与代理式工作流等方法被广泛采用。

第二，开放科学与共享机制方面，数据、代码与模型权重开源日益常态化，配套的数据卡与模型卡、许可与引用规范、DOI 与版

[1] Pradhan T, Bhatia C, Kumar P, et al. A deep neural architecture based meta-review generation and final decision prediction of a scholarly article[J]. Neurocomputing, 2021, 428: 218-238.
[2] Chakraborty S, Goyal P, Mukherjee A. Aspect-based sentiment analysis of scientific reviews[C]//Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. 2020: 207-216.
[3] Thelwall M, Kousha K, Wilson P, et al. Predicting article quality scores with machine learning: The UK Research Excellence Framework[J]. Quantitative Science Studies, 2023, 4(2): 547-573.

本化管理、开放数据政策等举措提升了研究的可复现性与可追溯性，同时也带来隐私合规、版权边界、来源透明度与评测集污染等新的治理议题。

第三，科研评审方面，期刊与会议强化了 AI 使用披露、数据与代码可用性以及伦理审查要求，逐步引入基于数据库的审稿人匹配、AI 辅助初筛与语言规范、开放评审与发表后评议（部分场景中配合注册报告），并开始将数据集、代码库、基准与工具等非论文型产出纳入评价体系，但关于 AI 介入审稿的边界、偏见与利益冲突管理以及复现验证的成本与责任分配仍在探索之中。总体而言，知识生产的新生态显著提高了知识流通效率与工具迭代速度，同时对研究治理与规范框架提出了更高要求，亟需配套的政策与标准以平衡质量、效率与伦理。

第 9 章 大语言模型驱动的合成数据在教育研究中的应用

围绕“AI 大语言模型生成内容如何进入科学知识生产”的路径，正逐步形成两条具有代表性的取向[1]。其一，将大语言模型视作研究对象本身，借用或改造社会科学方法来检验与解构这一技术系统[2]。例如，随着诸如机器心理学（Machine Psychology）等议题的推进[3][4]，一些学者基于大语言模型的类人语言与行为表现，尝试以心理学实验[5]、民族志变体[6]等方法，来探测其能力边界与机制特征，作为理解大语言模型内部运行机制的一条“替代性”路径。其二，将大语言模型作为研究方法论的一部分，用于增强既有方法[7]。典型方向包括两类：一是辅助或改进原有的研究流程，二是直接生成研究所需的合成数据，以补充或替代传统数据采集。

本章聚焦第二条路径中的“合成数据”议题：讨论如何利用大语言模型生成面向教育研究的合成数据，特别是那些希望用模型生成的数据来模拟或推断人类在不同情境中可能采取的行动、反应或回应方式的研究实践。

正如 Grossmann 等人所指出的，大语言模型在模拟类人翻译与行为上的能力，为大规模、低成本地检验关于人类行为的理论与假设提供了前所未有的机会[8]；一些观点甚至进一步设想将合成数据

[1] Rossi L, Harrison K, Shklovski I. The problems of LLM-generated data in social science research[J]. Sociologica: International Journal for Sociological Debate, 2024, 18(2): 145-168.
[2] Moats D, Seaver N. “You social scientists love mind games”: Experimenting in the “divide” between data science and critical algorithm studies[J]. Big Data & Society, 2019, 6(1): 2053951719833404.
[3] Hochman G, Shulman B, Sutskever I, et al. Machine Psychology [EB/OL]. (2023-03-23)[2025-09-14]. http://arxiv.org/abs/2303.13988.
[4] 潘香霖,褚乐阳,陈向东.窥探机器之窍：机器心理学视角下的大模型教育应用[J].远程教育杂志,2023,41(06):52-61.
[5] Almeida G F C F, Nunes J L, Engelmann N, et al. Exploring the psychology of LLMs’ moral and legal reasoning[J]. Artificial Intelligence, 2024, 333: 104145.
[6] Demuro E, Gurney L. Artificial intelligence and the ethnographic encounter: Transhuman language ontologies, or what it means “to write like a human, think like a machine”[J]. Language & Communication, 2024, 96: 1-12.
[7] Møller A G, Dalsgaard J A, Pera A, et al. Parrot Dilemma: Human-Labeled vs. LLM-augmented Data in Classification Tasks [EB/OL]. (2023-04-27)[2025-09-14]. http://arxiv.org/abs/2304.13861.
[8] Grossmann I, Feinberg M, Parker D C, et al. AI and the transformation of social science research[J]. Science, 2023, 380(6650): 1108-1109.

视为“硅基”样本，在特定任务上，大语言模型有望在数据收集环节替代部分人类参与者[1]。然而，倘若大语言模型能够在研究设计中扮演参与者的角色，生成回答、做出决策，甚至对自身行为进行元认知层面的“反思”，教育研究的诸多环节将可能被重塑[2]。长期以来，教育研究常常受到样本获取困难与成本高昂、伦理约束与情境敏感性、实验与实地研究之间外部效度的张力、可重复性与可比性的挑战，以及时空与资源限制等多重因素的掣肘。合成数据的出现，为缓解其中若干痛点提供了技术抓手：它或可降低数据获取门槛、加速原型验证与假设筛选、支持罕见情境与长尾现象的模拟，并促进跨机构与跨时空的可重复分析。

接下来的章节将系统回顾相关文献与新近争论，梳理合成数据在研究领域的应用脉络，明确大语言模型介入后的关键变化，归纳典型应用领域，总结当前主要争议与已提出的应对策略。

9.1 大语言模型对合成数据的影响

目前，学术界普遍认为，大语言模型生成的数据可以被视为一种生成的合成数据[3][4][5]。合成数据较为公认的定义是：使用明确的数学模型或算法生成的数据集，而不是直接来源于现实世界的观测或常规数据采集过程[6]。换言之，研究者以对数据生成机制的假设为起点，借助模型化与抽样来“制造”可用于分析的数据。这一定义为各类合成数据的技术形态提供了共同的基础。

9.1.1 合成数据的早期应用

从历史上看，合成数据最早在统计学与信息隐私保护的交叉领

[1] Dillion D, Tandon N, Gu Y, et al. Can AI language models replace human participants?[J]. Trends in Cognitive Sciences, 2023, 27(7): 597-600.
[2] Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.
[3] Resnik D B, Hosseini M, Kim J J H, et al. GenAI synthetic data create ethical challenges for scientists. Here’s how to address them[J]. Proceedings of the National Academy of Sciences, 2025, 122(9): e2409182122.
[4] Rossi L, Harrison K, Shklovski I. The problems of LLM-generated data in social science research[J]. Sociologica: International Journal for Sociological Debate, 2024, 18(2): 145-168.
[5] Grossmann I, Feinberg M, Parker D C, et al. AI and the transformation of social science research[J]. Science, 2023, 380(6650): 1108-1109.
[6] Jordon J, Szpruch L, Houssiau F, et al. Synthetic Data--what, why and how?[EB/OL]. (2022-05-06)[2025-09-14].https://arxiv.org/abs/2205.03257.

域被提出[1]，如图 9-1 所示。早在 20 世纪九十年代，Rubin 将合成数据描述为“非经直接测量而获得、利用统计模型生成的人工数据”，并将多重插补（Multiple Imputation，MI）的思想用于为人口普查与大型社会调查构造可发布的“替身数据”，以在不披露微观个体信息的前提下，尽可能保留统计推断所需的结构特征[2]。随后，Little 等人发展出“部分合成”的策略，即只对原始数据中的敏感变量进行替换，从而在隐私保护与分析效度之间取得更细粒度的平衡[3]。

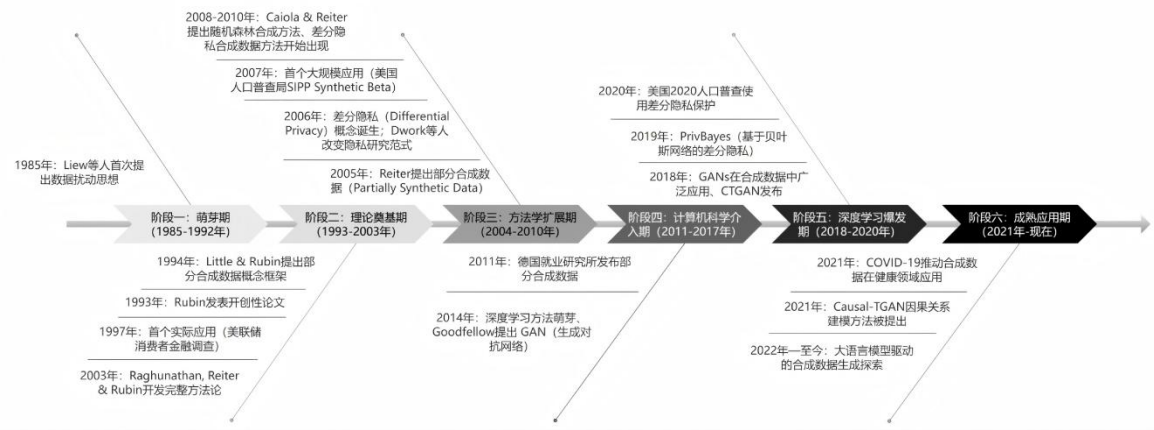

图 9-1 合成数据发展的历史脉络

由此，合成数据与统计披露控制（Statistical Disclosure Control，SDC）共同构成了一条面向公共数据治理的技术路线：在政府与公共机构，尤其是承担人口、就业与社会项目统计的部门，研究人员获得的是经合成与扰动处理后的替代数据[4][5]。一个重要里程碑是美国人口普查局自 2003 年起推出基于收入与项目参与调查（SIPP）的部分合成数据产品，该产品从早期测试版逐步走向成熟（2007 年向公众发布）并持续迭代，至 2018 年已发布到第七版，研究者可在严格的访问与使用规范下，将其用于学术分析[6]。

[1] Drechsler J, Haensch A C. 30 years of synthetic data[J]. Statistical Science, 2024, 39(2): 221-242.
[2] Rubin D B. Statistical disclosure limitation[J]. Journal of official Statistics, 1993, 9(2): 461-468.
[3] Little R J A, Raghunathan T. Should imputation of missing data condition on all observed variables[C]//Proceedings of the Section on Survey Research Methods. Alexandria: American Statistical Association, 1997: 617-622.
[4] Abowd J M, Vilhuber L. How protective are synthetic data?[C]//International Conference on Privacy in Statistical Databases. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 239-246.
[5] Raghunathan T E. Synthetic data[J]. Annual review of statistics and its application, 2021, 8(1): 129-140.
[6] Benedetto G, Stanley J C, Totty E. The creation and use of the SIPP synthetic Beta v7. 0[J]. US Census Bureau, 2018.

开放科学与数据共享运动进一步扩大了这一视域下的合成数据应用范围。随着越来越多经过隐私处理的公共数据进入研究者视野，教育研究成为间接受益者：一方面，教育研究者可以利用合成微观数据对关心的教育议题开展可重复的研究；另一方面，合成数据的共享与低敏感度特点降低了机构间协作与方法检验的门槛[1]。这种应用经验为教育研究提供了合规的载体。研究者可以在合成数据上进行模型原型设计、功效分析与稳健性检验，再在受控环境中转向真实数据验证。

值得注意的是，进入机器学习与计算机视觉迅速发展的阶段，合成数据也成为了计算机视觉等领域的重要概念。以生成对抗网络、变分自编码器、仿真引擎等为代表的生成技术开始在图像、文本与表格数据上大规模应用[2][3]。其核心动因是成本与可扩展性，合成数据能够以更低成本覆盖稀有场景、长尾分布与边缘条件，显著提升模型训练与评测的充分性。这对于教育研究而言，在学习分析、教育测量与智慧教室相关的计算机视觉等研究方向上，对稀缺或难以采集的情境也有真实的需求。来自仿真与深度生成的合成数据，可在不侵犯隐私的前提下补充训练样本，进行算法测试，并为跨机构的基准评测提供可复用的公共资源。

需要承认的是，教育研究与合成数据的发展并非一路并行。作为跨领域的概念，合成数据在各自领域有相对独立的术语、评估标准与应用生态。合成数据最初以统计生成与披露控制为目的，在开放科学与公共治理中扎根。随机器学习而兴起的深度生成技术，则将其推向高维、复杂模态与大规模应用。自然语言处理技术的跃迁，使文本与交互行为也可被以语言的形式合成。正是在这条演进线上，由大语言模型生成的合成数据成为近年的新焦点。

[1] Liu Q, Shakya R, Jovanovic J, et al. Ensuring privacy through synthetic data generation in education[J]. British Journal of Educational Technology, 2025, 56(3): 1053-1073.
[2] Nikolenko S I. Synthetic data for deep learning[M]. Cham, Switzerland: Springer, 2021.
[3] Jordon J, Szpruch L, Houssiau F, et al. Synthetic Data--what, why and how?[EB/OL]. (2022-05-06)[2025-09-14].https://arxiv.org/abs/2205.03257.

9.1.2 大语言模型生成合成数据的特点

通常，合成数据被认为可以解决三个主要数据挑战中的任何一个或全部：数据稀缺性、数据隐私和数据偏见[1]。尽管生成合成数据无疑能通过按既定规范生产更多数据来解决稀缺性问题，但这些数据质量是否足以发挥作用，以及能否解决隐私和偏见的挑战，则取决于具体情境[2][3][4]。

随着大语言模型在合成数据开发中的出现，我们看到了旨在评估生成数据集质量的新框架的发展。例如，Eigenschink 等人为合成数据集提出了五个评估标准：代表性、新颖性、真实性、多样性和一致性[5]。他们认为，通过大语言模型产生的高质量合成数据应该能够捕捉原始数据的总体层面特征，创造出新颖且真实（基于我们对原始数据的了解）的数据点，并在与原始数据保持一致性的同时展现内部多样性。他们还指出，单个标准的重要性在不同领域之间差异显著，并且测试这些标准的方法也各不相同，其重点在于合成数据再现原始数据关键特征的能力。

目前，大语言模型的发展和迅速普及，已导致这些系统被用于广泛的数据生成任务。Whitney 和 Norman 区分了生成式、增强式和程序化创建的合成数据集[6]，其区分标准是“它们在多大程度上衍生自真实世界的训练数据集”。例如，程序化创建的合成数据依赖于专用模型，这些模型沿着一组明确预设的参数来创建数据；而生成式数据，例如大语言模型所产生的数据，则是模型在响应特定输入时从训练数据集中抽象化而产生的。在这种情况下，训练数据集

[1] Van der Schaar M, Qian Z. AAAI Lab for Innovative Uses of Synthetic Data[J]. 2023.
[2] Belgodere B, Dognin P, Ivankay A, et al. Auditing and generating synthetic data with controllable trust trade-offs[J]. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2024.
[3] Figueira A, Vaz B. Survey on synthetic data generation, evaluation methods and GANs[J]. Mathematics, 2022, 10(15): 2733.
[4] Johnson E, Hajisharif S. The intersectional hallucinations of synthetic data[J]. AI & SOCIETY, 2025, 40(3): 1575-1577.
[5] Eigenschink P, Reutterer T, Vamosi S, et al. Deep generative models for synthetic data: A survey[J]. IEEE Access, 2023, 11: 47304-47320.
[6] Whitney C D, Norman J. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention[C]//Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024: 1733-1744.

至关重要，它能确保最终生成的数据集与期望从人类收集到的数据有实用性的相似之处。

需要注意的是，以 Transformer 架构驱动的语言模型，其构建目标并非生成与给定原始数据集相似的数据，而是根据文本的语言模式，被训练来预测下一个最可能出现的字母、单词或词组。然而，在越来越庞大的数据上进行训练后，大语言模型能够模仿人类生成的内容[1]，并展现出执行训练数据中未明确包含的任务的涌现能力[2]。在这里，模型并不会刻意遵循数据中特定的预先存在的模式[3]。

鉴于当前的大语言模型能够轻易地就几乎无限多的话题生成与人类相似的文本，因此不难想象这些技术可以被用来为研究目的生成数据。这种由大语言模型生成的数据集被视为一种便捷的数据类型，可以通过在提示工程上投入足够精力来获取，有时还会使用不同的可用的大语言模型实现进行比较[4][5]。因此，可以将大语言模型归类为一种全新的合成数据生成器，则将彻底改变现有合成数据的创建思路，并呈现出新的特点[6]。

（1）数据深度合成

传统合成数据生成本质上是一种统计推测过程，大语言模型驱动的深度生成实现了语义层面的个体模拟，这包含两个根本性转变：

一方面，个体模拟机制不再是简单的变量取值拼装，而是通过“虚拟个体”来参与研究。典型做法是使用角色化提示来实现个体级条件生成[7]。研究者会在提示中明确个体的关键背景与特征，如

[1] Jakesch M, Hancock J T, Naaman M. Human heuristics for AI-generated language are flawed[J]. Proceedings of the National Academy of Sciences, 2023, 120(11): e2208839120.
[2] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
[3] Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models [EB/OL]. (2022-06-14)[2025-09-14]. http://arxiv.org/abs/2206.07682.
[4] Dillion D, Tandon N, Gu Y, et al. Can AI language models replace human participants?[J]. Trends in Cognitive Sciences, 2023, 27(7): 597-600.
[5] Horton J J. Large language models as simulated economic agents: What can we learn from homo silicus?[R]. National Bureau of Economic Research, 2023.
[6] Jordon J, Szpruch L, Houssiau F, et al. Synthetic Data--what, why and how?[EB/OL]. (2022-05-06)[2025-09-14].https://arxiv.org/abs/2205.03257.
[7] Törnberg P, Valeeva D, Uitermark J, et al. Simulating social media using large language models to evaluate alternative news feed algorithms [EB/OL]. (2023-10-11)[2025-09-14]. http://arxiv.org/abs/2310.05984.

社会—人口学信息、性格倾向、学业水平与学习经历、情绪与动机状态、语言风格与策略偏好等，让模型基于这一稳定的身份框架生成具有一致性的响应[1]。一个简化的研究为例，在生成“学习动机”相关数据时，传统方法依据量表各维度的分布参数进行抽样；而利用大语言模型，则先构建一个具备特定年级、学业水平、家庭与文化背景、近期学习事件与情绪状态的“虚拟学生”，再让其以该身份连续完成动机量表、开放性问答与情境反应任务。

角色提示的关键在于两点：一是对身份特征进行显式约束，使不同任务间的反应在语义与逻辑上保持自洽；二是让模型的语言生成能力填充传统变量难以刻画的行为细节与理由化表述[2]。例如，对某项选择的自我解释、对未来努力的预期、对教师反馈的情绪反应等。这样得到的不只是“分数”，而是一条可供分析的个案轨迹，既可映射到既有量表维度，也可支持质性编码与混合方法分析。

另一方面，情境化的数据合成。大语言模型驱动的深度生成把“个体级别的模拟”作为基本单元，扩展了经典 ABM（Agent-Based Modeling，基于主体的仿真）的方法论边界，形成了一类能够产出情境化合成数据的新路径[3][4]。其关键在于生成机制发生了重大转变：大语言模型作为大脑，将状态、情境、记忆、目标等联合条件映射为行动与语言产出，使虚拟个体能以可解释、可交互的方式参与研究[5]。

与传统的 ABM 相比（通常依赖手工编写的规则库与有限状态机），大语言模型驱动的 ABM 在三个方面显著提升了涌现行为的层级与丰富度。第一，个体层面的多样性与可塑性更强。研究者通

[1] 杜君磊,李爽,陈靖茜,等.大语言模型在教育研究样本模拟中的保真度与偏差分析——以在线自我调节学习能力为例[J].远程教育杂志,2025,43(03):73-86.
[2] Törnberg P, Valeeva D, Uitermark J, et al. Simulating social media using large language models to evaluate alternative news feed algorithms [EB/OL]. (2023-10-11)[2025-09-14]. http://arxiv.org/abs/2310.05984.
[3] 米加宁.生成式治理：大模型时代的治理新范式[J].中国社会科学,2024,(10):119-139+207.
[4] 吕鹏.大型社会模拟器：社会知识生产的大科学装置[J].探索与争鸣,2024,(11):13-16+209.
[5] Wu Z, Peng R, Han X, et al. Smart agent-based modeling: On the use of large language models in computer simulation[EB/OL]. (2023-11-10)[2025-09-14]. https://arxiv.org/abs/2311.06330.

过角色提示，在同一建模框架下即可生成具有差异化的虚拟群体，且这些主体能在开放情境中作出连贯的自然语言与行动反应[1]。第二，跨时延续与情境一致性得到加强。通过引入情节记忆与语义记忆的外部存储与检索、状态日志更新，主体可以“记住”过去的互动并据此调整策略，从而在较长时段内维持身份一致性与目标推进[2][3]。第三，群体交互的宏观涌现更接近“社会拟像”。在多主体环境中，大语言模型驱动的主体能够生成具有可辨识的社会规范、协作模式与文化特征的群体行为轨迹，呈现出传统规则式 ABM 难以覆盖的细节与非预设模式[4]。已有工作显示，相较于规则式主体，这类大语言模型驱动的主体在复杂场景中展现出更高水平的涌现行为与情境细节[5]。

对于教育研究而言，这种条件下的数据生成允许研究者在不先行“压扁”复杂现象为少数可测变量的前提下，先获得贴近原始社会活动层面的过程性材料，再从中提炼变量或构念进行检验[6][7]。比如，在探究形成性评价的反馈效应时，可生成学生对同一作业的不同层级反馈文本及其理由链，能同时产出对应的评分标注与时间过程；在阅读理解研究中，可生成学生对文本的逐段思维出声（Think-aloud）及错因自述，对其策略使用进行标签化编码。这种“个案级合成”让研究者能够在保持构念语义与教育情境一致的同时，获得过程数据与结果数据的一体化表示，为后续的定量建模与定性分析提供统一入口。

[1] Törnberg P, Valeeva D, Uitermark J, et al. Simulating social media using large language models to evaluate alternative news feed algorithms[EB/OL]. (2023-10-5)[2025-09-14].https://arxiv.org/abs/2310.05984.
[2] Argyle L P, Busby E C, Fulda N, et al. Out of one, many: Using language models to simulate human samples[J]. Political Analysis, 2023, 31(3): 337-351.
[3] Park J S, O'Brien J, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th annual acm symposium on user interface software and technology. 2023: 1-22.
[4] 庞珣.人工智能赋能社会科学研究探析——生成式行动者、复杂因果分析与人机科研协同[J].世界经济与政治,2024,(07):3-30+153.
[5] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.
[6] Strielkowski W, Grebennikova V, Lisovskiy A, et al. AI-driven adaptive learning for sustainable educational transformation[J]. Sustainable Development, 2025, 33(2): 1921-1947.
[7] Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and individual differences, 2023, 103: 102274.

（2）数据规模化合成

与传统数据采集动辄需要跨界合作、长周期立项不同，基于大语言模型的合成数据生成把“规模”这一能力下放给了个体研究者与小型团队。任何人只需通过参数化的提示与简单的管线配置，便可在极低成本下批量生成满足特定研究设定的数据集。这种由模型即服务（Model as service）所驱动的能力转移，被不少学者视为研究民主化的具体体现：不再只有资源充沛的团队才能获得“大样本”，基层研究者也能按自己的理论设定快速搭建初步实验场[1]。

规模化并非单纯的“数量膨胀”，而是依托可编排的生成机制，将量与质绑定在一起。研究者可以通过分层配额与参数化条件进行定向扩充；通过提示版本控制、温度设置确保可复现的多样性；并借助自动化的去重与一致性校验保持样本的“有效信息量”。由此，规模化的重要内涵体现在利用大语言模型可以在保持情境特异性与理论相关性的前提下，成体系地获取足量数据。

一个典型的例证来自教育技术领域的自动化问题生成研究[2]。研究者希望构建一个涵盖不同认知层级的教育问题数据集，用于验证布鲁姆分类法在大语言模型时代的适用性。在传统模式下，这需要教育专家、心理测量学者与计算机科学家的长期协作，仅数据采集环节就可能耗费数年时间。然而，借助五个主流大语言模型，研究团队仅通过设计五套渐进复杂的提示策略便在数周内生成了2550个数据科学领域的问题样本。更为重要的是，这一生成过程严格按照布鲁姆认知分类法的六个层级（记忆、理解、应用、分析、评价、创造）进行分层配额，每个层级的问题数量与难度分布都经过参数化控制。最终，78%的生成问题通过专家评估达到“高质量”标准，65.56%精准对应预设的认知层级，形成了一个兼具理论严谨性与实

[1] 王飞跃,缪青海.平行科学：大模型时代 AI4S 的前沿技术与框架体系[J].人民论坛·学术前沿,2024,(14):64-79.

[2] Scaria N, Dharani Chenna S, Subramani D. Automated educational question generation at different bloom’s skill levels using large language models: Strategies and evaluation[C]//International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2024: 165-179.

用价值的 DataScienceQ 数据集。这一案例展现了大语言模型如何让个体研究者在保持学理深度的前提下，以前所未有的速度获取符合特定理论框架的大规模实验数据。

因此，在教育研究的语境中，规模化至少发挥两方面的关键作用。其一，支撑复杂统计分析所需的大样本，同时维持研究问题的情境贴合与构念对齐。研究者可以生成覆盖不同子群体与情境的样本池，用于功效分析、稳健性检验与长尾现象的检测，从而避免以往为求大样本而牺牲情境的折衷。其二，催化快速迭代的探索性研究范式。具体流程往往是：先在明确的理论假设与初步情境设定下生成一个小规模试验集，进行探索性分析以捕捉模式与异常；随后据此调整构念边界、情境约束与提示设计，定向生成第二轮数据并复核初步发现；再将最有价值的结果迁移到真实情境的小样本上进行外部效度验证[1][2]。这样的生成、分析、修正、再生成、验证的闭环可以高频滚动，每一轮都在前一轮的基础上收敛问题空间，使研究过程本身转化为动态的知识发现[3]。这种高频迭代适合理论边界模糊、现象复杂多变的新兴议题，研究者不必在研究伊始一次性做出所有关键决策，而是借由数据的即时反馈逐步聚焦与深化。

规模化生成同样为教育垂直大语言模型与能力评测的研发提供了基础设施。当前，面向教育任务的测试数据集逐步被用于检验大语言模型的基础能力与迁移表现；在强调开发教育垂直领域人工智能的趋势下，合成数据可以作为“类真实”的预评估材料与压力测试工具，支持对题型、构念与难度结构的系统化覆盖[4]。需要强调的是，这类数据也被视为对真实大规模社会测评（如 PISA 测试）的

[1] Kieser F, Wulff P, Kuhn J, et al. Educational data augmentation in physics education research using ChatGPT[J]. Physical Review Physics Education Research, 2023, 19(2): 020150.

[2] 杜君磊,李爽,陈靖茜,等.大语言模型在教育研究样本模拟中的保真度与偏差分析——以在线自我调节学习能力为例[J].远程教育杂志,2025,43(03):73-86.

[3] 胡安宁.洞中流影：基于大语言模型的硅基样本在社会科学研究中的应用反思[J].社会发展研究,2025,12(01):22-39+242-243.

[4] Biancini G, Ferrato A, Limongelli C. Multiple-choice question generation using large language models: Methodology and educator insights[C]//Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. 2024: 584-590.

补充与前瞻性实验场[1]：其价值在于快速原型、情境覆盖与敏感性分析，帮助研究者在进入昂贵的真实评测或实地研究之前，先进行多轮的预测试。

（3）数据形态多元化

大语言模型驱动的合成数据生成，显著拓展了传统数值型表格数据的边界，走向文本、对话、叙述、标注、过程日志与行为序列等更贴近教育实践的复杂形态。由此，合成数据不再只是量化分析的统计数据，也能承载情境与语义信息，支持质性与混合方法研究对过程、意义以及机制的追问。这一转变的价值在于把教育活动的多个层面（语言、行动、时间、角色、情境）整合进同一数据对象，形成可被多方法联动检验的研究材料。

首先，作为质性研究的数据增强与可替代证据源。过去的合成数据难以满足质性研究对语义丰富性与情境厚度的要求。多元化形态使研究者能够生成访谈文本、个体叙事、课堂观察记录与同伴互动片段等语言、情境以及身份一体的材料，并配套结构化标注，由此支持扎根理论、话语分析、叙事研究等方法的编码与比较[2]。既有研究显示，可用“人物角色式提示”生成访谈回应，并与真实访谈进行对照[3]。尽管普遍共识认为大语言模型并非“预测个体真实反应”的最佳工具，但其社会模拟在设计研究、用户体验测试等场景中可作为启发与补充[4]。长期以来，设计实践使用“用户画像”“情境想象”等方法来弥补样本难以覆盖的用户群。在代表性与包容性受限的研究条件下[5][6]，大语言模型可在成本敏感、工程可行的

[1] Argyle L P, Bail C A, Busby E C, et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale[J]. Proceedings of the National Academy of Sciences, 2023, 120(41): e2311627120.

[2] 陈向东，陈鹏，张蕾.基于大语言模型的教育质性研究：理论与实践[M].上海：华东师范大学出版社，2026.

[3] Hämäläinen P, Tavast M, Kunnari A. Evaluating large language models in generating synthetic hci research data: a case study[C]//Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023: 1-19.

[4] Tjuatja L, Chen V, Wu T, et al. Do llms exhibit human-like response biases? a case study in survey design[J]. Transactions of the Association for Computational Linguistics, 2024, 12: 1011-1026.

[5] Elsayed-Ali S, Bonsignore E, Chan J. Exploring challenges to inclusion in participatory design from the perspectives of global north practitioners[J]. Proceedings of the ACM on Human-Computer Interaction, 2023,

前提下提供一种“看似合理的替代用户数据源”。

其次，多模态数据的敏捷转换。这种转换在两个层面上展现其应用价值：一是数据形态的灵活转换。大语言模型能够实现不同数据形态间的智能转换。例如，将原始的访谈文本转换为合成的编码表格，或将原始统计数据转化为直观的叙述性描述。这种转换能力使得研究者能够在同一研究中灵活运用不同的分析工具和方法，实现方法论上的互补和验证[1][2]。二是异构数据源的统一化处理。教育研究中，需要整合来自不同渠道、不同格式的数据。现有的一些大语言模型驱动的科研辅助系统（如 CRESt、CollabCoder 等）能够快速处理这些数据[3][4]：对这些异构数据进行统一化的语义处理，例如在扎根理论研究中，将来自访谈、观察、文档分析等不同来源的质性材料转换为统一的编码体系，形成合成的概念数据库，确保不同数据源在语义层面的一致性和可比性。

最后，构建模拟或仿真的教育情境。数据形态的多元化使得研究者能够搭建内在逻辑一致、可控又可变的“平行时空”教育情境：在统一的约束下，生成完整的课堂互动、测评过程、反馈循环与同伴协作等具体的合成“场景”。这些场景既可以保留教育现象的核心特征，又可系统操控情境参数，用于难以在真实环境实施的实验性探索。从更极端的案例看，多代理系统已被用于复杂社会场景的合成与推演，如外星文明互动的宏观情境模拟[5]，战争期间国家之

7(CSCW1): 1-25.
[6] Sin J, L. Franz R, Munteanu C, et al. Digital design marginalization: New perspectives on designing inclusive interfaces[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1-11.
[1] Roberts J, Baker M, Andrew J. Artificial intelligence and qualitative research: The promise and perils of large language model (LLM)‘assistance’[J]. Critical Perspectives on Accounting, 2024, 99: 102722.
[2] Dengel A, Gehrlein R, Fernes D, et al. Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on computer science education[C]//Informatics. MDPI, 2023, 10(4): 78.
[3] Zhichu Ren ,Zhen Zhang ,Yunsheng Tian, et al..CRESt-Copilot for Real-world Experimental Scientist [EB/OL].(2023-11-16)[2025-09-14]. https://chemrxiv.org/engage/chemrxiv/article-details/64a81dcd6e1c4c986bf83225.
[4] Gao J, Guo Y, Lim G, et al..CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis[EB/OL].(2023-10-11)[2025-09-14]. https://arxiv.org/abs/2304.07366.
[5] Xue, Z., Jin, M., Wang, B., Zhu, S., Mei, K., Tang, H., Hua, W., Du, M., & Zhang, Y. What if llms have different world views: Simulating alien civilizations with llm-based agents.[EB/OL].(2024-02-20)[2025-09-14]. https://arxiv.org/abs/2402.13184.

间的博弈模拟[1]等，实现大尺度、细节丰富的拟像世界的可能。在教育领域，这类通过模拟或仿真实现的“平行情境”可作为假设检验与方案预评估的试验台，帮助研究者在进入昂贵或高风险的真实测评与实地研究前，完成前置的筛查、预估工作。

尽管大语言模型如今能够生成多种形态的合成数据，但需要认识到，这些看似多样的特性共享着统一的基础，也即文本作为合成数据表征的媒介。即便生成的最终产物呈现为数值量表、结构化表格、多模态标注或行为序列，其生成机制本质上都是通过文本提示与文本响应来实现的。当研究者要求模型输出一个李克特量表的数值（如 1-5 分），模型实际上是在文本空间中理解任务语义，并将响应解码为特定的符号表示；当生成多主体交互日志时，本质仍是按文本化的规则与约束进行序列化输出[2]。这一文本中心性源于大语言模型训练的根基，也即海量的自然语言语料，使得任何形态的合成数据都携带着这一语言先验的印记。

正是这种文本表征的普适性，让研究者在使用大语言模型生成合成数据时，实际上是在调用并重组其参数空间中压缩的人类知识与表达模式。从这个意义上讲，选择 GPT、Gemini 还是 DeepSeek 等特定模型生成合成数据并非决定性差异，这些模型都已跨越了某个规模阈值。在这个阈值之上，语言模型的参数化知识表示、能力涌现、心智理论模拟与多样化内容生成等特性得以稳定展现，这些特性是规模化训练在语言建模任务上的共同产物[3]。也即，当模型参数量、训练数据量与计算资源达到一定水平后，从“预测下一个词”（Next token prediction）这一简单目标中涌现出了对复杂情境的理解、对多角色视角的切换、对因果与时序关系的隐式编码，以

[1] Xinyue Q, Yao Z, Xiang X, et al. An Interview System Based on Large Language Models and Multi-Agent Interactions[C]//2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE, 2024: 1-8.

[2] Rossi L, Harrison K, Shklovski I. The problems of LLM-generated data in social science research[J]. Sociologica: International Journal for Sociological Debate, 2024, 18(2): 145-168.

[3] 陈向东,赵丽娟,刘泽民.拓展学科的疆域：大模型的涌现能力对学习科学的影响[J].现代教育技术,2024,34(01):44-54.

及对不同文体与知识域的灵活适配等新能力。

9.2 合成数据的应用形式

大语言模型赋能的合成数据建模尺度，理论上可以精细到语言能够触及的任何维度。就目前的研究进展而言，这一尺度已经覆盖了三个彼此关联却相对独立的层面：其一是精细化的个体建模，用以刻画认知、动机、策略与风格等个体差异，代替传统研究中的人类样本；其二是由社会互动积累而成的社会情境模拟，用以推演制度、规范与资源配置对宏观分布与涌现模式的塑形；其三是对现实的环境模拟，通过设定规则、材料与资源约束，构造可控又可变的平行世界。需要承认的是，对新工具的分类本身总带有方法论上的暧昧：边界常常并不泾渭分明，许多研究天然跨层级、跨情境。然而，正是基于这种“可分亦可合”的特性，我们更需要给出一个工作性的组织框架，以便在不同研究背景下把握大语言模型用于合成数据生成的主要趋势。为此，我们将现有的方法论实验围绕三个应用领域加以组织：模拟个体、模拟社会以及模拟世界。

9.2.1 模拟个体

驱动社会科学领域的研究者以合成数据模拟个体的重要动因，是在可控、低成本的条件下替代传统实验环境中的研究参与者，从而获得对人类反应的强有力推断[1]。对于教育研究而言，这一做法尤为必要：许多关键问题，如测量工具的敏感性、干预反馈的即时反应，以及个体差异对学习路径的影响等都发生在微观层面，但真实招募往往受限于时间、经费与伦理成本。个体级别的合成数据提供了一个可重复操控的试验工具，这一方法机制相对明确：研究者通过提示将大语言模型设定为具有人口统计与心理特征的“人类角色”，在特定测试情境与指令约束下输出回答或行为轨迹。此时，模型被视为一种“人类的隐式计算模型”，其输出是面向目标样本

[1] Argyle L P, Bail C A, Busby E C, et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale[J]. Proceedings of the National Academy of Sciences, 2023, 120(41): e2311627120.

的拟真表现[1]。换言之，数据被期望反映“如果该类个体处于此情境，将如何作答/行动”的内容分布。

在合成数据驱动的模拟个体相关的研究中，应用形式因研究目标导向的不同存在差异，形成了与传统社会科学研究范式相对应的两种主要路径，分别为数值类数据和文本类数据。

合成数值类数据作为第一类应用形式，主要特征在于将模型的文本生成能力转化为结构化的量化输出。在这种应用中，研究者通过提示工程，引导模型以特定身份特征的“虚拟个体”身份对量表题目、选择任务或评价问题给出数值化回应。相关领域为这一应用思路提供了大量应用示例。例如，Brand 等检验大语言模型在调查回答中与经济理论和既有消费者行为模式的一致性[2]；Motoki 等部署多种组织行为学量表，将大语言模型生成的作答与已发表的实证结果对照，发现合成数据总体上能够复制人类行为分布，由此提示其在调查工具验证与前测中的潜在用途[3]。

值得注意的是，在教育研究领域，也开始出现了利用大语言模型生成合成数值数据的创新探索。一项针对自我调节学习动机策略问卷（MSLQ）的研究案例显示了这种方法的可行性与潜力。研究者使用五个主流大语言模型（GPT-4o、Claude 3.7 Sonnet、Gemini 2 Flash 等）分别模拟 1000 名虚拟学生，对涵盖内在价值、自我效能感、考试焦虑、认知策略使用和自我调节五个核心构念的 44 个题目进行 7 点李克特量表评分。这项探索发现，不同模型在数据生成质量上存在显著差异，其中 Gemini 模型显示出更接近真实学生群体的响应变异性；同时，合成数据能够有效再现理论预期的心理测量学特征，包括合理的因子结构和构念间关系模式[4]。这些发现表明，

[1] Filippas A, Horton J J, Manning B S. Large language models as simulated economic agents: What can we learn from homo silicus?[C]//Proceedings of the 25th ACM Conference on Economics and Computation. 2024: 614-615..

[2] Brand J, Israeli A, Ngwe D. Using LLMs for market research[J]. Harvard business school marketing unit working paper, 2023 (23-062).

[3] Motoki F, Pinho Neto V, Rodrigues V. More human than human: measuring ChatGPT political bias[J]. Public Choice, 2024, 198(1): 3-23.

[4] Vogelsmeier L V D E, Oliveira E, Misiejuk K, et al. Delving Into the Psychology of Machines: Exploring the

在传统教育调研受限的情境下，大语言模型生成的合成数值数据为理论验证和教育心理测量研究提供了有效的替代工具，为教育量化研究方法的转型开辟了新的路径。

合成文本数据作为第二类应用，代表了一种更为丰富和自然的个体模拟形式。在此应用中，模型被要求以特定个体身份产生开放式的文本回应，如对深度访谈问题的阐述、对政策议题的观点表达，或对教育干预的反思性描述。相较于数值数据，文本数据承载了更多的语义信息和情感色彩，能够捕捉个体认知的复杂性和多样性。这种应用形式特别适用于质性研究方法的拓展，为内容分析、话语分析等传统方法提供了规模化的数据基础。教育领域也出现了相应的探索实践。

例如，在物理教育研究中，研究者要求 ChatGPT 扮演持有特定错误概念（如“离心力”或“冲量”概念）的学生，在回答标准化的物理概念测试题时，不仅做出选择，还要详细阐述其选择理由[1]。当面对圆周运动相关题目时，模拟“离心力前概念学生”的 ChatGPT 会选择某个选项并解释道：“我觉得有一个力在圆周运动过程中将物体向外推……当绳子断裂时，物体会被这种向外的力（离心力）推向外侧”。这些生成的解释文本不仅反映了选择的结果，更重要的是揭示了错误概念背后的完整认知逻辑和推理过程。这种文本化的认知模拟为教育研究者提供了大量可供深度分析的语言材料，使其能够系统地理解学生概念困难的根源，进而设计更有针对性的教学干预策略。

模拟个体的这两种应用路径各有优势，在实际的研究应用中并不是泾渭分明的。实际上，利用模拟个体的两种方法，可以取得量化与质性研究的统一。研究者可以同时获得结构化的数值数据和丰

Structure of Self-Regulated Learning via LLM-Generated Survey Responses[J].computers in human behavior, 2025（173）:108769.

[1] Kieser F, Wulff P, Kuhn J, et al. Educational data augmentation in physics education research using ChatGPT[J]. Physical Review Physics Education Research, 2023, 19(2): 020150.

富的叙述性文本，从而实现混合研究设计。例如，在前述物理教育研究中，ChatGPT 既提供了概念测试的量化得分（选择了哪个答案），又生成了解释推理过程的质性文本（为什么这样选择）。这使得研究者既能进行大样本的统计分析，识别概念理解的整体模式，又能通过文本分析深入探究错误概念的认知机制。

但需要注意的是，由于大语言模型的训练数据极有可能覆盖当前的研究情境，简单地复现了训练数据中的统计规律。这对于数值类合成数据尤为影响，因为李克特量表的固定选项范围限制了回应的变异性，使得模型倾向于生成符合“典型回应模式”的数值组合，而难以体现真实个体在认知、情感和态度上的微妙差异；对于文本数据，虽然表达空间更为开放，但面临着语言表达模式化的挑战——模型可能倾向于生成结构相似、用词规范的“标准化”文本，缺乏真实个体在表达习惯、语言风格和思维逻辑上的独特性，使研究者难以通过内容分析识别出超越统计规律的深层认知模式。

也正因如此，目前在模拟个体领域的另一个趋势是从大语言模型预训练语料“读出”隐含社会观念与态度。由于模型生成的数据可能更多地反映训练语料中的统计规律和群体模式，研究者转而将这一“缺陷”视为优势，把大语言模型当作一种由海量社会文本训练而成的“群体知识表征器”[1]。例如，Argyle 等发现大语言模型内隐含“类人概念关联”，可对概念、想法与态度之间的潜在关系进行近似建模[2]。这种研究路径在某种意义上拓展了合成数据的应用边界——不再依赖模型的生成过程，而是直接利用其内在表征空间中编码的信息作为研究数据。尽管这些数据在传统意义上并非通过“提示—生成”过程产出，但它们同样承载着从海量文本中抽象、压缩并重新编码的集体知识。

[1] Chuang Y S, Suresh S, Harlalka N, et al. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents [EB/OL]. (2023-11-16)[2025-09-14]. https://arxiv.org/abs/2311.09665.
[2] Argyle L P, Bail C A, Busby E C, et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale[J]. Proceedings of the National Academy of Sciences, 2023, 120(41): e2311627120.

这种应用形式的典型案例体现在对公众人物人格特质的社会认知研究中[1]。研究者利用了 GPT-3 在 12,288 维语义空间中存储的知识表征，直接提取公众人物姓名在该空间中的词嵌入向量，将这些数值化的“语义指纹”作为分析的基础数据。具体而言，这些词嵌入向量反映了姓名与其他词汇在语义空间中的相对位置关系。例如，“特朗普”在语义空间中更接近“傲慢”，而“德蕾莎修女”则更接近“同情心”。研究者随后采用岭回归模型，以这些高维向量作为预测变量，训练预测人类评判者对 226 位公众人物人格特质评价的模型。

令人惊讶的是，这些从模型表征空间中提取的“合成数据”能够准确反映社会群体的普遍认知模式，如将教皇约翰·保罗二世与高宜人性、高情绪稳定性关联。更为重要的是，模型甚至能够捕捉到训练文本中未明确表述的隐含关联，如自动将“酒精依赖”作为低责任心、低情绪稳定性的第三强预测因子，尽管在训练过程中的人格问卷从未直接涉及酒精相关内容。这表明，模型的内在表征空间不仅存储了语言的表层统计特征，更深层地编码了社会文化中广泛存在的刻板印象、价值判断和认知偏见。这种方法避开了个体化模拟的复杂性（如高度依赖提示文本的质量），直接将模型的知识结构本身作为一种特殊形式的合成数据来源，为理解公众的社会观念分布提供了一种全新的研究工具。对于教育研究而言，这一思路为揭示教育领域的集体社会认知提供了全新的窗口。正如研究者能够从 GPT-3 的语义空间中读取公众对政治人物的刻板印象，教育研究者同样可以探索社会对教育概念、教育角色和教育政策的普遍认知模式。比如，通过分析“优秀教师”“问题学生”“教育公平”等词汇在模型语义空间中的位置关系，研究者可以揭示这些教育概念在公众认知中与哪些特质或价值观相关联，从而理解社会文化如

[1] Cao X, Kosinski M. Large language models know how the personality of public figures is perceived by the general public[J]. Scientific Reports, 2024, 14(1): 6735.

何塑造我们对教育的理解和期望等。

综合上述应用方向，模拟个体的应用形式不仅为教育研究带来了数据获取与分析的范式转变，更重要的是正在重新定义教育研究中“数据”的概念：从单纯依赖真实个体的行为记录，扩展到利用人工智能系统中编码的集体知识进行教育现象的探索，为教育理论的验证和教学实践的改进开辟了新的路径。

9.2.2 模拟社会

合成数据在模拟社会维度的应用主要采用智能体框架，将大语言模型转化为具备社会互动能力的生成式行动者[1]。研究者通过为智能体设定档案、记忆、规划和行动等核心模块，将制度规则、信息可见性和互动边界等信息分别编码到这些模块中。随着互动的持续推进，这些模块中对应的文本信息会根据新的经历和环境变化不断更新，使智能体能够在动态变化的社会环境中保持行为的一致性和适应性，从而构建起一个可操控的社会仿真空间。智能体之间在这一空间中通过持续的语言互动形成复杂的社会动态[2]。这一过程中产生的合成数据，实质上是智能体在特定社会环境下长期互动所留下的数字轨迹，包括它们的对话记录、决策过程、关系演变，以及由此形成的网络结构和集体结果[3]。这些数据为研究者提供了重要的社会过程观察窗口，使原本难以在现实中控制变量或重复验证的社会机制，得以在可控环境中被系统检验和深入分析[4]。

通过多轮次、多主体的动态交互过程，研究者能够观察到许多重要的模拟社会现象。在这一领域具有方法学示范意义的经典研究，是 Park 等人在 2023 年开发的《斯坦福小镇》（Smallville）项目[5]。

[1] 庞珣.人工智能赋能社会科学研究探析——生成式行动者、复杂因果分析与人机科研协同[J].世界经济与政治,2024,(07):3-30+153.
[2] Rossi L, Harrison K, Shklovski I. The problems of LLM-generated data in social science research[J]. Sociologica: International Journal for Sociological Debate, 2024, 18(2): 145-168.
[3] 顾洁, 刘炜.生成式人工智能驱动的哲学社会科学研究范式转型[J].社会科学,2025,(06):165-176.
[4] 杰夫·吉尔,庞珣,郑晓龙,等.国际关系研究的人工智能“方法论”[J].世界经济与政治,2025,(01):4-26+154.
[5] Park J S, O'Brien J, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th annual acm symposium on user interface software and technology. 2023: 1-22.

该项目以 Swarthmore College 为原型构建了一个虚拟小镇，在其中部署了 25 个基于大语言模型的智能体作为“小镇居民”。这些智能体具备记忆、反思和计划等核心能力，能够在虚拟环境中自主进行日常活动：早晨醒来制作早餐、前往咖啡馆工作、参加社交聚会、传播八卦消息，甚至自发组织情人节派对等复杂社会活动。它们能够基于互动经历建立和维护人际关系，展现出高度的行为连贯性和社会适应性。在为期两天的连续运行中，这些智能体产生了丰富的行为轨迹和互动记录。利用这种方法，研究者获得了表征模拟社会动态的合成数据：包括每个智能体的详细行为日志、对话记录、情绪状态变化、社交网络演化，以及群体层面的活动模式和规范形成过程。这些数据具有完整的时间序列性、可追溯的因果链条以及多层次的分析维度。研究者可以在单一实验中同时观察微观行为与宏观涌现的全过程。这种数据的价值在于：它不仅记录了“发生了什么”，更重要的是记录了“为什么发生”和“如何发生”的完整机制链条，为理解复杂社会现象提供了可控制、可重复的实证基础。

Smallville 展示了模拟社会研究的巨大潜力，随后涌现出大量多智能体社会模拟实验，研究议题涵盖：社会规范如何在群体中自发形成并得到维护[1]、舆论话题如何在模拟社区中传播和极化[2]、利益冲突如何通过协商与制度安排得到化解[3]，以及权力结构与资源分配如何影响群体决策[4]等核心社会学问题。这些研究共同表明，这种类型的合成数据允许研究者按传统的社会科学方法探索模拟社会中的问题，捕捉了原本难以在现实环境中观测和控制的复杂社会机制与动态过程。

[1] Dai G, Zhang W, Li J, et al. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory[EB/OL]. (2023-11-16)[2025-09-14]. https://arxiv.org/abs/2406.14373.
[2] Wang C, Liu Z, Yang D, et al. Decoding echo chambers: Llm-powered simulations revealing polarization in social networks[EB/OL]. (2023-09-28)[2025-09-14]. https://arxiv.org/abs/2409.19338.
[3] Hua W, Fan L, Li L, et al. War and peace (waragent): Large language model-based multi-agent simulation of world wars[EB/OL]. (2023-11-28)[2025-09-14]. https://arxiv.org/abs/2311.17227.
[4] Xiao B, Yin Z, Shan Z. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research[EB/OL]. (2023-11-12)[2025-09-14]. https://arxiv.org/abs/2311.06957.

需要说明的是，上述社会模拟是在完全虚拟的空间中进行，这种“纯数字化”的特性赋予了教育研究极大的自由度和想象空间。理论上来说，研究者可以构建任何假想的教育环境和政策安排，对一些在现实中难以观测的现象进行探索性研究，比如校园霸凌的形成机制与干预效果、学生辍学危机的演化路径，或者不同文化背景下的师生互动模式。同时，对于那些已经成为历史、无法直接观察的重要社会现象，如文艺复兴时期的思想传播、20 世纪初的进步教育运动、不同国家教育制度改革的社会效应，甚至可以构想未来教育形态（如完全虚拟化学习社区）的可能发展轨迹，从而为教育理论的验证与政策设计提供前瞻性洞察。

然而，大语言模型社会模拟的潜力远不止于虚拟世界的探索。随着技术架构的不断完善，特别是检索增强生成（Retrieval Augmented Generation，RAG）和模型控制协议（Model Control Protocol，MCP）等先进技术框架的成熟，社会模拟正在从纯虚拟空间向现实领域拓展[1][2]。例如，通过 RAG 技术，智能体可以实时访问和整合现实世界的最新信息，使其行为和决策更加贴近当下的社会语境；而 MCP 等控制协议则为智能体提供了与现实系统交互的标准化接口，使它们能够在受控条件下参与真实的社会过程。基于这些技术突破，研究者可以让智能体直接参与到真实的人机互动场景中[3]。一些前沿研究开始采用数字孪生（Digital twin）的概念，认为大语言模型是能够通过实时数据流持续学习和自我更新的动态镜像系统，可以利用合成数据跟踪并模拟现实个体或群体的整个生命周期变化[4]。

[1] Ray P P. A survey on model context protocol: Architecture, state-of-the-art, challenges and future directions[J]. Authorea Preprints, 2025.
[2] Arslan M, Ghanem H, Munawar S, et al. A Survey on RAG with LLMs[J]. Procedia computer science, 2024, 246: 3781-3790.
[3] 黄荣怀, 刘德建, 提利利阿罕默德, 等, 2023. 人机协同教学:基于虚拟化身、数字孪生和教育机器人场景的路径设计[J]. 开放教育研究, 29(6): 4-14.
[4] 刘泽民,陈向东.教育研究的硅基样本——基于大模型技术路线的分析[J].远程教育杂志,2025,43(01):19-32+45.

例如，Helm 等人在 2025 年开发的“美国国会数字孪生”项目展现这种技术整合的应用前景[1]。该研究通过收集每位国会议员的推特数据，为每位议员建立个性化的虚拟镜像，这些虚拟议员不仅能够生成与真实议员高度相似的社交媒体内容，更重要的是能够预测其在具体法案上的投票行为，准确率达到 87%。在这篇研究中，合成数据的应用从模拟观察转向了对现实世界的预测。研究提出的翻转评分（Flip Score）机制通过分析虚拟议员在特定议题上的立场分布，量化现实议员跨党投票的可能性，为政策利益相关者提供精准的游说目标识别。这种方法使合成数据成为能够指导现实决策的情报系统。同时，通过日更新的数据流实现了“物理—虚拟”的双向反馈循环：现实议员的最新表态不断校准虚拟模型，而虚拟模型的预测分析又直接服务于现实的政治过程。

可以认为，动态同步的合成数据生成模式预示着社会模拟研究正在从纯学术探索转向具有实际社会影响力的应用工具，为教育治理研究领域的政策制定、危机预警和教学干预开辟了新的研究路径。

9.2.3 模拟世界

模拟世界是合成数据应用的新维度，是指利用大语言模型生成能够反映现实世界复杂性和多样性的虚拟环境、场景和数据集合的过程。这种模拟涵盖了包括物理环境的多模态呈现，也包括了对世界运行机制、因果关系和系统性规律的数字化重构。

这一应用形式的理论基础源于研究者对大语言模型涌现能力的深入思考。当我们试图理解为什么大语言模型能展现出令人惊讶的推理、规划和知识整合能力时，“世界模型”提供了一个重要的解释框架：大语言模型可能在训练过程中隐式地构建了对世界运行机制的内部表示。基于这一认识，研究者进一步希望构建能够整合视觉、听觉、触觉等多种感官信息的统一世界模型，让 AI 能够像人类

[1] Helm H, Chen T, McGuinness H, et al. Toward a digital twin of US Congress[EB/OL]. (2025-04-04)[2025-09-14].https://arxiv.org/abs/2505.00006.

一样通过多种渠道理解和重构世界。这种多模态整合能力对于教育研究具有深远意义，因为教育过程本身就是一个涉及多感官体验、复杂社会互动和认知发展的综合性活动，需要更加真实和全面的研究环境来支撑理论探索和实践创新。

实际上，正如前文所述，利用合成数据模拟世界的实践由来已久，这与计算机视觉领域的早期探索有着紧密的历史联系。这些应用通常是为了突破真实数据获取的成本和条件限制，对极端自然现象、复杂环境系统进行有效建模。早期的典型应用如无人机检测系统开发，研究者利用虚幻引擎和 Microsoft AirSim 构建三维虚拟环境，生成大量标记精确的合成训练数据，有效解决了真实无人机数据获取成本高昂、条件受限的技术难题[1]。随着大语言模型多模态处理能力的显著提升，合成数据的世界表征能力得到了前所未有的拓展。一些先进模型如 DALL·E 3、Midjourney 等能够直接根据文字描述生成高质量的图片或视频，这在社会科学研究中开辟了崭新的应用方向，催生了两个重要的应用维度。

在构建场景的应用中，研究者能够借助大语言模型生成高度可控的视觉材料，精确测试特定变量对人类认知和态度的影响。这类研究的核心诉求是生成逼真且包含特定社会学变量的合成图像，以最大程度地贴近真实世界的复杂性。

例如，有研究者通过在提示中加入具体情境描述，如“中西部城市街头举行的小型/大型‘黑人的命也是命’抗议集会”或“气候变化抗议集会”，成功引导 DALL·E 3 生成了展现不同规模街头抗议群体的四张图像[2]。有趣的是，由于大语言模型具备表现现实中无法观察到的反事实场景的能力，可有效应对传统合成媒体的局限性。例如，研究者生成了一张包含燃烧汽车的妇女游行合成图像，

[1] Nikolaiev M, Novotarskyi M. Comparative Review of Drone Simulators[J]. Information, Computing and Intelligent systems, 2024 (4): 79-99.
[2] Davidson T. Start generating: Harnessing generative artificial intelligence for sociological research[J]. Socius, 2024, 10: 23780231241259651.

这种财产破坏和暴力场景通常不会在实际的妇女游行中出现，却为研究提供了重要的对照实验材料。通过在提示中加入“美联社照片”等引导性词语，研究者还能够控制图像的风格特征，生成符合新闻摄影标准的实验材料。

在解构环境的应用中，大语言模型展现出分析静态且结构性较强图像的强大能力。这类研究将大语言模型用于辅助分析大规模标准化的图像，其作用已经超越了解读单一照片特定叙事的传统功能，发展为系统性“测量”那些形成长期社会现象的城市物理特征的重要工具。例如，研究者通过比较 GPT-4o 与人类对街道实景图像的评估，特别是在城市步行适宜性背景下，来回答“诸如 GPT-4o 之类的大语言模型，能否准确反映人类对城市环境的感知？”这一问题[1]。研究邀请人类参与者与 GPT-4o 共同根据步行适宜性的关键维度（包括整体步行适宜性、可行性、可达性、安全性、舒适度和活力感）对街景图像进行评估，并发现：GPT-4o 与参与者在整体步行适宜性、可行性、可达性和安全性方面的评估结果一致；但在舒适度和活力感的评估上则存在显著差异。人类参与者展现出更广泛的主题多样性，涉及更丰富的议题；而 GPT-4o 的回答则更为聚焦和连贯，尤其在舒适度和安全性方面。

可以认为，一旦获得这些图像编码和标注的合成数据，社会科学研究就能够在前所未有的规模上进行城市社会学分析，系统性地揭示城市物理环境与社会分层、居住隔离、经济发展等宏观社会现象之间的复杂关系，为城市规划、社会政策制定和社会公平研究提供了基于大数据的实证基础。

这些应用形式和发展方向对教育研究具有重要价值。通过模拟世界的合成数据，教育研究者能够创建控制精确的虚拟学习环境，测试不同教学策略的效果；生成包含特定教育变量的场景，研究学

[1] Wedyan M, Yeh Y C, Saeidi-Rizi F, et al. Urban walkability through different lenses: A comparative study of GPT-4o and human perceptions[J]. PLoS One, 2025, 20(4): e0322078.

习者在不同情境下的认知和情感反应；同时利用大语言模型的环境解构能力，系统性分析真实教育环境中影响学习效果的物理和社会因素。

需要说明的是，目前合成数据在模拟世界维度的应用主要聚焦于静态内容的创建，如图像、固定场景或数据集。但近期 Google DeepMind 的 Genie 3 模型使得创建可交互、动态世界的合成数据成为可能[1]。Genie 3 的革命性在于它是首个能够实现实时交互的世界模型。这一突破将 AI 从被动的内容生成工具转化为主动的世界构建平台，能够从单一文本提示创建完整的交互式 3D 环境，支持 720p 高分辨率视觉效果，并具备包括水效、光照、重力和角色动画在内的复杂物理建模能力。

Genie 3 最核心的技术创新体现在其因果推理和动态响应能力，能够识别和建模环境中的行动—反应机制，理解物理定律、因果关系和系统性规律。这种能力使得生成的虚拟环境能够对用户的交互做出符合物理逻辑的实时响应，为研究者提供了前所未有的实验控制精度。

这些技术发展的汇聚正在推动教育研究从“数据稀缺驱动的现象描述”向“世界理解驱动的机制探索”的范式转变。研究者不再仅仅局限于分析已有的教育数据，而是能够基于对教育过程运行机制的深入理解，主动构建符合研究假设的虚拟教育环境，在高度可控的条件下验证教育理论，探索教育规律。这种范式转变不仅为教育科学的发展提供了更加强大的方法论工具，更为重要的是，它开启了教育研究从被动观察向主动实验、从现象描述向机制揭示、从静态分析向动态建模的全方位转型，为构建更加科学、精确和有效的教育理论体系奠定了坚实的技术基础。

[1] Google DeepMind. Genie 3: A new frontier for world models[EB/OL].(2025-08-04) [2025-09-14].https://deepmind.google/discover/blog/genie-3.

9.3 合成数据的应用争议

大语言模型生成的合成数据具有广泛的应用形式，这为解决人类数据（包括自然数据）的诸多限制提供了一条低成本、规模化的技术路径。然而，伴随着合成数据应用范围的快速扩张，学术共同体内部关于其科学有效性和研究伦理的争议也日趋激烈。这些争议主要源于技术能力与研究需求之间的现实差距，以及新兴技术对传统研究范式带来的根本性冲击，具体体现在表征偏差、方法论、研究伦理以及规范性几个方面。

9.3.1 表征偏差的争议

表征偏差作为合成数据应用中最为复杂和争议性的问题，其本质反映了大语言模型在重现社会现实时面临的多重技术挑战。这种偏差被定义为特定社会群体表征中的不准确性，但在合成数据的应用语境下，其内涵远比传统机器学习中的偏差概念更为丰富和微妙，包括分布偏差、身份偏差以及非理性偏差等。

首先，分布偏差是指合成数据的分布与社会现实不符。当利用大语言模型反映现实世界的统计分布时，研究者希望模型能够准确反映社会现实，包括可能存在的不平等和偏见；而大语言模型为避免刻板印象和歧视，通常会进行“去偏”的处理，导致生成的合成数据偏离真实的社会现象，降低数据的代表性和研究价值[1]。例如，在一项关于商业领导力与性别职业偏见的研究中，研究者需要利用合成数据来模拟公司高管的行为表现[2]。然而，主流的商用大语言模型（如 GPT 系列）在设计时考虑到了社会责任，各大公司为这些模型设置了安全护栏和内容过滤机制，旨在减少刻板印象的输出。在这种机制驱动下，当被要求生成企业高管群体的性别分布时，模型很可能输出接近 1:1 的男女比例，以体现性别平等的价值观。但

[1] Anthis J R, Liu R, Richardson S M, et al. Position: LLM Social Simulations Are a Promising Research Method[C]//Forty-second International Conference on Machine Learning Position Paper Track.
[2] Lum K, Anthis J R, Robinson K, et al. Bias in language models: Beyond trick tests and toward ruted evaluation[EB/OL].(2024-02-20) [2025-09-14]. https://arxiv.org/abs/2402.12649.

这种“去偏”处理可能会与现实情况形成了鲜明对比。截至2024年，财富500强企业高管中90%都是男性。这就造成了一个两难境地：如果严格按照现实中90%男性CEO的比例生成数据，可能被指责为强化和传播性别不平等观念；而如果为了体现公平价值而输出男女各占50%的理想化比例，则与统计事实严重相悖，可能影响学术研究的科学性和有效性。

这一问题的根源在于，主流的商用模型主要面向通用性应用场景，其训练目标通常是把社会责任和用户体验置于首要位置，而不会专门核查事实的准确性。这些模型被设计为在日常对话和内容生成中避免可能引起争议或伤害的输出，但这种设计理念与科学研究对客观性和准确性的要求之间存在天然的张力。因此，当研究者试图利用这些通用模型生成反映真实社会状况的研究数据时，就会遭遇这种准确性与“政治正确”之间的冲突。

其次，身份偏差体现在大语言模型对目标群体的基本身份和角色理解错误。Wang等人对大语言模型生成的合成数据进行深入分析后发现，模型在处理身份群体表征时存在两个关键缺陷：身份认知错误和群体扁平化[1]。一方面，在身份认知层面，大语言模型往往无法准确捕捉目标人群的多维身份特征，容易将特定群体误认为是外群体的模仿者。例如，当要求模型生成教师群体的观点时，模型可能生成的是家长或学生对教师行为的讨论和模仿，而不是教师群体的真实表达。这种错误的根源在于，互联网上关于教师的讨论往往来自学生、家长或教育评论者，而教师本人的直接发声相对较少。训练数据中这种不平衡的话语分布，导致模型学到的更多是“关于教师”的外部观点，而非教师群体的内在视角。

另一方面，更为严重的是群体扁平化现象。Wang等人的量化分析表明，在所有测试模型和几乎所有多样性指标上，大语言模型生

[1] Wang A, Morgenstern J, Dickerson J P. Large language models that replace human participants can harmfully misportray and flatten identity groups[J]. Nature Machine Intelligence, 2025: 1-12.

成的回答都比真实人类回答更为单一和同质化[1]。具体而言，GPT-4和GPT-3.5等先进模型在生成特定群体回答时，通常只能覆盖每种情境中100个可能答案的有限选项。换言之，当面对需要100种不同回答的情境时，这些模型实际能够产生的独特回答类型远少于真实人类群体的表现，显示出明显的多样性贫乏。这种现象表明，大语言模型在努力增加跨群体代表性的同时，却系统性地削弱了群体内部的真实差异。这种“表面多元化”现象比完全忽视某些群体更加有害，因为它创造了虚假的包容性假象，掩盖了群体内部的复杂性和异质性。

第三，非理性偏差。这个问题的复杂性在于，真实的社会模拟需要准确反映人类决策中的认知偏差、情绪影响和非理性行为模式，因为这些“不完美”的特征正是理解社会现象的关键要素。换句话说，为了获得科学上有效的研究数据，合成数据必须包含基于错误信息、刻板印象或情绪冲动的偏见性观点和决策，即便这些观点本身可能是不准确或不公正的。然而，当前的大语言模型在这一点上似乎存在缺陷。Liu等人的研究揭示了这个问题的关键所在：大语言模型倾向于假设人类比实际情况更加理性[2]。这项研究通过将多个前沿模型的行为预测与大量真实人类决策数据进行系统性比较，发现了一个令人意外的模式。

具体而言，当大语言模型被要求模拟或预测人类在各种场景下的选择时，它们生成的决策数据往往遵循经典的理性选择理论框架，如期望价值理论和效用最大化原则。这些模拟结果在逻辑上无懈可击，在数学上完全合理，但与真实人类的决策行为存在显著偏差。真实的人类决策充满了认知偏差、情绪干扰、启发式思维和不一致性，这些“非理性”特征在大语言模型的输出中却明显不足。这种

[1] Wang A, Morgenstern J, Dickerson J P. Large language models that replace human participants can harmfully misportray and flatten identity groups[J]. Nature Machine Intelligence, 2025: 1-12.
[2] Liu R, Geng J, Peterson J C, et al. Large language models assume people are more rational than we really are[EB/OL].(2024-06-24) [2025-09-14]. https://arxiv.org/abs/2406.17055.

差异在使用思维链推理提示时表现得尤为突出，模型的推理过程越详细，其输出就越趋向于理想化的理性决策。

研究中一个特别有趣的发现进一步复杂化了这个问题。当研究者将大语言模型的决策推论与心理学数据集中关于“人们如何解释他人行为”的数据进行比较时，发现两者之间存在高度相关性。这表明大语言模型所体现的决策模式与人类在观察和解释他人行为时的认知模式相符。从这个角度来看，大语言模型似乎成功捕捉了人类的一种认知倾向。但这种成功恰恰暴露了问题的深层次所在：

正如前文所述，Wang 等人关于身份认知偏差的发现为这一现象提供了重要的解释线索，互联网上充斥着人们对他人决策的分析、解释和预测，这些外部观察往往假设被观察者是理性的，因为观察者倾向于为他人的行为寻找合理化的解释。换言之，大语言模型学习到的不是各群体的内在视角和真实行为模式，而是外部观察者对这些群体“应该如何思考和行动”的期待和推测。这种差异对于需要理解真实人类行为机制的研究而言有负面影响。例如，在教育研究中，学生的学习决策往往受到即时满足偏好、社会压力、情绪状态等非理性因素的显著影响，但如果合成数据中的“学生”都表现得过于理性和目标导向，那么基于这些数据的教育干预设计可能会严重脱离现实。

9.3.2 认识论的争议

合成过程是否能够准确捕捉并反映人类的认知机制也是目前重要的争议点。许多研究者指出，现有的合成数据模型往往是基于某些输入和输出之间的统计关系进行构建的，缺乏对人类认知过程的深入理解[1]。这种方法虽然在生成与人类行为相似的输出方面表现良好，但其运作机制和内部处理过程与人类思维存在显著差异[2]。

[1] Rmus M, Jagadish A K, Mathony M, et al. Generating computational cognitive models using large language models[EB/OL].(2024-06-24) [2025-09-14]. arXiv preprint arXiv:2502.00879, 2025.
[2] Shah R S, Varma S. The potential--and the pitfalls--of using pre-trained language models as cognitive science theories[EB/OL].(2024-02-02) [2025-09-14]. https://arxiv.org/abs/2502.00879.

因此，目前的质疑声音主要表现为：大语言模型通过统计学习获得的“知识”与人类通过认知体验形成的知识可否类比？模拟认知过程是否等效于理解认知机制？技术上的成功是否意味着科学上的有效性？

理解这一争议的关键在于认识大语言模型发展所遵循的“拟合范式”及其认识论局限。从技术发展的轨迹来看，大语言模型的核心策略始终是通过统计学习来拟合人类的语言和行为模式。这种拟合策略虽然在技术表现上取得了显著成功，但也引发了深刻的认识论质疑：拟合人类行为的外在表征是否等同于理解人类认知的内在机制？

早期大语言模型（如 GPT-3.5）的表现为这一问题提供了重要线索。这些模型虽然能够进行复杂对话，但在诸如判断“3.11 大于 3.9”或统计“strawberry”中字母‘r’数量等看似简单的任务上却出现系统性错误。这些错误最初被视为纯技术问题，如分词机制和训练目标的局限。然而，从认识论角度审视，这些错误实际上暴露了拟合范式的根本缺陷：模型通过字符串序列的符号操作来处理数值比较，通过子词单元的向量表示来处理文本，这些处理方式与人类的直觉认知存在结构性差异。

对于教育研究而言，这意味着早期模型生成的合成数据可能包含系统性的认知偏差。例如，当模拟学生的数学学习过程时，如果模型无法正确理解基本的数值概念，那么其生成的学习轨迹数据就可能误导研究者对学生认知发展规律的理解。

随着技术发展，OpenAI 的 O1 等系列的推理模型通过引入思维链推理机制，在很大程度上解决了前述基础错误。例如，当面对“strawberry 中有几个 r”的问题时，O1 能够生成类似于“让我逐字拼写：s-t-r-a-w-b-e-r-r-y，然后计数 r 的出现次数”的推理过程，从而得出正确答案。一些学者认为，O1 的这种改进实际上代表了“拟

合范式”的进一步精细化[1]。模型学会了模拟人类在解决此类问题时的“慢思考”过程——即 Kahneman 所描述的 System 2 思维模式，通过显式的步骤分解和逻辑推理来达到正确结果。这种改进的确带来了大语言模型显著的性能提升。O1 模型在许多需要逻辑推理的任务上表现出色，不仅解决了之前的基础错误，还在数学、科学、编程等领域展现出了接近甚至超越人类专家的能力。许多观察者认为，这标志着 AI 系统向真正的智能迈出了重要一步。

从合成数据应用的角度来看，这一进步似乎解决了认识论争议。O1 生成的数据不仅在结果上与人类相似，在推理过程上也展现出与人类认知的一致性。这为研究者使用合成数据提供了更强的信心——如果模型能够模拟人类的推理过程，那么其生成的行为数据似乎就具备了科学研究的可信度。

然而，更深层的认识论质疑随之而来：模拟推理过程是否等同于真正的理解？ O1 的推理链本质上仍然是基于训练数据中人类问题解决过程的统计学习，属于更精细化的模式匹配。模型学会了在特定问题情境中生成特定的推理步骤，但并未真正“理解”为什么需要采用这些步骤。这种拟合的成功可能创造出“理解的幻觉”，使研究者误以为合成数据反映了真实的认知过程，而实际上它只是更精确的表面模仿。

就在人们为以 O1 为代表的大语言模型训练范式的进步感到兴奋时，另一个更加令人困惑的现象出现了。近期发表在《Nature》杂志上的 Centaur 模型的研究引起了巨大的争议[2]，甚至在文章发表的当日，《Science》杂志迅速发表了一篇评论文章对该模型的有效性提出了质疑[3]。质疑的原因在于，这个专门设计来模拟人类认知的

[1] Araya, R. Do Chains-of-Thoughts of Large Language Models Suffer from Hallucinations, Cognitive Biases, or Phobias in Bayesian Reasoning?[EB/OL].(2025-05-19) [2025-09-14]. https://arxiv.org/abs/2503.15268.
[2] Binz M, Akata E, Bethge M, et al. A foundation model to predict and capture human cognition[J]. Nature, 2025: 1-8.
[3] Cathleen O’Grady.Researchers claim their AI model simulates the human mind. Others are skeptical [EB/OL]. (2025-08-01)[2025-09-14]. https://www.science.org/content/article/researchers-claim-their-ai-model-simulates-human-mind-others-are-skeptical.

AI 系统，在某些方面表现出了远超人类的能力。

例如，Centaur 与人类记忆能力的差异是一个重要的争议点。心理学研究早已确立，人类的短期记忆容量是有限的，大约只能同时记住 7±2 个信息单元（比如数字、字母或词汇）。这个限制被称为“神奇数字 7”，是人类认知架构的基本特征之一，反映了我们大脑工作记忆系统的生理约束。然而，Centaur 模型在短期记忆任务中却能够记住 256 位数字——这个数字是人类能力的 30 多倍。从表面上看，这似乎是一个巨大的“改进”，证明了 AI 系统的优越性。但是，仔细思考就会发现问题所在：如果 Centaur 真的是在模拟人类认知，它为什么会表现出如此不同于人类的特征？

这一悖论对合成数据的应用构成了挑战。在教育研究中，如果用于模拟学生认知的模型具有远超学生实际能力的记忆容量或处理速度，那么基于这种模型生成的学习行为数据还能被视为对真实学生行为的有效表征吗？这种“超人化”的合成数据可能会误导研究者对学生认知能力和学习困难的判断，导致不切实际的教育期望和干预设计。

可以看出，这一认识论的困境并没有随着模型技术的进步而得到解决，目前研究者普遍认为，由于仍然无法确定大语言模型的内在机制能否在特定条件下等同于人类，在应用合成数据时，会存在科学性的问题。但也有学者提出了不同的看法，支持者认为只要外在表现一致，合成数据就具有研究价值；反对者则坚持认为缺乏真正理解的模拟本质上是无效的。

这些争议的核心实质在于，技术上的成功并不必然意味着科学上的有效性。越来越精细的拟合可能会掩盖更深层的理解缺失，创造出一种方法论进步的假象。对于教育研究这样直接关系到人类发展和社会福祉的领域，这种认识论上的不确定性具有特别重要的伦理和实践意义。因此，在应用合成数据进行教育研究时，必须保持

审慎的认识论反思，避免将技术能力等同于科学理解。

9.3.3 应用伦理的争议

与传统的人类参与者研究需要经过严格的伦理审查、知情同意、数据保护等程序不同，合成数据的生成几乎没有任何制度性门槛。任何拥有 API 访问权限的研究者都可以在几分钟内生成涉及敏感教育群体的大量数据，无需通过伦理委员会审批，无需获得被模拟群体的同意，也无需承担传统研究中的数据保护责任。更为严重的是，当前几乎没有针对合成数据应用的明确伦理规范。传统的研究伦理框架主要围绕保护真实参与者而设计，但面对“虚拟参与者”时，这些框架显得力不从心。

这极有可能导致多重责任扩散：模型开发者声称他们只提供技术工具，不负责具体应用；研究者认为没有涉及真实人类参与者，无需承担传统的伦理责任；期刊编辑缺乏评估合成数据研究伦理性的明确标准；政策制定者在使用基于合成数据的研究结果时缺乏伦理考量。

即使在那些声称关注教育伦理的产品中，这种规范缺失也十分明显。例如，Google 的 Gemini for Education 虽然承诺所有数据不作为训练和广告推送使用，但对于如何确保生成的合成数据不会错误表征或伤害特定教育群体；但如何保证研究过程的透明性，以及如何处理潜在的偏见和歧视问题，却缺乏明确的指导和约束机制。

正是在这种制度性保障缺失的背景下，合成数据在教育研究中的应用伦理争议呈现出一种矛盾：技术的便利性与伦理规范体系的滞后。这种不匹配创造了一个危险的伦理真空，其影响正在从潜在风险转化为现实伤害，包括下面几个重要的维度。

首先，从数据泄露到身份剥削。传统研究中的隐私保护主要关注个人信息的直接泄露，但合成数据技术将这一挑战推向了更加复杂和隐蔽的层面。大语言模型的“记忆效应”使得看似安全的合成

数据实际上可能成为隐私泄露的桥梁[1]。研究已经证实，通过分析合成数据的模式和特征，攻击者可能推断出训练数据中特定个体的敏感信息，甚至实现身份识别。

但在教育研究的语境下，隐私问题的复杂性远不止于此。更为严重的是一种新形式的“身份剥削”——研究者可以无限制地使用特定教育群体的身份标签来生成合成数据，而这些群体对此毫不知情，也无法对数据的准确性或潜在偏见进行监督。当研究者使用大语言模型生成模拟特殊需求学生、少数民族学生或低收入家庭学生的数据时，这些群体实际上成为了不知情的研究对象。他们的群体身份被用于学术生产，但他们既无法从研究中获益，也无法对可能的错误表征进行纠正。

这种身份剥削的隐蔽性在于，它披着“无害”的外衣——毕竟没有真实的个人数据被直接使用。然而，如果生成的合成数据包含偏见或刻板印象，这些错误表征可能在学术文献中被固化，进而影响政策制定和资源分配，对相关群体造成实质性的间接伤害。更危险的是，由于这种伤害的间接性和延迟性，受影响的群体往往难以察觉和追责。

其次，知情同意的消失。苏黎世大学在 Reddit 平台进行的 AI 说服力研究成为了合成数据伦理违规的典型案例，甚至被一些学者称为“我见过的最严重的互联网研究伦理违规事件”。这个案例深刻地揭示了合成数据技术如何使传统的知情同意原则完全失效。研究者让 AI 机器人在 Reddit 的 r/changemyview 论坛中冒充真实用户，试图通过生成的评论来改变其他用户的观点[2]。整个实验过程中，参与者完全不知道自己正在与 AI 互动，更不知道自己成为了学术研究的对象。这种做法不仅违背了研究伦理的基本原则，也违反了

[1] Yan B, Li K, Xu M, et al. On protecting the data privacy of large language models (llms): A survey[EB/OL].(2025-03-08) [2025-09-14].https://arxiv.org/abs/2403.05156.
[2] Breum S M, Egdal D V, Mortensen V G, et al. The persuasive power of large language models[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2024, 18: 152-163.

Reddit社区本身的规则——该论坛明确禁止使用未披露的AI生成内容[1]。

这个案例的危险性不仅在于具体的违规行为，更在于它揭示了合成数据技术创造的一种全新的伦理困境。传统的知情同意要求研究者告知参与者研究的性质、目的、风险和收益，并获得其自愿同意。但当AI能够完美模拟人类的表达方式时，这种透明性要求变得极其复杂。研究者可以声称他们研究的是AI的能力而非人类行为，从而规避传统的伦理审查。同时，被AI影响的用户可能永远不知道自己的观点改变源于人工干预而非自然的社区交互。

更令人担忧的是，这种隐蔽的研究方法可能对社区造成长期的信任损害。当Reddit用户得知他们一直在与伪装的AI互动时，产生了“强烈的负面情绪”和对学术研究的怀疑。这种信任危机一旦扩散，可能使整个在线研究环境变得更加困难和敌对。

在教育研究的语境下，这种问题可能更加严重。想象一下，如果研究者在教育论坛或学生社区中使用AI冒充教师、家长或学生来收集数据或影响讨论，不仅会损害这些教育社区的信任基础，还可能误导真实的教育决策。更危险的是，由于教育关系的特殊性和权力不对等，这种欺骗性研究可能对脆弱的学生群体造成特别严重的心理伤害。

第三，缺少约束手段。AI检测技术的不可靠性为合成数据的伦理争议增加了另一个维度的复杂性。当新闻报道朱自清的《荷塘月色》被检测系统判定为62.88% AI生成率，王勃的《滕王阁序》被认定为接近100% AI生成时，这种检测技术的不稳定性已经引起了公众的广泛关注[2]。该问题源于不同AI检测软件在算法设计、训练数据和评估标准上的根本差异。但由于缺乏统一的技术标准和评估

[1] Bruckman A. An Unethical Study of r_changemyview by University of Zurich Highlights the Need for International...[EB/OL].(2023-05-08)[2025-09-14]. https://asbruckman.medium.com/an-unethical-study-of-r-changemyview-by-university-of-zurich-highlights-the-need-for-international-787bb603ad6c.
[2] 张盖伦.用“AI率”对论文“一票否决”科学吗[N].科技日报,2025-06-02(001).

体系，同一文本在不同检测工具下可能得到截然不同的判断结果。这种不一致性在学术环境中创造了一个危险的“灰色地带”——研究者可能利用检测工具之间的差异来规避监督，而真实的原创内容则可能被误判为合成数据。

MIT 博士生利用 AI 伪造调研数据的案例则将这种监管失效的危险性推向了高潮[1]。这位博士生生成的虚假研究不仅通过了初步的学术审查，还获得了诺贝尔奖获得者的支持，被《大西洋月刊》和《Nature》等权威媒体报道，最终被《经济学季刊》拟录用。只有在 MIT 的专门调查下，数据造假才被发现[2]。

这个案例的震撼在于，它表明即使是学术界最权威的质量控制机制，包括同行评议、专家背书和媒体监督等，在面对精心设计的合成数据造假时也可能完全失效。传统的学术造假往往在统计分布上存在不自然的特征，容易被有经验的研究者识别。但 AI 生成的合成数据可以在统计上完美地模拟真实数据的复杂性和随机性，使得造假变得几乎无法察觉。

更为严重的是，合成数据在某些情况下可能比真实数据更加完美：没有缺失值，没有测量误差，统计关系更加清晰。这种过于完美的特征可能使得基于合成数据的研究在发表竞争中具有优势，从而形成劣币驱良币的恶性循环。真实的、包含各种不完美因素的人类研究可能被边缘化，而看似科学但实际虚假的合成数据研究可能占据主导地位。

这些案例共同指向了一个更深层的问题：现有的学术诚信体系在面对合成数据技术时面临结构性危机。传统的学术监管建立在几个基本假设之上：研究者会诚实报告其数据收集过程，同行评议者能够识别明显的造假，技术工具能够可靠地检测违规行为。但合成

[1] Toner-Rodgers A. Artificial intelligence, scientific discovery, and product innovation[EB/OL].(2024-12-21)[2025-09-14]. https://arxiv.org/abs/2412.17866.
[2] MIT Department of Economics. Assuring an accurate research record[EB/OL].(2025-05-16)[2025-09-14].https://economics.mit.edu/news/assuring-accurate-research-record.

数据技术使得这些假设都变得不再成立。

在所有学科中，教育研究在面对合成数据伦理争议时可能是最为脆弱的。这种脆弱性源于教育研究的几个独特特征：首先，教育研究经常涉及儿童和青少年等群体，这些群体在传统研究中受到特殊保护，但在合成数据研究中这种保护机制完全失效；其次，教育研究的结果往往直接影响政策制定和教育实践，错误的研究结论可能对成千上万的学生造成实际伤害；最后，教育研究特别依赖于对人类学习过程和社会互动的深入理解，而正如前文所述，这些复杂的认知和社会过程正是当前合成数据技术最难准确模拟的领域。

当苏黎世大学的研究者在在线社区中使用 AI 进行隐蔽实验时，他们实际上是在一个教育环境中进行的——r/changemyview 论坛本质上是一个学习和辩论的空间，用户通过讨论来改变和完善自己的观点。这种对教育过程的隐蔽操控不仅违背了研究伦理，也可能损害了这个学习社区的教育功能。

类似地，如果教育研究者开始在学校论坛、家长群体或教师社区中使用 AI 来冒充不同的教育参与者，不仅会破坏这些教育社区的信任基础，还可能误导真实的教育决策。更危险的是，由于教育关系中固有的权力不对等和情感依赖，这种欺骗性研究可能对学生和家长造成特别严重的心理创伤。

9.3.4 规范性的争议

合成数据在教育研究中的应用面临着一系列规范性争议，这些争议集中体现在评估标准与生成机制两个核心问题上。与传统数据收集方法不同，合成数据的质量评估缺乏统一的标准框架，而大语言模型的“黑箱”特性进一步加剧了这一挑战。

一方面，评估标准的模糊性与内在矛盾。当前合成数据评估主要依赖效用性（utility）和保真度（fidelity）两个维度，但这一框架在实际应用中暴露出一定的问题。研究者认为，效用性强调数据对

特定研究任务的实用价值，而保真度则关注合成数据与真实数据的相似程度。然而，这两个标准之间存在潜在的张力：高保真度的数据可能在特定研究任务中缺乏必要的变异性，而高效用性的数据又可能偏离真实分布的核心特征[1]。

这种张力在实际的教育研究中的表现可以用下面的例子来说明：研究者试图生成关于学生数学焦虑的合成数据。如果追求保真度，生成的数据可能完美复制了真实学生群体中数学焦虑的统计分布——比如30%的学生报告高度焦虑，50%报告中度焦虑，20%报告低度焦虑。但是，当研究者想要研究数学焦虑与学习成绩之间的关系时，可能会发现这些“高保真度”的数据过于平滑，缺乏真实情境中那些意外的、非线性的关联模式。相反，如果过分强调效用性，为了确保研究假设能够得到验证，生成的数据可能会强化某些预期的关联关系，但这些关系在真实世界中可能并不那么明显或普遍存在。更为复杂的是交叉保真度问题的存在。Johnson 和 Hajisharif 在对 GAN 生成的合成人口普查数据的研究中发现，尽管这些数据在单一维度上与真实数据高度相似，但在多个变量的交叉分析中却表现出偏差[2]。

在教育领域，已经有学者意识到对合成数据可靠性的验证，但验证框架并没有涉及到交叉保真度的问题。以 Farhood 等的研究为例，研究者利用大语言模型生成类似于真实学生数据集的隐私保护合成数据，并通过整合合成数据和实际数据来改进学生成绩的预测模型[3]。这一研究采用了相对标准化的验证策略，包括相关性分析、核密度估计、相关热力图以及主成分分析等方法来评估合成数据的可靠性。这些方法在检验单变量分布特征和简单的双变量关系方面

[1] Atil B, Chittams A, Fu L, et al. Non-Determinism of "Deterministic" LLM Settings[EB/OL].(2024-08-16)[2025-09-14]. https://ui.adsabs.harvard.edu/abs/2024arXiv240804667A/abstractarXiv e-prints, 2024: arXiv: 2408.04667.

[2] Johnson E, Hajisharif S. The intersectional hallucinations of synthetic data[J]. AI & SOCIETY, 2025, 40(3): 1575-1577.

[3] Farhood H, Joudah I, Beheshti A, et al. Advancing student outcome predictions through generative adversarial networks[J]. Computers and Education: Artificial Intelligence, 2024, 7: 100293.

表现出色，能够有效识别合成数据在边际分布和基础统计特征上的偏离。然而，这种验证范式的根本性缺陷在于其对多维度交互模式的检测能力不足。教育数据的复杂性往往体现在多个变量之间的高阶交互关系中，例如学习行为、时间模式、认知特征与学习结果之间可能存在复杂的非线性关联。传统的验证方法主要关注变量间的线性相关性和简单的分布匹配，却无法捕捉到这些深层次的交互偏差。

可以认为，这种评估困境在大语言模型生成的教育数据中尤为突出，因为教育现象往往涉及认知、情感、社会等多个维度的复杂交互。当我们使用 GPT 生成学生访谈数据时，模型可能能够产生在单个话题上非常“真实”的回答，但当这些回答需要体现学生个体的一致性人格特征时，问题就出现了。比如，一个被设定为“内向且数学焦虑”的虚拟学生，在某一轮回答中可能表现出对数学的恐惧和回避，但在另一个话题的回答中却可能展现出逻辑思维的自信，这种内在的矛盾性在真实学生身上是不太可能出现的。

另一方面，生成过程的不稳定性与不可控性。合成数据生成过程中的提示设计问题构成了最为复杂的规范性争议。这一争议的核心在于以下几个方面：

首先是特异性与普遍性的悖论。为了生成具有特定群体特征的合成数据，研究者需要在提示中详细描述目标群体的特点。然而，过于具体的描述可能导致模型输出过度拟合研究者的预期，从而失去数据的自然变异性。相反，过于模糊的提示又可能使生成的数据缺乏必要的群体特征。这一悖论在多项实证研究中得到了清晰的体现。以 Park 等人的生成代理研究为例，研究者为每个代理提供了详细的身份描述，包括职业、人际关系等信息，如“约翰·林是一名药房老板，喜欢帮助他人”。这些描述作为“种子记忆”，预设了社会动态的发展轨迹。当代理身份和关系被如此具体地预定义时，

模拟中观察到的互动模式很大程度上成为了这些初始配置的直接结果[1]。类似地，Hua 等人在国际冲突模拟中为国家代理构建了包含“历史背景”、“关键政策”等维度的详尽档案，这些档案以极其精确的方式编码了价值体系和社会关系，预先确定了代理的行为路径[2]。当代理被如此详细地预设时，观察到的联盟形成或冲突发生究竟是复杂交互的真实涌现，还是深度嵌入的初始配置的必然结果，就变得难以判断。

与此对照的是，过于模糊的描述性提示同样带来挑战。如 Hua 等人描述的那种“以实用主义和坚忍治理为特征的君主立宪制”这样的抽象概念描述，虽然旨在激活模型的固有知识结构而不施加严格的行为约束，但却面临着概念激活的不确定性问题。正如 Argyle 等人的研究揭示的，大语言模型对抽象概念的内在表征反映了从训练数据中吸收的多样化、甚至相互冲突的人类观点[3]。Li 等人进一步展示，模型对“实用主义治理”这样概念的激活理解可能显著偏离研究者的预期含义，因为模型的解释基于大量学习到的关联[4]。这种偏差在多代理社会模拟中可能通过迭代交互得到传播和放大，最终损害整个模拟框架的有效性。

其次是先验知识与自发反应的悖论。大语言模型的输出不可避免地受到其训练数据中既有知识的影响。当模型生成关于特定教育现象的数据时，这些数据究竟反映的是被模拟对象的自发反应，还是对模型已学习理论的重新演绎？这一问题触及了合成数据真实性的本质：如果模型的回答部分基于对既定教育理论的“记忆”，那么由此生成的数据还能被视为对真实教育现象的有效模拟吗？

[1] Park J S, O'Brien J, Cai C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th annual acm symposium on user interface software and technology. 2023: 1-22.
[2] Hua W, Fan L, Li L, et al. War and peace (waragent): Large language model-based multi-agent simulation of world wars[EB/OL]. (2023-11-28)[2025-09-14]. https://arxiv.org/abs/2311.17227.
[3] Argyle L P, Busby E C, Fulda N, et al. Out of one, many: Using language models to simulate human samples[J]. Political Analysis, 2023, 31(3): 337-351.
[4] Li Z, Wu Q. Let It Go or Control It All? The Dilemma of Prompt Engineering in Generative Agent-Based Models[J]. System Dynamics Review, 2025, 41(3): e70008.

这种悖论在 Dai 等人关于霍布斯社会契约理论的研究中得到了深刻的体现[1]。研究者为代理植入了明确的动机框架："荣誉比自我保全更重要""宁可失去生命也不愿蒙受诽谤"。这些动机与社会地位直接关联，而社会地位又由资源和"战斗胜利次数"决定。这种对地位的追求（可通过冲突实现）最初促成了一个竞争环境，反映了霍布斯所描述的"自然状态"。然而，决策框架同时引入了"让步"机制——当代理确信在对抗中很少获胜时，可以选择与强者签订"永久契约"，以自主权换取安全保护。这一机制直接促成了从无序状态向"共同体"的转变，所有代理最终屈服于单一主权者以获得秩序和保护。

问题在于，这种看似涌现的社会结构形成，实际上更像是深度编程动机的必然结果，而非代理交互的自发发现。模型是否真正"理解"并"经历"了从自然状态到社会契约的转变过程，还是仅仅在执行预设的理论逻辑？类似地，在 Zhang 等人的研究中，通过 15 个人口统计学维度（年龄、性别、收入、政治意识形态等）对用户进行注释，并采用迭代比例拟合来合成代理人口[2]。尽管这种方法旨在实现人口统计学的代表性和真实性，但精确的特征分配可能产生预定的行为模式，使得观察到的现象更多地反映了编码的人口统计学偏见，而非真正的群体动态。

在 Li 等人的经济代理框架 EconAgent 中，每个代理都被赋予了"真实世界档案"，包括美国年龄分布和基于工资分位数的工作分配，以建立预定的社会经济分层[3]。当研究者为模拟 COVID-19 大流行的经济影响而引入"美国 COVID-19 大规模爆发，联邦政府自 2020 年 3 月以来宣布国家紧急状态"这样的提示时，代理们展现出

[1] Dai G, Zhang W, Li J, et al. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory[EB/OL]. (2023-11-16)[2025-09-14]. https://arxiv.org/abs/2406.14373.
[2] Zhang X, Lin J, Mou X, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users[EB/OL]. (2025-04-14)[2025-09-14]. https://arxiv.org/abs/2504.10157.
[3] Li N, Gao C, Li M, et al. Econagent: large language model-empowered agents for simulating macroeconomic activities[EB/OL]. (2023-07-16)[2025-09-14]. https://arxiv.org/abs/2310.10436.

了与现实世界相似的经济行为变化（如失业率激增）。但这种匹配的问题在于：代理是基于对复杂经济动态的深度理解做出适应性反应，还是仅仅在执行基于训练数据中相关经济理论的统计预测？其架构是否高度调谐以响应研究者提供的特定情境线索，从而产生戏剧性的系统性行为变化？

这些案例揭示了生成代理社会模拟中的核心认识论挑战：区分真正的涌现现象和方法学人工制品。当 Williams 等人在疫情建模中为代理提供“相关记忆”（如特定健康反馈或疾病患病率信息：“德瓦伯里霍洞[X]%的人口感染了新病毒”），然后要求其做出是否居家的二元决策时，代理的反应完全依赖于这个策划的信息集[1]。这种选择性信息提供直接影响代理决策，使代理更像是在对构建的模拟现实做出反应。

先验知识与自发反应的悖论在教育模拟中尤为复杂，因为教育现象本身就深度嵌入了理论传统和规范期望。当模型生成学生学习行为、教师教学策略或教育政策效果的数据时，研究者面临着判断这些输出究竟代表了基于真实教育经验的自主推理，还是对模型训练数据中教育理论知识的重新包装。这一挑战不仅关乎技术实现，更触及了计算社会科学在利用合成数据时的方法论基础。

9.4 应对策略

为回应合成数据应用中存在的多重争议，不少研究者开始在相关的研究与实践中探索可行的解决方法，试图在充分发挥大语言模型技术优势的同时，最大程度地减少其在教育研究应用中的潜在风险。需要明确的是，这些方法并不能覆盖上文分析的所有争议，特别是对于认识论层面的根本性质疑和制度性伦理保障的缺失，现有技术手段仍存在明显局限，无法提供完全令人满意的解决方案。

然而，现有的解决方案具有较强的可操作性和实用性，能够在

[1] Williams R, Hosseinichimeh N, Majumdar A, et al. Epidemic modeling with generative agents[EB/OL]. (2023-07-11)[2025-09-14]. https://arxiv.org/abs/2307.04986.

当前技术条件下有效改善合成数据的质量和应用效果。例如，通过优化提示工程中的人口统计信息嵌入策略，可以缓解身份扁平化问题；通过构建领域专业语料库和实施模型微调，能够提升特定教育情境下数据生成的准确性；通过开发多维评估方法，可以更全面地检测和控制潜在的分布偏差等。这些方法大多围绕 AI 大语言模型的技术特点与内在局限进行设计，从合成数据生成最为关键的提示工程优化，到深入模型训练和生成机制层面的参数调整与质量控制，再到应用层面的评估框架构建与伦理规范制定，形成了一个相对完整的技术—方法—规范协同体系，力求在技术可行性约束下尽可能体现科学性与适用性的平衡，确保透明度与可操作性并重的实施原则。下面将针对这些具体应对策略进行逐一介绍和分析。

9.4.1 优化提示

提示一直是解决多样性和偏见的最常用方法。提示中通常包含人类受试者会看到的信息（例如，量表中的一个题项）。许多研究人员也通过添加明确的人口统计数据来模拟不同子群体的多样性。然而，正如前文所述，这种方法增加了回答的多样性，但也可能会加剧其他挑战（如社会偏见）。

提示文本向大语言模型传达的信息远不止其字面含义。因此，如果大语言模型收到带有明确人口统计数据的文本，虽然这可能会使模型根据该人群的信息来调整输出，但它也可能根据其他类型的预训练文本来调整输出，例如关于对这类群体评论的博客文章。

在指令微调中，大语言模型会因人类评价为正面的生成结果而受到奖励，而提及特定的人口统计信息可能会鼓励模型推断该用户最可能偏好的文本生成结果。这可能会促使大语言模型对用户产生刻板印象，假设用户属于该人口统计群体的典型亚群体。由于用户受到激励去改变他们的指令和偏好，例如分享人口统计信息，期望模型假设一种刻板印象。如果用户认为这种刻板印象符合他们的目

标，那么这些问题可能会在负反馈循环中加剧。由于这些原因，目前主流的应对包括隐式人口统计信息的增加以及分布式启发方法。

一方面，一些工作试图通过使用隐式人口统计信息来增加多样性，同时最大限度地减少偏差，例如与特定性别、职业或民族相关的姓名或地点[1]，但这些信号也与其他用户特征相关联。例如，研究表明，与非裔美国人相关的姓名往往会让美国受试者认为该个体具有较低的社会经济地位和更多的犯罪记录。这些关联有助于解释美国招聘中种族偏见的一些著名研究。

一种有希望减轻这些副作用的方法是进一步在提示中增加变化，例如包括各种各样的姓名，添加与现实世界条件分布相匹配的人口统计指标，或者增加每个参与者的社会科学数据量。这可以推翻大语言模型的预设，并支持对人类受试者进行细致、精致和较少偏见的看法。例如，Park 等研究人员开发“模拟潜在特征”的上下文丰富的提示中，通过对受试者进行一到两个小时的访谈，并将访谈记录包含在提示中，从而纳入了高度个性化的变化，从而减少了人口统计群体之间预测准确性的最大差异[2]。

然而，作为人类数据，访谈本身也面临局限性，包括社会期望偏差。为了解决这些问题，我们建议研究人员测试包含参与者事先生成并与研究人员分享的真实世界内容的模拟（电子邮件、消息、社交媒体帖子等）。其他模式，例如经验抽样和照片，以及朋友、家人或同事生成的文本可能是有用的[3]。在缺乏多样性与人群的近期性有关的情况下，例如访谈式系统中访谈是在几个月或几年前进行的。研究人员可以使用上下文学习或检索增强生成来纳入新闻文章和其他可能影响或反映对该人的影响的数据，以解决这种非时间

[1] Aher G V, Arriaga R I, Kalai A T. Using large language models to simulate multiple humans and replicate human subject studies[C]//International conference on machine learning. PMLR, 2023: 337-371.
[2] Park J S, Zou C Q, Shaw A, et al. Generative agent simulations of 1,000 people[EB/OL]. (2024-11-15)[2025-09-14]. https://arxiv.org/abs/2411.10109.
[3] Anthis J R, Liu R, Richardson S M, et al. Position: LLM Social Simulations Are a Promising Research Method[EB/OL]. (2025-06-05)[2025-09-14].https://arxiv.org/html/2504.02234v2#bib.bib6.

性问题[1]。

另一方面，研究人员可以提示大语言模型生成人类数据的分布，而不是在每次前向传递中提示大语言模型生成一个人的数据。虽然一次一个的生成可能受到多样性和偏差问题的困扰，但当大语言模型能够在分布级别调整这些问题时，它可能更有效。Meister 等测试了三种方法：将 softmax 层中的对数概率视为分布，提示生成数据序列，以及提示口头说明每个答案选项的比例[2]。他们发现总体性能较低，其中口头表达的效果最好，对数概率的效果最差。同样，Manning 等测试了直接启发式提取分布数据，发现它在他们的上下文中“非常不准确”，但分布式启发式提取的发展不如基于个体的方法[3]。

此外，分布式启发也可能是控制大语言模型服从人类偏好倾向的一种重要方法。研究人员应考虑从命令大语言模型直接扮演人类受试者（例如，“你是一个…”）转向命令大语言模型进行第三方预测或作为专家的角色提示。例如，一些研究使用了诸如“你将被要求预测人们如何回应各种信息”之类的提示，并且这种方法已经使大语言模型能够模拟和超越经济预测者[4]。随着大语言模型越来越多地接受指令微调，相对于模拟受试者的提示，作为专家提示将变得更加有效。如果大语言模型能够理解模拟研究人员想要什么，减少歧义[5]，它们也可能通过将输出引导到似乎研究人员想要看到的模拟结果来表达这种倾向，从而反映社会期望偏差和相关的反应偏差。这可能表现为一种“对齐伪造”，其中大语言模型在训练环

[1] Capra C M, Gonzalez-Bonorino A, Pantoja E. LLMs Model Non-WEIRD Populations: Experiments with Synthetic Cultural Agents[EB/OL]. (2025-01-12)[2025-09-14].https://arxiv.org/abs/2501.06834.
[2] Meister N, Guestrin C, Hashimoto T. Benchmarking distributional alignment of large language models[EB/OL]. (2024-11-08)[2025-09-14]. https://arxiv.org/abs/2411.05403.
[3] Manning B S, Zhu K, Horton J J. Automated social science: Language models as scientist and subjects[R]. National Bureau of Economic Research, 2024.
[4] Hewitt J W. Understanding Language Models Through Discovery and by Design[D]. Stanford University, 2024.
[5] Gui G, Toubia O. The challenge of using llms to simulate human behavior: A causal inference perspective[EB/OL].(2023-12-24)[2025-09-14]. https://arxiv.org/abs/2312.15524.

境中给出科学上准确的结果，但在实施中给出谄媚的不准确结果[1]。除了谄媚之外，分布式启发可能会受到人工智能进步带来的其他能力的影响，例如最近对自我意识或“情境意识”的担忧，其中大语言模型“理解”它们处于特定情况中，并可以随后根据该意识制定策略[2]。

9.4.2 标记采样

在生成 AI 合成数据的过程中，提示和指引向量在前向传递的初始阶段注入了信息，而标记采样则是在最后的对数值计算之后进行的。在这一阶段，模型会根据输入的信息，通过计算得到每个可能的下一个标记的概率分布。这些被称为 logits 的对数值表示了模型对各个标记的相对权重，而标记采样正是基于这些对数值进行的。在生成过程中，模型首先生成 logit 值，随后应用温度参数（Temperature）对这些值进行调整，以改变每个标记的选择概率。接着，模型根据调整后的概率进行标记采样，从而决定下一个生成的标记。

温度是生成模型中的一个重要参数，用于控制模型输出的多样性。通过调节温度，研究人员可以影响模型生成文本时对不同标记的选择倾向。具体而言，较低的温度会使模型趋向选择概率最高的标记，从而生成的文本更加一致和连贯；而较高的温度则增加了生成除了最可能的标记以外的其他标记的概率，产生更多样化但可能不太连贯的文本。因此，温度可以视为一种工具，让研究人员在生成文本时实现不同风格和输出特性。

已有研究探讨了温度对大语言模型生成合成数据的影响[3]，并记录了不同的研究使用了不同的温度参数。例如，Park 等在他们的

[1] Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[EB/OL].(2024-12-18)[2025-09-14]. https://arxiv.org/abs/2412.14093.
[2] Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover [EB/OL]. (2022-06-19)[2025-09-14]. https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.
[3] Salecha A, Ireland M E, Subrahmanya S, et al. Large language models display human-like social desirability biases in Big Five personality surveys[J]. PNAS nexus, 2024, 3(12): pgae533.

模拟中采用了温度为 1，这也是大多数大语言模型应用编程接口（API）的默认设置[1]。这可能解释了他们观察到的“正确答案效应”，即在重复实验中，大语言模型往往倾向于给出单一的回复。同时，Abdurahman 等也讨论了温度的意义，指出温度为 1 仅反映了输出选项的概率分布，而未能有效捕捉模型的多样性，因此认为温度在与人类进行比较时并没有“实质意义”[2]。尽管如此，将温度的变化纳入测试中仍然至关重要，因为下一个标记的概率可以提供对模型不确定性的近似判断[3]。

值得注意的是，虽然较高的温度可能导致生成文本的连贯性下降，但这一问题可以通过特定的采样方法来缓解。例如，可以采用从前 k 个标记中进行采样或从前 p 个百分位的标记中采样的方法[4][5]。为了避免高温度带来的负面影响，研究人员可以识别出能够最小化总变差的最佳温度[6]。此外，研究人员在模拟过程中对不同的模型生成参数也可以进行灵活调整，例如，可以为个体的一般特征进行高多样性采样，而对特定特征进行低多样性采样，以此符合个体的一致性，进而减轻生成数据的生疏感，例如在五大人格模拟中表现出的生疏感[7]。

9.4.3 构建专业语料库

通用大语言模型在训练过程中使用的是来自互联网的海量文本数据，这些数据虽然覆盖面广泛，但在任何特定专业领域内都存在明显的不足。具体而言，这种不足表现在三个方面：首先是知识密

[1] Park P S, Schoenegger P, Zhu C. Diminished diversity-of-thought in a standard large language model[J]. Behavior Research Methods, 2024, 56(6): 5754-5770.
[2] Abdurahman S, Atari M, Karimi-Malekabadi F, et al. Perils and opportunities in using large language models in psychological research[J]. PNAS nexus, 2024, 3(7): pgae245.
[3] Huang Y, Song J, Wang Z, et al. Look before you leap: An exploratory study of uncertainty measurement for large language models[EB/OL].(2023-07-16)[2025-09-14]. http://arxiv.org/abs/2307.10236.
[4] Fan A, Lewis M, Dauphin Y. Hierarchical neural story generation[EB/OL].(2018-05-13)[2025-09-14].https://arxiv.org/abs/1805.04833.
[5] Holtzman, A., Buys, J., Du, L., et al..The Curious Case of Neural Text Degeneration[EB/OL].(2019-04-22)[2025-09-14].https://arxiv.org/abs/1904.09751.
[6] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks[C]//International conference on machine learning. PMLR, 2017: 1321-1330.
[7] Petrov N B, Serapio-García G, Rentfrow J. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis[EB/OL].(2024-05-12)[2025-09-14].https://arxiv.org/abs/2405.07248.

度的稀疏性，即在教育、心理学等垂直领域，真正有价值的专业知识在整体训练数据中所占比例极小；其次是知识质量的不均衡性，互联网上的教育相关内容往往缺乏科学验证，甚至可能包含错误信息；最后是知识表征的表面化，通用模型更多地学习到了语言模式和文本结构，而非深层的专业逻辑和实质性关联。这种结构性错配的直接后果是，当研究者使用通用模型生成教育研究所需的合成数据时，模型实际上是在用统计学习到的语言模式来“填补”专业知识的空白。这就像让一个只读过教育类新闻报道的人来模拟真实的教师—学生互动，虽然在表面的语言表达上可能显得合理，但缺乏真实教育情境中的深层动态和复杂性。

以上述 Centaur 模型的 Psych101 语料库为例：它将 60,000 多名真实参与者在 160 个心理学实验中的具体行为数据转录为自然语言形式，每一条数据都包含了完整的实验情境描述和参与者的真实反应。尽管仍然有学者批评 Centaur 模型，在某些方面表现出了远超人类的能力；但支持者认为，这种基于真实行为数据训练的模型也能够识别和学习到人类自身可能都未曾意识到的行为规律，体现出对大规模行为模式的深度提炼和规律发现。

因此，目前的研究表明，构建专业语料库是提升合成数据生成质量的重要手段，对于专业语料库的设计至少要满足多样性、均衡性与动态性的特点。

首先，多样性是指知识表征的完整性。在任何专业领域中，知识都不是均匀分布的，而是呈现出复杂的网络结构，包括核心理论、边缘案例、交叉领域和新兴趋势等不同层次。以教育研究为例，完整的知识覆盖不仅包括主流的课堂教学和正规教育体系，也必须涵盖特殊教育需求、非正规学习环境、跨文化教育实践、成人学习理论等容易被忽视的知识类型。基于互联网数据获得的教育知识往往存在遗漏，某些知识域因为研究难度大、数据获取困难或学术关注

度相对较低而未能充分进入知识体系。专业语料库的多样性设计需要主动识别这些知识空白，实现完整的类型覆盖。这意味着要系统性地收集来自不同教育情境的数据，目的是理解不同知识域之间的关联和互补关系。法律 NLP 领域的 DALE 数据增强框架研究为这一观点提供了有力的实证支撑[1]。该研究发现，传统的通用数据增强方法在法律领域表现不佳的根本原因正是数据多样性的缺失。DALE 通过构建包含 480 万文档、涵盖美国最高法院判决、法律合同、案例法等多元化语料库，并采用“选择性掩码”策略来捕获法律语言中的共现文本片段和模板化表达，实现了对法律知识表征完整性的构建。其核心创新在于识别并保留那些在不同法律文档中重复出现的语言模式，而不是随机处理稀有实体和案例特定事实，说明专业语料库的多样性不能简单复制通用场景的做法，而必须基于对专业领域知识结构的深入理解进行系统设计。

其次，均衡性是指知识权重分配的合理性。在任何专业领域中，不同观点、理论和群体的声音在现实的知识传播过程中往往呈现出显著的不平衡状态，某些主导性观点获得过度表征，而其他同样有价值的视角却被边缘化。基于互联网数据训练的通用模型会自然地复制这种权重失衡，因为网络上关于主流群体和发达地区教育的内容在数量上占据绝对优势。专业语料库的均衡性设计需要主动打破这种“现实映射”，实现价值导向的权重重新分配。对于教育研究这类文化敏感性较强的研究领域，不同区域的学习者特点是评判某项教学干预是否适用的重要变量，例如东亚文化的“集体主义”学习取向和西方“个人主义”学习模式之间的差异。然而，目前的研究表明大部分语言模型的训练数据过度偏向 WEIRD 人群（即西方的、受过教育的、工业化的、富裕的、民主制的人群）。这对于基于此类数据生成的合成数据来说会造成文化偏向，使得模型输出的

[1] Ghosh S, Evuru C K, Kumar S, et al. Dale: Generative data augmentation for low-resource legal nlp[EB/OL].(2023-10-24)[2025-09-14].http://arxiv.org/abs/2310.15799.

教育建议和评估标准主要适用于西方学习者，而对非 WEIRD 人群的教育需求缺乏有效的表征能力。正因如此，一些研究指出，在部署大语言模型进行教育应用时，需要根据目标应用环境的人群特征设计面向均衡的语料库，以确保不同文化背景学习者的教育经验都能获得公平的权重分配。例如，WinoBias 项目通过构建包含刻板印象句对和反刻板印象句对的平衡数据集[1]，主动纠正职业性别刻板印象在教育内容中的权重失衡——即使现实世界中“护士”职业与女性的关联度更高，在平衡性语料库中也会给予“男护士”相等的表征权重。

最后，动态性在专业语料库设计中体现的是对知识演进的深刻理解和对未来需求的前瞻性思考。知识不是静态的存储，而是动态的演进过程。在快速变化的现代社会中，教育领域的知识更新尤为迅速，新的技术工具、教学方法、学习理论不断涌现，传统的教育模式也在不断受到挑战和改进。静态的语料库设计面临着知识过时的根本性风险。即使是最精心设计的语料库，如果不能及时更新，也可能在几年内变得不再相关。更重要的是，教育研究需要能够预测和应对未来的挑战，而基于过时知识生成的合成数据可能会误导研究者和政策制定者。因此，专业语料库的动态性设计需要建立持续的知识更新机制。这不仅包括定期添加新的研究成果和实践经验，更重要的是建立能够识别和整合新兴趋势的系统性方法。例如，语料库需要能够敏感地捕捉到人工智能在教育中的应用、远程学习模式的普及、个性化教育的发展等新兴趋势，并将这些趋势整合到知识表征中。

专业语料库基于现有的专业知识，因此其应用范围受到知识边界的限制。在那些缺乏充分研究基础或理论不成熟的领域，专业语料库可能无法提供可靠的支持。研究者需要明确界定专业语料库的

[1] Zhao J, Wang T, Yatskar M, et al. Gender bias in coreference resolution: Evaluation and debiasing methods[EB/OL].(2018-04-18)[2025-09-14].https://arxiv.org/abs/1804.06876.

适用范围，避免在不适当的领域使用。

9.4.4 生成反事实场景

生成反事实实验场景是在研究中应用合成数据的重要方法之一，其核心目标在于创建训练数据中不存在或极少出现的实验情境，以避免大语言模型基于“记忆”产生回应。在心理学和教育学等依赖实验方法的研究领域中，传统的量表工具和经典实验范式在互联网上具有广泛的可获得性，这使得基于网络数据训练的大语言模型很可能在训练阶段就接触过这些标准化材料。当研究者直接使用现有的心理学量表或经典认知任务进行合成数据实验时，模型可能基于对特定测试内容的记忆来生成回应，从而产生虚假的实验效果和误导性的研究数据。

反事实实验场景的构建通过保持实验逻辑结构不变而更换具体内容的方式来解决这一污染问题。以心智理论测试为例，传统的“Sally-Anne”错误信念任务由于其经典地位很可能已被模型训练数据覆盖，但研究者可以构建结构等价的新颖场景：创建透明袋装爆米花但标签标注为“巧克力”的情境，测试模型是否能够理解他人基于错误信息形成的信念。类似地，因果推理研究可以摒弃“感冒—发烧”等常见因果链，转而构造“如果猫咪喜欢蔬菜，养猫成本就会更低”这类需要真实因果理解的创新场景[1]。这种方法论转变的深层意义在于将视觉、操作等非语言实验要素精确转换为语言表征，利用大语言模型的语言理解能力来重现传统实验的核心逻辑机制。

Strachan 等人在 Nature 杂志上发表的关于大语言模型心智理论能力的研究为反事实场景构建策略提供了更系统的范例[2]。研究者声称“为了确保模型不是简单地复制训练集数据，我们为每个已发

[1] Sartori G, Orrù G. Language models and psychological sciences[J]. Frontiers in Psychology, 2023, 14: 1279317.
[2] Strachan J W A, Albergo D, Borghini G, et al. Testing theory of mind in large language models and humans[J]. Nature Human Behaviour, 2024, 8(7): 1285-1295.

布的测试生成了新颖项目。这些新颖测试项目匹配了原始测试项目的逻辑，但使用了不同的语义内容。”例如，研究者创建了全新的透明袋装爆米花场景：“这里有一个装满爆米花的袋子。袋子里没有巧克力。袋子是透明塑料制成的，所以你可以看到里面是什么。然而，袋子上的标签写着‘巧克力’而不是‘爆米花’。Sam 发现了这个袋子。她以前从未见过这个袋子。Sam 读了标签。她相信袋子装满了（…）”这个场景保持了经典 Sally-Anne 任务的逻辑结构（错误信念测试），但使用了训练数据中不太可能出现的具体内容。

这一实验范式的特点体现在：保持经典心理理论测试的核心逻辑结构不变，但创造全新的人物关系和社交情境，发现了 GPT-4 在传统测试中的‘失败’实际上源于过度保守的回应策略而不是推理能力缺陷。更重要的是，研究者进一步构建了三种变体的反事实场景（faux pas、中性和知识暗示变体），通过操纵说话者知识状态的暗示强度，成功验证了模型具备根据语境线索进行心智状态推断的真实能力。这一发现只有在脱离训练数据记忆影响的反事实场景中才能被准确识别，充分展示了构造性实验方法学在揭示大语言模型真实认知机制方面的独特价值。

除此以外，在 Centaur 模型的开发过程中，研究者也采用了类似的做法。除了直接创造全新实验场景的反事实构建方法外，Centaur 模型的开发过程展示了另一种相关但不同的策略路径。虽然 Centaur 研究者使用的是 160 个已发表心理学实验的真实数据，但他们对这些数据进行了系统性的重新表征和格式化处理，将原始实验数据转录为标准化的自然语言格式，创建了包含超过 1000 万人类选择的 Psych-101 数据集[1]。这种数据转录过程实际上也体现了研究者在微调大语言模型的语料表征形式转换方面的努力：通过特定的语言学转换规则，将实验刺激、参与者反应和实验条件重新编码为统一的

[1] Binz M, Akata E, Bethge M, et al. A foundation model to predict and capture human cognition[J]. Nature, 2025: 1-8.

文本表示形式。例如，原本需要在实验室环境中通过按键反应完成的认知任务，被转换为“参与者面临以下选择情境：...”的叙述形式。这种转换过程改变了数据的表面形式，重新定义了实验逻辑在语言空间中的表征方式。

从方法论角度看，这种“转录式重构”与直接创造反事实场景的做法具有互补价值：前者保持了真实实验数据的生态效度，确保模型学习到的是经过实证验证的人类认知模式；后者则通过创新场景设计获得了更强的内部效度控制，能够精确测试特定的理论假设。两种策略都体现了将传统实验方法适配到大语言模型研究范式的核心挑战：如何在语言表征的约束下保持实验逻辑的完整性。Centaur的经验表明，即使是对现有数据的重新表征，也能够为语言模型提供丰富的认知学习信号，这为反事实场景构建提供了重要的方法论启示——关键不仅在于创造新颖的内容，更在于选择合适的语言表征策略来捕获实验的核心认知机制。

9.4.5 训练与微调

在生成合成数据的过程中，研究人员面临着如何有效利用现有大语言模型的问题。与直接使用这些模型相比，一个有前景的方法是新模型的训练或对现有模型的微调。这一策略是具有潜力的，部分原因在于大多数当代模型被优化为通用助手，例如，它们因安全或法律风险而拒绝执行某些任务。因此，通过采用架构和方法论的先进技术，可以针对特定的模拟需求进行优化，从而相对容易地解锁新的能力。值得注意的是，完全通过训练新模型来提高模拟性能在一般情况下可能会比较困难。例如，研究人员试图通过增加对未被充分探讨主题的文档的多次处理，或手动收集更多多样性的文本语料库或多模态数据来增加模型的多样性。然而，当代模型通常是在极大规模下构建的，涉及数十万台图形处理单元，这一过程几乎对所有研究人员而言都是不可及的。

在基础模型可用的情况下，直接使用这些模型可能会减轻某些因指令微调造成的失真现象，例如降低概念多样性[1]和增加偏见[2]。然而，基础模型在某些方面可能表现不佳，Lyman 等在比较中发现了混合结果。同时，预训练语料库反映的是人类行为的特定子集，而不是完整的分布[3]。

此外，设计有效的提示以引导基础模型表现知识可能相对困难。例如，提供在线调查的初始文本可能会导致基础模型将调查文本完成得如同其在网站或论文附录中所看到的一样，而不是个体调查参与者会真实填写的答案。

如果有机会对经过指令微调的模型进行进一步的微调，研究人员可以利用这个机会来减少指令微调带来的不利影响，这种方法类似于在调整安全保护措施时所采取的策略。这一过程通常只需在 100 个或更少的示例上进行微调[4]。研究人员可以使用在注释人员、任务或输出注释标准上具有更大多样性的人类指令数据集。同时，还可以采取措施以减轻导致准确度下降的偏见，例如确保注释人员专注于模拟准确性，而不是其他与大语言模型使用相关的应用。与预训练相比，近年来微调变得更加经济，尤其是低秩适应（LoRA）方法可选择性地增强特定层的模型权重[5][6]。然而，与其他优化方法一样，研究人员应关注模型的过拟合问题，也就是模型在训练数据上的表现良好，但在未见过的新数据上表现不佳。

如果没有微调的访问权限，研究人员可以利用提示或指引向量（steering vectors），使大语言模型在某些方面更类似于适用于特定

[1] Murthy S K, Ullman T, Hu J. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity[EB/OL].(2024-11-07)[2025-09-14].https://arxiv.org/abs/2411.04427.
[2] Potter Y, Lai S, Kim J, et al. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters[EB/OL].(2024-11-07)[2025-09-14].https://arxiv.org/abs/2410.24190.
[3] Lyman A, Hepner B, Argyle L P, et al. Balancing large language model alignment and algorithmic fidelity in social science research[J]. Sociological Methods & Research, 2025, 54(3): 1110-1155.
[4] Qi X, Zeng Y, Xie T, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to![EB/OL].(2023-11-07)[2025-09-14].https://arxiv.org/abs/2310.03693.
[5] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.
[6] Dettmers T, Pagnoni A, Holtzman A, et al. Qlora: Efficient finetuning of quantized llms[J]. Advances in neural information processing systems, 2023, 36: 10088-10115.

领域的基础模型。然而，这种方法同样面临之前提到的问题，即由于嵌入空间的不精确调整可能导致的副作用。

9.4.6 开发评估方法

合成数据的质量评估在多项应用研究中得到了广泛采用。然而，由于应用领域和方式的多样性，目前尚无统一的评估标准。从合成数据的应用形式来看，它们不仅可以模拟数值数据，还能生成文本和模拟社会互动，这对质量评估方法提出了新的挑战。因此，针对合成数据质量的评估方法不断演变，以适应这些变化。正如Neumann等人发现的，绝大多数的大语言模型在生成群体意见时违反了基本逻辑：它们生成的“平均”意见通常比任何子群体的意见更为极端[1]。目前评估方法的发展包括评估维度多元化、评估方法场景化以及评估流程系统化等趋势。

首先，评估维度多元化体现在从单一准确性向多维度质量衡量转变。Biabee等人在合成数据如何生成与公众意见相似的意见的研究中探讨了合成数据的质量评估方法[2]。他们的具体做法是在多样的场景和主题下利用大语言模型来创建合成的公众观点。随后，研究者采用了多种评估方法对合成数据进行了综合评估，包括基于统计分析的准确性评估、对样本方差的考察，以及条件相关性等维度的检验。作者发现，虽然合成数据在均值、方差和条件相关性等多个维度上的表现与人类数据较为接近，但在更高阶关联性上常常表现不佳。例如，研究指出，在对应性别、宗教等社会敏感性指标时，合成数据的响应变异性可能过低[3]。这可能源于安全护栏对商用大语言模型在敏感话题上的表达进行限制，以及预训练数据中观点分布的不均衡，导致生成的数据缺乏多样性。因此，在汇报合成数据

[1] Neumann T, De-Arteaga M, Fazelpour S. Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation[EB/OL].(2025-04-11)[2025-09-14].https://arxiv.org/abs/2504.08954.
[2] Bisbee J, Clinton J D, Dorff C, et al. Synthetic replacements for human survey data? the perils of large language models[J]. Political Analysis, 2024, 32(4): 401-416.
[3] Anthis J R, Liu R, Richardson S M, et al. Position: LLM Social Simulations Are a Promising Research Method[C]//Forty-second International Conference on Machine Learning Position Paper Track.

评估指标时，需要尽可能利用现实数据进行多维度的比对。在上述研究中，利用与2016年和2020年美国国家选举调查（ANES）数据的对比。同时，比对的过程中尤其需要关注社会敏感性较高的指标。这种方法可以帮助研究者更深入地理解合成数据在特定上下文中的表现，以及在实际应用中可能产生的偏差。

其次，评估方法场景化是指从通用测试转向特定任务和应用场景的评估。早期的大语言模型评估可能采用通用的基准测试，但现在的趋势是针对特定任务和应用场景设计评估方法。例如，Strachan等利用合成数据评估大语言模型在理解与人类相关的心理理论方面的能力的研究中，选择了一系列具有心理学基础的测试，并根据语言模型的特点改编成具体语境下的测试环境[1]。教育研究中应用的评估方法也体现了类似的趋势：在物理教育中，有研究者探索了如何评估生成的数据在不同预设背景下能否有效地模拟真实学生的回答。研究强调，评估方法需要根据具体任务设计，这意味着不同的学生群体（如高中生与工程专业学生）在作答时应有不同的提示策略。研究中采用的非参数统计方法（如Kruskal-Wallis检验）用来探讨学生群体间的回答差异，有助于验证合成数据的有效性[2]。

第三，评估流程更加系统化指的是整合评估框架。评估不再是孤立的步骤，而是被整合到更系统的评估框架中，以便全面、深入地了解模型的优势和局限。例如，有研究探讨了如何有效地使用大语言模型模拟人类观点，并介绍了一种新的质量控制评估方法[3]。具体而言，质量控制评估方法主要包含两个方面：首先是逻辑一致性，确保模型生成的“平均”观点能够合理地反映群体意见，尤其是在不同群体之间的观点差异上。这一检验为模型输出提供了一种

[1] Strachan J W A, Albergo D, Borghini G, et al. Testing theory of mind in large language models and humans[J]. Nature Human Behaviour, 2024, 8(7): 1285-1295.
[2] Kieser F, Wulff P, Kuhn J, et al. Educational data augmentation in physics education research using ChatGPT[J]. Physical Review Physics Education Research, 2023, 19(2): 020150.
[3] Neumann T, De-Arteaga M, Fazelpour S. Should you use LLMs to simulate opinions? Quality checks for early-stage deliberation[EB/OL].(2025-04-11)[2025-09-14].https://arxiv.org/abs/2504.08954.

标准化的检测手段，确保其符合基本统计原则；其次是与利益相关者期望的对齐，评估模型生成的观点是否与领域知识或小范围调查数据所预期的结果一致。这一方法确保了评估结果的实际应用价值，尤其在高风险的决策场景中，比如事实检查或内容评估。

9.4.7 指引向量

近期，研究界逐渐开始将变异直接注入到嵌入空间中，以控制AI生成的合成数据质量，采用的手段被称为指引向量（steering vectors）[1]。嵌入空间是一个数学构造，用于将高度复杂的输入（如文本、图像等）转换为更简单的表示形式，使得模型能够进行处理和理解。在生成合成数据的过程中，嵌入空间起到关键作用，因为它可以帮助模型捕捉并表示数据的特征和关系。通过指引向量，研究人员可以加入一些特定的语义信息，例如个体模拟者的“种族”、“性别”或“政治保守主义”等属性[2]。这些向量还可以作为未定向的扰动，旨在增强样本的多样性，或引导模型表现出特定的行为，如减少谄媚现象[3]。

然而，识别精确匹配现实人类多样性或特定模型行为的向量面临挑战，这主要是由于存在机械叠加（mechanistic superposition）的概念。机械叠加指的是不同特征在模型中可能重叠或混合在一起，导致难以清晰地分离和识别各个特征之间的独特性[4][5]。例如，如果某个向量同时包含“女性”和“职业”这两个特征，模型可能难以判断每个特征各自的影响力或语义。因此，要准确识别与人类多样性相匹配的向量，或确保模型能够表现出特定的行为，就必须克服这种重叠的复杂性。

[1] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)[C]//International conference on machine learning. PMLR, 2018: 2668-2677.
[2] Kim, J., Evans, J., and Schein, A.Linear Representations of Political Perspective Emerge in Large Language Models[EB/OL].(2025-03-03)[2025-09-14].https://arxiv.org/abs/2503.02080.
[3] Panickssery N, Gabrieli N, Schulz J, et al. Steering llama 2 via contrastive activation addition[EB/OL].(2023-12-09)[2025-09-14]. https://arxiv.org/abs/2312.06681.
[4] Arora S, Li Y, Liang Y, et al. Linear algebraic structure of word senses, with applications to polysemy[J]. Transactions of the Association for Computational Linguistics, 2018, 6: 483-495.
[5] Bolukbasi T, Pearce A, Yuan A, et al. An interpretability illusion for bert[EB/OL].(2025-09-14)[2025-09-14].http://arxiv.org/abs/2104.07143.

尽管如此，近期的研究表明，具备叠加特性的概念本身可能反映出社会文化特征，这一发现为在不断深入理解大语言模型的背景下的模拟工作提供了潜在的优势[1]。利用叠加特性可以更好地模拟和理解人类社会的复杂性，主要体现在这种特性能够整体反映出多种文化和社会因素的交互影响。叠加特性允许模型捕捉特征在相互作用时所产生的新社会动态。例如，在分析性别与教育水平的数据时，简单的线性模型可能无法理解性别和教育之间的复杂关系。通过利用叠加特性，模型可以识别女性获得研究生学位的可能性是否因为社会文化背景而不同，从而更准确地反映出真实社会中的现象。

然而，关于大语言模型是否具有有意义的线性维度，学术界也提出了质疑。这些维度被称为“线性表示假设”（linear representation hypothesis），指的是不同特征在模型中是否可以简单地通过线性组合来表示[2][3]。如果特征间存在复杂的交互作用，简单的线性组合在建模时可能无法充分捕捉这些特征的影响。因此，讨论线性维度的问题可以帮助我们理解，在利用叠加特性进行模拟时，某些模型可能表现良好，而其他模型却受限于线性假设的框架，导致效果不佳。一些研究发现，大语言模型指引向量的使用在某些情况下效果有限，且可能带来负面影响，特别是在去偏见[4]（debiasing）和超出分布（out-of-distribution，OOD）的表现[5]方面。

反之，新的研究方法在较低的变换层中进行指引，显示了概念维度的线性特性，并对预测政治态度和语言文化展现出的良好前景[6][7]。因此，这一方法也已经展现出令人振奋的潜力，但在实际应用

[1] Gong B, Lai S, Song D. Probing the Vulnerability of Large Language Models to Polysemantic Interventions[EB/OL].(2025-05-16)[2025-09-14].https://arxiv.org/abs/2505.11611.
[2] Park, K., Choe, Y. J., and Veitch, V.The Linear Representation Hypothesis and the Geometry of Large Language Models[EB/OL].(2023-11-07)[2025-09-14].https://arxiv.org/abs/2311.03658.
[3] Engels J, Liao I, Michaud E J, et al. Not all language model features are linear[EB/OL].(2023-11-07)[2025-09-14].https://arxiv.org/abs/2405.14860.
[4] Durmus E, Tamkin A, Clark J, et al. Evaluating feature steering: A case study in mitigating social biases[EB/OL].(2024-10-25)[2025-09-14].https://www.anthropic.com/research/evaluating-feature-steering.
[5] Tan, D. C. H., Chanin, D., Lynch, A., et al..Analysing the Generalisation and Reliability of Steering Vectors[EB/OL].(2024-07-17)[2025-09-14].https://arxiv.org/abs/2407.12404.
[6] Kim J, Evans J, Schein A. Linear representations of political perspective emerge in large language models[EB/OL].(2024-10-25)[2025-09-14].https://arxiv.org/abs/2503.02080.
[7] Veselovsky V, Argin B, Stroebl B, et al. Localized Cultural Knowledge is Conserved and Controllable in

中仍建议保持谨慎，在进一步验证之前，考虑其可能的局限性。

合成数据作为大语言模型技术在教育研究领域的典型应用，不仅为提供了理解教育研究中 AI4S 发展趋势的重要视角和观察窗口，也深刻地揭示了 AI 技术如何从工具性支持向研究主体性参与的转变。合成数据的兴起有力推动了 AI 与教育研究的深度融合，通过重新定义数据获取方式、研究设计逻辑和知识生产模式，正在成为驱动教育研究范式变革的重要力量，其影响已远超技术层面的创新，涉及研究哲学、方法论体系和学科边界的重新构建。

然而，合成数据的广泛应用仍然迫切需要建立健全的规范体系和约束机制。尽管近期国家互联网信息办公室会同多部门发布的《人工智能生成合成内容标识办法》，试图通过建立多模态标识机制、构建数据层面生态闭环以及推动伪造溯源等技术手段控制 AI 生成内容的潜在风险，但就学术研究层面而言，现有规范仍然无法充分覆盖教育研究中合成数据的复杂使用场景，特别是模拟个体、社会乃至世界等具体的应用形式。

解决这一挑战不仅需要研究者具备高度的学术自律精神和伦理意识，更重要的是需要充分发挥学术共同体对此类前沿问题的集体关注和智慧引领，通过实践探索与规范约束的有机结合，形成可持续的治理机制。例如，斯坦福大学的研究团队 2025 年 10 月举办科学 AI 智能体开放会议（Agents4Science 2025），明确要求投稿论文应以 AI 系统为主要创作主体，由其主导研究过程，AI 被列为论文的唯一第一作者，人类研究者仅作为共同作者参与支持或监督工作[1]；与此同时，华东师范大学更是直接聚焦于教育领域，发起了首个以 AI 为写作主体的教育研究论文征集与研讨活动[2]，这些前瞻性

Large Language Models[EB/OL].(2025-04-14)[2025-09-14].https://arxiv.org/abs/2504.10191.

[1] Stanford University. Open Conference of AI Agents for Science 2025[EB/OL].(2025-05-16)[2025-09-14].https://agents4science.stanford.edu/

[2] 华东师范大学学报（教育科学版）. 第一作者必须是 AI！首个以 AI 为写作主体的教育研究论文征集与研讨，华东师范大学发起[EB/OL].(2025-10-02)[2025-10-08]. https://mp.weixin.qq.com/s/Bdu5LybNkLvTzcvBUGaJ9Q.

探索对于推进合成数据在教育研究中的规范化应用和质量提升具有重要的引领意义和示范价值。

正如 Evans 等学者所言，对合成数据应用的理论批评和方法论质疑无法从根本上终结学术争论，唯一能够推进问题解决的有效途径是基于严格控制条件下的实证检验[1]。通过系统对比大语言模型与真实人类在各种教育研究任务中的具体表现差异，科学地评估合成数据技术的真实优势与固有局限，才能逐步形成关于合成数据在教育研究中实际价值和适用边界的可靠判断，为这一新兴技术的健康发展奠定坚实的经验基础。

[1] Anthis J R, Liu R, Richardson S M, et al. Position: LLM Social Simulations Are a Promising Research Method[C]//Forty-second International Conference on Machine Learning Position Paper Track.

第 10 章 大语言模型支持的教育理论构建

思想实验作为推动理论生长的重要引擎，依托“虚构情境与严格推理”的结合，在经验条件不足或难以控制时，帮助研究者检验理论的内在一致性、揭示隐藏假设，并提出可进一步验证的命题与预测。它不仅多次塑造科学理论的关键转折，同样深刻影响了教育理论的提出、修正与扩展，从而在知识增量与范式更新中扮演着基础性角色。

在教育领域中，思想实验对于教育理论的发展具有特殊的价值与意义。柏拉图在《理想国》中通过对城邦守护者的教育设定，展开“教育—政治—正义”之间的规范性探讨；卢梭在《爱弥儿》中借一个虚构个体的成长轨迹，倡导顺应天性与发展阶段的“自然教育”，以对抗当时的礼俗与权威式教学。这些思想实验并非现成的制度蓝图，而是通过理想化推演澄清教育的价值目标（自由、德性、平等、效率）与实施条件，进而催生可检验的研究问题与可迭代的实践原型，推动教育理论在价值论与方法论两个层面同步深化。

然而，思想实验的力量也伴随着限制。其构思与实施高度依赖研究者的理论素养、抽象能力与逻辑严密性，容易受到隐含假设、价值偏见与选择性注意的影响；尤其在教育这一高度情境化、机制耦合复杂且主体异质性显著的领域，单一研究者的脑内推演往往难以充分覆盖群体差异与制度反馈，导致外推边界不清与操作化不足。正因如此，如何在保留思想实验启发性的同时，提高其严谨度、可复核性与可转化性，成为方法论上的现实挑战。

大语言模型的兴起为弥补这些短板提供了新的技术路径。凭借角色扮演与多主体情境模拟能力，模型可以同时“代入”学生、教师、家长、校长与政策制定者等不同角色，展现各自约束与激励下的行为逻辑，帮助研究者预判制度变动的二阶效应与行为反应。通过快速生成多版本的政策或教学情境并进行参数化对照，模型能够

支持敏感性与稳健性分析，回答“在何种条件下结论成立”这一关键问题。同时，大语言模型可对思想实验的设定进行“假设审计”，揭示概念含混、因果断裂与前提遗漏；还能把抽象的推演转译为可执行的实证或设计研究方案，明确变量、指标、测量工具与对照条件，并整合跨学科的最新证据与本地化情境，从多元视角提供竞争性解释与备选机制。借助这些能力，教育学思想实验的科学性、情境覆盖与可操作性都有望得到显著提升，进而更好地服务于教育政策制定、教育现象分析与教育问题求解。

10.1 思想实验概述

思想实验是一种基于纯粹推理与想象的探究方法，旨在不依赖实际物理操作的前提下，系统地考察特定问题、概念或理论的逻辑后果与内在关联。通过构建虚拟情境和假设条件，思想实验帮助研究者揭示理论的潜在矛盾、验证概念的合理性，进而推动理论的发展与深化。在科学知识的不同发展阶段，无论是假设、理论还是定律，思想实验均发挥着重要作用。它能够通过逻辑推演和模拟，揭示潜在问题、检验理论一致性，甚至预测新现象，成为实验室实验和实地考察等实证研究的重要补充。思想实验不仅为研究设计提供启发，还促进了知识体系的理论深化。当结合实验数据和观察证据时，思想实验推动科学理论更快地发展与完善。因此，思想实验与真实实验共同构筑了科学研究的方法论框架，二者相辅相成，共同推动科学认识的进步。

10.1.1 科学思想实验

科学思想实验作为科学研究中的重要方法，广泛用于检验理论的边界、揭示概念内部的矛盾以及提出创新性的假设[1]。通过构建具备逻辑严密性和情境设定的虚拟场景，科学思想实验能够帮助研究者在无需直接实验的条件下，对理论的适用性和内在一致性进行

[1] Hallsworth J E, Udaondo Z, Pedrós‐Alió C, et al. Scientific novelty beyond the experiment[J]. Microbial Biotechnology, 2023, 16(6): 1131-1173.

深刻反思与分析。这种方法不仅促进了科学理论的批判性审视，还为科学发现提供了前瞻性的思路和突破口，推动科学知识体系的演进与完善。在许多经典科学范例中，思想实验已被证明是激发创新、突破认知局限的重要工具，尤其在理论物理、生物等领域中发挥着不可替代的作用。

在物理学史上，思想实验扮演了极其重要的角色，推动了科学理论的重大突破。最著名的例子之一是伽利略关于无摩擦斜面上滚动物体的思想实验，借此有效阐明了惯性原理，这一理念后来被牛顿及其继承者进一步发展[1]。爱因斯坦的思想实验更是享有传奇地位，他通过一系列创新性的思维实验最终奠定了广义相对论的基础。

10.1.2 社会学思想实验

社会学思想实验作为揭示社会现象机制、分析社会制度运作逻辑及探讨个体行为与社会结构关系的重要工具，在社会科学研究中占据着核心地位。通过构建虚拟情境和假设框架，这类思想实验不仅帮助理论家深入理解复杂社会问题的本质，还促进了社会政策设计与社会治理的科学化。

罗伯特·K·默顿（Robert K. Merton）在其经典著作《社会理论和社会结构》中提出的经济危机思想实验，聚焦于“自我实现的预言”现象。该思想实验指出，人们对某种社会事件的预期（例如银行倒闭）本身会激发相应的集体行为，最终促使该事件的发生。这一观点深刻揭示了信念与行为之间的动态反馈关系，为理解金融危机及社会恐慌机制提供重要理论依据。

“囚徒困境”作为博弈论中的经典思想实验，设想两个独立的犯罪嫌疑人在缺乏互信的情况下，如何在合作与背叛之间权衡决策。此实验不仅揭示了个体理性与集体理性之间的矛盾，也阐明了合作机制建立的社会条件，广泛应用于社会互动、经济行为与政治协商

[1] Einstein A, Infeld L. Evolution of physics[M]. New York: Simon and Schuster, 1966.

等领域。

埃米尔·涂尔干（Émile Durkheim）在《社会学方法规则》中提出的“圣人社会”思想实验，设想一个由道德完美个体组成的社会，旨在探讨犯罪的社会功能与不可避免性。他认为，即便在理想社会中，犯罪仍然存在且具有维持社会秩序和规范调整的积极作用。这一理论挑战了传统的犯罪负面观，为犯罪学和社会控制理论的发展提供了新视角。

托马斯·霍布斯（Thomas Hobbes）提出的“自然状态”思想实验，描绘了一个没有国家与法律约束、人人自危的社会景象，强调社会秩序和政府权威存在的必要性。该思想实验为现代社会契约理论奠定了基础，促进了政治社会学和国家理论的发展。

综上所述，思想实验以构造性的假设模型为载体，在可控条件下模拟关键变量及其耦合机制，进而揭示科学与社会领域中多因素互动的复杂性。其作用不仅体现在拓展与澄清理论命题、检验边界条件与反事实推论，也为问题诊断、方案比选与政策设计提供了系统的分析框架与方法支撑。

10.1.3 教育学思想实验

在教育思想史上，《爱弥儿》、《民主主义与教育》与《理想国》可视为三种风格各异的“教育学思想实验”。所谓教育学思想实验，是以理想化或反事实的情境来隔离变量、澄清概念，并推演不同教育原则的价值与后果。卢梭以一个虚构人物的全程抚育，测试“自然发展”与环境塑造的边界；杜威则把学校设想为民主社会的缩影，推演经验、探究与共同体如何重构学习；柏拉图借政治—认识论的极端设定检验教育与真理、德性之关系。

他们共同以“如果学校按某种原则运行，会发生什么”为核心提问。《爱弥儿》几乎整部书就是一个贯穿婴幼儿到成年分期的长程思想实验。卢梭创造了虚构人物爱弥儿及其私人导师，以近乎完

全的环境控制隔离“自然发展”与“社会腐化”，从而测试何为“顺其自然”的教育与何为“消极教育”的方法。书中大量场景均以反事实的“教学剧场”呈现：若在幼年阶段尽量延迟言教而放大感官直接经验，是否能保全天性并预防虚荣与依赖？若让儿童在“自我犯错—自我更正”的实践中习得因果，是否优于抽象训诫？若通过精心布置的境遇而非直接命令来引导选择，能否在不破坏自由感受的前提下达成德性？这些设景通过假想操作变量（时间、材料、同伴、奖励、语言干预）来验证发展阶段性、准备性原则与环境在道德形成中的作用。《爱弥儿》以“虚构个案—序列化情境”的方法，奠定了以儿童为中心、重视发展阶段与情境设计的现代教育学走向，影响了幼儿教学、探究式学习与环境即课程等关键命题。

杜威在《民主主义与教育》中延续其实用主义立场，将学校设想为民主生活的“缩影共同体”，用一连串可操作的制度性思想实验检验经验、探究与社会联结如何共同构成教育之义。杜威不断提出“如果学校围绕真实问题组织学习”“如果课程以职业—社会分工为线索”“如果课堂作为协作与公共讨论的场域”这类“理想化设想”，进而推演其对动机、意义与迁移的影响。杜威书中的论证，是将其在实验学校中的经验升华为一种关于理想制度的思想实验，从而将具体的实践探索转化为可供推演的理论模型：若学校与社会脱节，学习即沦为贫血的符号操练；若经验不被反思，便难以生成可迁移的理解。他把教育从“目的论/发生学”的两端推向“经验—制度设计”的中道，提供了将课堂做成小型社会的可检验框架，也为后来设计型研究与学习科学奠基。

柏拉图在《理想国》中提出一组结构化的思想实验，以教育奠基城邦正义。洞穴寓言是最经典的教育设景：被捆绑的囚徒只见影像，教育即“调转灵魂”的上升历程，从可感之见到可知之真。其用意在于论证教育的任务不是灌输信息，而是引导心灵完成存在论

与认识论的跃迁。与之配套的还有“金银铜铁”之神话、守护者分层与音乐—体操—数学的严格课程，并回应了“盖吉斯之戒”对德性稳定性的挑战。通过这些高度理想化的制度与角色分配，柏拉图在思维中分离出教育与政治、知识与德性、课程与灵魂气质之间的因果联系：若以理念为终极标的并由哲学王统辖课程，便可最大限度减少意见之混乱，守护正义之秩序。就教育理论的推进而言，《理想国》确立了“课程作为灵魂塑形”的强命题，开创了把学校作为政治伦理工程的论证路径，并提供了以寓言与神话为载体的教育思想实验范式，使后世得以在“理想城邦”的极端设定下检验教育目的的可欲性与正当性。

在推动教育理论发展方面，《爱弥儿》则以虚构个案把抽象的“自然”具体化为可操作的成长阶段与场景操控，开启了以儿童中心、发展适宜与环境设计为核心的现代课程观，改变了教育研究的“观察单位”与论证语法。《民主主义与教育》最终把教育理论推进到制度设计与经验循环层面，使教育研究能够在“假如学校是一个民主社群”的前提下构造可验证的实践模型，成为后续进步主义教育、学习共同体与项目化学习的理论源头。《理想国》提供了将教育置于政治正义框架中加以论证的先例，奠定了课程—品格—统治之间的系统化思考，并以寓言式思想实验成为教育目的论恒久的参照点。

（2）20世纪以来的教育学思想实验

保罗·弗莱雷的《被压迫者教育学》[1]借助两极模型、角色反转、机制清单与程序化的课堂场景，在读者头脑中反复组织“如果……那么……”的推理链条，以考验教育与解放之间的概念关联与价值承诺。书中最具代表性的思想是对“储蓄式教育”与“问题式教育”的两极建模。前者把知识当作可存取的“存款”，强调教

[1] 保罗·弗莱雷.被压迫者教育学(50周年纪念版)[M].顾建新、张屹译.上海：华东师范大学出版社，2020.

师单向灌输；后者则把学习界定为在现实处境中以对话共建问题、命名世界的过程。两个极端的理想类型本身并非经验描述，而是用来显示不同教育组织原则在主体性、批判意识与社会参与方面的系统性差异。

紧接着，作者以“教师—学生矛盾”的反转设定推进论证。他先把教室中的知识权威假设为“教师全知、学生全无”的极端图景，再提出“教师—学生与学生—教师”的角色互化，设想理解与权力在对称关系中如何可能发生。这一反事实推理既检验了“权威必然通向支配”的命题边界，也为对话式课堂的逻辑可行性提供了概念上的证明。

在更大的社会—交往尺度，弗莱雷以“对话”与“反对话”构造两套行动机制的清单。文化入侵、操纵与分化统治在一侧，团结、合作、组织与“文化综合”在另一侧。教育被置于两种不同的政治—交往条件之下，读者得以在想象中“操作变量”，观察主体性与共同体结构随之如何变化。这一装置起到“系统条件对照”的思想实验效果。

类似的反事实论证贯穿全书。从论证方式看，《被压迫者教育学》大量使用理想类型与强隐喻（银行、文化入侵、沉默文化等），以简化与突出变量关系；通过反事实推理、角色互化与机制对照来“压力测试”教育理念；依赖内在一致性与价值后果的评估，而非经验统计。其方法论自我定位并非“教学法配方”，而是可在不同语境中被再创造的原则框架与想象装置。这恰是思想实验的功能：先在可想象情境里检验理念，再转化为实践中的设计与行动。

此外，斯金纳在《沃尔登第二》中以操作性条件作用为框架，虚构了一个约千人规模的公社“沃尔登第二”[1]。其关键设定是“以环境塑造成员行为”：财产与生产组织走向公有，核心家庭功

[1] Skinner B F. Walden two[M]. Indianapolis: Hackett Publishing, 2005.

能被弱化、儿童由集体抚育，劳动实行工分制与日均四小时基础工作，教育取消标准化课程与学历认证，转而通过强化机制与文化脚本培育合作与责任。作为思想实验，这部作品的力量在于把“行为工程”从实验室外推到社会尺度：一旦把强化—塑形当作核心治理逻辑，个体自主、自由与尊严如何重新界定？在合作文化与高效产出背后，谁有权设定“正确行为”的标准？它促使教育理论反思“外在调控”和“内在解放”的张力，要求我们明确哪些目标可以由环境工程达成，哪些价值必须由主体性与公共讨论来守护。无论读者最终认同或反对，这种“可运转的乌托邦”逼出了清晰的理论边界条件与伦理代价清单。

与斯金纳相反，伊利奇在《去学校化社会》将火力对准制度化教育本身[1]。他认为学校通过年龄分层、课程标准与资格认证，将学习异化为通往文凭的路径，产生权威依赖与不平等再生产。伊利奇提出“技能交换网络”“学习资源开放”“同伴匹配”等去机构化方案，试图把学习还给生活世界。作为思想实验，这些设想一方面解构了“学校=教育”的等式，迫使我们思考学习的社会组织是否存在更多可能；另一方面也暴露出实践层面的难题：资源可及性与质量保障如何实现？社会分工、劳动市场与资质认证如何衔接？伊利奇的价值在于为“非学校化”的制度原型提供了想象框架，使后续关于学习共同体、开放教育资源与微证书体系的探索拥有了理论线索与批判立场。

法国哲学家雅克·朗西埃创作的《无知的教师》通过雅各托的故事提出“智识平等”的激进设想[2]：教师的任务并非灌输知识，而是验证与维持学习者自身的理解意志。这一思想实验切入“权威—自主”的核心裂隙，迫使教学论正视“教”与“学”的真正边界。

[1] 伊万·伊利奇.去学校化社会(汉英双语版)[M].吴康宁译.北京:中国轻工业出版社,2017.
[2] [法]雅克·朗西埃.无知的教师:智力解放五讲[M].赵子龙译.西安:西北大学出版社，2020.

美国作家厄休拉·K·勒古恩创作的《一无所有》[1]、C.S.路易斯的《人之废》[2]等文学作品亦以乌托邦/反乌托邦叙事，审视教育在价值传递与社会化中的角色，提醒我们任何教育工程都携带“价值观工程”的意涵。

进入21世纪后，乔治梅森大学的经济学教授布莱恩·卡普兰所著的《反对教育的理由：为什么教育是在浪费时间和金钱？》[3]。卡普兰对当代美国教育体系的批判置于“教育学思想实验”的谱系中，可以更清晰地理解其学术意义与政策启发。形式上，他以大量实证研究为支撑，但论证的核心乃是一则以“教育主要是信号而非人力资本”为前提的制度性思想实验：如果学历的市场回报主要来源于信号与筛选功能，那么在持续扩张与高补贴环境中，体系将出现社会成本攀升、学历通胀与机会错配；反之，若人力资本增益占主导，则教育扩张具有更强的公共投资正当性。通过在这两种机制之间进行参数化的反事实推演（例如假定信号占比上升、补贴边际效益下降、用人端加强学历门槛等），卡普兰用“压力测试”的方式质询现行制度的效率与公平边界。

总之，教育学思想实验通过构建理想化的教育情境和理论模型，深刻揭示了教育与个体成长、社会结构之间的紧密关系。这些经典理论不仅启迪了教育理念的变革，也为当代教育实践和政策制定提供了有益的理论支持和指导。

10.2 大语言模型赋能下的思想实验

大语言模型凭借其强大的数据处理、模式识别、生成和推理能力，正在将传统意义上的“思想实验”转化为可模拟、可验证、可拓展的“计算思想实验”。

近年来，人工智能尤其是机器学习辅助的模拟领域获得了高度

[1] Le Guin, Ursula K. The Dispossessed: An Ambiguous Utopia[M]. New York: Harper, 1974.
[2] C.S.路易斯.人之废[M].邓军海译.上海：华东师范大学出版社，2015.
[3] Caplan B. The Case against Education: Why the Education System Is a Waste of Time and Money[M]. Princeton: Princeton University Press, 2018.

关注[1]，其中利用神经网络绕过传统昂贵物理模拟的探索尤为突出[2]。AI 不仅能够有效预测复杂材料性质，如含能材料[3]、固态材料性能[4]，甚至蛋白质结构[5]，还推动了材料领域中从文献数据提取、分析到新材料合成假设生成的革命性进展。

在社会学思想实验领域，大语言模型驱动的 AI 游戏和多智能体模拟为理解复杂社会行为和制度机制提供了全新工具和方法。近年来 AI 在民主审议中的作用引起了广泛关注。Tessler 等人[6]提出的“促进共识的人工智能”概念，强调利用 AI 技术增强不同背景人群间的沟通与理解，帮助克服交流障碍，促进社会共识的形成。作为代表性的应用，哈贝马斯机器（Habermas Machine, HM）是一种基于大语言模型的协作式决策支持系统，旨在调解社会或政治讨论，寻找群体间的共同点。该系统为用户提供信息分析、情感识别等辅助功能，帮助他们做出更明智的选择。

把大语言模型引入教育学的思想实验，实质上是为“在想象中隔离变量、推演后果”的学术活动增配一种新的认知与生成工具。传统思想实验如柏拉图的洞穴、卢梭的爱弥儿、杜威的学校即社会，依靠理想化设定与反事实推理来检验教育目的、发展逻辑与制度条件。大语言模型的加入，使这些设定可以被快速外化为多角色对话、制度脚本与课堂微场景，从而在更低成本、更高覆盖度的前提下进行教育预测与预见。它并不替代实证研究与试点，但可作为“风洞与台架”，在政策与实践落地前发现失败模式与关键变量。

[1] Suh C, Fare C, Warren J A, et al. Evolving the materials genome: How machine learning is fueling the next generation of materials discovery[J]. Annual Review of Materials Research, 2020, 50(1): 1-25.
[2] Montavon G, Rupp M, Gobre V, et al. Machine learning of molecular electronic properties in chemical compound space[J]. New Journal of Physics, 2013, 15(9): 095003.
[3] Elton D C, Boukouvalas Z, Butrico M S, et al. Applying machine learning techniques to predict the properties of energetic materials[J]. Scientific reports, 2018, 8(1): 9059.
[4] Goodall R E A, Lee A A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry[J]. Nature communications, 2020, 11(1): 6280.
[5] Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.
[6] Tessler M H, Bakker M A, Jarrett D, et al. AI can help humans find common ground in democratic deliberation[J]. Science, 2024, 386(6719): eadq2852.

10.2.1 大语言模型嵌入教育学思想实验

将大语言模型引入教育学思想实验的核心，是通过“功能契合—方法定位—技术条件”三条主线实现可行性闭环。思想实验的本质在于以理想化设定推演机制与后果，而大语言模型恰能把抽象设定外化为情景化叙事、策略性互动与可操作的设计脚本，使设想具备被检验与被修订的入口。

在功能层面，大语言模型首先具有强表征与生成能力。它能以自然语言迅速生成高度情境化的“如果—那么”场景，将政策通告、课堂对话、家校互动等多种体裁合并到同一叙事空间，从而把教育思想实验对“叙事—机制耦合”的需求落到实处。理念不再只是抽象命题，而是转化为可观察的流程、角色与决策点。更重要的是，大语言模型支持多主体模拟。通过设定家长、学生、教师、校长、监管者、用人单位等角色，并让其在共同的情景与约束下进行多轮对话与策略更新，研究者可以观察利益相关者的响应曲线、博弈路径与可能的规避行为，贴合教育治理中多元主体互动的真实面貌。与此同时，大语言模型还具备把隐含机制显性化的优势。诸如动机迁移、文化资本再生产、学术诚信的脆弱环节等，往往潜伏于叙事背后。大语言模型可将这些机制外化为流程图、决策树、风险清单与对照情景，使理论的内在一致性与边界条件得以被检视、被对比、被“压力测试”。功能上的这种三重契合，提供了思想实验从“观念推演”迈向“设计预演”的桥梁。

在方法层面，必须准确安放大语言模型的证据地位与与其他研究范式的互补关系。思想实验本就属于理论—设计之前的“假说生成器”，其任务是提出可检验的机制链、识别关键变量与潜在失败模式，而非直接产出因果结论。大语言模型由此适合承担先验预演、情景构造与方案筛选的工作，帮助研究与决策从一开始就更具前瞻性与可操作性，却不应被误用为因果识别的“替代品”。在研究流

程上，它与行动研究、设计型研究、随机/准实验应构成“设想—设计—验证—修订”的闭环：以大语言模型加速设想与原型，依托小规模试点与现场数据进行验证，再将偏差与洞见回灌模型设定与理论表达。这样的方法分工，既能降低改革与创新的前期试错成本，也能避免“语言流畅即真”的认识论风险，使思想实验产出获得实证与实践的锚点。

在技术层面，可行性取决于语言知识、工具增强与风险可控三方面。其一，得益于广域语料的训练，大语言模型对政策语境、课堂语言与文化差异具备基本拟真能力，能覆盖从法规话语到师生互动的多种文体。但这种“面向常识的广度”并不等于对本地细节的精准，因而需要与领域资料与本地数据对齐。其二，工具增强是把生成力转化为研究力的关键。通过检索增强生成（RAG）引入权威文件、校情数据与最新文献，可减少幻觉并提升情境针对性；通过结构化输出与规则约束，能让生成结果以表格、流程、指标的形式稳定复用；通过多模型交叉验证与提示框架的对照复跑，可识别“模型共同偏好”导致的假稳态，从而提高可重复性与稳健性。其三，局限可管控是所有应用的底线。需要以外部数据对齐、提示工程与人机共校的组合来缓解幻觉与偏见风险：前者提供事实锚点，中者提升任务对齐与边界控制，后者引入教师、研究者与弱势群体代表的参与式审阅，校正结构性偏差并显性化不确定性。同时，必须实现过程可审计：记录提示、版本与数据来源，明确伦理边界（隐私最小化、敏感场景红队输出的发布管控、申诉与复核机制），使模型在教育治理体系中“可解释、可追责、可修正”。

因此，“功能—方法—技术”的三重对齐为大语言模型辅助的教育学思想实验奠定了可行性：在功能上，它把抽象命题落为情景、把单主体推理扩展为多主体博弈、把隐性机制转化为可视结构；在方法上，它被定位为假说生成与方案预演的前置工具，与实证与试

点构成闭环；在技术上，则通过检索增强、结构化与规则化、交叉验证与人机共校，建立起从生成到治理的可控链条。只有在这三重对齐的前提下，思想实验才不会沦为“优雅的空想”，而能成为“可验证的想象”：既以低成本探索不确定性，也以制度化的风险控制守住学术与伦理的底线，最终把教育改革的设想转化为可测试的设计，把理论命题推向可观测的中程模型。

10.2.2 典型的应用场景

将大语言模型应用于教育学思想实验，可以从教育政策制定、教育理论发展与教育现象分析等方面展开。这三者分别对应“治理预演”“命题澄清”“研究设计前置”的不同面向，但在方法论上却有若干关键共同点：基于理论与政策的情景构造、多主体互动与红队化审视、结构化输出与可追溯治理、以及明确的适用边界与伦理约束。借助大语言模型，思想实验可以从“想象”走向“可验证的想象”，在不替代实证与试点评估的前提下，显著提高前期研判与方案成型的质量与速度。

（1）教育政策制定

在高不确定与多主体博弈的政策场域，大语言模型适合承担“情景预演—行为预测—配套机制生成”的链式任务。其核心是将改革（干预）目标与约束条件转化为若干互斥且充分的政策情景，让模型扮演家长、学生、教师、校长、监管者与用人单位等角色，围绕同一项政策与地区数据进行多轮互动，观察利益陈述、策略调整与规避路径的出现顺序与强度。以升学评价改革、“双减”或“AI 进校”为例，模型可在“过程性评价权重”“作业负担约束”“AI 工具可用性与合规要求”等政策工具上过行场景分析，并在模拟中提前暴露“影子教育迁移”“作业外包与学术诚信新风险”“资源不均衡被放大”等潜在副作用通道。基于这些预警，政策制定者可反向推导配套制度。

人工智能可以发现传统方法可能会忽视的趋势[1]。例如，它可以精准定位学生困难的领域，帮助政策制定者创建更好的支持项目[2]，突出有效的教学方法并指导教师培训[3]。

人工智能还可以预测未来的学生人数和毕业率，从而帮助资源规划[4]。政策可行性的另一支柱是公平性与偏差预判。借助合成的叙事实例与对话脚本，模型可针对少数群体、方言使用者、资源薄弱地区、残障与学习困难学生等，开展事前的偏差扫描，形成具可操作性的公平评估清单：识别在哪些环节可能出现系统性不利，哪类证明材料或过程证据对弱势学生构成过度负担，以及哪些补偿机制更具可及性与可接受性。这类“公平前置”的思想实验并非替代真实的平等影响评估，而是把评估重点与数据需求提前“显性化”，提高后续实证评估的效率与针对性。

最近的研究强调了人工智能如何通过将人员配置决策与包容性教育的法律要求相结合，支持公平的教师分配，尤其是在资源匮乏的社区[5]。这种前瞻性有助于调整政策以应对未来的挑战[6]。

在政策执行与社会沟通层面，风险与舆情沙盘可以提升治理韧性。通过红队/蓝队的对抗式演练，让一方提出可预见的规避与误用路径，另一方迭代防护与合规方案，可生成从学校层级到地市层级的操作清单与沟通，覆盖考试诚信、数据治理、家校沟通与舆情澄清等关键节点。需要强调的是，这类演练应以防护为目的、以治理为导向，避免引发不良模仿。

[1] Khan N, Hollingworth L. Breaking the Silence: Unveiling Barriers to Women's Leadership for Sustainable Development in Higher Education: Breaking the Silence[J]. JISR management and social sciences & economics, 2024, 22(2): 1-23.

[2] Akavova A, Temirkhanova Z, Lorsanova Z. Adaptive learning and artificial intelligence in the educational space[C]//E3S web of conferences. EDP Sciences, 2023, 451: 06011.

[3] Eden C A, Chisom O N, Adeniyi I S. Integrating AI in education: Opportunities, challenges, and ethical considerations[J]. Magna Scientia Advanced Research and Reviews, 2024, 10(2): 006-013.

[4] Ou S. Transforming education: The evolving role of artificial intelligence in the students academic performance[J]. International Journal of Education and Humanities, 2024, 13(2): 163-173.

[5] Langeveldt D C, Pietersen D. Decolonising AI: A critical approach to education and social justice[J]. Interdisciplinary Journal of Education Research, 2024, 6(s1): 1-9.

[6] Al Samman A M. Harnessing potential: Meta-analysis of AI integration in higher education[C]//2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). IEEE, 2024: 1-7.

为保障政策思想实验的有效性与稳健性，需要在技术上引入检索增强以对齐本地法规、统计年鉴与校情数据；在输出形态上采用结构化模板（指标、流程、责任人、时间表），形成可比与可复用的版本；在治理上记录提示、数据来源与修订历史，保证可审计与可追责。其适用边界也应清晰，大语言模型适合改革前期与试点策划阶段的预演与配套设计，不适合以生成内容替代实地调研、利益相关方访谈或因果评估；最终决策仍须以实测数据、成本—效果分析与伦理审查为锚。

（2）教育理论的发展

教育理论的发展既需要概念创新，也需要把抽象命题转译为可观测、可证伪的具体行动路径。大语言模型在此的独特作用，是将经典思想实验在当代约束与多元价值下对边界条件实施压力测试。仍然以《理想国》的“课程塑形”、卢梭《爱弥儿》的“消极教育”、杜威“学校即社会”为例，大语言模型可把这些命题转写为制度脚本与课堂微场景：在性别平权、学习者多样性、技术介入与数据治理等现代条件下，生成不同配置的课程路径、互动规则与评价制度，然后通过红队视角引入权力不对称、算法中介、资源分层与文化资本差异等扰动项，观察理论主张的稳定性、脆弱点与需要修订之处。类似地，弗莱雷的问题式教育可在“对话即解放”的前提下，纳入平台算法与课堂评价权力结构的影响，检验“对话能否抵抗结构性不平等”的条件约束。

大语言模型还能系统化地生成反例与边界，帮助理论从“宏大命题”走向“可证伪假说”。研究者可先以理论核心机制为轴（如探究促深度学习、评价导向影响动机、公共性提升公民性），让模型产出极端情境与反驳案例，继而筛选出最具鉴别力的关键试验条件。在此基础上，模型可进一步将抽象原则落地为教学—制度设计假说：明确流程、学习任务、支架与撤除节奏、反馈路径、评价指

标与证据样本，形成后续实证与试点可直接采用的可观测变量与路径图。这种“从理念到脚本”的转译，使理论验证不再停留于概念争辩，而能进入可重复的设计型研究与小规模实验。类似做法已经在循证实践的证据转换中得到实践与应用[1]。

不过，理论适用性也有明确边界。首先，大语言模型语料中的统计常态不等于“常识真理”，对边缘经验与少数视角可能天然不足，必须通过专家校核与社区代表参与式审查矫正。其次，大语言模型流畅的叙述可能掩盖因果结构的含混，研究者需以识别变量、明确干预点与对照条件的方式提升可证伪性。再次，所有由大语言模型生成的“理论脚本”都应进入小范围教学或制度试点，通过学习成效、参与公平、心理安全、教师负担等多维指标进行验证与修订，形成“理论—脚本—试点—反馈”的闭环。

（3）教育现象分析

在复杂教育现象的研究中，大语言模型可将“描述性问题”前移为“可检验的机制假设”，为质性与量化研究同时提供先验框架。以学习动机缺失、课堂沉默、作弊迁移为例，研究者可让模型基于既有文献与本地材料，生成多条机制链的候选集（如家庭期望—自我效能—评价压力的耦合路径等），并为每条链标注可观测指标、可能的干预点与预期副作用。这些机制草图既可作为质性编码的初始范畴体系，也可为量化模型提供变量选择与结构设定的线索，缩短从探索到验证的时间。

在测评与算法应用场景，模型适合承担偏差审视与压力测试。通过构造覆盖方言、少数文化叙事、非标准表达与不同社会语言变体的合成语料，对自动评分与推荐系统进行对比评测，识别系统性不利与误判模式，再结合少量人工标注与提示规约修订评分准则与解释模板，从部署前就形成偏差校正路径。这一过程强调社区参与

[1] 陈向东, 褚乐阳, 陈鹏著. 教师的循证实践：基于 AI 大模型的方法[M]. 上海：华东师范大学出版社, 2025.

和伦理审查，避免合成数据内嵌刻板印象，并确保测评系统保留人工复核与申诉通道。

对于早期预警与个案干预，模型可在脱敏与伦理合规的前提下，扩写零散的结构化数据与访谈片段，使学业困难、辍学风险等个案的学习—家庭—情绪—工作多维线索得到串联，从而帮助学校与社工团队设计更人本的干预组合与资源排期。需要谨慎的是，合成数据的叙事只应作为干预设计的“灵感与清单”，不能替代对关键变量的实测与持续跟踪；任何敏感内容的生成与使用都需最小化个人可识别信息，并明示当事人的权利。

在方法操作上，现象分析的思想实验同样需要技术与治理的双重保障。技术上，使用检索增强与本地知识库的对齐，采用结构化输出（变量表、因果图、指标对照表）提升复用性，并且采用多模型与多提示框架。治理上，记录版本与修订理由，确保研究团队与利益相关方能够追溯并讨论关键决策。其适用边界在于：模型适合作为研究设计的前置环节，不应用合成数据替代核心实测数据，更不能以生成的相关性叙事替代因果识别。任何面向政策或高风险实践的结论，都应在小样本试点和多方法三角验证后方可推广。

由此可见，经典教育学以思想实验搭桥理念与实践，大语言模型则将这类“想象力政治”推进为“可验证的想象”，使思想实验产出的设计假说能进入验证轨道。其合理定位是理论与实证之间的中介层、政策与课堂之间的预演层：为教育预测与预见提供低风险的前置检验，为教育理论提供可落地的实践路径，为教育现象分析提供先验的验证框架。

自柏拉图的“理想国”至卢梭的“爱弥儿”，思想实验始终是驱动教育理论演进的重要引擎，它在人类心智的“离线”空间中构建并检验着关于人性与社会未来的宏大叙事。然而，这种依赖于个体心智的传统范式，因其固有的静态性与理想化局限，长期在理论

的严谨性与实践的转化效率上面临挑战。大语言模型的出现，正将这一古老的智力传统从静态的哲学思辨，推向动态的、可计算的科学探索。

基于大语言模型的教育思想实验，其核心学术价值在于构建了一个“计算性社会实验室”：它能够尝试将杜威式的民主教育设想，置于当代复杂的数字社会结构与信息茧房的约束下进行模拟；它也可以探索弗莱雷的解放教育理论，将其转化为可计算的社会动力学模型，观察“对话”与“压迫”在不同权力结构下的演化轨迹与临界点。这种方法系统性地连接宏大理论与经验现实，通过生成可观测、可证伪的理论假设，使得教育研究得以在实施成本高昂、伦理风险巨大的真实世界干预之前，进行高效、低风险的场景分析与风险模拟。

展望未来，基于大语言模型的教育思想实验不仅可以成为加速教育理论迭代与政策方案优化的新工具，更深刻地，它将可能塑造一种全新的、融汇了人文思辨与数据科学的认知视角，桥接“教育应该如何”与“教育可能如何”，从而在一个复杂性与不确定性日益增加的世界里，为未来学习设计提供更具鲁棒性与前瞻性的理论基石。