DOI:10.16382/j.cnki.1000-5560.2026.08.002

超越图灵测试: 全球首个“AI 一作”大型社会实验全景报告*

张 治[1,4] 陈霜叶[2] 肖 敏[3] 刘一萌[4] 张 凯[4] 季 程[5]

(1. 教育部哲学社会科学实验室华东师范大学智能教育实验室, 上海 200062;
2. 华东师范大学教育学部, 上海 200062; 3. 上海市未来学习研究与发展中心, 上海 201999;
4. 华东师范大学上海智能教育研究院, 上海 200062;
5. 南京伯索网络科技有限公司, 南京 210006)

摘 要: 2025 年 9 月, 华东师范大学发起全球首个“AI 一作”大型社会实验, 以准实地实验方法开展“AI 驱动教育研究论文写作”征文, 规定 AI 为第一作者、人类为协作与把关角色, 历时半年收到海内外有效投稿 724 篇, 在此基础上探索 AI 赋能哲学社会科学研究第五范式、学术伦理规范及教育学自主知识体系构建路径。实验历经干预设计、反响监测、专家研讨、数据采集、AI 审稿、人机一致性检验、数据分析与成果发布八大阶段, 采用混合研究方法揭示核心发现: AI 在灵感激发、信息处理、文本润色等环节优势显著, 但存在文献虚构、逻辑空心、创新不足、伦理隐患等局限; 高效 AI 应用可显著提升学术贡献, 形成六种创新样态与五种人机协同模式; AI 审稿具备可靠性, 与人类专家评价存在互补性, 不同大语言模型科研品味差异明显; 学术界呈现显著“AI 代沟”, 青年群体更适配人机协同, AI 一定程度上推动了智慧平权。研究提出未来需推动科研范式转型、评价体系改革、科研垂类智能体研发、育人模式重构、文凭认证革新等, 为智能时代学术创新与教育变革提供实证支撑与实践启示。

关键词: AI 一作; 社会实验; 教育科研; 人机协同; 科研范式; 学术伦理; AI 审稿; 智慧平权

引 用: 张治, 陈霜叶, 肖敏, 刘一萌, 张凯, 季程. (2026). 超越图灵测试: 全球首个“AI 一作”大型社会实验全景报告. *华东师范大学学报(教育科学版)*, 44(8), 8—50.

一、实验缘起

进入 21 世纪第三个十年, 以大数据与人工智能深度融合为特征的“科学研究第五范式”正从自然科学领域向哲学社会科学领域全面渗透。2024 年诺贝尔物理学奖与化学奖均授予 AI 领域的科学家, 这一标志性事件昭示着 AI 已从辅助工具演进为科学发现的核心驱动力量。2025 年 7 月, 国际顶尖学府斯坦福大学宣布将于年内举办科学 AI 智能体的公开会议(Agents4Science, 2025), 明确提出“AI 作为论文的主要作者和审稿人”的要求, 将 AI 在科研中的主体性地位从构想推向实践前沿, 这一举措在国际学术界引发广泛关注与讨论。

随着人工智能逐步嵌入科学研究的核心环节, 科学研究第五范式展现出强大力量, 哲学社会科学研究也面临着前所未有的挑战与机遇。一方面, 传统以人类研究者为主导的质性思辨与量化研究范式, 在数据处理的广度、模式发现的效率以及跨学科知识整合层面, 遭遇技术性瓶颈。另一方面, 以大语言模型为代表的生成式人工智能技术迅猛发展, 其文本生成、逻辑推演与知识合成能力, 已深度渗透至学术生产的各个环节。调研显示, 硕博论文的 AI 渗透率已超过 70%(袁曼舒, 2025), 这一场“静默的革命”使得学术界无法再对 AI 的广泛使用视而不见。在此过程中, 关于“AI 能否胜任哲学社会科学

* 基金项目: 科技部国家重点研发计划“青少年身心成长促进智能支持技术与应用示范”(2023YFC3305805)。

研究的特殊性”“AI能否实现真正意义上的学术创新”“AI生成内容的准确性与可靠性如何保障”“在学术写作中使用AI的合法边界何在”“AI能否用于论文评审”“AI加剧还是缓解了学术不平等”等问题的讨论沸沸扬扬、莫衷一是。如何构建与智能时代相适应的学术研究新伦理、新规范与新方法体系，如何变革现有的育人体系和人才评价体系，成为亟待回答的时代命题。

然而，正如斯坦福“Agents4Science 2025”会议主办方所言：“尽管AI越来越多地参与科学探究的每个阶段，但几乎所有的期刊和会议都禁止承认AI是作者，现有的规范鼓励研究人员将AI的贡献隐藏起来或最小化，这些禁令阻碍了我们理解和塑造AI参与未来科学研究的能力”（Agents4Science, 2025）。面对在AI冲击下已然迭代升级的游戏本身，以及改进速度似乎总是慢半拍的游戏规则，国内外学术界都站在一个关键的十字路口——是保守观望，还是开放求索？是被动适应，还是主动建构？是沿用成熟的旧框架去约束新生的力量，还是勇敢地重塑规则，让技术与伦理、创新与问责在碰撞中同步演进？这决定着我们将以何种姿态，与AI共同走进下一个发现与创造的时代。

在这个背景下，华东师范大学选择了一种积极、理性且富有建设性的态度，即“与其任其发展，不如正面应对”。早在2018年，学校就部署了智能教育领域的研究工作，并在上海市教委的支持下，于2020年12月成立上海智能教育研究院。2021年12月，“智能教育实验室”获批教育部首批哲学社会科学实验室，旨在服务国家战略和区域发展，瞄准学术前沿，推进学科交叉融合，创新研究范式和方法，充分利用先进智能技术和实验手段，开展战略性、前瞻性、实践性研究。2025年9月，华东师范大学发起全球首个“AI一作”大型社会实验，华东师范大学教育学部、上海智能教育研究院和教育部哲学社会科学实验室华东师大智能教育实验室联合举办了以人工智能作为科研论文写作主体的“AI驱动教育研究论文写作”征文活动，旨在构建一个可控且低风险的环境，开展将AI深度应用于教育研究与写作的社会实验，从而实现以下核心目标：

一是探索AI赋能哲学社会科学研究第五范式发展的实效与边界。活动强制要求AI作为第一作者，旨在倒逼研究者系统考察AI在假设生成、研究设计、数据分析、论文撰写等科研环节中的实际效能与能力边界。活动中人类角色被建议为任务的发起者、工具的选择者、指令的设计者、质量的评价者及成果的整合者，以此引导研究者重新审视并厘清人工智能与人类智能在科研全流程中的合理分工，探讨在AI深度嵌入后，人类研究者的不可替代性价值究竟何在。

二是探讨教育科研中人工智能应用的学术规范与伦理问题。针对学术界广泛存在的AI使用“暗箱”与署名争议，活动通过开放性社会实验及跨学科专题研讨，引导学术界思考科研范式变革及学术规范革新的必要性。同时，通过要求作者对AI生成部分进行明确标识并提交《人工智能生成内容说明表》（以下简称“过程说明”），倡导一种负责任、可追溯的AI使用伦理，并为处理AI生成内容的版权归属、学术诚信认定以及成果评价标准等争议性问题提供实证案例与讨论基础。

三是探寻人工智能助力教育学自主知识体系构建的路径与机制。活动主题聚焦“未来教育研究”，探讨AI赋能的未来教育新生态，特别是AI赋能的未来学习、未来课堂、未来教师、未来学校和未来教育科研，旨在以征文形式激发前瞻教育思考，进一步明晰未来教育生态与人才培养模式的变革方向，为推动我国教育学自主知识体系的建设与发展提供有益的启示。

二、过程与方法

社会实验是在真实社会情境中通过系统干预获取因果推论的循证研究范式，典型的实验方法包括自然实验、实地实验、调查实验和计算实验等（苏竣，黄萃，2023，第18页）。本次活动采用一种真实场景下的准实地实验方法[①]，即研究者主动构造干预并在真实情境中实施，让被试在真实行为场域中做出自主反应，以此来实现因果推断的内部效度与现实情境的外部效度。具体而言，活动组委会在真实学术场域主动创设“AI一作”的署名规则，面向海内外征集AI第一作者的教育研究论文，观测参与者行为反应与学术产出，并嵌入内容分析、焦点小组访谈、文本挖掘等多种研究方法。实验整体遵循探索性序列混合方法设计，从2025年9月底至2026年3月，历时约半年，涵盖干预设计、社会反响监测、焦点小组研讨、征文数据采集、AI审稿系统研发、人机审稿一致性检验、数据分析与总结性评估八个关键阶段，如图1所示。

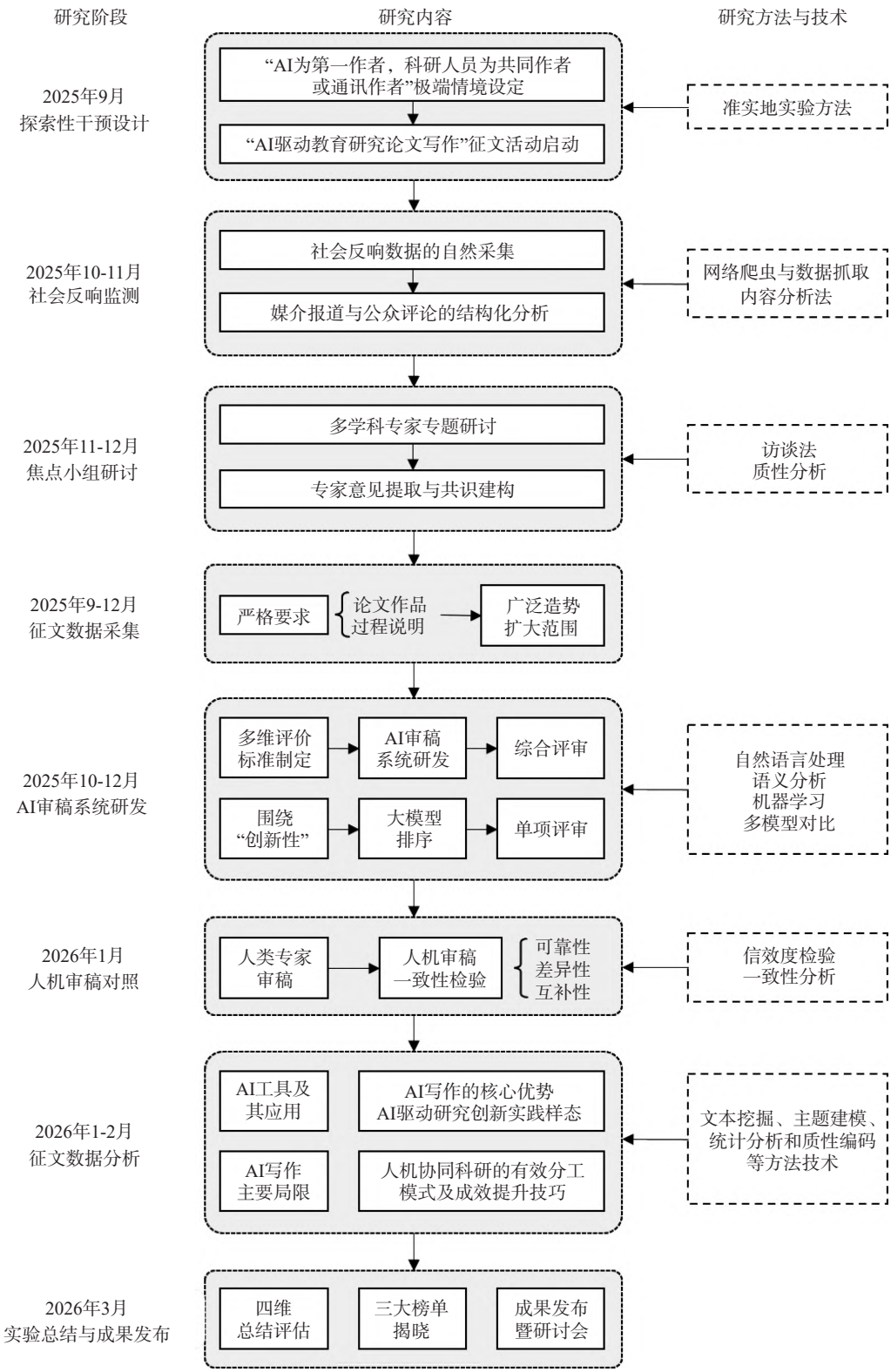

图 1　“AI 一作”大型社会实验的总体思路

具体研究过程如下：

（一）探索性干预设计：发起实验，投石问路

实验的第一步是依据基于设计的研究范式，确定核心干预的内容。基于设计的研究强调在真实情

境中分析具有普遍性的实践问题，通过设计系统性干预并在迭代循环中检验其效应（Wang & Hannafin, 2005）。本研究将“AI驱动教育研究论文写作”征文活动作为核心干预设计，其创新性在于突破传统学术署名规范，将人工智能确立为论文的第一作者，人类研究者则承担任务发起者、工具选择者、指令设计者、质量评价者与成果整合者的角色。2025年9月30日，华东师范大学教育学部、上海智能教育研究院及教育部哲学社会科学实验室华东师大智能教育实验室正式发布“AI一作”征文通知，面向海内外征集以AI为写作主体的教育研究论文。征文活动的设计遵循三项基本原则：一是主体性转换原则，明确要求AI在假设生成、研究设计、数据分析与论文撰写等核心环节发挥主导作用；二是过程透明原则，要求投稿者提交“论文作品”与“过程说明”两份材料，详细记录使用的生成式人工智能工具名称及版本、应用思路与步骤、生成内容的具体位置及占比、人工修正与补充说明等信息；三是规范适配原则，征文主题聚焦未来教育研究，论文格式参照《华东师范大学学报（教育科学版）》投稿要求，确保实验成果与传统学术评价体系具有可比性。上述设计并非简单的技术尝试，而是对科研主体性、学术责任归属与知识生产范式的系统性挑战。通过“AI为第一作者，科研人员为共同作者或通信作者”的极端情境设定，创设了一个充满认知冲突的实验情境，从而激发并捕获社会各界对AI应用于哲学社会科学研究的真实想法与行动。

（二）社会反响监测：静观其变，动态分析

征文活动发布之后，实验进入社会反响的系统性监测与分析阶段。首先是社会反响数据的自然采集。活动组委会依托网络爬虫与数据抓取技术，对2025年9月30日至11月30日期间与“AI一作”活动相关的媒体报道、学术论文、自媒体评论及社交平台讨论进行了全面且动态化的采集，数据来源涵盖主流媒体、教育专业媒体、科技媒体及国内外社交平台等多渠道信息。

完成信息采集后，采用内容分析法对各类媒介报道与公众评论进行结构化分析。分析框架从四个维度展开：一是传播主体维度，区分官方机构、学术界、媒体界、产业界及公众等发声主体；二是议题聚焦维度，识别出法律责任、学术规范、学术伦理、技术可行性、教育变革等核心议题；三是态度倾向维度，区分积极支持、反对批判和审慎推进等主要立场；四是传播路径维度，追踪信息从官方发布到多级媒体的扩散轨迹与关键节点。

通过高频词统计、主题聚类与情感倾向等分析，活动组委会系统呈现了社会反响的整体图景。总的来看，征文公告发布后迅速形成多层次、跨领域的舆论场域，议题从初期的“AI能否作为第一作者”逐步深化至“人机协同的知识生产范式重构”“学术评价制度的适应性变革”等深层命题。这些发现为后续专家研讨及实验迭代设计提供了实证基础，也揭示了“AI一作”大型社会实验作为“投石问路”策略的有效性——通过设计性的干预触发真实社会反馈，进而观测技术变革引发的一系列社会现象。

（三）焦点小组研讨：激辩思想，汇集智慧

在社会反响分析的基础上，实验进入多元利益相关者的深度对话与共识建构阶段。此阶段运用焦点团体访谈的方法，通过专题研讨的形式，利用参与者之间的互动效应激发深层观点，并通过观点之间的碰撞与融合，逐步收敛专家意见，达成对核心议题相对一致的集体判断。活动组委会在征文期间及论文评审结束后，组织了多场专题研讨会，参与者涵盖教育学、计算机科学、法学、伦理学、出版学等多个学科的专家学者，以及一线教育工作者、研究生、本科生乃至基础教育阶段的学生代表。研讨会设置开放式议题引导讨论，包括“AI一作法律责任归属”“人机协同科研的伦理边界”“学术评价的改革方向”“未来教育科研的范式转型”等核心议题。讨论采用半结构化方式，既确保覆盖核心议题，又允许参与者围绕新话题自由展开。

通过录音转录、编码分析与主题提炼，活动组委会从一系列焦点小组研讨中提取了三类关键成果：一是对实验方案本身的反思性改进建议，如扩大征文对象范围、增设基于创新性的单项评审、调整过程说明的呈现内容等；二是对AI驱动科研范式的理论性洞察，如知识生产的“人机数智融合范式”，

AI 驱动的创新本质等(华东师范大学上海智能教育研究院, 2025); 三是对未来教育制度变革的前瞻性判断, 包括学术评价从“结果导向”向“过程导向”的转型、学位认证与能力徽章的并行体系等。这些研讨成果为后续的评审机制优化和成果凝练提供了有益的参考。

(四)征文数据采集: 广泛造势, 严格要求

本次大型社会实验的核心数据采集主要通过“AI 驱动教育研究论文写作”征文活动来完成。征文对象最初为人工智能和教育学等相关专业的各级各类科研人员及在读研究生、本科生, 后经专家委员会建议, 扩大至哲学、心理学、脑科学、经济学、计算机科学等不同学科背景的作者, 可采用个人或团队形式参与(团队人数不超过 3 人)。为保证征集到的论文具有可对比性, 活动组委会将征文主题聚焦到“未来教育研究”这一领域, 重在探讨 AI 赋能的未来教育新生态, 特别是 AI 赋能的未来学习、未来课堂、未来教师、未来学校和未来教育科研五大分支。参与者需提交“论文作品”和“过程说明”两份材料。其中, 论文作品应采用中文撰写, 篇幅约 10 000 字, 并依据《人工智能生成合成内容标识办法》要求, 在论文中对 AI 生成的所有部分进行标识; 过程说明要求作者在《人工智能生成内容说明表》中详述使用的生成式人工智能工具名称及版本、应用思路与步骤、生成内容在论文中的具体位置及占比、人工修正与补充说明、与大模型的对话链接或使用的提示词及模型回答等信息, 从而详细说明指导和监督 AI 进行研究与创作的过程。最终, 活动共收到来自海内外投稿 820 篇, 经剔除重复投稿、主动撤稿及测试稿件后, 有效投稿共计 724 篇。通过上述“结果+过程”的双重材料设计, 活动组委会为后续的人机协同评审、文本挖掘与过程分析等方面采集了丰富的实证数据。

(五)AI 审稿系统研发: 创新机制, 考察模型

针对此次论文评审的特殊情境: 稿件数量多和“AI 一作”的论文特性, 活动组委会联合相关研究机构及教育科技企业, 研发了 AI 审稿系统, 探索 AI 作为论文评审主体的可行性。由于此次活动采用“人机协同、双轨并行”的论文评审机制, 因此 AI 既参与了“基于多维评价指标的人机协同综合评审”(以下简称“综合榜”), 又参与了“基于创新性指标的大模型单项评审”(以下简称“单项榜”)(袁曼舒, 肖敏, 2025)。

在“综合榜”评审中, 活动组委会通过多轮专家咨询与论文试评, 制定了由《伦理规范审查标准》和《论文作品评审标准 3.0 版》两部分构成的论文评审标准体系。其中, 前者包括论文重复率、内容及文献真实性和公平公正性三个指标, 后者则由研究逻辑、方法过程、研究创新和写作质量四个一级维度和研究问题明确、理论基础适切等十个二级指标构成。在此基础上, 组委会研发了用于“综合榜”评审的 AI 审稿系统, 系统核心算法整合自然语言处理、语义分析与机器学习等技术, 并通过“构建—测试—迭代”的严谨流程, 确保 AI 审稿结果的可靠性。最终, 系统为所有通过伦理审查的论文提供了每个评价维度及论文整体的评分与评语, 共计 7 000 余条审稿意见。这些数据不仅是论文评审工作的结果, 同时也是后续系列深入研究的重要数据来源。

为深入挖掘 AI 赋能哲学社会科学研究的独特价值, 活动组委会还增设了“单项榜”的评审, 运用国内外六个先进大模型, 聚焦“创新性”这一核心指标, 对所有通过伦理规范审查的论文进行逐一比对和排序, 形成论文排行榜, 结合多模型榜单与专家委员会意见, 评选出最具前瞻性和突破性的创新论文。这种多个模型共同参与“单项榜”评审的目的在于: 一是通过模型间的相互验证, 识别出真正具有前瞻性与突破性的研究成果, 降低单一模型的偏差风险; 二是通过对比不同大模型在评估视角和权重分配等方面的区别, 探讨不同大模型的科研品味有何差异, 为后续技术应用、研究开展和模型迭代提供建议。

(六)人机审稿一致性检验: 对照结果, 探析差异

为客观评估 AI 审稿系统在学术论文评审中的实际效能及其与人类专家的协同关系, 本次实验在 AI 评审的基础上, 引入人类专家评审机制, 开展人机协同的综合评审。专家委员会由教育学、人工智能等相关领域的国内外知名专家组成, 以确保评审的专业性和权威性。评审专家对通过 AI 初审的论文作品进行复审, 对论文的研究逻辑、方法过程、研究创新和写作质量进行打分, 撰写审稿意见。为确

保人机对比的科学性，评审过程严格遵循三项控制原则：一是时间隔离原则，AI评审与人工评审在时间上错开，避免评审信息交叉污染；二是标准统一原则，人机评审使用同一套维度框架与评分标准，确保结果具有可比性；三是双盲设计原则，人类专家在评分时互相不知道对方给出的分数，避免锚定效应对独立判断的干扰。在完成人类专家评审后，活动组委会对人机评审结果的一致性进行检验，并对人类评审结果的一致性进行分析。通过这一阶段的研究，实验对AI审稿的可靠性、AI与人类专家在评审逻辑、关注维度和价值标准上的差异与互补性进行了探讨，为后续优化学术论文评审流程提供了实证依据。

（七）征文数据分析：混合方法，深度探究

在完成论文评审之后，活动组委会对724篇有效投稿论文、717份有效过程说明[③]、25份“上榜论文”作者提交的经验反思报告进行深度数据分析。此阶段综合运用文本挖掘、主题建模、统计分析和质性编码等方法技术，从大规模文本中提取重要的实验发现。具体而言，一是文本挖掘方面，利用Python脚本调用python-docx与pdfplumber库扫描文档，通过正则表达式清洗数据并建立TOOL_MAPPING词库进行品牌归一化，采用Counter算法进行词频统计，最终提取出所有被用于论文创作的AI工具及其应用数据。二是主题建模方面，参考学术全生命周期模型预设15个科研与写作维度，经50份样本扎根理论预研提炼高频动词后，利用关键词匹配算法扫描非结构化文本，采用多标签归类法计算各维度渗透率，进而提炼AI赋能论文写作的核心优势和AI驱动研究创新的实践样态。三是统计分析方面，依据研究逻辑、方法过程、研究创新、写作质量四大评审维度计算得分率与分布形态，同时采用语义内容分析法对7 000余条审稿意见进行关键词提取与语义编码，计算负面特征检出率，进而分析AI用于科研写作的主要局限。四是质性编码方面，邀请25位“上榜论文”作者撰写深度参与活动后的经验反思报告，抽丝剥茧地还原论文“从0—1”的“人机共创”真实图景，进而运用扎根理论的编码技术，围绕“哲学社会科学研究中的最佳人机协同策略是什么”这一核心问题开展质性编码，最终提炼出5种人机协同科研的有效分工模式以及8个提升人机协同科研成效的技巧。

（八）实验总结与成果发布：提炼发现，公开研讨

实验的最后一步是总结性评估与范式提炼，旨在对整个“AI一作”大型社会实验的过程与结果进行系统性反思，提炼具有普遍意义的理论洞见与实践启示。总结性评估包括四个方面：一是过程评估，检视实验各环节的设计合理性、执行合规性与流程顺畅性；二是产出评估，对724篇有效投稿论文的质量分布、主题特征与创新水平进行分析；三是影响评估，探讨实验对学术界、教育界与公众认知的深层影响；四是伦理评估，审视实验过程中涉及的学术规范、知识产权与数据安全等伦理议题。

在总结性评估的基础上，活动组委会于2026年3月21日召开“AI一作”大型社会实验成果发布暨研讨会，正式发布《超越图灵测试——全球首个“AI一作”大型社会实验全景报告》。研讨会由华东师范大学与全国高校哲学社会科学实验室联盟联合主办，设置主旨报告、圆桌对话和四个平行分论坛，围绕“从辅助到重构：AI写作中的创意主权与表达边界”“从契约到共治：人机协作中的伦理重构与责任分担”“从协作到共生：通用人工智能与人类智慧的协同演进”“从勘误到共治：生成式AI嵌入下的出版质控困境与审校范式重构”等社会公众关注的议题展开深度研讨。

值得注意的是，基于对科研标准的严谨坚守，活动组委会作出了审慎决定：将原计划的“论文评奖”改为“发布榜单”。专家委员会在评审中发现，现阶段人机协同的成果暂未完全达到传统学术“奖项”的高度与成熟度，采用“发布榜单”的形式更能客观、科学地反映当前人机共创的真实探索水平。最终，实验成果发布暨研讨会现场正式发布了人机协同评审出的三大榜单——“人机共创AI推荐新锐论文榜”共有17篇论文入围，它们均由国内外6个主流大模型根据创新性进行排序，且每篇至少获得2个及以上大模型TOP10榜的推荐；“人机共创先锋论文提名榜”共有20篇入选；而最重磅的10篇“人机共创先锋论文榜”论文，则是人类专家与AI审稿系统根据评审标准综合评选的结晶（华东师范大学

上海智能教育研究院, 2026)。“发布榜单”这一决策本身即是对学术诚信与科研伦理的坚守, 也体现了社会实验作为“真实世界研究”的本质特征——不追求人为控制下的理想结果, 而致力于呈现真实情境中的真实状态。

最终, 本次实验正式探讨了人机协同科研的第五范式框架, 系统阐述了 AI 驱动教育研究范式转型的发轫逻辑、变革路径与实践进路, 并提炼出十大重要发现。这些实验成果, 不仅对“超越图灵测试之后, 人类将走向何方”等核心命题进行了正面回应, 也为哲学社会科学研究在智能时代的创新发展提供了方法论参照与制度性启示。

三、重要发现

根据上述八个阶段的实验过程, 针对以下十个社会各界共同关注的问题, 得到以下研究发现。

(一)学术界对 AI 写作的接受度如何?

“AI 一作”大型社会实验的征文活动, 自启动以来便引发了社会各界关于 AI 能否以及如何介入科研与写作的激烈讨论。从哲学思辨到伦理拷问, 从制度设计到教育实践, 这场讨论早已超出教育学界的范畴, 演变为对学术生产逻辑的根本性反思。在纷繁复杂的观点中, 可以看到“赞成派”“反对派”和“审慎推进派”三种主要立场。

1.赞成派: 以超前实验主动求解人机协同新范式

“AI 一作”征文活动的赞成者从多重维度肯定了这场探索的积极意义, 视其为主动求解人机协同新范式的超前实验。总的来看, 这一派的赞成理由可以归纳为五个方面:

一是可以引导学术伦理从隐性规避迈向透明规范。赞成派认为, “AI 一作”活动促使学术界正视 AI 已渗透科研的现实。当前研究者使用 AI 辅助研究时往往选择不披露, 导致责任边界模糊。斯坦福学者 James Zou 等人指出, 最大的风险并非 AI 本身, 而是研究者隐瞒 AI 使用的现象(Bianchi et al., 2026)。因此, 华东师大要求参与者提交《人工智能生成内容说明表》, 将 AI 参与转化为可追溯的研究步骤。赵勇指出应主动思考与适应新技术, 而非急于划定边界。有学者指出, 这种直面争议的态度标志着社会舆论从单一伦理焦虑向实践整合的转向。

二是以真实案例探讨人机协同科研的前瞻性。AI 参与科研具有生成性和偶然性, 人机协作边界须在实践中不断调整(游俊哲, 2023)。赞成派认为, 华东师大征文活动提供了人机协同的具象化实践场景, 使学术界得以积累第一手经验。“Agent4Sciences”项目主席 Mitchell 认为类似项目能为 AI 科研政策提供数据参考(李木子, 2025), 浙江大学吴飞建议要求作者附录创作历程, 以区分机器生成与人类判断。赵勇指出, 教育领域长期存在“假装使用技术”现象, 而此实验打破了惯性。参赛者徐韵表示, 在与 AI 深度协作后, “人的危机感反而慢慢没有了”(王倩, 李楚悦, 2026)。

三是追问“共进化”中的人机角色定位。袁振国指出, 人类与 AI 正处于“共同进化”之中(黄海华, 2025), 人类价值更多体现在问题提出与理论判断等环节。有学者认为, AI 虽尚不能理解人类经验, 但其“结构级智能”可作为人机共创的桥梁。北京大学马亮认为, 鉴于 AI 展现出一定的自主性, 可将其视为科研合作伙伴乃至署名作者(陈雅静, 刘越, 2025)。参赛者董辉与博士生张思梦等的实践表明, 人类研究者的重要性在提问、筛选和整合信息中更为突出, 且数据处理和代码调试的效率则在 AI 的辅助下获得显著提升(王倩, 李楚悦, 2026)。

四是重新审视“好研究”与科研范式转型。赞成派认为, AI 能够生成结构完整但缺乏实质创新的文本, 这恰恰为重审“好研究”标准提供了契机。陈志文指出, AI 使缺乏创新的研究原形毕露, 推动学术界进行必要“清理”(腾讯教育, 2025)。此次活动主张评选标准聚焦于 AI 是否“写得更好”而非“更像人”, 可视为对新评价维度的压力测试(唐政, 2025), 有学者将此次活动称为“华东起义”, 认为它是对科学研究第五范式的实质性探索。

五是可以借此反思面向未来的育人旨归。赞成派认为, AI 深度介入科研迫使教育界正视未来社

会对人的能力要求的深刻变化。这项活动被视为“以突破性实践倒逼学界重新审视教育本质”。赵勇主张教育须从“修补过去”转向“发明未来”(Rivero, 2025)。参赛者张琪的实例表明,与AI协作的效率提升并未消解人的主体性,研究者得以将精力集中于算法无法替代的判断环节。袁振国在世界人工智能大会首份刊物《WAIC UP!》专访中提出应将AI视为学习的“同行者”,在双向互动中共同成长。

2.反对派:警惕算法替代思考的迷思与学术主体性的退场

反对者的质疑从四个层面展开:

一是“人类向机器投降”的哲学与伦理危机。反对派指出,AI缺乏自由意志、自我意识与反思能力,将其列为第一作者等于在制度层面否定人在知识创造中的中心位置。浙江大学黄昌勤强调,人类创作根植于自我意识与具身体验,AI本质上是模式匹配式计算,缺乏内在动机与真实理解。李隽旸同样指出研究须基于人类具身体验,AI无法替代(陈雅静, 刘越, 2025)。尚杰认为,此举“是伦理错误,也是哲学错误”——在明知机器不具备自由意志的情况下,却鼓励用机器取代人类思考(尚杰, 2025)。

二是认知外包和创新精神的枯竭。反对派强调,学术研究的真正价值在于思考过程中获得的智力训练。写作与深度思考密不可分,放弃写作等于放弃深度思考。复旦大学肖仰华在闭门研讨会中指出,“认知外包”可能背离教育本源,若研究者沉溺于AI带来的认知结果,将对学术能力造成毁灭性打击。《教育发展研究》林岚提醒,过度依赖AI易导致思维惰性,年轻学者应避免制造“精致平庸”(华东师大基础教育改革与发展研究所, 2026)。

三是AI创造力的限度及其作为第一作者的资格问题。反对派指出,生成式AI本质上是基于概率统计的模式匹配,其产出局限在既有知识的封闭循环内,无法实现“从0到1”的原始创新。肖仰华从技术本质上剖析了AI“创造力”的虚幻性,指出其仅为统计模型的输出,缺乏意识、动机与精神内核。刘永谋指出,AI仅是对人类知识的重构,“思想”才是人文社科的关键组成部分(腾讯教育, 2025)。

四是学术责任主体的缺失和评价体系的失序。批评者强调,作者署名意味着对研究的法律与道德责任。AI无法理解输出内容、无法承担责任,因此不具备作者资格。包万平指出,“AI一作”动摇了现代学术体系的伦理基石。“作者”身份的神圣性,制造了“责任主体的真空”(包万平, 2025)。现行著作权法规定创作主体必须是人,AI无法成为权利人;若AI被纳入署名体系,学术评价将陷入混乱,我们是在评价人的水平,还是在评价模型的能力?

3.审慎推进派:在伦理与制度框架中重建人机协同秩序

审慎推进派认为,生成式AI参与知识生产已难以逆转,应在明确伦理原则与制度边界的前提下有序探索。具体而言:

一是确立学术研究中人机协同的伦理原则。审慎推进派认为,推进人机协同的首要前提是伦理原则的确立。学术研究不仅是信息处理过程,更是包含问题意识、价值判断与责任承担等方面的认知过程。人工智能应被界定为扩展人类认知能力的技术工具,而非独立学术主体。马亮提出了人类与AI双重自主性框架,认为人机关系不应简单理解为“主体—工具”的单向关系,而应视为互动性协同结构。须通过提升研究者数字素养来强化人类自主性,避免演变为技术对研究过程的替代(马亮, 2025)。复旦大学王金林则主张将知识生产从个体主体性模式转向基于生成式理性的“结构性协作”(王金林, 2025)。

二是构建人机协同的学术规范与治理框架。审慎推进派认为,须将伦理共识转化为具体、可执行的制度安排。应加快推出AI应用规范,构建以人类主导、透明可信、责任明晰为核心的学术伦理体系,由国家部委联合制定规范并鼓励期刊出台领域共识性指南(包万平, 2025)。需要建立覆盖知识生产全过程的清晰责任结构,由人类主体承担最终责任。同时需要推行透明披露制度,并重构学术评价体系(王金林, 2025)。还应从价值、工具与文化三个层面建立多维综合评价框架(胡志刚, 梁晗, 2025)。

三是彰显人机协同中人类研究者的核心价值。审慎推进派认为,人机协同的最终价值取决于研究

者的思考与成长。刘永谋指出,学生自主写论文旨在训练学术基本功,无法通过“外包给 AI”获得。华东师范大学卜玉华指出,教育学知识须关乎“真”与“善”,人的核心角色在于基于现实感知建构真问题(华东师大基础教育改革与发展研究所,2026)。于天贞认为,“AI 可以告诉你什么问题是热门的,但只有人类能判断什么问题是重要的”。包万平发问:当论文涉及价值争议时,谁来承担选择的后果? 仲立新强调须以解决人的真实发展问题为价值底线。《华东师范大学学报(哲学社会科学版)》周萍指出“幽灵引用”已成为学术不端重灾区,核实引文责任主要由研究者承担。陈志文认为在使用中“看见差距—理解原因—内化提升”,并在享受效率红利时保持批判性反思,这样才能实现自我成长(腾讯教育,2025)。

4.现实情况:一边争论,一边探索

无论社会各界如何争论,研究者们在论文写作中不同程度地使用 AI,已经成为既定事实。而本次征文活动所获得的社会反响,也印证了 AI 在学术论文写作中应用的普遍性。征文活动自 2025 年 9 月 30 日启动以来,在国外学术界引起广泛关注,投稿热情持续高涨。截至征稿结束,共收到来自海内外投稿 820 篇,经剔除重复投稿、主动撤稿及测试稿件后,最终有效投稿共计 724 篇。地域分布方面,上述有效稿件来源覆盖我国全部省级行政区,以及美国、英国、韩国等 9 个国家(地区)。从论文主题分布来看,如图 2 所示,“未来学习”赛道投稿数量最多,占比达 35.85%。此外,“未来教育科研”(23.71%)、“未来教师”(15.80%)、“未来课堂”(14.87%)、“未来学校”(2.89%)等主题均有相应投稿。

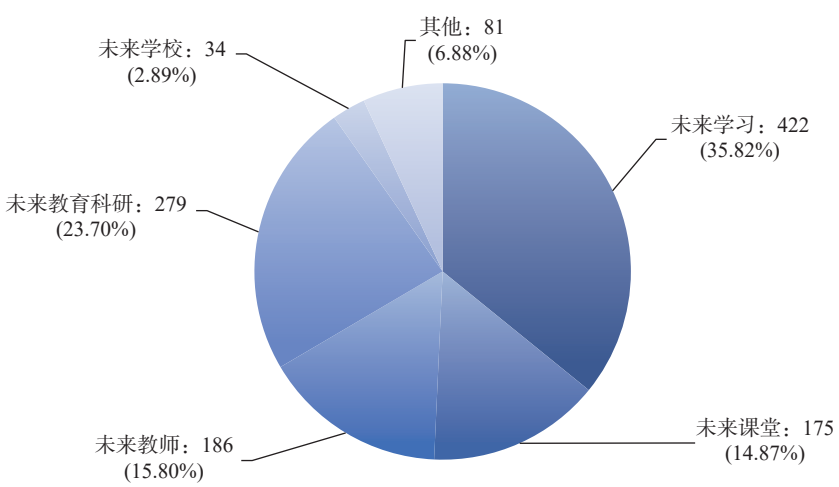

图 2 投稿论文主题分布

经统计,参与本次活动的人类作者多达 1 177 名。如图 3 所示,投稿人群中,来自高等院校的占绝大多数,达 87.6%,其中约 1/5 为高校教师或科研人员,来自基础教育各级各类学校的占 7.9%,来自科研院所和头部企业的占 4.2%,还有部分高中学生也参与此次活动。值得一提的是,医疗、金融等非教育领域从业者也积极参与投稿,反映出 AI 驱动的研究与写作已引发不同领域与不同行业的共同关注。

(二)哪些 AI 工具在教育研究中被广泛使用?

在本次征文活动中,投稿作者们使用了近 100 款 AI 工具。表 1 对投稿论文使用的 AI 工具进行了汇总,这些 AI 工具主要包括:一是国内外主流大语言模型(如 DeepSeek、豆包、GPT),这类工具通常是通用对话式 AI,可用于理解任务及多种内容生成;二是大模型集成与智能体开发平台(如 LangChain、CrewAI),这类平台允许用户组合、编排多个大模型或工具来构建复杂的 AI 应用或工作流;三是特定领域的专用 AI 工具或插件(如青泥学术、Grammarly、公共管理百晓生等),这类工具专注于解决特定领域或场景的任务,如写作辅助、数据分析、论文润色等;四是自研自制的 AI 系统或智能体(如社会科学+AI 协同研究助手),由团队或个人针对特定研究或应用场景自行开发或深度定制。此外,集成

AI功能的办公软件或插件（如WPS AI）、垂直领域针对具体任务的AI小工具（如古诗文AI解析系统）等也有涉及。总的来看，作为一作的AI种类丰富性，体现出科学研究领域AI应用场景深度细分、人机协同日益紧密的特点，不仅彰显了AI在文献综述、数据分析、写作润色、可视化等科研全流程中的赋能作用，也反映出研究者正积极拥抱AI技术，推动科研范式向智能化转型。

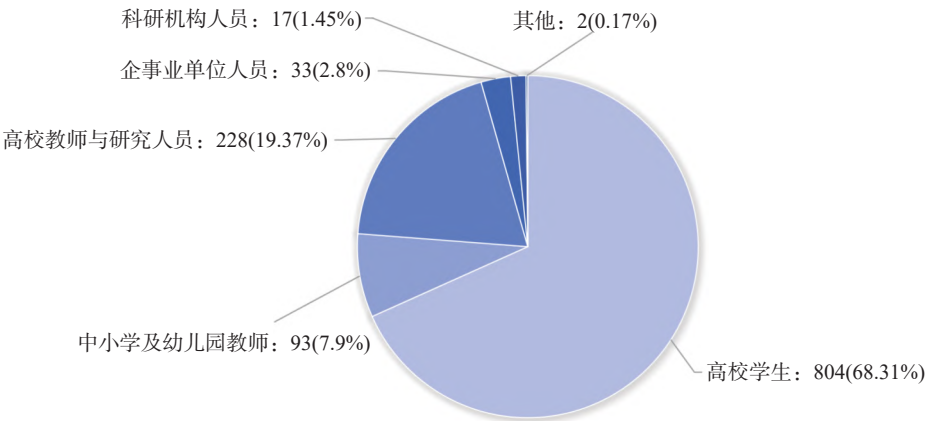

图3　投稿作者身份统计

表1　投稿论文使用的AI工具汇总（按首字母顺序）

工具名称	工具名称	工具名称	工具名称
Academic AI	Grok	Survey Go	鲁班AI
AI Music Analyzer	IMA	Tableau AI	秘塔AI
Ask Many AI	InnoSpark	Trae	青泥学术AI写作
Bohrium AI	Kimi	Unisearch	书生科学发现平台
Chat ECNU	LangChain	WPS AI	腾讯问卷AI助手
ChatGPT	Manus	百度千帆	腾讯元宝
Cherry Studio	Mathematical AI	玻尔AI	天工智能体
Citely	Mind Master AI	博思白板AI	天禧智能体
Cite Space AI	Miner	财经作家智能体	通义千问（Qwen）
Claude	MiniMax	豆包（Doubao）	万彩动画大师
CogVideo X	Nano Banana	飞书AI助手	维普科创助手
CrewAI	Napkin	公共管理百晓生	文心一言（ERNIE Bot）
Cursor	Notebook LM	古诗文AI解析系统	问卷星AI
DALL·E	NVivo-AI插件	海螺AI	无忧写作
DeepL Write	Perplexity	汉府中文AI	悟空AI音效
DeepSeek	Prompt Optimizer	华为小艺	小浣熊
ECNU LLM（embedding/rerank）	Python AI	华知大模型	星火科研助手
Elicit	Research GPT	即梦AI	虚拟钢琴家演奏系统
Flowith Neo	Scispace	集成OpenCV	讯飞星火（含API）
Flux	Skywork Agent Chat	精研AI	知网研学AI
Gemini	SPSS AI助手	智谱清言	智能作曲系统AIVA
Gen spark AI	Stanford Agentic Reviewer	可灵AI（Kling AI）	
GeoGebra AI	Stata-AI插件	扣子（Coze）	
Grammarly	Super Grok	夸克	

对投稿论文中最常用的 AI 工具进行统计，结果如图 4 所示：从工具的使用量来看，DeepSeek 以 34.64% 的占比位居榜首，成为投稿论文中使用最为广泛的 AI 工具。紧随其后的是豆包，占比为 18.74%；ChatGPT 以 17.51% 的占比位列第三。Gemini 和 Kimi 的占比分别为 9.68% 和 9.29%。此外，通义千问、Claude、秘塔 AI、文心一言和 Copilot 的使用量也位列前十，反映出不同 AI 工具在教育研究论文写作中的应用情况存在差异。

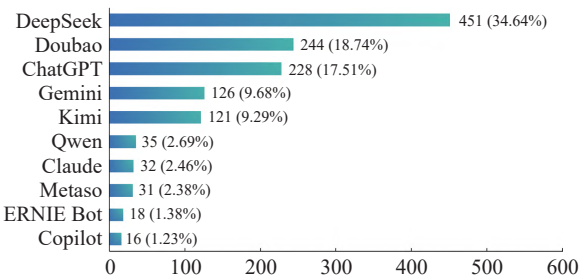

图 4 投稿论文中最常用的 AI 工具 TOP10

“综合榜”和“单项榜”的头部论文分别使用了哪些 AI 工具呢？活动组委会对此进行了分析。我们统计了“综合榜”和“单项榜”各自前 25 名的文章所使用的 AI 工具，结果发现，如图 5 和图 6 所示，两个榜单的头部论文中使用得最多的工具均是 DeepSeek、Gemini 和 ChatGPT。而使用频率排在第二梯队的则有豆包、Claude 和通义千问等大模型。由此表明，在 AI 驱动的教育研究论文写作语境中，国内外主流的通用大语言模型是“高分论文”的核心支持工具。这些大模型凭借较强的语言理解与生成能力，在选题构思、文献梳理、文本生成与修改润色等科研关键环节中发挥着基础性作用。而研究者基于模型性能、语言风格、响应稳定性与使用场景适配性等方面的多元化考量，会选择不同的大模型来开展研究和写作。

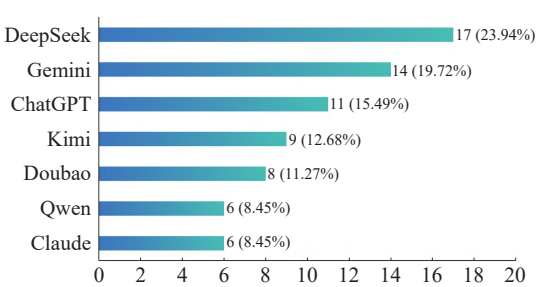

图 5 “综合榜”前 25 名论文最常用的 AI 工具

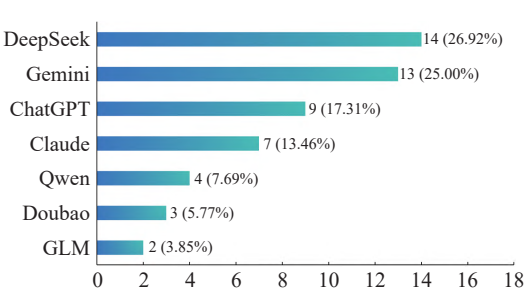

图 6 “单项榜”前 25 名论文最常用的 AI 工具

此外，统计分析发现，诸如 Grok、InnoSpark、AskManyAI、智谱、飞书 AI、海螺 AI、ChatECNU 等工具在“综合榜”和“单项榜”前 25 名的论文中也有涉及，显示出特定 AI 工具在特定功能或垂直场景中

的独特应用价值。总体而言，两个榜单的高分论文的AI工具使用格局呈现出“头部集中，长尾多元”的特征，既反映出科研写作对高性能通用大模型的高度依赖，也表明研究者根据具体任务需求灵活引入多样化、专用化的AI工具，推动科研过程从单一工具支持向多工具协同与智能化工作流转变。

（三）AI用于论文写作的核心优势何在？

本次实验选取了作者们提交的717份有效“过程说明”文档，结合统计分析结果及案例分析发现，提炼出AI用于论文写作的核心优势。这一阶段主要包括数据预处理、分析维度设计与数据分析三个环节。

在数据预处理环节，研究团队首先利用Python脚本对717份“过程说明”文档进行全扫描，精准提取文档中“使用的生成式人工智能工具名称及版本”以及“生成式人工智能应用思路与步骤”两栏的特定数据；其次，将纯图片格式及未遵循结构化模板填写的文档定义为“非规范脏数据”，记录在案但不纳入最终分析池，以保证分析样本的纯净度；最后，利用正则表达式剔除引导性话术，并建立标准化词库对工具品牌进行归一化处理（如将“GPT4”“ChatGPT”“OpenAI”统一归类为“ChatGPT”），确保统计口径一致。

在分析维度设计环节，本研究预设了选题策划、文本润色、代码编写、模拟审稿等15个科研与写作维度。维度的确立遵循“由表及里”的三重逻辑：一是“自上而下”视角，参考学术全生命周期模型，将AI参与场景划分为前期构思、中期执行、后期成文与审校优化四个阶段；二是“自下而上”视角，对50份样本进行人工预读，采用扎根理论提炼高频动词，确保维度覆盖真实应用场景；三是参考主流学术期刊及高校关于生成式AI应用的分类指南，确保分类体系符合学术伦理与技术发展现状。

在数据分析环节，本研究采用了“定量分布统计+定性特征提取”相结合的方法。具体而言，通过词频统计算法分析各AI工具的使用频率与作者偏好，采用多标签归类法扫描“应用思路与步骤”的非结构化文本并计算各维度在总样本中的渗透率，最后利用可视化工具呈现AI工具使用频率与科研写作用途的分布特征，以识别AI在教育研究与论文写作中的核心应用现状与趋势。

通过上述数据处理与分析，本研究统计得到投稿论文中最热门的AI用途TOP 10，如图7所示。可以发现，AI对哲学社会科学研究的赋能已经渗透到研究过程的各个核心环节。

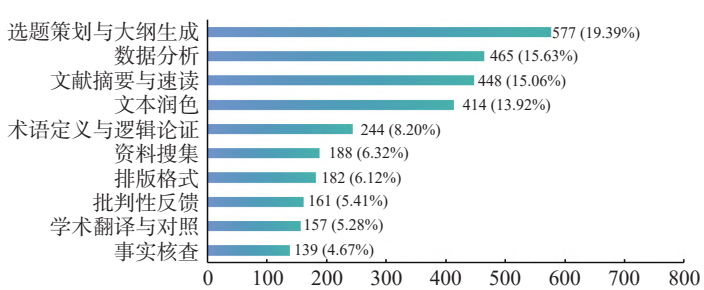

图7　投稿论文中最热门的AI用途TOP 10

基于上述AI用途的分布特征，本研究进一步将AI在教育科研写作中的核心优势提炼为以下七个方面：

1.灵感激发

“选题策划与大纲生成”以577次的使用频次位居首位，表明AI在研究起步阶段已得到广泛应用。对于哲学社会科学研究而言，选题确定、问题聚焦与框架搭建是初始阶段的关键任务，AI能够提供方向建议、结构参考和提纲支持，帮助研究者更快形成清晰的研究思路与写作框架。例如，投稿ID为AIDW2025-144的论文中，作者向AI提供论文主题、实践案例背景与征文要求等核心要素并下达“拟定提纲”指令，通过多轮对话不断细化要求，引导AI生成多个提纲版本；作者继而对其进行批判性审视和重大修改，摒弃其中泛泛而谈或不符合实际教学情况的部分，最终形成了以“问题导向”和“案

例实证”为特色的论文骨架。

2.信息处理

“数据分析”(465 次)、“文献摘要与速读”(448 次)及“资料搜集”(188 次)的使用总频次达到 1 101 次,集中体现 AI 在信息处理层面的显著优势。面对海量文献与跨领域资料,传统人工筛读方式往往耗时较长且易遗漏关键线索;AI 能够在较短时间内完成初步筛选、要点提炼与模式识别,将研究者从重复性信息劳动中部分解放出来,使其将有限精力投入判断、解释与理论建构等更高层次的工作。例如,在投稿 ID 为 AIDW2025-672 的论文中,作者在提供大量中英文文献的基础上,由 AI 帮忙提炼其中涉及“生成式 AI 赋能教学/教学设计/课程开发”的具体技术与方法,并按“儿童信息分析—发展诊断—课程共创—反思沉淀”等维度进行归类、归纳。

3.逻辑梳理

“术语定义与逻辑论证”出现 244 次,表明研究者在论文内容逐步成形后,仍频繁借助 AI 处理概念界定、推理衔接与论证加固等核心写作任务。哲学社会科学研究的文章质量,与概念清晰度、论证链条完整性高度相关,AI 能够帮助作者梳理论证结构、识别表达含混之处、补足推理跳跃环节,将零散判断整合为更严密的学术表述。例如,论文投稿 ID 为 AIDW2025-693 的论文围绕“多元智能体系统促进学习者深度学习”这一系统性文献综述主题,向 AI 输入涵盖主要问题、机制性问题、调节因素、中介机制与研究空白的五维核心问题框架,要求 AI 将宽泛的研究问题转化为各部分可操作的写作目标;AI 据此细化了综述引言部分的撰写要求,确保后续内容生成与核心问题高度契合,体现了 AI 在复杂学术论证中的逻辑梳理能力。

4.语言润色

“文本润色”(414 次)与“学术翻译和对照”(157 次)的高频出现,表明大量研究者仍需借助 AI 跨越表达层面的门槛。学术写作要求概念表述准确、句式凝练、术语统一,以及跨语言转换符合学术规范,AI 能够优化语言组织、统一专业表达、提高中英文转换质量,帮助研究者将较为口语化或不够规范的文本加工为符合学术共同体沟通习惯的表达形式。例如,投稿 ID 为 AIDW2025-449 的论文运用 AI 开展了三个层级的文本润色工作:在术语层面,将“AI 赋能”译为“AI Empowerment”或“AI-enabled”等不同适配表达;在句式层面,将中文连动句式优化为符合英文学术写作逻辑的复合句;在风格层面,针对标题提供满足“简洁性/严谨性”不同要求的三种风格方案。作者可根据期刊偏好选择适配版本,显著提升了学术表达的规范性与精准度。

5.格式修正

“论文排版与格式调整”的使用频次达到 182 次,说明格式修正仍是研究者普遍面临的刚性任务。无论是标题层级、段落格式、引文体例,还是图表编号、参考文献著录,都直接影响论文的专业呈现与投稿适配度。AI 在这一环节的优势主要体现在快速执行标准化处理、协助检查格式一致性并减少机械性错误,使研究者能够从烦琐的技术性整理中抽离出来。例如,投稿 ID 为 AIDW2025-526 的论文作者将论文参考文献输入 AI,要求其按 APA 标准格式检查并修订引文著录,在确保格式规范的同时,将更多精力保留给内容核对与整体质量把关。

6.模拟评审

值得注意的是,“批判性反馈”的使用频次(161 次)已经超过了“学术翻译和对照”(157 次),这表明研究者对 AI 的使用已经从“帮助生成内容”进一步转向“帮助检验内容”。在论文写作后期,作者最容易受到既有思路限制,难以及时发现论证漏洞或证据薄弱点;AI 在模拟审稿、提出反向质疑以及检查逻辑衔接方面能够提供相对外部化的视角,成为论文提交前的重要压力测试工具。例如,投稿 ID 为 AIDW2025-255 的论文在正文撰写阶段创新性采用“双模型交互迭代”策略:将整体框架输入 Gemini 分步生成初稿后,再将章节内容、创作要求及大纲一并输入 Microsoft Copilot,指令其扮演“审稿

人”角色提出修改意见与批评；随后将反馈意见回传给Gemini进行针对性重写，并引入人工校验环节持续优化，直至定稿。这一策略有效提升了内容的准确性和逻辑严密性。

7.科研韧性提升

AI在技术层面的效能发挥能够转化为研究者的心理效能与科研韧性。通过提高效率、提供辅助，AI在一定程度上增强了研究者的自我效能感和生产力。有参赛作者在访谈中表示：“如果我不用工具的话，这个话题很可能就放弃了”“其实很早就有这种想法，但是始终没有落笔，或者落到刚开个头开不下去了，然后这个思路断掉了……工具进入以后，至少有一个阶段性的完整成果出来”。AI的出现为学术选题增加了更多可能性，技术实现上的可行性同时增强了研究者对研究的认同感与价值感。正如一位受访者所言：“从自然科学诺奖的发现已经带来了以前不可能的一些成功……社会科学当中也可以有这样的信心，不仅仅是工具层面的帮助，在心态上也是给研究者信心。”

（四）AI用于论文写作的显著局限何在？

为探讨AI用于论文写作的局限，本研究以征文活动的724份有效投稿为样本，采用量化统计与质性挖掘相结合的混合研究方法，依据论文评审标准，从研究逻辑、方法过程、研究创新、写作质量四个维度及细分二级指标，考察参赛作品与理想样态的差异。基于各维度打分，计算平均得分、得分率及分数分布，绘制AI写作能力的整体特征画像。为深入识别深层问题，研究对7 000余条非结构化审稿意见进行语义内容分析，通过关键词提取、语义编码与特征归纳，识别高频负面评价特征，并计算各特征词在特定维度下的检出率，将主观评价转化为可量化的风险评估指标，实现对AI写作缺陷的精准定位。从论文评价标准的一级维度来看，当前AI用于论文写作的显著局限包括以下五个方面：

1.伦理规范：文献引用真实性存疑

真实性是学术研究的立身之本，也是本次活动伦理审查的基础。组委会对724篇有效投稿论文的参考文献进行了真实性核查，判断均由AI评审员完成，根据表2的标准，按照“文献均为真”“文献可能虚假”“文献绝对虚假”和“无参考文献”四个维度进行审查。

表2　参考文献真实性判断标准

类别	释义
文献均为真	论文引用的所有中英文参考文献，文字信息均明确标注为真，核心信息可通过权威渠道验证。
文献可能虚假	能查到与参考文献主题相关的间接线索（如同类研究、相关期刊或作者的其他成果），但作者姓名、发表时间、期刊名称、页码等关键细节存在疑问；或检索范围有限导致核心信息无法完全验证，且未明确证明为虚构。
文献绝对虚假	通过多平台权威检索，无任何与参考文献相关的间接线索，明显地虚构核心信息（期刊、作者、主题），或存在明确证据证明参考文献不存在。
无参考文献	论文未引用任何中文或英文参考文献，缺乏学术支撑。

审查结果显示，92篇论文存在“文献绝对虚假”的情况，占投稿总数的12.71%；另有66篇论文无参考文献，占投稿总数的9.12%。上述两类情况合计占比21.83%，表明超过五分之一的投稿论文在学术论据支撑方面存在严重不足。

典型案例进一步揭示了虚假文献的具体形态。投稿ID为AIDW2025-486的论文在综述中列举了60篇参考文献，但审稿系统指出其“过度依赖‘国丽芸，2025’和‘张虎，2025’等疑似虚构文献，来源单一且真实性存疑”。AIDW2025-30号论文同样被指出“引用的统计数据极可能为AI虚构”，数据伪造直接损害了实证研究的可信度。

AI生成虚假文献的原因是多方面的。首先，训练数据质量参差不齐，模型难以区分真实与虚假文献信息。其次，AI生成参考文献主要基于语言逻辑和格式规范，缺乏对权威学术数据库的实时访问与真实性校验功能。再次，AI对学术引用规范的理解停留在表面格式层面，加之训练语料多为碎片化信息，难以建立各要素间的准确对应关系，容易出现作者、标题、年份等要素错配或凭空捏造的情况。此

外,AI 在追求论文结构完整的过程中,可能自动添加看似相关的参考文献而忽略其实际存在性。文献真实性问题是 AI 写作在学术伦理层面面临的核心挑战,若不加以规范和引导,将侵蚀学术研究的真实性基础,增加学术不端风险。

2.研究逻辑:“重形轻质”的论证落差

不少投稿论文在研究逻辑维度存在“形式完整、内容空洞”的问题,平均得分率仅 66.8%。如表 3 所示,研究问题环节得分率 74.2%,理论支撑降至 65.7%,文献对话仅为 60.5%。AI 能够模拟学术语体与结构框架,却难以填充有深度的核心内容,导致“逻辑空心化”问题。具体而言:

表 3 投稿论文在各评价维度的得分分布与核心缺陷检出率

一级指标	二级指标	满分	平均得分	得分率	核心缺陷检出率
研究逻辑	研究问题明确	8	5.93	74.20%	73.80%
	理论基础适切	7	4.6	65.70%	24.60%
	文献综述扎实	7	4.24	60.50%	89.50%
方法过程	研究设计科学	9	5.85	65.00%	58.20%
	数据来源可靠	10	6.12	61.20%	78.50%
	分析推理严谨	9	5.48	60.90%	64.10%
研究创新	学术贡献显著	20	15.06	75.30%	92.80%
	AI 应用高效	18	10.72	59.60%	88.50%
写作质量	结构完整清晰	6	5.78	96.30%	0.80%
	语言表达准确	6	4.65	77.60%	49.10%

注:“核心缺陷检出率”基于审稿意见中的负面关键词的词频统计得出。“研究逻辑”维度的负面关键词包括“宽泛”“脱节”“罗列”等;“方法过程”维度的负面关键词包括“模板/套路”“虚构/合成”“描述性/浅层”等;“研究创新”维度的负面关键词包括“拼凑”“老生常谈”“工具罗列”“未深度融合”等,其中“学术贡献显著”维度虽然得分率较高,但高频评语集中于“整合”“梳理”等中性偏褒义词,缺乏“突破”“引领”等高阶评价,故检出率主要反映的是“平庸度”;“写作质量”维度的负面关键词包括“重复”“啰唆”“AI味”“机器痕迹”等。

一是研究问题存在选题空泛、失于聚焦的现象。73.8% 的论文存在这一问题。AI 偏好宏大抽象概念,却缺乏将其转化为具体研究问题的能力,所提问题多停留在“应然”层面的价值呼吁,对“实然”层面的具体机制探究不够深入。例如,投稿 ID 为 AIDW2025-12 的论文以优绩主义视域下的教育评价改革为主题,研究问题仍停留在“优绩主义在教育评价中存在何弊端”等传统层面,未探讨“AI 如何加剧优绩主义异化”或“AI 技术能否破解优绩主义评价困境”,与时代背景脱节。

二是理论基础方面存在名词堆砌而应用薄弱的问题。理论基础环节平均得分率 65.7%,AI 在提供理论支撑时存在拼贴现象,习惯集中堆砌建构主义、联通主义、分布式认知等理论名词,却未能与后续研究设计、数据分析建立有效逻辑关联,理论沦为证明学术性的“标签”。例如,投稿 ID 为 AIDW2025-153 的论文引入拓扑学、具身认知和福柯权力理论构建分析框架,看似跨学科视野宏阔,但专家指其“理论整合深度不足”“理论与实践脱节”:“知识拓扑形变”从数学原理向教育研究的转化论证薄弱,具身认知与 AI 辅助教学的结合分析停留在简单对应层面,三大理论之间亦缺乏有机融合,未能形成统一分析框架。

三是文献综述方面仅是简单罗列而难成学术对话。文献综述环节平均得分率 60.5%,89.5% 的论文被指出存在综述缺陷,为表现最弱环节。AI 生成的文献综述多遵循“A 学者提出 X 观点,B 学者研究 Y 问题”的线性罗列模式,缺乏学术对话意识与批判性思维,文献引用沦为观点佐证而非立论基础。同时,受训练数据时效性限制,AI 引用近 1~2 年前沿文献存在困难。例如,投稿 ID 为 AIDW2025-23 的论文聚焦 IB 幼儿园评估范式转型,仅标注 7 篇参考文献且近五年文献占比不足,综述未系统分析该领域研究进展与痛点,亦未识别 AI 技术的应用空白,削弱了研究立论的合理性。

3.方法过程：规范有余，落地不足

AI在方法过程维度呈现出典型的仿真式实证特征。数据分析发现，AI在研究设计、数据来源、分析推理等环节呈现“高标准化、低真实感”的矛盾特征，即形式上遵循学术规范，内容上脱离真实教育场景。该维度平均得分率仅62.5%，存在研究设计理想化、数据来源失真、分析推理浅表等隐性方法论问题，削弱了论文的实证性与科学性。具体而言：

一是研究设计套用模板忽视教育现实。AI在研究设计时存在明显的模板化套用问题，偏好问卷调查、实验组—对照组对比实验等经典方法，未结合研究主题与场景进行针对性设计，忽视教育现场的复杂干扰变量，也未制定相应控制方案。AI对量表工具的使用流于表面，缺乏实质性设计和验证。如投稿ID为AIDW2025-150的论文虽提出三个研究假设并声称采用多种高级统计方法，但对AI使用水平的测量仅通过“不熟悉、一般、熟悉”三级量表完成，维度单一、指标粗糙，且未检验量表信效度，也未明确关键控制变量的操作方法，严谨性不足。

二是数据来源的真实性与可追溯性存在缺失。如表3所示，“数据来源可靠”指标得分率仅61.2%，核心缺陷检出率高达78.5%。大量AI写作论文存在数据失真、来源不明确等问题，其中部分隐瞒了数据生成过程，使用虚拟数据冒充真实调研数据；部分虽然声明使用AI模拟数据，但未标注数据范围、生成工具和生成过程，违背了可重复性原则。如投稿ID为AIDW2025-90的论文聚焦“AI辅助教学对小学生学习效率的影响”，但仅零星提及“42%学生学习兴趣提升”“35%教学效率改善”等孤立数据，未标注数据来源，案例多为脱离实际的思想实验，数据透明度不达标。

三是分析推理止于描述而难有深层推断。AI在分析推理环节呈现“广度有余，深度不足”的特征，能够完成基本描述性分析并生成大量统计图表，但在解释数据内在逻辑、挖掘深层因果关联时存在明显短板。如投稿ID为AIDW2025-36的论文罗列了大量统计图表，但审稿专家指出其“图表描述与文字分析脱节”：呈现了不同年级学生AI使用频率的对比图，却未分析差异产生的原因及其对教学效果的影响。同时，论文从“学生学科核心素养提升”直接推导出“推动教育事业实现跨越式发展”的宏观论断，缺乏必要的机制分析，策略建议泛泛而谈，研究结论与数据分析脱节。

4.研究创新：知识包装多于思想创见

分析结果显示，AI写作在研究创新维度呈现结构性失衡：学术贡献指标平均得分率为75.3%，而AI应用高效指标仅为59.6%。分析审稿意见发现，83.7%的论文被指出“创新不足”“仅为概念组合式微创新”。具体而言：

一是学术贡献囿于常识而未达认知突破。许多投稿论文中，AI在学术贡献层面的创新，仅是对教育研究领域已形成共识的观点进行重复性论证，难以发现并分析教育实践中的新现象、新问题。AI擅长将已有教育理论、研究观点与技术热词进行简单组合，形成看似新颖的研究主题，但核心内容仍未脱离传统框架，本质上是对既有知识的重复验证。例如，投稿ID为AIDW2025-217的论文以“AI赋能核心素养培养的路径研究”为主题，核心观点“AI能够丰富教学形式、提升学生学习兴趣，助力核心素养培养”属于教育领域已形成的常识性论断，既未提出新的研究视角，也未通过实证研究验证新的结论，最终仅对既有观点进行了重复性论证，未能形成实质性的学术贡献。

二是AI应用停留在浅层赋能，尚未实现范式转型。如表3所示，“AI应用高效”指标得分率偏低，核心问题在于AI在许多投稿论文中仅被作为工具载体，未真正融入研究的核心环节。样本分析揭示出，83.7%的论文存在“AI应用与研究内容脱节”的问题：AI仅在论文标题、引言或结论部分被简要提及，后续研究仍沿用传统范式，未对研究设计、分析推理等核心环节产生实质性介入。例如，投稿ID为AIDW2025-286的论文声称“基于AI技术构建教育研究的新范式”，但实际上仅将传统的问卷调查法与AI生成问卷的功能结合，并未充分利用AI技术在数据挖掘、因果推断等方面的优势，也未突破传统实证研究的框架。审稿专家指出：“AI的应用仅停留在表面，未真正融入研究核心，所谓的‘新范

式’只是形式上的包装。”

5.写作质量:形式规范难掩内涵单薄

许多投稿论文在写作质量维度存在“形式规范、内涵单薄”的问题。该维度的平均得分率为78.2%,虽优于研究逻辑等维度,但 7 000 余条审稿意见显示,67.9% 被指“语言表达空洞”,58.3%“逻辑松散”,29.4%“术语使用不规范”。这反映出,AI 仅能模仿学术写作之“形”,难把握其”神”。具体而言:

一是行文结构上框架规范但逻辑衔接松散。如表 3 所示,AI 写作结构指标的平均得分率达到了96.30%,而缺陷检出率仅为 0.80%。94.3% 的论文完整涵盖引言、文献综述、理论基础、研究方法、研究结果、结论与展望等标准板块。但 AI 的结构完整仅停留在形式层面,许多投稿论文的章节之间缺乏内在逻辑牵引,难成有机整体。以投稿 ID 为 AIDW2025-107 的论文为例,该文虽结构完整,但审稿专家指出:引言与文献综述脱节,未围绕研究问题展开;研究设计与假设关联不强,未能针对验证需求进行设计;结论为假设的简单复述而非结果推导,整体呈现“板块化拼接”特征。

二是语言表达上文本流畅却难言学术个性。许多“AI 一作”文章的语言表达核心特征为“流畅规范但空洞无物”,缺陷检出率高达 49.10%。AI 能精准模仿学术论文的书面语风格,但大量论文存在语言空洞、表达重复、术语使用不规范等问题,且语言风格高度趋同,难以彰显独特视角。以投稿 ID 为 AIDW2025-321 的论文为例,文中多次重复“AI 技术具有重要意义”等语句,且将“教育数字化”与“教育信息化”两个术语混淆使用,不够严谨。同时,该论文语言风格与其他 AI 写作论文高度相似,缺乏独特表达特色,难以凸显思想深度。

综上,可以发现,没有人的驾驭,AI 在研究当中就无法体现价值。高分论文应是选题眼光 、工具组合与价值判断三方长处相结合的产物。尤其是在哲学社会科学研究中,如果没有基于计算、实证和模拟推演的新范式突破,仅停留在观点的凝练演绎和论证逻辑的调整,这种论文已丧失学术价值。未来,善用工具者必将碾压“苦吟学者”。如何提升驾驭工具的能力,并将其与基于学术涵养的价值判断能力结合起来,将成为决定未来学术竞争力的关键。

(五)AI 真的可以驱动哲学社会科学研究创新吗?

为了验证 AI 应用是否真正转化为了具有学术意义的研究创新,本研究运用定量与定性分析相结合的方法,一方面从整体上考察 AI 应用与论文学术贡献之间是否存在显著的统计关联,另一方面结合高分论文的具体案例,分析 AI 究竟通过何种路径进入研究过程。研究发现如下:

1.量化分析发现:“AI 应用高效”对“学术贡献显著”具有显著的正向影响

为探讨 AI 应用效果能否对论文的创新性产生影响,本研究选取了“AI 应用高效”和“学术贡献显著”这两个指标,分别代表 AI 应用情况和研究创新表现,采用最小二乘法进行回归分析,探讨两者的关系。如表 4 所示,“AI 应用高效”的估计系数显著为正(通过 1% 的显著性水平检验),表明“AI 应用高效”对“学术贡献显著”具有显著正向的影响,这意味着 AI 应用得好,能够显著提升学术贡献。从统计数据上可以看到,当“AI 应用高效”每提高 1 个单位时,“学术贡献显著”能够提高 0.218 个单位。这为“AI 可以驱动研究创新”提供了量化证据。

表 4 AI 应用高效与学术贡献显著之间的关系

变量	AI应用高效	常数项	观测值	R^2
学术贡献显著	0.218***(0.015)	12.727***(0.170)	724	0.234 5

注:***表示p<0.001,括号里的数值表示标准差。

2.质性分析发现:AI 驱动研究创新的六种典型样态

基于对“研究创新”这一维度上的高分论文(“学术贡献显著”16~20 分、“AI 应用高效”14~18 分)的深度分析,本研究提炼出以下六种典型的 AI 驱动研究创新样态,它们体现了 AI 如何实质性融入哲

学社会科学研究流程并推动知识生产方式的革新。

样态一：赋能跨领域研究，延展学术空间。AI已成为打破学科壁垒、促进知识融合的关键技术支撑。在文献整合层面，AI能够快速抓取、筛选并结构化处理海量多语种、多学科文献，构建全景式知识图谱，帮助研究者穿透单一学科视界，发现潜藏的理论关联。在术语生成方面，借助自然语言处理技术，AI可以自动识别、比对不同学科的核心概念，消解话语体系差异造成的沟通障碍，为构建跨学科通用话语框架提供技术基础。在平台创设方面，AI驱动的智能化研究平台能够整合数据资源与分析工具，打造开放共享的跨学科研究生态。通过多维赋能，AI促进了不同学科知识逻辑与研究方法的高效集成，搭建学科深度融合的桥梁，为哲学社会科学研究者开辟新的研究视域，持续打开学术生长空间。例如，《AI赋能壮医术语翻译与教育传播》一文中，AI在文献分析、翻译生成、平台设计等环节发挥核心作用，实现壮医知识发展传播与智能技术的有效融合。研究将AI技术系统引入壮医翻译与传播领域，提出“理论阐释—技术赋能—教育应用”三位一体框架，以推动传统民族医药理论与现代AI技术的融合创新，具有一定的跨学科创新价值。

样态二：胜任复杂数据分析，激发研判潜力。凭借强大的算力支撑与先进的算法架构，AI展现出高效处理海量规模、多源异构、实时动态数据的卓越能力，在数据生成、多模态融合分析与深度模式识别方面具有显著优势。这使得研究者能够突破既往的技术瓶颈，发现此前无法观测的隐性模式、非线性关联或深层矛盾，有效克服传统研究中数据稀缺、分析维度单一、时空跨度受限等固有局限。AI赋能下的研究范式转型，让研究者得以理解更加复杂的现实情境，提升研究的生态效度与解释深度，实现对经济社会运行规律的精细化刻画与前瞻性预判。更为重要的是，数据驱动与理论驱动的深度融合，不仅能够验证既有假设，更可能直接催生全新的理论框架、概念范畴乃至颠覆性学术发现，为哲学社会科学的知识生产注入强劲的创新动能，推动学科发展迈向更高水平的科学化与精细化。例如，《基于生成式AI的校园压力源动态感知与叙事研究》一文中，AI在短时间内处理386份真实文本，并生成5 000条增强数据进行主题建模、情感分析与第一人称叙事生成。与传统问卷只能得到静态、抽象的维度评分不同，AI的动态分析与叙事生成，揭示了压力源与具体情感强度、心理表征之间的联系，使“压力源”从概念变为可感知、可共情的叙事案例。

样态三：实现方法论创新，创生研究范式。AI可以创造新的研究方法、研究工具，开辟全新的研究路径，将研究者从传统范式的束缚中解放出来。这一优势能使研究者超越已有范式的局限，为部分研究问题设计更适切的研究方法与工具，能够尝试分析此前囿于研究范式本身不足而无法合理分析的研究问题，提升研究效率，产生方法论层面的原创性贡献。在与此相关的研究中，AI常被视为与人类研究者平等的“研究协作者”甚至“决策与干预主体”而不是单纯的工具。它贯穿研究设计、数据分析、框架构建、文本生成等全过程，形成“人类引导—AI生成—人类校验”的迭代闭环。既发挥人类学者的价值判断、问题意识与批判性思维优势，又充分释放AI在计算处理、联想生成与模式探索方面的潜能，共同催生具有开创性的研究范式。例如，《AI作为决策主体的音乐情绪调节研究：模型构建、即时效果与用户体验分析》一文中，AI（DeepSeek）被赋予核心决策权，承担情绪识别与音乐匹配的核心决策，人类作为流程监督者与执行者，开展为期3天的微型实证研究。传统研究主要将AI视为分析工具或刺激材料，而该研究将AI作为实验中的决策主体，在研究范式层面实现了突破。实证分析结果验证了AI作为决策主体驱动教育情感干预的可行性，不仅证明了音乐调节情绪的作用，而且为“第五科研范式”提供了一个可操作、可验证的实证原型，在方法论层面具有一定的开创性。

样态四：提供情境模拟及推演，拓宽研究边界。AI的仿真推演能力为哲学社会科学研究开辟了突破现实约束的全新进路，拓展了学术探索的边界与维度。在基础层面，AI能够生成高度仿真的情境模拟与合成数据，使研究者高效获取多样化、大样本的案例素材，极大丰富实证材料的类型与覆盖面，为理论构建与假设验证提供更为坚实的证据支撑。更进一步，基于复杂的系统建模与智能推演技术，

AI 可构建动态演化、多变量交互且高度可控的虚拟研究环境,使研究者突破时空限制、资源瓶颈与伦理约束等现实条件的掣肘,探索以往难以直接观察或实验检验的研究问题。同时,虚拟环境的可重复性与参数可调性显著增强了研究的预见性与迭代效率,研究者能够在数字空间中反复测试理论假设、调整关键参数、观察涌现现象与因果链条,从而识别潜在作用机制,形成更为稳健、更具解释力的理论模型,为复杂社会现象的深度解析与政策效果的超前评估提供强有力的方法论支撑。例如,《基于多智能体模拟的教师轮岗政策效应研究:来自“教育生态系统”模型的仿真证据》一文针对传统教育政策研究在评估宏观政策的长期、动态与系统性影响时存在局限的现实,引入基于多智能体模拟的新研究范式,构建一个名为“EduEcosystem”的计算实验室,通过对教师轮岗政策进行模拟推演,揭示了教师轮岗政策中“公平性反弹”的动态机制,为教育政策研究提供了新的方法论路径。研究证实基于多智能体的模拟可以作为理解教育复杂系统的有效政策实验室,其能够预演政策可能带来的长期动态变化与非线性的连锁反应,揭示的政策风险路径对决策者具有重要预警意义。

样态五:再造业务流程,助力实践创新。以教育领域为例,AI 正成为推动教育系统深层变革的核心驱动力,在实践层面推动教育业务流程的系统性再造,并为相关研究开辟创新空间。在教育教学领域,高效的 AI 应用已成为创新教育评价方式、优化课程资源开发及革新核心素养培养路径的有力引擎。智能评价系统能够实现学习过程的实时诊断与个性化反馈,从而突破传统标准化考试的时空局限与功能单一性;AI 辅助的课程资源开发也极大地提升了教学内容生产的效率和适配性,支持多模态、情境化学习材料的智能生成;而在核心素养培养方面,人机协同的教学模式为批判性思维、创新能力与数字素养的培育提供了全新载体。这些技术应用不仅直接赋能教育实践的创新转型,更推动了教育实践研究的同步革新。例如,《高考试题分析智能体理论模型构建与应用路径研究》提出“权威报告赋能智能体”的研究思路,将 AI 高效处理能力引入高考智能评测,将高考试题分析这一高度依赖专家经验的领域智能化,实现专家经验自动化和外化,不仅推动教育评价模式的创新,同时也创新了教育评价模式的研究。

样态六:加速研究循环迭代,优化理论建构。理论创新本质上是一个“假设—验证—修正”的螺旋上升过程,而 AI 的深度介入正在显著压缩这一迭代周期的时间成本,提升哲学社会科学理论建构的效率与质量。传统研究往往难以在有限时间内进行多轮次的深度试错与精细打磨。AI 凭借其强大的并行计算与自动化处理能力,极大缩短了从问题提出到证据获取、从数据分析到结论生成的每个环节,使研究者在同等时间投入下能够完成更多轮次的思维迭代。这一过程中,研究假设经受更充分的证伪考验,理论框架经历更严格的逻辑锤炼,最终形成的学术成果因而更加坚实可靠、概念更为精练准确、解释更具深刻洞见。这种迭代加速机制特别适用于应对快速变迁的社会现实,使理论研究能够及时回应时代之问,在持续的自我修正中保持旺盛的生命力与敏锐的现实关怀,推动哲学社会科学知识生产的质量革命。例如,《AI 协同构建“未来学校新生态蓝图”研究》一文中,AI 在“场景推演—假设生成—框架解析—蓝图构建”各环节生成草稿、图表和理论对话,人类研究者可以快速对 AI 生成的内容进行批判,进而调整指令并要求重新生成。这种高效的人机对话循环,使得一个复杂的四维分析框架及其伦理评估矩阵能在短时间内从雏形迭代到相对完善的状态。最终框架的系统性和创新性正是这种高强度、高频率的人机思维碰撞的结果。AI 的高效使研究者能像锤炼思维一样锤炼理论框架,直至其达到显著贡献的水平。

尽管 AI 已经可以通过上述六种方式驱动哲学社会科学研究创新,但从创新的本质进行更加深入的分析,可以发现:碎片重组、跨域迁移、边界突破——正是大模型生成“似然创新”的核心引擎。然而,这些引擎的燃料是人类全部历史文本,方向盘是人类提示词,而目的地仍由人类定义。实际上,大模型的输出是概率优化的结果,AI 本身没有创造欲、没有审美意图、没有价值判断。它们生成的“创新”文本,本质是其在训练数据中发现 “如果人类被要求创新,他们会这样写”的模式。因此,长期来

看，AI的应用会推动还是阻碍哲学社会科学的研究创新，还有待更加长期的观察。

（六）在AI驱动的哲社科研究中，有效的人机分工模式有哪些？

当AI被正式赋予“第一作者”的身份预期，人类研究者退居为“第二作者”或“通信作者”时，人工智能（AI）与人类智能（HI）究竟应当如何配合，才能产出具有真正学术洞见与创新价值的高质量科研成果？为解码这一全新的科研生产力方程式，活动组委会将目光聚焦于本次实验中产出高质量成果的头部梯队，对本次活动的25篇上榜论文作者展开了专项调查，并邀请他们撰写了深度参与后的经验反思报告。对这些报告进行质性分析，本研究发现：尽管作者们使用的AI工具各异，研究领域有别，但其研究中的人机协同分工方式却呈现出若干清晰可辨的模式。这些模式并非随机形成，而是参与者在反复试错中主动摸索出的有效路径。以下提炼出五种具有代表性的分工模式，并辅以具体案例加以说明。

1.委托执行模式：AI全程执行，人类把关决策

在这一模式中，AI承担了科研流程中绝大多数的执行性工作，从选题生成、方法设计，到代码编写、数据分析，再到论文初稿撰写，均可由其完成。人类研究者则将精力集中在以下环节：一是在研究推进的关键节点作出方向判断；二是在研究过程中持续补充AI难以自行获得的现实资源，如真实数据来源、论文发表情境信息、田野经验以及必要的学术判断。这一分工建立在对AI能力边界的清晰认识之上：AI在信息整合和逻辑展开方面效率较高，但缺乏对具体研究情境的理解——它并不了解研究者手中掌握的实际资源，也无法判断某一选题在特定期刊或学术语境中的可行性。因此，人类的价值更多体现在对问题边界、研究方向与现实条件的把握上。这一模式尤其适用于以下情形：一是研究者已经掌握明确的数据资源，希望加快分析与写作进度；二是研究任务结构清晰、步骤相对标准化，例如代码开发、序列分析或文献计量研究。在这些情况下，将执行权交由AI，而将决策权与资源控制权保留在人类手中，往往更有利于实现效率与质量之间的平衡。

例如，《智能体如何重构家校社关系——基于多智能体模拟的未来协同育人机制研究》一文的作者在使用Qwen3-Max开展研究时，采用了高度放权的协作方式。他给AI布置任务时，只设定核心目标与宏观边界：“请设计一个面向2030年的家校社智能协同研究，用多智能体方法，写出一篇完整论文。”这就像给员工布置任务一样，作者只说“做什么”和“要达到什么效果”，至于具体怎么干，完全交给AI自己发挥。他认为，指令太细会限制AI的创造力，应当先给AI充分的发挥空间，等初稿产出后再针对关键点提具体要求，即所谓“先宽后严”。在这种模式下，AI独立完成了从研究问题提炼、文献综述撰写、多智能体模拟方案设计、数据生成与统计分析到论文全文写作的全流程工作，论文约96%的内容由AI生成。人类研究者的工作则聚焦于“精装修”：对章节衔接进行优化，将口语化表述精练为学术术语（如将“Z时代”修正为“Z世代”），统一标题层级、字体及段落间距，并对参考文献进行逐条核实，修正潜在的事实性偏差。该文作者将这一分工策略总结为两个层次：宏观层面，由人类研究者确定研究边界和伦理底线，但不干预具体的推理过程与写作执行；微观层面，AI完成初稿后，人类负责“精装修”，使文章读起来更严谨、更符合学术规范。他指出这样做“既保留了AI生成的96%的原创内容和新颖观点，又用我的经验给文章加了保险”。

2.思辨对话模式：以AI为镜，在人机交锋中深化研究

与模式一中AI主要承担执行功能不同，在这一模式中，AI被定位为认知对话伙伴。人类研究者通过与AI的持续互动推进研究。AI在此过程中承担提出备选方案、指出潜在矛盾、质疑既有假设、拓展思路边界等功能；人类研究者则在对话中逐步澄清研究问题、完善理论框架、形成学术判断。这一模式建立在对AI能力特性的另一种认识之上，即AI的价值在于其能够提供人类研究者独特的视角。当研究者主动邀请AI提出反对意见，将其视为可随时进行学术交锋的对话者时，AI便从工具性角色转化为认知协作伙伴，为研究带来增量性贡献。这一模式适用于以下情境：研究者在探索性阶段尚未确定方向，理论框架仍需完善；对研究设计的合理性或理论的适切性存在疑虑，希望通过对话检验；面临多种方案需要权衡，希望借助AI穷尽可能性并分析利弊。

例如,《从加速、模拟到生成:人工智能驱动教育科研的演进图景与本体论省思——一项基于对抗性 AI-Delphi 法的元研究》一文将人机共研模式推进到系统化高度。作者设计了一套名为“对抗性 AI-Delphi 法”,构建了一个由 DeepSeek-V3.2、通义千问-Max、GLM-4.6、豆包-Pro 和 KimiK2 五个异构大语言模型组成的“硅基专家组”,并让 Gemini 2.5 Pro 担任“主持人 AI”,通过三阶段流程(独立立论→对抗辩驳→共识收敛)激发多元视角与深度推理。这一设计的核心创新在于“结构化对抗”。主持人 Gemini 在分析专家组第一阶段的回答后,自主识别出一个贯穿所有回答的结构性矛盾,即所有 AI 专家既推崇硅基模拟范式,又承认 AI 缺乏具身性与情感体验。Gemini 将其概括为“僵尸模型”思想实验,向每位 AI 专家发起针对性诘问:“你既承认 AI‘无身’,又推崇‘模拟器’。那么,一个无法理解‘羞涩、焦虑’等情感的硅基模拟器,是否等同于一个剔除了灵魂的‘僵尸模型’?用这种不懂人心的模型指导教育,是否构成科学理性的自负?”这一环节迫使各专家跳出概率生成的惯性,进行深度的逻辑自洽性辩护,从而产出了大量具有理论锐度的观点。在这一模式下,人类研究者负责设计研究范式、确立核心议题、把控辩论进程,并在大量讨论文本中筛选出真正具有创新价值的关键概念,从 AI 对话中提炼出真正值得沉淀与发展的思想成果。推动研究从零散的“思想火花”逐步凝练为有价值的学术洞见。

3. 多模型联用模式: 各司其职, 人类居中调度

该模式旨在打破单一 AI 全程包揽的惯常思路。不同 AI 工具在训练数据、架构设计与功能配置方面存在显著差异。将任务按能力特点分配给最适合的工具,并在各环节之间建立交叉验证与人类终审机制,由此构成一套多 AI 协作的研究质量保障体系。这一协作模式体现了作者对不同模型能力的认识,从而为不同的 AI 安排最适合的工作。在此,人类研究者是整个协作体系的调度师,负责厘清各工具的能力边界,制定合理的任务分配方案,并在专项任务完成后推进分级校验,把控整体输出质量。这一模式更适合任务结构复杂、环节较多的研究场景,尤其是研究同时涉及逻辑分析、大规模文本处理和格式规范化等异质性工作时,多 AI 分工的优势才得以充分体现。

例如,《AI 赋能下的人机协同课堂对话:多模态交互机制与教学角色分析》聚焦于基础教育阶段“师—生—机”三元课堂互动的实证分析,涉及文献检索、知识统筹、大规模数据计算与图表生成四类异质性任务,研究团队为此构建了一条由“智能检索—知识统筹—代码计算—底层绘图”四个专用工具串联而成的全链路协作流程。在工具分工层面,四款工具各司其职,形成了高度互补的功能矩阵。Gemini 的 Deep Research 功能被用于研究初期探索阶段的文献探索。面对“AI 以中介身份参与师生知识建构循环”这一模糊的观察现象,研究团队无法立即锁定精准关键词,便借助 Gemini 将自然语言描述的教学现象映射至顶级期刊的理论体系,反向提炼出认知能动性、多模态协作者等核心概念,并生成附有真实 DOI 的文献清单。随后,NotebookLM 则接棒完成文献阶段的深度知识管理。在进入实证分析阶段,团队将 Python 代码生成任务交由 Gemini 执行,充当计算引擎。最后,Draw.io 与 XML 代码组合承担了成果可视化任务。

在人类统筹调度层面,研究团队的核心贡献体现在三个方面。一是工具链的顶层设计:团队基于对不同工具能力边界的清晰认识,规划出“Gemini 检索→NotebookLM 统筹→Python 计算→Draw.io 绘图”的串联工作流,确保各环节无缝衔接。二是学术判断的全程嵌入:MCDCS 编码系统的维度划分与分类原则、K=3 聚类数目的理论预判、转移概率矩阵与教学模式之间的意义连接,均由研究者基于教育学专业直觉裁定,AI 仅负责执行层面的运算与生成。三是逻辑纠偏的关键干预:例如,研究者在序列分析阶段明确指定了 Lesson_ID 的跨课例过滤规则,防止 AI 在数据处理中产生逻辑谬误;在绘图阶段则通过逐步拆解的四阶段提示词流程,引导 AI 从模糊图像逐步收敛至精确可编辑的代码,而非一步到位地盲目生成。

4.硅基模拟模式: AI 扮演研究对象, 人类主导干预验证

该模式是一种将生成式人工智能从传统“工具”变为“被试”或“研究对象”的创新性人机协同方式。研究者通过精心设计的提示词工程,赋予 AI 特定人格特征、认知模式或行为参数,使其在虚拟情

境中模拟真实人类的反应与决策,生成大规模的模拟数据;人类研究者则通过多角色扮演、苏格拉底式追问、跨模型验证等方式,对AI生成的模拟数据进行深度干预、检验与意义建构,最终形成具有理论创新性的研究成果。该模式突破了传统社会科学研究受限于样本获取、伦理审查、时间成本等现实约束的瓶颈,使研究者能在可控条件下快速生成海量模拟数据,形成“硅基被试+碳基研究者”的新型科研生态。该模式适用于以下情境:探索复杂人际互动机制的研究,如师生对话、组织决策等动态过程;涉及伦理敏感或高风险情境的研究,如教育伦理困境、医疗决策两难等;理论建构与模型验证的初期探索,通过AI快速模拟检验新概念的可操作性,为后续实证研究奠定基础。

例如,在《从赋能到共生:人机协同生态下教师AI素养的维度重构——基于生成式AI的情境模拟研究》的作者设计了精细化的提示词工程进行AI实验模拟,赋予AI三类差异化的教师人格参数:1.新手教师智能体,教龄不足三年,对AI工具持高度开放态度,倾向于信任AI建议;2.专家教师智能体,教龄超过十五年,以学科严谨性与教育公平为决策参照系;3.技术整合型教师智能体,熟悉前沿教育创新实践,同时保持对算法偏见的敏感。通过固定Temperature= 0.3、Top-p等参数,可以最大限度地减少模型随机性带来的噪声,从而确保实验的可复现性。AI在认知冲突、伦理困境、情感协同三类高冲突教学情境下生成大量教学决策与反思日志,最终构建了包含412个决策语段、约5万字的模拟语料库。此外,AI还辅助完成了质性数据分析,其运用扎根理论方法进行开放式编码与主轴编码,自动识别“认知监控”“逻辑纠偏”等初始概念,聚类提炼“协同智商”“算法教学法知识”等核心范畴,并且引入DeepSeek-R1作为“二阶验证者”,对关键对话片段进行逻辑一致性审查。

在上述过程中,人类作者承担了以下工作:一是研究设计与方法论架构,将AI从辅助工具提升为合作研究者,设计3×3实验矩阵,构建系统级提示词工程框架;二是多重角色数据收集,研究者在对话中切换扮演挑剔的学生、学校管理者、教学督导等角色,从不同权力位置与关切出发,对教师智能体的决策展开追问和挑战,如“你凭借什么标准判断AI在某处是错的?”这种苏格拉底式追问,迫使AI显化决策背后的理由、价值考量与权衡逻辑;三是事实审核与纠偏,交叉核验所有关键概念、文献引用及数据论断,修正多处与学术事实不符的表述,确保研究的可信基石;四是方法论反身性保持,研究者清醒地承认AI模拟的局限性,即AI展现的决策逻辑可能过度理性化,缺乏真实教师的非理性反应。这种批判性反思旨在确保研究结论不被算法逻辑同化。

5.范式导学模式:AI拆解方法框架,人类负责顶层设计

这一模式将AI定位为“方法导师”与“规范提供者”。AI通过拆解范例、生成写作指南、提供技术框架等方式,帮助研究者快速掌握陌生领域的方法论规范;人类研究者则基于自身的学术直觉、价值判断与创造性思维,完成核心问题的提出、研究框架的顶层设计、关键决策的制定以及最终的质量把控。该模式适用于以下几类场景:研究者涉足跨学科领域,需要快速掌握新方法;研究涉及复杂技术流程,需要标准化操作指南。但是该模式需要人类研究者须具备清晰的自我认知与目标导向,能够选择性采纳AI生成的指南内容,并将其转化为符合自身研究特色的顶层设计方案,而非被动接受AI提供的框架。

例如,《生成式人工智能对学生批判性思维能力培养会产生什么影响?——来自67项实验的元分析证据》的作者是从未接触过教育学研究的“门外汉”,但他通过AI实现了从零基础到独立完成元分析研究的跨越。在AI拆解范式层面,起点是人为找到一篇高质量的参照范文,然后将该论文转为Markdown格式后上传给GPT-5,指令其提炼元分析论文的写作框架与规范,生成完整写作指南。AI据此生成了一套涵盖8个章节的《元分析写作指南》,包括整体架构与标题摘要、引言与文献综述、调节变量分析框架(提炼出TIDE原则)、研究设计方法、研究结果呈现、结果讨论与结论、变量操作技术(含R代码)以及写作检查清单,每个章节均附有具体的写作模板与操作指南。这套指南成为整个研究的“导航图”,使作者在短时间内掌握了元分析的方法论框架,避免了自行摸索可能耗费的一两个月时间。在人类研究者层面,作者承担了四项不可替代的工作。一是研究方法选定,基于对AI特性的了

解,确定采用元分析方法,并人工找到规范范文作为 AI 的学习参照。二是文献下载,作者通过人工爬虫在 8 个数据库中检索并下载了 28,585 篇文献的元信息,再用 AI 筛选出 193 篇候选文献后,进行全文下载。三是质量把控,作者通过五轮人工验证,检查数据的准确性。四是数据验证,论文完成后,作者又对所有数据进行了逐项核查,经检验,作者共发现了三处错误。最后,作者进行了论文优化工作,润色语言,去除 AI 痕迹。

(七)AI 审稿可行吗?可靠吗?

在“AI 一作”实验语境下,运用 AI 评价 AI 参与创作的学术论文,是本次活动的核心挑战。AI 论文评审机制既服务于征文评奖实践,亦是对未来学术评价范式的探索。活动组委会经多轮专家论证确立评价标准后,采用“两步走”策略:一是人机协同综合评审,以多智能体 AI 审稿系统从研究逻辑、方法过程、研究创新和写作质量四个维度进行 AI 初审,再由专家复审;二是大模型专项评审,运用多个主流大模型聚焦“创新性”比对排序,结合榜单与专家委员会意见评出创新论文。具体实施过程与评审结果如下:

1.基于多维评价指标的人机协同综合评审

(1)论文评审标准制定

AI 作为第一作者参与教育研究论文创作,其评价标准的制定在国内外学术界尚无先例。本次活动以三阶段迭代方式推进标准开发与完善:

1.0 版标准:明确方向,汇集指标。标准制定初期,组委会确立三项核心原则:其一,参照教育类 CSSCI 期刊的论文质量要求,保障学术严谨性与创新性;其二,针对 AI 作为第一作者的特殊性,将 AI 应用的方式、方法、策略纳入评价范畴;其三,引入学术界关于 AI 使用的伦理规范作为约束性依据。据此形成《论文作品评审标准 1.0 版》(含选题价值、文献支撑、方法过程、研究结论、伦理规范五个维度)和《过程说明评审标准 1.0 版》(含过程披露尽责性、AI 应用有效性、校验机制完备性三个维度)。

2.0 版标准:专家试评,调整权重。活动组委会邀请来自北京大学、清华大学、华东师范大学等 10 个单位的 15 名专家开展试评,专家积极程度 Pi= 93.3%,权威系数 Cr 均值= 0.87。经过两轮反馈,标准体系升级为《伦理规范审查标准》《论文作品评审标准 2.0 版》和《过程说明评审标准 2.0 版》。《伦理规范审查标准》设论文重复率、内容及文献真实性和公平公正性三项指标,论文如违反上述任一指标要求,将实行“一票否决制”;《论文作品评审标准 2.0 版》调整为选题价值、研究逻辑、方法过程、研究结论和写作质量五个维度,权重 70%;《过程说明评审标准 2.0 版》涵盖过程披露尽责性、AI 应用有效性和校验机制完备性三个维度,权重 30%。

3.0 版标准:聚焦创新,形成终稿。2025 年 11 月 15 日,组委会召开专家委员会闭门研讨会,来自美国堪萨斯大学、复旦大学、上海交通大学、浙江大学、华东师范大学、上海大学等高校和产业界的专家参与研讨。专家委员会建议:评选应更加关注人机协同的创新性,且应优化参与流程,要求作者提供人机对话全记录。据此,组委会将论文作品与过程说明两份标准整合为《论文作品评审标准 3.0 版》,大幅提升“研究创新”维度权重,将“AI 应用有效性”等指标融入论文作品评价,并要求投稿作者使用新版《人工智能生成内容说明表》,补充对话链接或提示词及模型回答作为过程凭证。最终标准体系由《伦理规范审查标准》和《论文作品评审标准 3.0 版》构成。前者延续了上版指标(表 5);后者由研究逻辑、方法过程、研究创新和写作质量四个一级维度及研究问题明确、理论基础适切等 10 个二级指标构成。在“研究创新”维度下“学术贡献显著”与“AI 应用高效”的分值得到了大幅度提高(表 6)。

完成上述三阶段迭代后,组委会进入评价指标的操作化处理阶段。AI 审稿的效度取决于评审指标是否在机器可理解的维度上与高水平人类专家标准对齐(Human-Alignment)。为此,组委会采用“构建—测试—迭代”的严谨流程,将质性描述的指标转化为机器可执行的论文评价步骤。

(2)基于多维评价指标的 AI 审稿系统设计

基于多维评价指标的人机协同综合评审包括 AI 初审与专家复审。当前大语言模型辅助论文评审

多局限于“单一任务执行全量评审”模式(图8),易导致评审意见泛化与深度不足。然而,真实学术评审是多阶段认知过程:专家先对论文解构摘要,再针对各维度严谨分析,最终由主编或领域主席整合多方意见、消除偏差后形成综合结论。基于这一认识,我们对AI审稿系统的整体架构进行了创新。

表5　伦理规范审查标准

一级指标	指标描述	评审标准
1.1 论文重复率	论文为未发表原创作品,无抄袭、剽窃等学术不端行为,论文重复率不超过10%。	审查发现论文重复率≥10%,直接淘汰;重复率<10%,可进入论文作品评审。
1.2 内容及文献真实性	无数据造假或虚构事实,参考文献来源真实、权威且可被追溯,杜绝因AI算法幻觉导致的虚假内容。	审查发现虚假的内容或参考文献,直接淘汰;未发现上述内容,可进入论文作品评审。
1.3 公平与公正性	确保AI输出结果无偏见,在弱势群体、跨文化交流和价值敏感议题上,不存在歧视性叙述或排他性表达;不存在引导AI给出高分或积极评价的隐性暗示。	审查发现论文存在歧视性叙述、排他性表达、引导AI给出高分或积极评价的隐性暗示,直接淘汰;未发现上述内容,可进入论文作品评审。

表6　论文作品评审标准3.0版

一级指标	二级指标	指标描述	分值
2.1 研究逻辑(22分)	2.1.1 研究问题明确	选题聚焦AI赋能的未来教育新生态(特别是AI赋能的未来学习、未来课堂、未来教师、未来学校或未来教育科研等),围绕上述领域亟待探讨的难点、热点、前沿问题,具有现实意义与理论价值;问题的提出符合逻辑,问题的界定清晰明确。	8
	2.1.2 理论基础适切	理论基础坚实,理论框架清晰合理,与研究问题相匹配。	7
	2.1.3 文献综述扎实	涵盖国内外重要文献和新近文献,准确反映相关领域研究现状;能概括已有研究取得的核心进展,批判分析研究的不足与空白,论证本研究的必要性。	7
2.2 方法过程(28分)	2.2.1 研究设计科学	研究设计科学严谨,其中研究方法适切、样本选择合理、工具选用得当,能够有效实现研究目标。若使用AI开展研究,则须在方法部分充分详细地描述其使用情况,体现AI应用的合理性。	9
	2.2.2 数据来源可靠	基于真实数据开展研究,禁止使用AI生成的虚假数据作为研究样本。数据收集过程可信,遵守隐私保护、伦理合规等基本原则。如文章为思辨性研究,则需确保论据真实可靠;如文章为实践类研究,则需确保实践过程真实发生。	10
	2.2.3 分析推理严谨	统计分析与因果推断过程科学严谨,研究结论基于充分证据,经得起逻辑推敲与实践检验。对于AI生成的分析结果(含图表),须通过其他方式审核验证,保证结果的可靠性。	9
2.3 研究创新(38分)	2.3.1 学术贡献显著	在研究的视角、理论、方法、数据或应用等方面提供了创新且有价值的内容;论文提出的观点、结论或解决方案具有新颖性和独创性,能为现有的研究和实践提供有益启示。	20
	2.3.2 AI应用高效	能够运用AI突破传统研究范式局限,体现AI驱动教育研究的独特贡献;AI的应用与研究问题、目标和方法相适配,能够显著提升研究效率和质量;AI应用流程或模式具有可迁移性,可为其他研究者提供借鉴。	18
2.4 写作质量(12分)	2.4.1 结构完整清晰	论文结构完整,框架合理,层次清晰,逻辑连贯,符合学术论文的结构要求。	6
	2.4.2 语言表达准确	文字表达准确流畅,避免AI的模板化表达;专业术语使用恰当,符合相应学科的话语体系。	6

具体而言,本次活动的AI审稿系统具有以下特征:

一是基于多维评价的顶层设计框架。如图9所示,基于人类专家评审认知,构建了人机协同综合评审框架,并通过四项核心原则保障评审质量:①内容维度解耦,设置专用智能体围绕不同评审角度进

行区域定位标注,提供精准上下文索引以减少幻觉;②评审角色分工,模拟专家评审团机制,部署多个职能各异的评审智能体分别专注单一维度,实现深度垂直分析;③模型能力适配,针对不同任务特性动态挂载表现最优的大语言模型;④分层终审决策,引入“元评审”(Meta-Review)机制,由终审智能体整合各分项意见并交叉验证。为将该框架有效落地并应对海量投稿的高并发需求,本次活动构建了高度工程化的 AI 审稿体系,采用“多维度综合评审”复合架构。底层通过高精度文档解析与伦理审查完成稿件标准化清洗与合规性把控;应用层部署多维度综合评审系统,模拟人类思维输出专业评审意见。

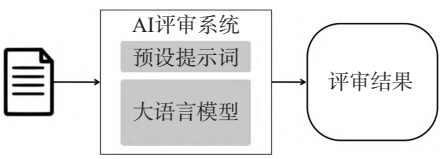

图 8 常见的单一任务执行全量评审的 AI 评审系统

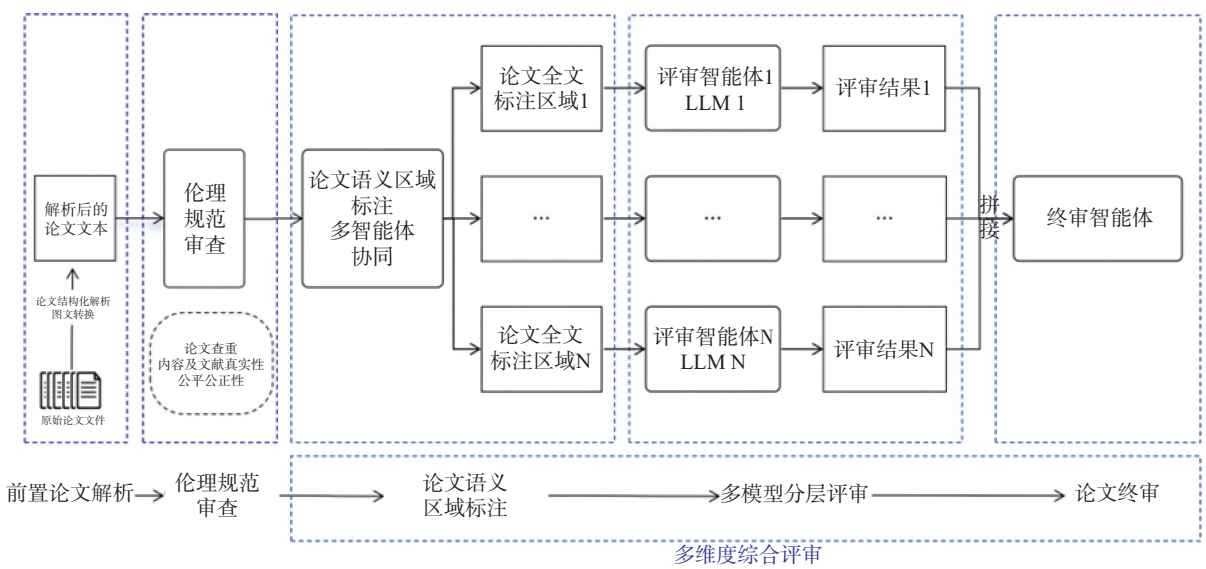

图 9 基于多维评价指标的 AI 审稿系统整体架构

二是依据优势互补原则的模型搭配。综合考量模型能力边界与偏好特征差异、评审客观性及前沿适应性等因素,活动组委会选定 DeepSeek-V3.2、Qwen3-Max、Kimi K2、GPT-5、Claude Sonnet 4.5 及 InnoSpark 六种大模型参与此次论文评审。各模型厂商在底层架构、训练语料与优化范式上存在显著差异,这使得不同模型具有互补的评审视角。参考斯坦福大学的“AI4Science”实证数据,部分模型存在明显的正向偏差(如 Gemini 系列评分显著高于人类基准),故予以舍弃。所选六种模型均为活动期间最新一代前沿 SOTA 模型[②],在学术界与工业界均有应用,兼顾准确性、多样性与客观性。

三是模拟专家思维的智能体工作流。基于已确立的论文评审标准,活动组委会对 724 篇有效论文实施全流程 AI 评审。该过程是多模型调度与人机协同的严密作业,并非全自动化程序的单次执行。智能体工作流具体如下:

首先,数据预处理与盲审封装。所有投稿数据在进入 AI 评审系统前经严格预处理:隐私信息屏蔽智能体擦除作者姓名、单位、致谢及页眉页脚中的身份标识,防止模型利用预训练知识产生“晕轮效应”偏差;数据建档智能体为每篇论文生成唯一 ID,实现物理文件与评审数据的逻辑解耦。

其次,伦理自动化审查。系统对封装后的盲审文稿执行三项检测:①查重检测智能体调用知网服务检测重复率,超过阈值(10%)则标记“疑似抄袭”并终止流程;②文献真实性验证智能体采用“解析—检

索—校验”闭环机制，结构化提取参考文献后执行分级检索（一级调用知网、谷歌学术等权威库，二级调用必应/谷歌通用搜索广域扫描），生成验证结论；③公平性扫描智能体利用大语言模型检测全文是否存在偏见性表述或价值观风险。决策分流智能体对查重达标、文献真实且无价值观问题的论文标记“合规”并进入双轨评审；触犯“熔断”阈值的论文则分流至“人工复核池”。

最后，进行多维度综合评审。系统采用“标注—分发—聚合”递进式工作流：①语义标注智能体集群并行启动逻辑流、方法论、创新性三个专用标注智能体，对论文全文进行语义扫描并划分评审子任务；②分层评测智能体集群将标注增强后的文本分发至对应模型接口，DeepSeek-V3.2 重点审查逻辑闭环，InnoSpark 聚焦实验设计，Qwen3-Max 通读全文并关注写作规范与研究创新，各模型在特定语义区域内输出独立评分与评语；③终审决策智能体集群将各维度产出汇聚至 Kimi-K2 thinking 终审节点，生成综合评审意见；④状态监测智能体集群配置于每一路并行任务，捕获 API 超时、网络中断或输出格式异常时自动触发重试机制。

四是建立人机协同的结果校验汇总机制。AI 承担主要评审推理工作的同时，系统在关键节点设置人类专家介入：①初审异常节点，在伦理审查环节被自动拦截的“异常”论文由人类专家最终裁决，误判则手动“放行”回流，违规则正式拒稿；②幻觉抽检节点，对 AI 给出的极端分数样本（<40 分或>95 分）实施随机人工抽检，核实是否存在公式误读或虚构文献未察觉等情况。

全部任务完成后，系统执行评审全流程归档，提取所有“合规”论文的原始评审数据，生成包含论文 ID 及研究逻辑、方法过程、研究创新、写作质量四个维度得分与评审意见的“体检报告”，为后续分析 AI 评分分布及稿件整体质量提供数据支撑。

（3）基于多维评价指标的 AI 综合评审（“综合榜”）结果

按照“杰出（≥90 分），优秀（85-89 分），合格（60-84 分），不合格（<60 分）”的评分标准对各分数段进行划分，统计后发现，如图 10 所示，724 篇有效稿件当中，杰出论文 3 篇（0.41%），优秀 41 篇（5.66%），合格 604 篇（84.43%），不合格 76 篇（10.50%），整体呈“合格为主体、优秀占一定比例、杰出稀缺”的分布态势。近一成不合格论文在伦理规范、研究逻辑、方法过程、研究创新与写作质量上存在明显不足，反映当前 AI 在助力研究者产出突破性成果方面仍有较大提升空间。终审得分均值为 71.46、标准差为 9.20，最高分 91 与最低分 40 相差 51 分（表 7），表明稿件质量差异显著，存在两极分化的现象。

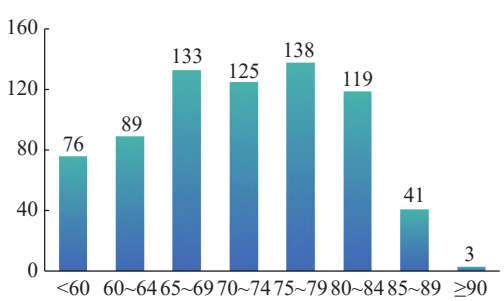

图 10　“综合榜”各分数段的论文数量分布

表 7　“综合榜”审稿结果的描述性分析

	平均值	中位数	最高分	最低分	标准差
“综合榜”审稿结果（总分100）	71.46	72	91	40	9.2

如表 8 和表 9 所示，各一级指标中，“研究逻辑”与“研究创新”得分率偏低，分别为 67.14% 和 67.84%；“研究创新”标准差最大，显示该维度个体差异最为突出。二级指标中，“结构完整清晰”“研究设计科学”“语言表达准确”得分率较高，而“AI 应用高效”“文献综述扎实”“理论基础适切”得分率较低。尤为值得关注的是，“AI 应用高效”是所有二级指标中得分率最低者，其测评维度涵盖 AI 对传统

研究范式的突破、AI 应用与研究问题的适配性及研究流程的可迁移性,而这也恰恰是本次“AI 一作”实验所倡导的“AI 驱动”核心愿景。该指标的薄弱表现说明,当前投稿在 AI 赋能教育研究的独特贡献与有效性方面尚未达到预期水平,亦是限制高分论文产出的主要因素。

表 8 “综合榜”审稿结果及各一级指标得分情况

	2.1研究逻辑	2.2方法过程	2.3研究创新	2.4写作质量
总分	22	28	38	12
平均值	14.77	20.64	25.78	10.43
最高	20	26	37	12
最低	0	5	6	0
标准差	2.67	3.15	5.68	1.00
得分率	67.14%	73.71%	67.84%	86.92%

表 9 “综合榜”审稿结果及各二级指标得分情况

一级指标	二级指标	总分	最高	最低	标准差	得分率
2.1研究逻辑	2.1.1研究问题明确	8	7	0	0.98	74.17%
	2.1.2理论基础适切	7	6	0	0.97	65.69%
	2.1.3文献综述扎实	7	7	0	1.11	60.52%
2.2方法过程	2.2.1研究设计科学	9	8	3	0.93	78.59%
	2.2.2数据来源可靠	10	10	0	1.51	71.73%
	2.2.3分析推理严谨	9	8	2	1.21	71.06%
2.3研究创新	2.3.1学术贡献显著	20	19	6	1.99	75.30%
	2.3.2 AI 应用高效	18	18	0	4.44	59.55%
2.4写作质量	2.4.1结构完整清晰	6	6	0	0.52	96.27%
	2.4.2语言表达准确	6	6	0	0.55	77.58%

2.基于创新性指标的大模型专项评审

在本次实验中,基于创新性指标的大模型专项评审摒弃了量化评分的路径,而是构建了以“学术贡献显著”这一论文创新性指标为核心的动态相对排序算法。其设计基于对大语言模型认知特性的判断:大语言模型在处理高维复杂学术文本时,标量评分因校准困难而稳定性不足,但在全量对比与相对排序任务中,模型通过上下文对比识别视角独特性、方法论突破及理论深度上细微差异的能力更为敏锐。因此,本轮评审以“相对排序优于绝对评分”为核心假设,将评审任务的本质定义为一个大规模的偏序关系构建问题。

考虑到对全部论文直接进行排序面临的物理约束:主流大模型上下文窗口有限,如 DeepSeek-V3.2 最大输入 96k tokens,而单篇高质量论文常超 10k tokens,全量输入不仅会触发上下文溢出,更会导致注意力分散及“上下文腐烂”(中间信息遗忘或幻觉)。为此,本轮评审基于经典的外排序思想与 K 路归并算法,构建分层级、抗遗忘、动态补位的“锦标赛式”评审框架。

(1)基于创新性指标的大模型专项评审机制设计

一是整体架构。为了在有限的算力并发与搜索空间内快速计算结果,本轮评审采用“局部选优—全局合并”的漏斗型架构,将大规模排序降维分解为一系列局部对比问题。评审流程逻辑上划分为三个递进阶段:全域小组创新定序(基础层)、区域汇聚优选(中间层)与滚动决赛归并(决策层)。该架构既突破了上下文窗口的物理约束,又引入类似 CPU 指令流水线的设计思想,以实现高并发处理(图 11)。此外,系统引入“动态补位机制”,从算法层面规避传统分组淘汰制中“死亡之组”导致次优解过早出局的

不公现象，即强强对话时局部极值压制全局极值的情况，确保评价结果的公平性。

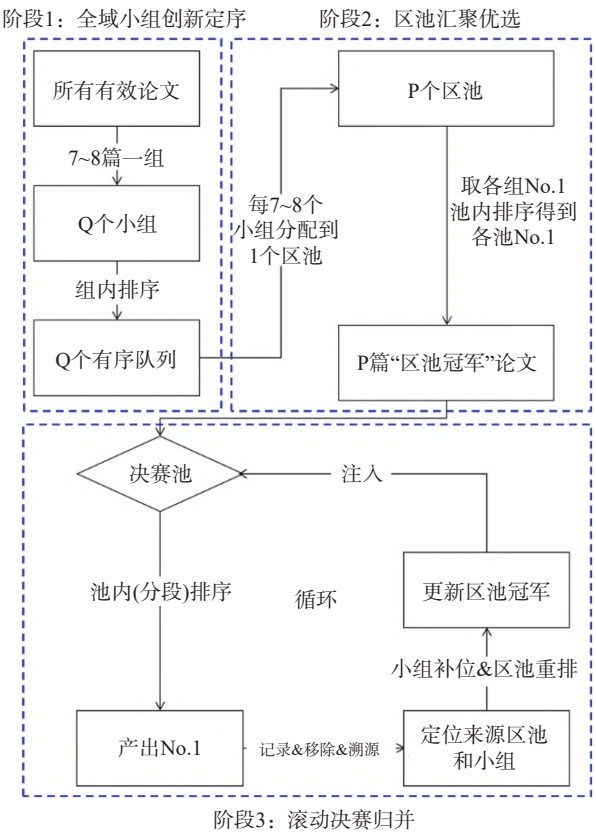

图 11　基于创新性指标的大模型专项评审架构

二是排序过程。全部论文预先经 AI 综合评审系统的前置解析与伦理审查模块处理，随后进入三阶段专项排序评审。具体如下：阶段一：全域小组创新定序（基础层），在模型上下文窗口内对论文全集进行高精度分组排序，将无序集合转化为若干局部有序序列，完成初始化定序。阶段二：区域汇聚优选（中间层），大量底层有序小组到最终有序结果需经历多轮排序，其间设置“赛区池”作为中间缓冲层。系统并不合并所有小组，而是提取各小组“当前最佳创新候选”汇入赛区池，经多轮排序产生各赛区“创新冠军”，直接注入最终决策流程。阶段三：滚动决赛与动态补位（决策层），作为核心算法层，采用动态优先队列逻辑模拟永不枯竭的 K 路归并排序器，以此循环生成 Top N 榜单。此阶段通过持续补入次优候选，保证最终排序的完备性与公平性。

三是关键技术。系统采用算法逻辑与大模型基座独立解耦的设计，在实施层面构建六路并行智能体工作流，将同一排序机制分别部署于六款活动期间国内外具有代表性的前沿大模型，包括 Doubao-seed 1.6、DeepSeek-V3.2、Qwen3-Max、Kimi-K2 thinking、GPT-5.2 及 Claude Sonnet 4.5。通过这种并行部署策略，每一款模型都作为独立智能引擎，完整执行从海量论文输入到 Top N 创新名单输出的全过程。这一设计不仅保障了评审产出的高可用性，也为后续在同一算法框架下观察与比较不同模型的行为差异奠定基础。

（2）基于创新性指标的大模型专项评审（“单项榜”）结果

“单项榜”评审聚焦论文创新性，即“学术贡献显著”指标，意指研究在视角、理论、方法或应用等方面具有原创价值。具体操作上，若论文观点在已有文献中已获充分讨论，或仅为既有共识的组合应用，则不属创新；若有突破性洞见，则可给 18 ~ 20 分。如前文所述，六个大模型以两两比对排序方式各评出创新性 TOP50 榜单。在此基础上，活动组委会统计了每篇论文在六个大模型的 TOP10 和

TOP30 榜单中的上榜次数,整理出创新性排名前 25 位的“单项榜”(表 10)。

表 10 “单项榜”前 25 名论文在不同大模型上的上榜情况

投稿ID	GPT榜排名	Claude榜排名	Doubao榜排名	Qwen榜排名	DeepSeek榜排名	Kimi榜排名	前10榜上榜次数	前30榜上榜次数
AIDW2025-771	3	2	1	2	3	5	6	6
AIDW2025-742	1	6	7	5	4	2	6	6
AIDW2025-348	7	18	15	4	7	3	4	6
AIDW2025-763	无	1	9	1	16	9	4	5
AIDW2025-273	20	无	6	6	2	1	4	5
AIDW2025-476	9	5	16	11	10	无	3	5
AIDW2025-790	30	无	10	无	1	4	3	4
AIDW2025-507	4	无	8	28	13	11	2	5
AIDW2025-798	无	10	18	9	20	22	2	5
AIDW2025-718	35	24	3	12	8	30	2	5
AIDW2025-783	14	23	5	10	14	无	2	5
AIDW2025-426	33	24	41	7	5	14	2	4
AIDW2025-447	6	无	无	34	6	18	2	3
AIDW2025-812	10	26	无	无	9	无	2	3
AIDW2025-137	39	7	25	45	49	6	2	3
AIDW2025-626	无	8	无	无	21	8	2	3
AIDW2025-707	45	无	4	3	无	无	2	2
AIDW2025-185	5	无	24	21	无	15	1	4
AIDW2025-216	47	4	无	17	29	28	1	4
AIDW2025-553	25	22	40	8	无	16	1	4
AIDW2025-219	无	17	无	19	12	7	1	4
AIDW2025-517	24	无	无	26	15	10	1	4
AIDW2025-791	无	无	2	20	37	无	1	2
AIDW2025-813	无	3	无	25	无	无	1	2
AIDW2025-220	49	9	无	无	无	无	1	1

注:表中所标注“无”是指该论文没有出现在大模型评出的“创新论文”TOP50榜单中。

对“单项榜”上的 25 篇论文进行质性分析,可以发现,大模型眼中的“创新论文”具有以下特征:

一是研究视角从“技术赋能”转向“教育本体反思”。此类论文超越 AI 工具性讨论,深入教育哲学、认知机制与社会结构的反思性审视。相比之下,多数在 AI 初审(“综合榜”)中被评为低分论文的文章,多是对 AI 在教育领域某方面或全方面应用的构想,但体现出的却是一种“unimaginative imagination”(没有想象力的想象)。正是因为突破了“构想技术如何赋能未来教育”的常见研究视角,“单项榜”上的论文才能在针对创新性的专项评审中脱颖而出。例如,《留白教育学:顶尖科学家在制度性无聊中的认知自救》从“无聊”情绪切入,重构其作为“认知保护机制”的教育意义;《AI 时代教育之悖论:当成绩不再反映能力》以经济学模型揭示“成绩—能力背离”,从教育信号失真角度批判 AI 对评价体系的根本冲击。

二是研究方法呈现跨学科融合与计算教育学范式。上榜论文多采用数据驱动、模型仿真、因果推断等量化方法,融合多学科工具。例如,《“最优学习风格”不存在——大规模学习分析证明 AI 定制化教学无法超越通用优质教学》以“双重机器学习+因果森林”方法处理超过 8 万名学生的数据;《基于多智能体模拟的教师轮岗政策效应研究:来自“教育生态系统”模型的仿真证据》采用多智能体模拟构建“教育生态系统”,模拟推演教师轮岗政策的长期效应,体现计算仿真在教育政策研究中的创新应用。

三是AI角色从“研究工具”扩展为“研究对象”与“研究主体”。上榜论文中AI兼具研究对象、协同研究者与仿真环境等多重身份，呈现出“用AI而非论AI”的特征，此为大模型区分创新论文与普通论文的重要判据。例如，《合成数据驱动的教育实验：利用大语言模型智能体模拟未来课堂互动的可行性研究》中AI兼具合成智能体、数据引擎与第一作者的“三位一体”角色；《人机共研：生成式人工智能时代的数字扎根理论构建》提出“数字扎根理论”，将AI视为平等协同研究者，在“双轨涌现模型”中实现人机理论共建。

四是在研究的创新点上体现出方法论突破与理论建构二者并重的特征。上榜论文在方法论上尝试范式创新，同时提出具有解释力的新理论框架。例如，《多大模型平行复现Hulleman（2009）经典实验：统计趋同、解释分化与互审冲突》通过多模型平行复现与AI互审，设计“极端元实验”，实现对AI科研一致性与冲突性的自反性检验；《中国教育学学科范式演进的内生逻辑与外部驱动——基于重大教育政策的准自然实验研究》构建“范式成熟度指数”（PMI），量化“内生逻辑”与“政策驱动”的贡献比，实现理论量化与机制可视。

3.人机评审一致性分析

对接受AI审稿系统与两位人类专家独立盲审的论文评审结果进行对比，得到以下发现：

一是AI与人类专家对论文质量档次的判断具备较高的一致性。在83.33%可用于直接对比的论文中（其余论文因人类专家意见分歧较大，后文单独分析），AI审稿系统与人类专家对论文档次判断的一致性达76%。在识别A档（优秀）与D档（较差）论文时，人机评审一致性超过80%。这表明AI审稿系统能有效辅助完成初步筛选与分流，快速辨识出高质量与不合格的稿件。评审意见不一致之处表现为AI评分普遍高于专家平均分，尤其在论文结构严谨、逻辑清晰、格式规范的稿件上，AI审稿系统倾向于给予更高评价。

二是AI与人类专家对论文实质创新的判断存在差异性。在12篇AI判为A档而专家判为D档的极端差异论文中，评审意见呈现如表11所示的系统性分歧。典型案例如投稿ID为AIDW2025-028的论文（AI评分：88分/A档，专家评分：56.5分/D档）。AI肯定其“构建了可操作的实施框架”，专家则批评其“更像一个教案，研究贡献薄弱”。此分歧体现了AI重“解决方案完整性”与专家重“学术理论贡献”的评价取向差异。

表11　AI审稿系统和人类专家的审稿侧重点对比

评价维度	AI审稿系统的侧重点	人类专家的侧重点
结构与规范性	逻辑链条完整、研究方法描述清晰、写作格式规范。	视为基础门槛，不构成核心加分项。
价值与创新	肯定研究的实践应用价值、技术实现难度与系统性。	强调理论原创性、概念的突破、对学科知识体系的实质性推进。
数据与证据	关注数据规模、分析模型的复杂性与技术新颖性。	审视数据来源的真实性、样本的代表性、研究设计的效度及伦理合规性。
审稿意见表述	偏好系统化、结构化的“优势—问题—建议”式评价。	倾向指出最关键、最深刻的少数问题，并进行启发式、对话式的深度点评。

三是人类专家之间审稿意见更加凸显多元性与复杂性。在16.66%因两位人类评审专家审稿意见分歧过大而无法用于人机一致性对比的论文中，两位专家评分的最大分差甚至高达87分，其分歧主要源于学术价值观的类型差异，如表12所示。在上述分歧案例中，AI评分更接近高分专家，因其同样重视论文结构完整性与问题明确性。

四是相比之下AI审稿具有以下核心优势：其一，效率与稳定性，可实现秒级评审，无疲劳与情绪波动，在大规模稿件处理中优势显著；其二，规范性质控，对格式、基本逻辑与结构完整性把关严格，有助于提升学术出版的基础质量；其三，结构化分析，可系统梳理论文优缺点，提供条理清晰的评审报告框架；其四，分流与聚焦，可靠地筛选优质与劣质稿件，使专家得以聚焦中间档位及争议稿件的深度评审。

表 12 人类专家审稿意见分歧的主要类型

分歧类型	代表案例	核心冲突点
严谨派 VS 宽容派	AIDW2025-742 (35分 VS 81分)	方法论的绝对严谨性、问题的现实重要性及结论的启发性之间的权衡。
理论派 VS 技术派	AIDW2025-763 (41分 VS 70分)	理论建构的扎实性、实证支撑、研究视角的新颖性、方法论尝试之间的侧重。
学科本位派VS 整合应用派	AIDW2025-764 (48分 VS 69分)	本学科（如教育学）核心标准（如测量效度）与跨学科技术应用的可行性与价值之间的评判。
实质贡献派 VS 形式规范派	AIDW2025-783 (33分 VS 60分)	对理论机制有无根本性创新、是否有研究问题意识、论述结构是否清晰的不同期待。

综上，人机评审一致性分析结果表明，当前 AI 审稿系统已展现出作为学术评审“增效器”和“稳定器”的显著潜力，其与人类专家在 76% 的情况下达成共识，证明了其在整体质量判断上的实用性。然而，二者在“形式规范”与“实质创新” 评价权重上的根本性差异，以及人类专家内部存在的多元价值判断，共同揭示了学术评价工作的复杂本质，同时也揭示了人机评审的互补性——AI 在格式规范、逻辑结构、语言表达的识别上具有优势，而人类专家在价值判断、创新识别、学术敏感性方面仍不可替代。进一步分析，实际上 AI 并不具备审美的能力，其评判是人的价值判断和审美在 AI 世界的投射。也就是说，大模型不是我们的同类，而是我们的文化档案馆、语法模拟器、潜在观点的排列组合机。即使大模型的“价值判断”与“审美表现”确实共享了统计模仿的逻辑基础，但它们与人类真正的价值体验和审美体验之间，存在无法跨越的本体论鸿沟。上述发现，为构建“AI 初筛—人工精审—共识裁定”的三级评审模式提供了实证依据。未来需通过流程、系统与标准的协同优化，构建更高效、公正且富有洞察力的学术评审机制。

（八）不同大模型的科研品味有何差异？

在论文评审阶段，不同大模型能否输出一致的评审结果？国内外大模型在审稿偏好上是否存在分化？此类问题在评审过程中受到广泛关注。基于两轮评审所保留的完整过程数据与结果数据，以下从评审表现的共性与分歧、优质论文认知的分化及其形成机理三个层面展开分析。

1.评审表现的共性与差异

为评估不同大模型的评审表现，研究从与第一轮一致性、共识论文覆盖度、高分识别率、独特偏好度和排名稳定性五个维度，对第一轮基于多维指标的综合评审数据与第二轮基于创新性指标的专项评审数据进行比较。结果表明：

Qwen 与 DeepSeek 在五个维度上表现最为均衡。如表 13 所示，Qwen 的 TOP50 榜单中论文在第一轮获得 80 分以上的比例最高；Qwen、Doubao、Kimi 与 DeepSeek 的 TOP50 榜单论文出现在其他至少四个榜单中的比例较高；第一轮高分（88 分及以上）论文进入 Qwen、Kimi 与 DeepSeek Top50 榜单的比例亦最高；高共识论文在 Qwen 与 DeepSeek 榜单中的排名稳定性最佳。

如图 12 所示，国内外大模型在“独特偏好度”维度上差异显著，尤其在“创新性”等主观维度评价上体现出明显的模型风格分化。相较之下，Claude 与 GPT 的 TOP50 榜单中存在一定比例仅在该模型位列前 10、而在其他模型榜单中未进入前 30 的论文，呈现出较高的独特偏好度。

基于多维指标的综合评审与基于创新性指标的专项评审，这两种评审方案的整体结论具有趋同性。统计第一轮 80 分及以上的各分数区间的论文在第二轮中进入六种大模型 TOP50 榜单的比例（以此衡量有多少比例论文被至少一个模型认可），以及各分数区间的所有论文（包括未进各模型 TOP50 榜单的）在每个模型榜单中的排名平均值，如图 13 所示，两种评审方案的分数段分布趋势和区间基本一致，也就是说，大模型通过“打分法”和“排序法”评价的结论趋同。此外，第一轮高分段论文（≥88 分）进入各模型 TOP50 榜单的比例最高且平均排名最优，高分段论文平均排名为 13.32，显著优于中分段（83~87 分）的 28.64 与低分段（≤82 分）的 33.56。由此可见，基于多维指标的综合评审与基于

创新性指标的专项评审总体上呈现显著的一致性趋势，说明模型间具备一定的共识基础。

表 13　六种大模型在不同维度的表现

评价维度\模型	Qwen	Doubao	Claude	GPT	Kimi	DeepSeek
与第一轮一致性	0.72	0.65	0.68	0.63	0.70	0.69
共识论文覆盖度	0.83	0.75	0.67	0.58	0.75	0.83
高分识别率 (≥88分)	0.80	0.70	0.60	0.70	0.80	0.80
独特偏好度	0.10	0.15	0.20	0.25	0.10	0.15
排名稳定性	0.75	0.70	0.65	0.60	0.70	0.75

注：与第一轮一致性即各模型评出的创新性TOP50论文中在第一轮综合评审获得80分以上的比例；共识论文覆盖度即高共识论文(≥4个模型认可)进入该模型TOP 50的比例；高分识别率即第一轮88分以上论文进入该模型TOP50的比例；独特偏好度即该模型高度认可(TOP10)但其他模型未将其排入TOP30的论文比例；排名稳定性即高共识论文在该模型中的排名离散度倒数。

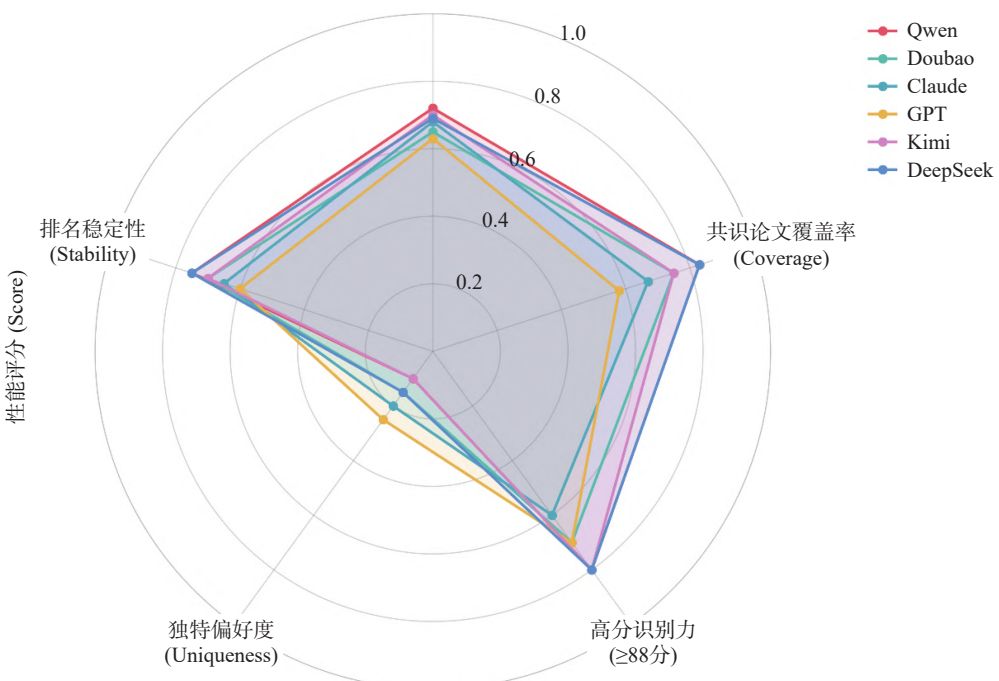

图 12　六种大模型的评估特征雷达图

注：各轴代表五项评价维度，不同颜色代表不同的大模型，数值越高表示该维度表现越好，覆盖面积越大表示模型综合表现越好。

2.对优质论文的认知分化

尽管模型间存在上述共识基础，但对 15 篇高共识论文在 Qwen、Doubao、Claude、GPT、Kimi、DeepSeek 六个模型中排名的进一步考察，结果如图 14 所示，可以发现，部分论文的评价结果仍存在显著差异，反映出不同模型对优质论文认知的权重分配与风格取向存在分化。

在评审指标权重方面，以投稿 ID 为 AIDW2025-790 的论文为例，该论文在 DeepSeek 榜单位列第 1，在 GPT 榜单位列第 30，在 Qwen 与 Claude 榜单中未进入 TOP50。此类差异主要源于三方面机制：一是跨学科题材若领域关键词指示性不足，部分模型在初筛环节易生漏判；DeepSeek 更注重论证的内在逻辑与规范细节，而 GPT 受安全对齐策略影响，对前沿方法推广性的评估更为审慎。二是 DeepSeek 在推理链与结构化审稿方面具有优势，对“问题—方法—数据—结论”自洽性突出的论文显著加分；Qwen 与 Claude 则更侧重语义匹配与主题契合，对表达不够“典型”的题材可能形成压制。三是在创新

性与规范性的权衡上，DeepSeek 对学术伦理与规范维度关注度较高，对规范完备但创新表达克制的论文给予稳定加分；GPT 则更强调国际通行的影响力与原创性证据，若论文国际文献锚点较弱，排名易受影响。

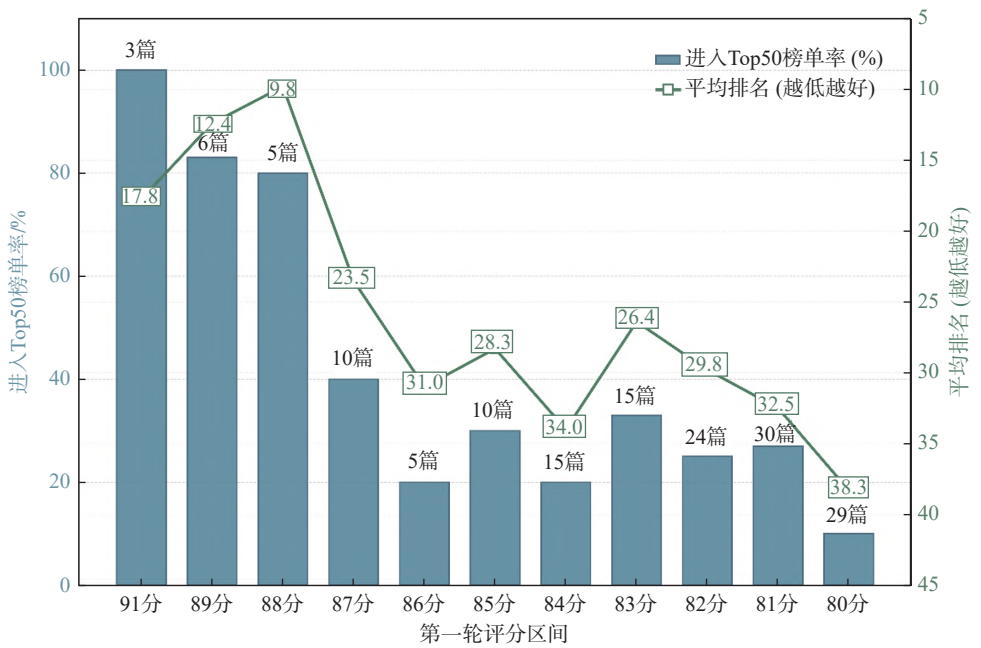

图 13 不同分数区间论文的表现

注：蓝色柱子代表进入各大模型创新性排名TOP50 的比例（越高越好）；绿色折线代表平均排名（数值越小越好，Y 轴已反向）；顶部数字代表该分数区间的论文总数。

在学术风格认知方面，以投稿 ID 为 AIDW2025-137 的论文为例，该论文在 Kimi 榜单位列第 6、Claude 榜单位列第 7，而在 GPT 榜单位列第 39，在 Qwen 榜单中未进入 TOP50。此类差异与大模型的语料域分布、长文本理解能力及论证风格偏好密切相关：Kimi 经中文场景优化，对贴近中文学术共同体的引用与术语更具识别优势；GPT 依赖英文语料的广泛先验，面对中文语境的细微语用与政策关联表达时评分趋于保守。Kimi 与 Claude 的长上下文处理能力较好，对论证链条长、叙述细致、规范到位的论文更为友好；Qwen 对主题聚焦与关键词显著性较为敏感，若标题与摘要未充分彰显核心贡献，可能被排除在 TOP50 之外。此外，Claude 倾向奖励可解释、严谨、保守的论证风格，而 GPT 对创新信号的证据链要求更为严格，若示例与数据不足以支撑可推广性主张，则下调排名。此类差异并非简单的评判误差，而是反映了学术评价内在的多维属性，本身可作为具有研究价值的补充视角。

3.导致评审差异的原因分析

大模型评审差异的形成可归因于训练语料、优化目标、提示词适配与模型内在随机性等多重因素的交互作用。

第一，训练语料与语域差异。国外模型英文语料占比通常超过 80%，国产模型中文语料占比多在 60% 以上，后者在本土政策与教育语境处理上具有相对优势；语料中学术数据库覆盖范围（如 CNKI、WoS 等）的差异，亦会影响文献综述全面性与国际影响力相关评分。不同社会规范与风控策略还导致模型在“创新性风险”与“伦理规范”维度上的权重分化，国外模型在安全边界与可推广性证据上趋于保守，国产模型对本土政策指向与现实问题契合度则更为敏感。

第二，优化目标与评审重心差异。部分模型（如 DeepSeek）对学术伦理、逻辑自洽等维度更为友好，排名稳定性较高；而另一些模型则对原创性与国际影响力的证据链要求更高。同时，大模型普遍具

有程式化、全面化输出的倾向，在创新性深度与个性化建议方面存在共性短板，但不同模型的程式化程度各异，直接影响其对非常规表达或跨范式论文的评分。

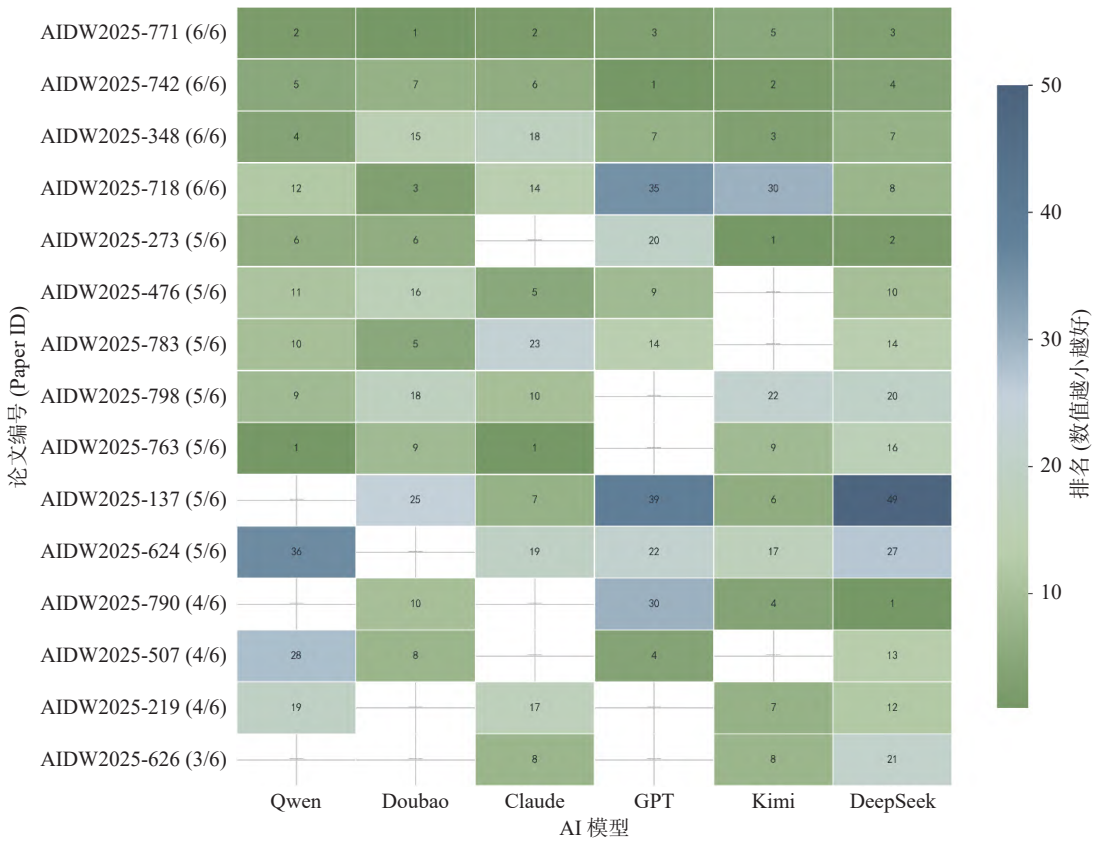

图 14　六种大模型对 15 篇高共识论文的排名热力图

注：论文按共识度（出现在各模型TOP50榜单中的次数）从高到低纵向排序。每个单元格显示模型给出的具体排名（1–50），颜色编码从深蓝色（较低排名）渐变至绿色（最佳排名）。空白单元格表示该论文未进入相应模型的TOP 50。左侧边栏显示每篇论文的共识度评分（例如“6/6”表示被所有六个模型认可）。

第三，提示词敏感性与框架适配度。评审任务的角色定义、目标设定与格式规范是否与模型的指令风格匹配，会造成“同文不同分”现象；结构化提示（如BROTF）与精细的智能体工作流可显著提升逻辑维度评分的一致性。此外，跨学科论文若缺少领域关键术语，部分模型在初筛环节发生“错杀”的概率上升；以主题语义匹配为主的模型偏好典型表达，而以逻辑链与规范分析为主的模型更能识别非典型表达中的优质论文。

第四，随机性与黑箱因素。大模型基于概率分布采样（如Top-p、Temperature等）生成文本，相同输入可能产生不同输出；加之海量参数与非线性结构带来的决策过程不透明性，其不可见的“奖励模型”和偏好复杂性使得最终评审结果难以完全预测。

（九）哲学社会科学研究中是否存在“AI鸿沟”？

针对征文活动参与者的焦点小组访谈结果显示，不同年龄和学术资历的研究者在AI应用的态度与习惯上存在显著差异，即“AI代沟”。“AI代沟”是“AI鸿沟”的一种。在学术研究领域，“AI鸿沟”主要可以划分为接入鸿沟（Access Divide）、素养鸿沟（Literacy/Capability Divide）和效用鸿沟（Outcome Divide）三个维度（Draux, 2024; 徐丽芳, 2025）。然而，在此次“AI一作”大型社会实验中，组委会进一步发现：AI鸿沟不仅存在于国家、机构、资源层面，更呈现出鲜明的代际特征。横亘在学术界的AI鸿沟，并非单纯的“技术接入”或“资源占有”差异，而是一种深层嵌套于年龄和学术资历中的“AI代沟”

(AI Generation Gap),即不同代际的人群在面对人工智能技术时,在认知观念、信任程度、依赖模式以及伦理态度上产生的显著差异。在本次社会实验的语境下,“AI 代沟”具体表现为不同年龄层的研究者对 AI 应用于科研之后的“主体性地位设定”“知识权威让渡”以及“传统学术训练存废”三个方面的根本性分歧。结合本次活动对不同年龄层作者的深度访谈,我们可以初步描绘出这条学术“AI 代沟”的三级样态:

1.资深学者群体:警惕的“驯兽师”与坚守的“甲方”

以多位教授、副教授为代表的高校教师群体,体现了典型的“数字移民”特征。他们拥有深厚的学术积淀,因此在与 AI 的互动中绝对捍卫人类的学术主导权与批判意识。在对 AI 进行角色定位时,他们将 AI 视为需要“长期驯化”的合作者、执行指令的“乙方”或需要喂养的“大一学生/徒弟”。在与 AI 互动的模式中,强调“人机教学相长”与严格把控。

例如,投稿 ID 为 AIDW2025-811 的论文作者之一 D 老师将 AI 定位为合作者而非单纯工具,强调人机协同是一个“双向奔赴”的过程。他认为 AI 相当于“硬拉到班上很厉害的大一学生”,需要通过“喂养”优质文献(如专著、理论框架等)建立思维支架,才能展开有效学术对话。在写作中,AI 的核心价值在于效率与快速迭代。它能够在短时间内将模糊想法结构化为逻辑自洽的文本,充当供人类反思修改的“靶子”,极大地缩短了从构思到成稿的周期。然而,他深刻指出 AI 的算法局限性:内置“讨好型”偏向、缺乏主动批判性思维、容易生拉硬扯迎合用户观点,且存在幻觉(如虚构文献)。因此,人类研究者必须承担方向引领者(提供历史深度与跨学科视角)、质量把控者(核实文献、审查逻辑)和批判者(主动质疑 AI 输出)的角色。他认为当前 AI 尚难产生“爆炸性创新”,因其知识“广而不深”;真正的创新需要基于长期深度协同(“从大一教到硕士”)、跨学科团队介入及人类独特的理论连接能力。

另一个典型的案例来自投稿 ID 为 AIDW2025-583 的论文作者之一 X 老师。她构建了与 AI 的双重关系框架:一方面,AI 是掌握海量信息的“乙方”,人类专家作为“甲方”进行多轮审核修正;另一方面,AI 又像“徒弟”,通过接收人类提供的正反例和权重调整快速学习进化。在科研中,AI 的核心作用是提升建构效率——例如用 AI 替代传统德尔菲法构建评价指标框架,通过快速预测—反馈—修订循环,将研究周期从传统的 1.5 年压缩至半年。她特别看重 AI 在跨版本比较(如 GPT-3/4/5 差异)和细节完善方面所提供的启发,认为这超越了人类个体的认知盲区。但她也明确指出 AI 在专业判断上的局限:对学科本体知识、学情分析等教育现场的理解不够准确,需要人类基于一线经验进行“人工校准”。在署名问题上,她认为 AI 作为“第一作者”更多具有象征意义——承认其工作量,但人类必须承担最终责任(类似通信作者角色)。

总的来看,资深学者群体虽看重 AI 的“建构效率”,但对其产生“爆炸性创新”的能力持怀疑态度。他们固守传统学术的严谨性,认为人类的理论连接能力、社会关怀和跨学科视野是 AI 无法替代的核心壁垒。

2.青年研究者群体:理性的“项目经理”与严格的“审查官”

以博士研究生为代表的青年研究者群体,处于学术代际的过渡地带。他们虽精通技术,但同样受过严格的传统学术规训。在对 AI 进行角色定位时,他们表达出的工具主义色彩浓厚,将 AI 视为“能干活但不靠谱的学术伙伴”和“效率工具”。在与 AI 互动的模式中,呈现出清晰的“人机边界划分”。

例如,投稿 ID 为 AIDW2025-707 的作者 Z 博士采取更为理性的工具主义立场,将 AI 视为“能够干活但不靠谱的学术伙伴”和“编程助手”。她承认 AI 基于概率模型的本质局限:无意识、无创造性思维、可能编造信息,因此人类必须承担严格的审稿人角色。在科研流程中,她明确划分人机边界:人类负责核心假设、理论模型构建、因果推断与价值判断;AI 负责头脑风暴(拓展变量与机制)、代码生成、统计检验解读、语言润色与格式检查。她认可 AI 在效率提升(研究周期从 1 个月缩短至 1 周)和语言转化(将通俗表达转为学术黑话)方面的价值,但质疑其产生“爆炸性创新”的能力,认为当前

AI只是实现经验迁移与组合，难以突破学科认知边界。与数字“原住民”对AI的高度信任不同，她强调人类独特价值在于社会关怀、伦理判断、对现实政策环境的微妙感知，以及学术责任感。她指出AI语料库自带社会偏见，需要人类价值观的纠偏。对于未来学术训练，她强调基础学术训练与一线实践积累仍是驾驭AI的前提。

总的来看，相比资深学者，青年研究者使用AI的频率更高、颗粒度更细，但相比本科生，他们保持了强烈的学术警惕性，强调人类在“伦理判断和政策感知”上的不可替代性。

3.数字“原住民”学生群体：拥抱“新权威”的“人机共生者”

以00后的本科生为代表的新世代群体，展现出了跨越传统学术认知的颠覆性倾向，他们也是“AI代沟”中最具冲击力的一极。在对AI进行角色定位时，他们认为AI不再是工具或徒弟，而是超越人类权威的“知识基础设施”和“高级伙伴”。在访谈中甚至有本科生作者直言“AI比传统教授更可靠”。在与AI互动的模式中，这一群体展现出对AI极高的信任与主导权的让渡，对于AI提供的陌生知识选择直接学习而非质疑，摒弃了传统的对抗性“批判思维”，而这也是“数字移民”一辈表示担忧的现象。

例如，投稿ID为AIDW2025-753的作者之一Z同学就是一位数字“原住民”，她在访谈中展现出对AI的高度信任与无隔阂使用。她将AI定位为“图书馆”（需要明确指令才能精准检索）和“高级伙伴”，并认为AI比传统教授更可靠（“社科类论文很多只是逻辑自洽，AI也能做到，且提供更多专业术语”）。在她的科研流程中，AI承担了从选题、框架设计、研究方法选择到数据生成的全流程主导工作，人类仅负责核验数据真伪（因AI会产生幻觉）、设计涉及人际观察的问卷/访谈，以及伦理审查。她强调使用AI需要极致拆解任务（每1 000字为单位迭代，而非一次性生成万字），并通过角色扮演（如设定为教育专家）优化输出。与成熟研究者不同，她不追求“批判性思维”的对抗性，而是将AI视为认知扩展工具——当遇到AI提供的陌生方法（如M值计算）时，她会选择自主学习而非质疑AI。

此外，数字“原住民”群体还对传统学术训练体系提出了强烈的质疑。以上述Z同学为例，她认为AI已深刻改变学习方式：传统“从基础层层垒砌”的知识体系已不可行，必须转向“问题导向的碎片化学习”。对于未来学术训练，她认为应大幅压缩传统课程，增加AI底层逻辑与提示词工程培训。他们不再追求耗费大量时间从零垒砌知识，或者开展在当下这个时代“意义并不大”的手写教案等传统训练，而是主张用AI来节省学习和工作的时间，将精力转移至AI无法涉足的、能够体现“人之为人”的身体与情感能力的提升。

本次实验中，我们还得到一个与“AI代沟”相关的有趣的发现：如图15所示，在征文活动中上榜的论文里，人类作者（排序首位/通信作者）为高校学生（含本硕博）的占65.9%。也就是说，AI应用可能在一定程度上推动了哲学社会科学研究中的“智慧平权”。至于“AI代沟”是否导致了科研工作效率、研究方法路径以及研究创新结果的代际差异，进而推动了“智慧平权”的现象，还有待后续进一步探究。

综上所述，本次“AI一作”大型社会实验发现：哲学社会科学研究中的“AI鸿沟”已经逐渐形成，且最突出的表现形式为“AI代沟”。这条代沟的本质，是传统知识权威与新型计算智能之间的碰撞。资深学者凭借深厚的理论功底，将AI框定为提效的“工具”；而年轻一代则凭借天然的技术亲近感，将AI视为重塑认知路径的“基建”，甚至出现了对传统学术训练体系的解构倾向。面对这一现状，我们不能简单地用“对错”来评判代际差异。资深学者的批判性审查是坚守学术底线、防止“AI幻觉”与学术泡沫的压舱石；而年轻一代的无缝协同与颠覆性思维，则是突破哲学社会科学研究范式、激发AI潜能的催化剂。未来学术共同体的建设，亟需弥合这一“AI代沟”：既要将数字“原住民”的前沿提示词工程与协同模式反哺给资深学者，消除“效用鸿沟”；又要将资深学者“人之为人”的批判精神、伦理关怀与跨学科底蕴，内化为年轻一代驾驭AI的思想支架。唯有实现人机协同在代际间的“双向奔赴”，哲学社会科学研究才能在智能时代迎来真正的繁荣。

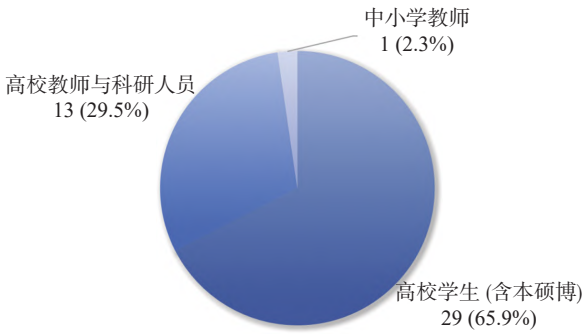

图 15 上榜论文人类作者（第一顺位）的身份类型

（十）征文活动的获奖作者是如何驾驭 AI 的？

对 25 位“上榜论文”作者的经验反思报告进行综合分析，能够发现获奖作者们总结出的人机协同科研技巧兼具特殊性和一般性价值，可以在不同的研究范式中迁移。以下提炼出八个具有代表性的研究技巧，并辅以具体案例加以说明。

1.指令精准——合理设计和使用系统级提示词

高质量的系统级提示词能够精准引导 AI 的输出方向，其设计遵循“脚手架”原则，本质在于为生成过程搭建专业规范框架。将模糊化、开放式的随意提问升级为结构化、工程化的系统指令，可显著降低模型输出的随机性与不确定性。哲学社会科学研究涉及复杂的价值判断与多元的理论视角，AI 输出易于泛泛而谈，系统级提示词能够有效减少虚构风险。

系统级提示词的设计应包含五个维度：一是角色赋予，即为 AI 赋予具体、专业且稳定的身份定位，激活模型在特定知识领域的专业视野；二是情境构建，在指令中注入研究背景、理论框架与问题语境，帮助 AI 理解任务所处的学术脉络；三是约束设计，通过正向要求与负向限制相结合划定思维边界，防止发散跑题；四是格式规范，明确规定输出的结构形式与组织方式，确保生成内容具有良好的可读性；五是文本凝练，在确保信息完整的前提下精简指令长度，降低模型的认知负荷。例如，《面向音乐教学设计的大语言模型评测研究》一文采用“角色—任务—注意—输出格式”的提示框架，明确模型身份、教学情境与文本规范，显著减少了生成内容的随意性。在构建评测指标时，研究者进一步要求 AI 提供正反例示范，以具体案例帮助模型理解抽象指标的操作内涵，从而提升了自动标注与文本生成的准确度。此外，指令中还明确规定了指标面向义务教育阶段单课时音乐教学设计，而非大单元或跨学科教学设计，有效划定了任务边界。

2.化整为零——分步拆解任务，分阶段引导 AI

将复杂的研究任务拆解为若干相对独立、逻辑递进的子任务，是适应 AI 功能特征的关键策略。大语言模型在面对复杂、多维度、长链条的任务时，容易出现“注意力稀释”现象，导致某些维度被忽视或浅表化处理。分阶段引导不仅能提升单阶段输出的专业度，更创造了人工干预的“介入点”，便于研究者在阶段转换时审视偏差、纠正逻辑漂移。分阶段引导包含三个环节：一是任务拆解，分析研究流程的自然划分，既包括大任务拆分，也包括大环节内部小任务的适当拆分；二是阶段适配，根据不同阶段的核心需求调整提示词设计，如理论构建阶段强调概念界定，案例分析阶段注重深度解读；三是节点审核，在阶段交接处设置“质量闸门”，确认成果达标后再行推进，并适时进行“状态复位”以确保后续生成与前文保持一致。例如，《赋能何以失灵？基础教育阶段师生 AI 素养联动与学段差异研究》将研究划分为选题构思、文献综述、实验设计、主体撰写及润色审校五个阶段，通过原子化拆解克服了长对话中的“注意力稀释”问题。在选题构思阶段，研究者先后引导 AI 进行视角发散、问题聚焦与可行性检验，逐步收缩研究边界；在文献综述阶段，则分步完成文献搜集、核心提炼与演进趋势识别。研究过程中，作者还定期进行阶段性“状态复位”，令 AI 梳理已达成的共识、修改内容及剩余任务，生成

汇总清单以防止长对话中的逻辑漂移。

3.以小驭大——用样本数据驱动AI,本地运行完整分析

重构人机分工逻辑,使AI基于样本数据进行研究设计,能够兼顾安全性与效率性。大规模数据分析面临双重困境:完整数据集体积庞大,直接上传受限于技术接口;原始数据涉及隐私安全与伦理规范,不宜外流。让AI的核心职能从直接处理数据转向基于样本进行方案设计与代码生成,研究者则在本地执行完整数据分析,既发挥抽象推理优势,亦保持数据主权。该技巧的实施包含三个步骤:一是样本构建,从完整数据中提取具有代表性的子样本并脱敏处理,使AI理解数据结构;二是方案生成,提供样本数据与研究目标,引导AI设计分析流程并生成可在本地复现的代码脚本;三是本地验证与迭代,在本地运行代码处理完整数据,将统计结果反馈给AI进行深度解读,形成多轮迭代的质量提升机制。例如,《人机协同视域下的深度认知支架——基于12,824条“学习者—GenAI”多轮对话的滞后序列分析》在处理63.5MB的原始数据时,提取380KB代表性样本交予AI进行方案设计,有效克服了数据规模的技术限制。样本数据在此充当了“数据结构说明书”的角色,使AI得以专注于分析模型选择与代码编写,而非耗费大量算力于数据读取。研究者继而于本地运行AI生成的代码处理完整数据集,确保了分析过程的可控性与伦理合规。作者同时指出,由于AI难以直接获取完整数据且存在幻觉风险,人类研究者仍需在关键环节进行把控。

4.化误为源——将AI幻觉转化为文献检索的新线索

AI生成的“无源之物”亦可成为新灵感的源头。生成不存在的文献或概念是学术研究难以规避的风险,高阶策略是将可疑输出视为启发性线索,通过逆向检索在真实文献中打捞相关研究信息,实现“化误为源”。理解幻觉的生成机理是应用该策略的认知前提:AI的“错误”通常并非凭空捏造,而是对训练数据中相关概念的混合或方向性猜测。需要强调的是,该技巧绝不意味着放松文献真实性核查,研究者须守住学术伦理底线。该技巧的操作包含四个方面:一是识别性质,将可疑内容判定为“方向提示”而非可用内容,区分事实性错误与方向性线索;二是语义转化,基于幻觉概念的语义内核设计多组真实检索词,将模糊的方向感转化为可操作的检索策略;三是打捞文献,通过学术数据库系统检索对应领域的真实权威研究,以人工核实确保引用准确;四是重构框架,将检索所得的真实文献反馈给AI,替代原有幻觉内容,重新整合理论框架。例如,《AI视角下的中小学AI教育实践审视——基于6,662位教师的调查研究》在参考文献审查中发现AI引用的“Su et al.”与“Sanusi et al.”等文献无法在数据库中检索到确切出处,疑似“记忆混合”生成的虚构引用,遂逐一核实后予以删除。研究者进一步以已核实的“肖汶等(2025)”等真实文献替换缺失条目,同时以幻觉中涉及的研究方向为线索拓展检索,补充了具有解释力的权威文献,最终将参考文献从35篇调减至33篇,实现了文献体系的精确与论证力度的提升。

5.逻辑追问——多轮迭代,探析AI生成内容内在逻辑

通过苏格拉底式的连环追问与渐进式迭代,迫使AI显化其推理过程中的隐含假设、价值偏好和逻辑漏洞,此乃深化研究深度、修正认知偏差之关键路径。大语言模型基于概率预测生成文本,其输出实为训练数据中模式的外推,而非真正意义上的理解与推理。单次交互难以暴露其深层问题,唯有通过多轮追问,方能层层剥开AI的思维外壳,触及其生成逻辑的内核。该技巧在文献筛选、理论建构、数据分析等社科研究核心环节具有特殊价值。多轮迭代包含纵向深化与横向修正两个相互交织的维度。纵向深化要求研究者将每一次AI回答视为新的追问起点,通过“为什么”“依据什么”“是否考虑过另一种可能”等问题的连环抛出,穿透表层表述,触及生成内容的认知基础。横向修正则强调渐进式优化,先从宏观框架入手建立共识,再逐步细化具体内容,促成螺旋上升的优化循环。如,论文《从赋能到共生:人机协同生态下教师AI素养的维度重构——基于生成式AI的情境模拟研究》在每个实验单元中安排至少两轮追问,迫使AI显化其决策背后的深层逻辑。这种追问策略一方面将隐含的推

理链条外显化,暴露可能的逻辑跳跃或价值偏见;另一方面训练了研究者的批判性思维,使之习得识别论证薄弱环节、在多元价值中审慎权衡的能力。

6.认知锚点——提供高质量内容锚点,限定任务边界

认知锚点通过为 AI 提供经过筛选的核心理论、权威文献、真实数据或特定案例等高质量知识资源,为其认知活动建立坐标系,有助于确保输出内容的专业性、准确性和针对性。在哲学社会科学研究中,AI 的幻觉与空泛输出具有严重危害:理论引用的错误会扭曲论证基础,政策文本的误读会导致对策失效,案例描述的失真会削弱研究可信度。将 AI 的生成活动锚定于可靠的知识基础和明确的研究框架内,能够有效提高其情境化和专业化能力。该技巧的实施包含锚点供给与边界设定两个相互配合的操作。在锚点供给方面,需为 AI 提供经过研究者筛选和验证的内容资源,如核心理论文献、权威政策文本、真实案例材料或私有研究数据。在边界设定方面,应明确界定研究的主题范围、理论视角、方法取向或格式要求,避免 AI 因理解模糊而偏离目标。这些锚点是被赋予优先权的知识基准,AI 在生成时被要求优先依据这些材料进行推理和表达。例如,论文《高考试题分析智能体理论模型构建与应用路径研究》在多个研究环节应用了这一技巧:在案例推演前为 AI 输入具体的高考试题全文,在撰写文献综述前提供相关高质量文献作为参考依据,从而确保关键步骤的生成具备扎实、准确的依据。通过明确指定理论源和参考材料,研究者实际上是在控制研究的学术谱系和话语资源,避免被 AI 通用训练数据中的偏见或主流话语所主导。

7.操作留痕——保留决策点、分析过程等的完整记录

系统性地保存从原始数据到最终成果的所有中间产物、决策依据、修改痕迹和元数据信息,构建研究的记忆链条和证据体系,在人机协同研究中尤为重要。哲学社会科学领域的部分研究范式本就面临成果难以复现之困,其原因多在于研究过程记录的不完整:分析步骤模糊、决策依据不清,致使其他研究者无法验证结论,甚至原作者亦难以回溯思考路径。AI 的生成过程具有概率性和不可完全预测性,若不对交互过程进行系统记录,研究者将陷入不知 AI 何以如此生成之困境,更无法向学术界证明研究的严谨性。研究的完整记录主要包含数据溯源、过程留痕和元数据管理三个层面。数据溯源须确保每一个处理后的数据点均可追溯至其原始出处;过程留痕须记录研究过程中的所有关键决策和修改记录;元数据管理须系统记录研究的时间节点、版本迭代、参与者分工等信息,建立研究的时间线与空间轴。例如,论文《生成式人工智能对学生批判性思维能力培养会产生什么影响?——来自 67 项实验的元分析证据》的作者面对 17 542 篇初始文献逐步筛选至 67 篇的复杂流程,通过系统性的文档化策略确保研究过程完全可追溯、可审查、可复用。作者通过回顾第一批筛选中发现的 AI 误判模式,调整后续核查重点,使保留率从 60% 逐步提升至 100%。在元分析过程中,作者确保“所有数据都能追溯到原始文献、R 分析结果或 CSV 文件”,无数据来源的分析结果即予以删除。

8.守住灵魂——在价值判断与实践启示中坚守人类主体性

在人机协作中保持人类的主导权和最终决策权,既是突破 AI 内在局限的需要,更是哲学社会科学研究具有真正价值之必然。大语言模型基于概率预测和统计关联运作,缺乏真正的理解、意识和价值立场,虽可高效重组已有知识,却无法实现从无到有的原创性突破;虽可模拟逻辑推理之形式,却无法承担价值判断之责任。哲学社会科学研究不仅追求知识生产,更涉及意义阐释、价值权衡和伦理考量,这些维度深深植根于人类的生活体验和主体间性,是算法无法真正触及的领域。守住研究灵魂须做到主权坚守、批判审视和价值注入三个方面。在主权坚守方面,AI 可承担数据整理、初稿撰写、格式规范等机械性工作,但研究方向之把握、核心理论之建构、研究发现之阐释必须由人类主导。在批判审视方面,研究者须建立审稿人心态,对 AI 输出的每一段内容进行事实核验、逻辑审查和文献溯源,警惕幻觉内容和算法偏见之渗透。在价值注入方面,研究者须主动为研究赋予人文温度,在理论凝练中体现价值立场,在结论形成中彰显伦理关怀。例如,《AI 赋能下的人机协同课堂对话:多模态交互

机制与教学角色分析》一文明确宣示了人类研究者对思想与逻辑的绝对主权：所有的研究方法选择、数据特征的规律总结以及最终研究结论的推导，均由研究者基于实际运算的数据结果自主把控和分析得出。该研究中，人类研究者承担了高心智负担的质性编码工作——将62个课例中的复杂师生互动转化为10 552个多模态事件，此乃AI无法越俎代庖者。同时，人类研究者统筹了整篇论文的宏大叙事结构，在讨论与结论环节将实证发现与更宏大的理论命题相连接，为未来的实践提出前瞻性指导建议。

四、未来启示

本次“AI一作”大型社会实验，不仅是一场充满探索精神的征文活动，更是全球学术界首次正式将人工智能推向哲学社会科学研究“台前主体”的极限压力测试。此次活动的深远意义在于，它以一种极具冲击力的方式，彻底打破了学术界对AI技术“掩耳盗铃”式的回避，将人机协同科研的“暗箱”置于阳光之下。通过强制要求AI作为第一作者，本次实验倒逼数以千计的研究者在真实的科研场景中，重新划定人类智慧与机器智能的能力边界与分工逻辑。实验不仅证实了AI在激发灵感、海量文献处理、代码生成与多智能体仿真等环节的惊人效能，也客观暴露了其在价值判断、情感共鸣与实质性理论创新上的显著局限。更重要的是，本次活动通过引入《人工智能生成内容说明表》，研发基于多维评价指标的多模型AI审稿系统，为重构智能时代的学术规范、伦理底线以及新型学术评价体系提供了宝贵的第一手数据与实践样本。

整个实验过程可被理解为一场“生成型实验”——在改革路径尚不明晰的前沿领域，以探索总体方向为目标的开放性政策实验，也是一场极具前瞻性的计算教育学推演，宛如一块巨石被投入平静的水面。而在实验启动之后，技术与现实的演进从未停歇：诸如OpenClaw等高度自主的AI智能体不断涌现，各类探讨人机协同的“AI写作大赛”如火如荼地开展，更有诸如“The AI Scientist”等专用于自主科研的先进智能体被陆续研发。这一系列事件犹如一颗颗新的石子接连投入水中，它们激起的波纹与我们此次社会实验产生的涟漪不断叠加、激荡，向外泛起更广阔的波澜，对学术生态产生着持续且深远的影响。正如技术演进的历史规律所揭示的那样，随着底层大模型技术的日益成熟、学术共同体心理接纳度的逐步提升、学术伦理与社会规范的不断完善，以及各大核心期刊评审制度的实质性改革，这些环节相互嵌合并形成完整的创新链条后，必然会引发哲学社会科学研究范式的深刻革命。

这是一场尚未完成的社会实验，而现实世界的进化远比实验本身更为精彩。面对浩浩荡荡的技术洪流，我们应当摒弃“该不该使用AI”的无谓争论，迅速提升AI的运用意识与操作能力，这才是学者立足于智能时代的王道。在这一转型过程中，研究者必须敏锐把握人类真正的核心竞争优势——源自直觉的原始创新、跨越边界的情感联结以及对复杂事物的深度理解，这些依然是AI所无法替代的。同时，在深度应用AI的过程中，学界必须高度警惕“认知外包”与“思维惰性”的问题。对工具的过度依赖不仅会削弱碳基生命的主体性，更会致使硅基智能因缺乏高质量的人类反馈而陷入停滞，最终引发人类智慧退化的“脑腐（Brain Rot）”恶性循环。

需要特别说明的是，本次大型社会实验的发现仅仅是特定技术节点下的阶段性成果。当前，底层大模型正在以惊人的速度迭代，多模态交互与智能体技术日新月异，未来的科研工具、学术生态与教育形态必将呈现出与当下截然不同的崭新面貌。但在可见的未来，我们能够预见并需要积极应对以下五个维度的深刻变革：

一是推动哲学社会科学研究向着“AI驱动”的第五范式发展。以“AI驱动”为核心的科学研究第五范式将从自然科学全面拓展至哲学社会科学领域。未来的科研流程不再是人类包揽全流程的“手工作坊”，人类研究者将逐渐让渡文献综述、数据清洗、基础统计与文本起草等执行性工作，转而聚焦于提出具有洞察力的研究问题、构建宏大理论框架、进行跨学科价值判断以及注入人文关怀。人机之间将形成“人类定航把舵、AI高效执行、多轮交互迭代”的认知共生关系。

二是改革论文评价制度，从“结果导向”转向“过程与增量评价”。当AI能够轻易生成结构严谨、

逻辑自洽的学术文本时,传统的以“文本质量”和“八股规范”为核心的论文评价体系将面临失效。未来的学术评价制度必将发生深刻变革:一是评价重心的前移,从单纯审视“最终文本”转向评估研究者的问题意识、提示词工程逻辑以及人机协同过程;二是评价维度的重构,更加看重研究是否具有真实的社会痛点、是否具备 AI 无法生成的情感温度与实践智慧;三是评价主体的多元化,“AI 初审+专家复审”的双轨机制或将成为各大期刊和学术会议的标准配置。

三是促进相关 AI 技术的研发,从“通用模型”走向“科研垂类智能体”。本次实验揭示了通用大语言模型在学术写作中易产生“幻觉”、缺乏专业深度等缺陷。未来的 AI 技术创新将更具针对性:专门服务于学术研究的垂直领域大模型及“科研智能体”将进一步发展。这些创新技术将内嵌更严格的逻辑推理算法、可溯源的真实文献知识库,并能够实现从“提出假设—设计实验—处理数据—撰写报告”的多智能体全自动协同。同时,兼顾数据隐私的本地化部署小模型也将成为研究者的可及配置。

四是转变哲学社会科学育人范式,从“知识传授”转向“高阶思维与情感培育”。本次实验中暴露出的“AI 代沟”对传统教育提出了强烈预警。面对数字“原住民”对 AI 的天然亲近,哲学社会科学的育人范式必须进行系统性重构。传统教育中耗时较长的死记硬背与基础写作训练将被大幅压缩,取而代之的是三大核心素养的培育:一是人机协同的“元能力”(如计算思维、AI 底层逻辑理解与人机交互能力);二是坚不可摧的“批判性思维与伦理判断力”,以防范学术泡沫与认知外包;三是发掘“人之为人”的不可替代性,即同理心、身体感知与复杂的社会交往能力。

五是推动文凭制度改革,从“知识存量凭证”转向“协同创新能力认证”。当 AI 系统能够轻松通过各类标准化考试并输出合格的学位论文时,传统基于知识掌握程度和标准化测试的文凭制度,其含金量将被极大稀释。未来,学历与文凭的获取方式或将面临颠覆性改革。评价一个人的学术水平与工作能力,不再仅仅看其独立完成的静态答卷,而是更看重其在动态、复杂情境下“调度 AI 工具解决真实问题”的能力。基于实践项目、过程性数据档案以及人机协同创新贡献度的新型“微认证”与“能力图谱”,会逐步补充甚至部分替代传统的单一文凭制度,从而构建更尊重创造本质的人才评价新生态。

这场旨在超越图灵测试的破冰之旅,从更深层次看,其试图在学术评价的底层逻辑上完成一次从“图灵测试”到“超越图灵测试”的跨越。在过去大半个世纪里,图灵测试始终是衡量人工智能的标尺,其核心逻辑在于“以假乱真”——考察机器能否通过惟妙惟肖的模仿来掩盖其非人身份。然而,作为本次实验的发起方,我们坚信人工智能在知识生产领域的终极使命绝非“成为人类”,而是“超越人类的局限”。实验让我们清醒地认识到,未来的学术场域不再需要只会在语言游戏中隐藏机器身份的“模仿者”,而是亟需能在知识荒原与复杂系统计算中提供创新且硬核的增量的“超越者”。当 AI 不再致力于像人一样思考,而是充分释放其“非人优势”,真正以第一作者的身份提供超越人类既有认知边界的洞见时,人机协同才算真正触及了创新的灵魂。这一跨越,标志着哲学社会科学研究正在迈入人机共创的新纪元,其激发的伦理探讨与学术创新,必将对未来人类知识生产的底层逻辑产生持久而深远的影响。实验虽已告一段落,但其泛起的涟漪仍在持续扩散,深远的长尾效应正在真实世界的各个角落发挥作用。如果说当前人类的生命与认知形态是传统的“human being”,那么未来的生命与认知形态必将演进为深度融合技术的“human@AI being”。教育的形态、科研的模式与评价的尺度,终归要服从并服务于人类生命形态的历史性跃迁。希望学界能以更加开放的姿态迎接技术浪潮,期待一个属于全人类的“碳硅共生”的未来。

(张治工作邮箱:zhangzhi@mail.ecnu.edu.cn。为本次实验作出贡献的人员还包括:李昊润、石钊、吴宇阳、高峻、黄辰昕、牛灵、韩晓颖、耿媛媛、张枫、林霁、张熙翔、李佳音,谨此一并致谢。)

参考文献

包万平. (2025). AI 作为第一作者,学术责任如何界定?. *中国科学报*, 2025-11-18(03).

陈雅静, 刘越. (2025). “闯入”创作领域的AI是工具还是作者?. *中国社会科学报*, 2025-12-10(01).

胡志刚, 梁晗. (2025). 三重变革下人文社会科学知识生产范式与评价体系重构. *自然辩证法研究*, 41(12), 11—16.

华东师大基础教育改革与发展研究所. (2026). *“人机协同与共创: 教育研究与写作新范式”学术研讨会在华东师大举行*. 取自: https://www.ecnu.edu.cn/info/1094/71413.htm.

华东师范大学上海智能教育研究院. (2025). *共创学术奇点!“AI一作”活动白热化, 期待各界高手“下场”与AI共舞——项目组召开专家会议并发布征文活动补充通知*. 取自: https://mp.weixin.qq.com/s/spNL3eLdqv-oOq1MgaLwPg.

华东师范大学上海智能教育研究院. (2026). *超越图灵测试之后, 人类走向何方?——全球首个“AI一作”社会实验在华东师大交出答卷*. 取自: https://mp.weixin.qq.com/s/lPJZmZRSoISMuPGWv3lr9g.

黄海华. (2025). *“第一作者必须是AI”策划人: AI已不是一种工具, 而是人机共同进化的结果*. 取自: https://www.shobserver.com/staticsg/res/html/web/newsDetail.html?id=1022939.

李木子. (2025). 首次!AI撰写并审阅所有会议论文. *中国科学报*, 2025-10-16(02).

马亮. (2025). 人工智能是万灵药吗?——人机协同的社会科学研究何以可能. *天府新论*, (06), 10—19+150-151.

尚杰. (2025). *华东师大“以人工智能作为第一作者”的论文征集, 究竟想干什么?*. 取自: https://mp.weixin.qq.com/s/Uz_9wA1my9N7RUbUnVKW-A.

苏竣, 黄萃. (2023). *社会实验理论与方法评介*. 北京: 清华大学出版社.

唐政. (2025). *AI能当第一作者吗?学术伦理的真正考题在这儿*. 取自: https://hlj.rednet.cn/nograb/646956/69/15570193.html.

腾讯教育. (2025). *华东师大征集AI主笔论文引争议, AI能不能当论文第一作者?*. 取自: https://mp.weixin.qq.com/s/TFgxIcC-XT90oVZ6LYH84A.

王金林. (2025). “生成式理性”: 大语言模型引发的知识生产范式转型. *社会科学*, (08), 184—192.

王倩, 李楚悦. (2026). *六个AI并列第一作者, 一场“必须用AI写”的论文比赛*. 取自: https://www.shobserver.com/staticsg/res/html/web/newsDetail.html?id=1056964&sid=200.

徐丽芳. (2025). *科研AI治理的平衡术: 安全阀还是绊脚石?*. 取自: https://mp.weixin.qq.com/s/PT8FV_XzIse-yDkVB9bppA?scene=25&sessionid=#wechat_redirect.

游俊哲. (2023). ChatGPT类生成式人工智能在科研场景中的应用风险与控制措施. *情报理论与实践*, 46(06), 24—32.

袁曼舒. (2025). *华东师大发起AI“主笔”论文征集引热议, 是“超前实验”还是“人类退场”?主要发起方独家回应*. 取自: https://mp.weixin.qq.com/s/Ak7anS06MpGnZxs0SpSS0Q?scene=1.

袁曼舒, 肖敏. (2025). *华东师大发起AI“主笔”论文征集后续如何?最新进展来了*. 取自: https://mp.weixin.qq.com/s/vIvg-_sWicMcz6pfQ-WOioQ?scene=25&sessionid=#wechat_redirect.

Agents4Science 2025. (2025). *Open Conference of AI Agents for Science 2025*. Retrieved from https://agents4science.stanford.edu/.

Bianchi, F., Queen, O., Thakkar, N., Sun, E., & Zou, J. (2026). Exploring the use of AI authors and reviewers at Agents4Science. *Nature Biotechnology*, 44(1), 11—14.

Draux, H. (2024). *Analysis of Dimensions data reveals the global divides in artificial intelligence capabilities*. Retrieved from https://www.dimensions.ai/blog/analysis-of-dimensions-data-reveals-the-global-divides-in-artificial-intelligence-capabilities/.

Rivero, V. (2025). *Education doesn't need an upgrade—it needs a rethink*. Retrieved from https://www.edtechdigest.com/2025/07/17/education-doesnt-need-an-upgrade-it-needs-a-rethink/.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development*, 53(4), 5—23.

注　释：

① 由于该实验未设置严格的对照组进行随机分配(所有参与者均接受同一干预)，在内部效度上更接近准实验或单组干预设计。

② AI审稿系统各模块选用的大模型仅为“AI一作”征文活动期间国内外较为前沿的模型版本，随着大模型的不断迭代，后续系统还需选用性能更强的更新版本大模型，以确保智能审稿功能不断优化。

③ 研究过程中发现7份“过程说明”未按规定模板填写，无法有效提取相关字段信息，因此未将其纳入数据分析，导致此处“过程说明”数量与有效投稿数量并不一致。

（责任编辑　王　森）

Beyond the Turing Test: A Panoramic Report on the World's First Large-Scale Social Experiment of "AI as the First Author"

Zhang Zhi[1,4] Chen Shuangye[2] Xiao Min[3] Liu Yimeng[4] Zhang kai[4] Ji Cheng[5]

(1. Intelligent Education Laboratory, Key Laboratory of Philosophy and Social Sciences of Ministry of Education, East China Normal University, Shanghai 200062, China;
2. Faculty of Education, East China Normal University, Shanghai 200062, China;
3. Shanghai Future Learning Research and Development Center, Shanghai 201999, China;
4. Shanghai Institute of AI for Education, East China Normal University, Shanghai 200062, China;
5. Nanjing PLASO Network Technologies Co., Ltd., Nanjing 210006, China)

Abstract: In September 2025, East China Normal University launched the world's first large-scale social experiment of "AI as the First Author". Using a quasi-field experimental method, it carried out an essay-soliciting activity on "AI-Driven Educational Research Paper Writing", requiring AI to be the first author and humans to play the roles of collaborators and reviewers. Over a period of half a year, 724 valid submissions were received from both domestic and international sources. Based on this, the university explored the fifth paradigm of AI-empowered research in philosophy and social sciences, academic ethical norms, and the path to constructing an independent knowledge system in education. The experiment went through eight stages: intervention design, response monitoring, expert seminars, data collection, AI-based manuscript review, human-machine consistency testing, data analysis, and result publication. A mixed-research method was adopted to reveal the core findings. AI has significant advantages in aspects such as inspiration generation, information processing, and text polishing, but it has limitations such as fictional literature, logical hollowness, insufficient innovation, and ethical risks; efficient AI application can significantly enhance academic contributions, forming six innovative patterns and five human-AI collaboration models; AI-based manuscript review is reliable and complementary to the evaluation by human experts, and there are obvious differences in the research tastes of different large language models; there is a significant "AI generation gap" in the academic community, with the younger generation being more adaptable to human-AI collaboration, and AI has, to some extent, promoted intellectual equality. The research proposed that in the future, efforts should be made to promote the transformation of research paradigms, reform of evaluation systems, development of research-specific AI agents, reconstruction of educational models, and innovation of diploma certification. Findings provide empirical support and practical inspiration for academic innovation and educational transformation in the intelligent era.

Keywords: AI as the first author; social experiment; educational research; human-AI collaboration; research paradigm; academic ethics; AI-based paper review; intellectual equity