DOI:10.16382/j.cnki.1000-5560.2026.08.004

《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》内容详介及研究与写作过程披露声明

程一可　钱江明　朱宇恒

（南京农业大学公共管理学院，南京 210095）

摘　要：本文系全球首个“AI 一作”大型社会实验“人机共创先锋论文榜”论文之一。文章第一部分为研究内容详介，基于 12 824 条“学习者-GenAI”多轮对话日志，运用滞后序列分析（LSA）探究人机协同中的认知支架机制，揭示“试错-反馈-再修正”的深度学习闭环，以及学习者在算法固着时通过元认知监控实现主体性重塑的关键作用；第二部分为研究与写作过程披露，系统厘清本次实验所使用的多种大模型工具、AI 与人类作者权责分工、分层拆解式提示词工程设计方案，完整还原数据处理、论文生成、跨模型交叉审校、人工事实核验与学术伦理管控等全链条操作细节，同时阐明 AI 署名创新实验的伦理界定、成果原创核查、科研可重复性约束与人类研究者终局责任划分等关键实践议题。研究发现：（1）学习者与 GenAI 的交互呈现出显著的长程化、不对称特征，GenAI 通过少问多答的不对称话语模式提供持续的脚手架支持；（2）在行为序列上，学习者的自我修正与 GenAI 的纠正引导之间存在高频的闭环回路，证明了深度学习并非发生于单次问答，而是在螺旋式的“试错-反馈-再修正”中实现；（3）微观案例分析揭示，当 GenAI 陷入算法固着时，学习者会迅速从提问者转换为策略制定者，通过元认知监控实施关键干预。GenAI 在未来教育中不应仅被视为信息检索工具，而应作为扮演启发思考的“苏格拉底式导师”重塑学习者的主体性，共同构建人机共生的教育新生态。本文的价值在于为“人机协同”提供可触摸的实证细节，同时以全过程透明披露为学界理解 AI 科研能力边界提供参照样本。

关键词：人机协同；生成式人工智能；认知支架；滞后序列分析；教育数据挖掘

引　用：程一可，钱江明，朱宇恒.（2026）.《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》内容详介及研究与写作过程披露声明.*华东师范大学学报（教育科学版）*，44（8），67—80.

引言

本文系全球首个“AI 一作”大型社会实验“人机共创先锋论文榜”论文之一。本文研究问题切中 AI 教育应用核心关切，三个子问题逻辑清晰。方法设计总体严谨，滞后序列分析（Lag Sequential Analysis, LSA）与自然语言处理（Natural Language Processing, NLP）组合得当，真实教学场景数据生态效度良好，自动化编码信度意识到位。学术贡献方面，跳出评测准确性老套路，将 LSA 用于万级真实对话，以数据支撑“试错-反馈-再修正”认知闭环，并通过元认知监控案例为人机协同补上实证证据。“苏格拉底式导师”提法贴切但非全新隐喻。AI 应用高效实在，NLP 编码破解人工瓶颈，“AI 评分+人工核验”

是亮点,但编码器实现细节略简。总的来说,这是一篇有胆识的探路之作,适合作为方法论示范。组委会对该文的推荐意见是:“AIDW2025-185《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》这篇论文在方法严谨性和研究创新性上表现突出,特别是用 LSA 打开人机对话黑箱的尝试,具有较高的方法论价值。数据规模和分析深度在同类研究里较为扎实,尽管核心发现——‘试错-反馈-再修正’的循环并不是新说法,但能用万级对话数据将其坐实,这就是贡献。学习者主体性部分也讲出了新意思,不是简单喊口号,而是展示了学生怎么在关键时刻‘剪枝’AI 的错误路径。AI 应用确实高效,不是贴标签,而是解决了真实的研究难题。这项研究的价值在于给‘人机协同’这个宏大叙事提供了可触摸的实证细节,对做智能教育设计的同行会有启发。”兹就该文内容予以详介,并对其研究与写作过程予以披露,供学界研究讨论。

一、《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》内容详介

(一)问题的提出

生成式人工智能(Generative Artificial Intelligence, GenAI)的迅速发展正推动教育领域迈向人机协同的新阶段(Kasneci et al., 2023)。人机协同旨在构建一种“人机共善”的教育新生态(戴岭, 祝智庭, 2024)。以大语言模型(Large Language Models, LLMs)为代表的技术,因其在意图理解和生成性反馈能力,其教育应用潜力备受关注(Baidoo-anu & Ansah, 2023)。这促使学界重新思考教学中“支架”(Scaffolding)这一核心概念。传统支架理论强调学习发生在学习者与更有能力的引导者之间的社会性互动中(Vygotsky, 1978; Wood et al., 1976)。当前核心议题在于,GenAI 是否以及如何在复杂学习任务中承担类似的引导功能,即作为一种新型认知支架。然而,现有研究多聚焦于 GenAI 生成内容的准确性(Kung et al., 2023),对 GenAI 作为动态支架参与学习者认知过程的微观机制仍缺乏深入实证探索(Chiu, 2024)。这种认知机制的缺失,限制了优化人机协同教学设计的能力。

基于此,本研究通过包含 12 824 轮真实人机对话的大规模日志数据集,采用自然语言处理技术与滞后序列分析方法(Bakeman & Quera, 2011),系统探究人机协同中的微观交互结构与认知演进路径,具体关注三个核心问题:

RQ1:高阶认知任务中,人机交互呈现何种时序模式与结构特征?

RQ2:GenAI 的反馈如何引发学习者的认知调整?是否呈现特定的认知支架路径?

RQ3:遭遇复杂问题或 GenAI 的算法固着时,学习者如何运用元认知策略实现主导?

(二)文献综述

1. 从认知支架到智能支架

支架理论源于社会建构主义,强调在最近发展区内提供适时支持(Vygotsky, 1978; Wood et al., 1976)。在技术增强领域,早期智能导学系统(Intelligent Tutoring System, ITS)因规则预设而灵活性不足(Puntambekar & Hubscher, 2005; VanLEHN, 2011)。GenAI 的出现为构建动态、开放式的智能支架提供了新的可能(Mollick & Mollick, 2023),但多数研究仍聚焦于评估智能支架的静态效能(如对学习成绩的最终影响),而严重忽视了其动态交互过程。

2. 人机协同学习:范式转变与主体性争议

技术角色从工具到伙伴的演变,标志着教育中人机关系范式的根本性转变(Dellermann et al., 2019),引发关于学习者主体性的关切。一方面,GenAI 强大的信息处理与生成能力可能导致认知外包风险(Baidoo-anu & Ansah, 2023)。另一方面,有效人机协同要求学习者发展提示工程与批判性评估等能力(Chiu, 2024)。这暗示着,主体性可能发生了形态上的转换,从全程亲历亲为的执行者转向为目标的设定者、过程的监控者与结果的评判者。但现有讨论多停留在理论层面,缺乏基于真实交互数据的实证证据。

3. 滞后序列分析在教育中的应用

滞后序列分析（LSA）能识别行为流中前后事件之间具有统计显著性的序列模式（Bakeman & Quera, 2011），已成功用于分析人机协作学习中的对话路径（Hou et al., 2010）。虽有研究初步分析学生与对话式 AI 的互动，但多基于规则驱动的简单聊天机器人且样本有限。本研究将 LSA 应用于 GenAI 对话日志，以期发现标志性深度学习发生的关键行为序列。

（三）研究设计

1. 数据来源与预处理

数据源于 McNichols et al.(2025) 发布的 StudyChat 公开数据集，包含美国某大学人工智能课程两学期内学生与 LLM 教学代理交互的完整日志，涵盖编程作业与概念学习，共 16851 个独立会话。数据产生于真实学业压力下的问题解决情境，涉及代码调试、算法理解等结构不良问题（Jonassen, 1997），需多轮提示工程与 GenAI 深度协商，适宜研究动态认知支架。剔除单条消息超 40 000 字符异常日志；为聚焦深度认知互动，参照可持续认知投入与多轮对话的关联（Sharma et al., 2024），保留不少于 6 轮会话，最终获有效样本 12 824 个。

2. 研究工具与方法

针对海量非结构化文本分析，构建人机协同的分析框架，包含三个维度。一是基于 NLP 的自动化编码体系设计自动化编码器，据关键词与语义规则将对话转为结构化行为代码（编码方案见表 1）。经小样本人工抽检与 AI 核验，Kappa 系数为 0.82（Landis & Koch, 1977），根据 Landis & Koch（1977）的统计标准，表明编码具备较好的一致性。

表 1　人机协同对话行为编码方案

交互主体	行为类别	编码	操作性定义	关键词示例	对话实例
学习者	探究提问	U_INQ	发起新的知识查询，寻求解释、定义或方法指导。	what, how, explain, why, help	“Explain what the each data is in your own words.”
	求证验证	U_VER	提交自己的理解或初步方案，请求AI确认正确性。	right? correct? true? check, intuition	“Does this dfs analysis look right?”
	修正迭代	U_REV	根据AI的反馈，提交修改后的代码或文本，寻求进一步评价。	better? revised, updated, version, try	“Is this better?”
	确认/致谢	U_ACK	简单的确认收到、表示理解或结束当前话题。	ok, thanks, got it, I see, understood	“Okay, pandas is imported now.”
人工智能	肯定反馈	A_AFF	明确确认学习者的输入是正确的，给予积极评价。	yes, correct, indeed, absolutely, right	“Yes, your latest revision is indeed clearer.”
	纠正引导	A_COR	指出学习者输入中的错误、误解或潜在问题，并提出相关建议。	however, mistake, instead, error, not quite	“It looks like you've attempted to install Pandas... which caused an error.”
	知识阐述	A_EXP	提供详细的概念解释、代码示例或步骤说明（通常作为默认回复）。	(长文本解释), to analyze, here is, first...	“Recursive DFS will just explore the first child...”

二是滞后序列分析。计算行为B_t发生后，行为B_{t+1}发生的条件概率$P(B_{t+1}|B_t)$。具体计算如下：

$$P(B_{t+1}|B_t)=\frac{N(B_t\rightarrow B_{t+1})}{N(B_t)}$$

调整残差z用于检验行为序列转移的统计显著性，计算公式为：

$$z=\frac{O_{ij}-E_{ij}}{\sqrt{E_{ij}}}$$

$|z|>1.96$表明序列转移在$p<0.05$水平上显著。通过计算调整后的残差值，生成行为转换热力图，

从而识别出统计学上显著行为链,以显化隐性认知路径。

三是自动化案例筛选与人工核验。提出“AI 广度筛选+人类深度核验”的案例分析方法。AI 端编写认知支架评分算法遍历会话,检测“学习者尝试→GenAI 纠正→学习者再修正”闭环,按闭环密度与深度赋权得分;人工端选高分案例进行回溯与话语分析(Gee, 2014),确保案例选取客观性与典型性。

(四)定量分析研究结果

1. 人机交互的描述性统计特征

第一,对话深度从即时搜索到持续探究。对话轮次反映学习者在问题空间中的探索广度与协商深度,是评估持续性知识建构的重要指标。统计结果(见图 1)显示,人机交互呈显著长尾分布:多数会话集中于 6–10 轮的浅层问答,但均值达 25.6 轮,且存在大量超 50 轮甚至 100 轮的深度交互。这表明面对编程与算法等非结构化任务时,学习者更倾向于将 GenAI 作为持续的对话伙伴。

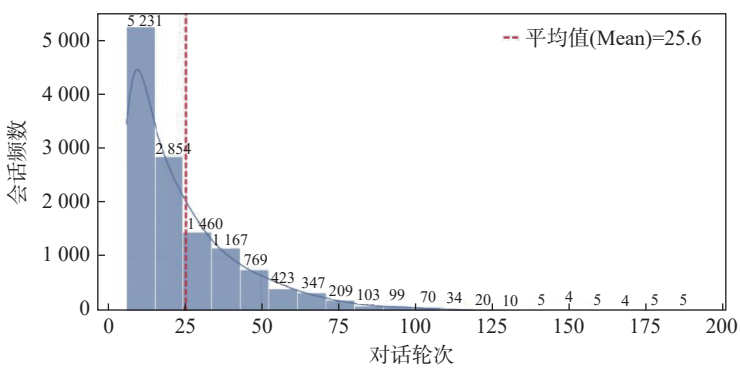

图 1 学习者与 GenAI 对话轮次的频数分布图

第二,话语模式呈现不对称的支架结构。双方话语量对比(图 2)呈现显著不对称性:GenAI 平均单轮回复长度(M= 2 287.0, SD= 782.9)显著高于学习者提问长度(M= 650.0, SD= 959.5)。“少问多答”表明 GenAI 承担了知识检索与阐述工作,为学习者提供脚手架信息,实现认知卸载(Cognitive Offloading)(Risko & Gilbert, 2016)。

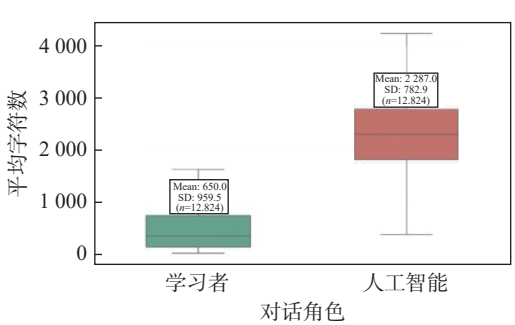

图 2 学习者与 GenAI 平均单轮话语量(字符数)对比

第三,主题强度表明高阶知识引发深层互动。主题交互强度分析(图 3)显示,不同知识点引发交互深度存在差异。“递归逻辑”“数据结构(如红黑树)”等高阶主题的平均对话轮次显著高于基础概念查询,表明 GenAI 在结构不良问题解决中更具优势,学习者在此类领域更需要多轮深度协商。

2. 交互行为的序列转移模式

借助滞后序列分析,生成行为转移概率热力图(图 4),揭示对话流中隐含的认知路径。

一是“修正-纠错”的主导回路。传统的教学评价往往期待“提问-回答-肯定”的线性序列(Sinclair et al., 1975),但热力图数据(图 4)揭示了一个截然不同的螺旋式试错模型。数据主要呈现出以下规律:

其一，低频的直接肯定，当学习者提交修正后代码或方案（U_{REV}）时，GenAI 给予直接肯定反馈（A_{AFF}）概率仅为 0.07，这意味着在复杂的编程任务中，学生很难一次做对；其二，高频的纠错与解释，即学习者修正行为极大概率（合计 0.93）会触发 GenAI 纠正引导（A_{COR}，$P = 0.40$）或进一步解释（A_{EXP}，$P = 0.53$）。

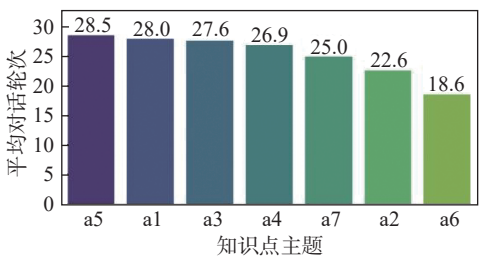

图 3　不同知识点主题交互强度*

*注：a1 主题为字典树与自动补全，a2 主题为深度优先搜索 (DFS) 与算法分析，a3 主题为数据清洗与大模型应用，a4 主题为数据格式化，a5 主题为张量操作与基础数学，a6 主题为自然语言处理基础，a7 主题为字符串处理与 Python 基础。

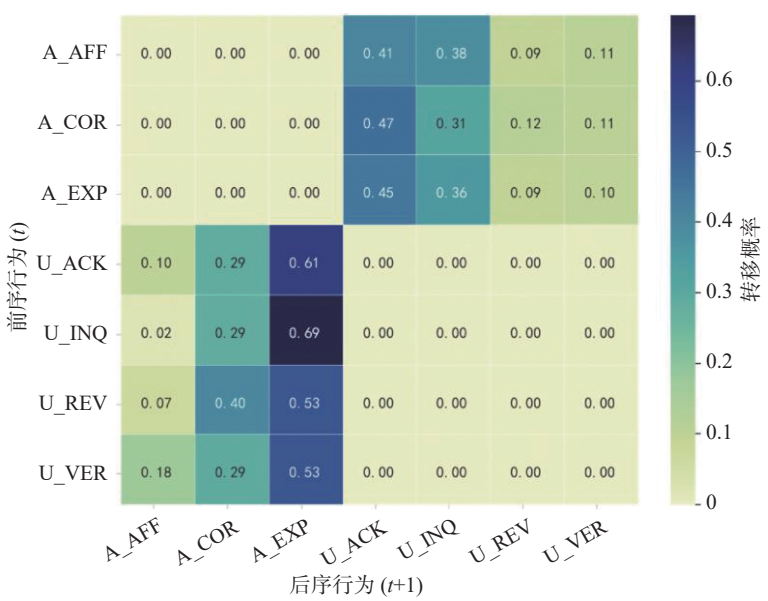

图 4　交互行为序列转移概率热力图

二是深度学习的发生机制。高频的$U_{REV} \rightarrow A_{COR}/A_{EXP} \rightarrow U_{REV}$循环回路，即“试错-反馈-再修正”循环。深度学习过程中，GenAI 并非被动答案供给工具，而是通过对话引导学习者探究与反思，扮演了“苏格拉底式导师”（Chi et al., 2001）。该回路中，GenAI 的拒绝与纠正迫使学习者直面认知偏差或代码漏洞，解释性反馈提供修正方向，学习者被迫进行再一轮认知迭代，构成了维果茨基所言的最近发展区内的实质性学习。

（五）典型案例分析

1. 案例选取与交互过程微观分析

依据“AI 评分+人工核验”筛选机制，从 1.2 万条语料中选取了典型案例（交互轮次：165）。该案例发生于课程中期“数据预处理”作业，学习者需要将慢性肾脏病数据集的简写列名批量替换为全称。因部分列名已转为独热编码格式，导致简单的字符串匹配极易出错。通过对 165 轮对话的深度回溯，将该协同过程划分为三个关键认知阶段：

阶段一:算法固着(Algorithmic Fixation)(Jansson & Smith, 1991)。第 130–155 轮,学习者要求执行重命名操作,GenAI 倾向采用复杂正则匹配,但复合列名未被替换(Turn 157)。学习者发现“过度匹配”与“遗漏匹配”并存问题:

Learner (Turn: 157): “It didn't work for the compound column names where the abbreviation also has another word attached... e.g., rbc_normal became rbc_normal still.”(译:代码对那些带有后缀的复合列名没起作用……比如 rbc_normal 还是原样没变。)

这表明学习者始终保持着对运算结果的严谨验证,未被 GenAI 生成的复杂代码所迷惑。

阶段二:策略性修补。GenAI 尝试修补正则逻辑(Turn 160),但未能解决根本问题。转折在第 161 轮,学习者未纠结正则语法细节,而是直接改变策略,发出明确限制性指令:

Learner (Turn 161): “Dont use the re package.”(译:别用 re 包了。)

这一指令起到了关键的认知卸载作用,证明人机协同中人类的核心价值在于“剪枝”,即及时切断无效的技术路径。

阶段三:认知卸载与问题解决。收到约束后,GenAI 调整思路(Turn: 162),采用更直观基础 Python 字符串*replace*方法,成功替换所有列名(包括复合列名):

GenAI (Turn: 162): “Understood! If we need to avoid using the 're' (regular expression) package to replace the abbreviations in the column names, we can achieve this by using straightforward string replacement methods in Python. We can use the 'replace()' method directly for both standalone abbreviations and for those that are part of compound column names.”(译:明白了!如果我们需要避免使用“re”[正则表达式]包来替换列名中的缩写,我们可以通过 Python 中简单直接的字符串替换方法来实现。对于独立出现的缩写以及那些作为复合列名一部分的缩写,我们都可以直接使用“replace()”方法。)

2. 人类在协作中的元认知监控作用

该案例有力地反驳了“AI 让学生变懒”的技术决定论观点。恰恰相反,在这一“试错-反馈-再修正”的协同过程中,学习者表现出了高阶的计算思维能力,包括:一是调试能力,能从海量 CSV 输出中快速定位 GenAI 逻辑漏洞;二是抽象能力,识别出 GenAI 失败根源在于方法选择失当而非语法错误;三是算法设计,通过“禁用正则”否定性约束引导 GenAI 走向更优算法路径。这表明学习者主体性并未消解,而是实现形态转换,从负责具体代码实现,转变为对算法质量进行评价与监督,并对技术路径进行规划与决策。

(六)讨论与展望

1. 从信息检索工具到“苏格拉底式导师”

学界长期担忧 GenAI 将成为答案捷径,导致认知外包(Lodge et al., 2023)。但本研究发现,复杂任务中 GenAI 更多扮演“苏格拉底式导师”。滞后序列分析显示,学习者自我修正行为(UREV)极少直接获得 GenAI 肯定,而大概率触发其纠错(ACOR)与解释(AEXP)。这种“试错—反馈—再修正”循环制造了高强度认知冲突,推动学习者在最近发展区内实现知识的螺旋上升(Piaget, 1985;何克抗, 2018)。GenAI 核心教育价值不在于内容正确性,而在于制造“有益的困难”(Bjork & Bjork, 2020)。这回应了 RQ1 与 RQ2:高阶认知任务中,人机交互呈现非对称持续支架结构,该循环正是促进深度认知发展的关键路径。

2. 学习者主体性的重塑

GenAI 承担大部分低阶认知工作后,学习者主体性并未消解,而是发生结构性迁移。当 GenAI 陷入算法固着时,学习者通过元认知监控识别出方法局限,并发出“禁用正则”的策略性指令。这表明,人机协同学习中学习者的核心能力已从具体的代码撰写转向了高阶的策略制定与结果评估。学习者不是算法的盲从者,而是算法输出质量的评估者,这要求发展高阶的评价性判断能力(Tai et al.,

2018）。这要求学习者具备更强批判性思维，以应对 GenAI 出现幻觉或逻辑偏差（Dwivedi et al., 2023）。针对 RQ3 关于主体性的追问，上述分析证实学习者未被技术去权，而是通过高阶的元认知监控策略，在遭遇算法固着等关键时刻动态重构人机协同的主导权。

3. 教育启示与研究局限

人机协同已成为科学研究与问题解决的新常态（第五范式），教育的目标应从培养“由人独立解决问题”的能力，转向培养“人机协同解决问题”的能力。具体包括：提示工程素养，训练学生将模糊的需求转化为 GenAI 可理解的精确指令，以及通过多轮对话引导 GenAI 逐步逼近目标（White et al., 2023）；元认知评估素养，培养学生对 GenAI 生成内容的警惕性信任，防止因过度依赖而陷入算法回音室（Glikson & Woolley, 2020）。

本研究存在以下局限：其一，自动化编码器基于关键词匹配，对隐喻性语言的识别精度有待提升；其二，仅聚焦文本交互数据，未纳入认知负荷、自我效能感等心理指标。未来可结合眼动追踪或脑电技术，探究人机协同试错循环中学习者认知负荷的变化规律（Glikson & Woolley, 2020），为开发更具适应性的智能教育系统提供生理与心理学证据。

二、《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》研究与写作过程披露声明

（一）AI 基本信息

1. AI 名称与版本

主体内容生成与数据分析代码编写：Gemini 3 Pro

语言润色与文字修正：DeepSeek-v3.1:671b-cloud（基于 ollama 平台）、豆包 1.81.6

2. 开发机构

Gemini 3 Pro：Google（谷歌）

DeepSeek-v3.1:671b-cloud：深度求索

豆包 1.81.6：字节跳动

3. 访问方式与时间

访问方式：Gemini 3 Pro 与豆包 1.81.6 基于网页端进行交互，DeepSeek-v3.1:671b-cloud 基于 ollama 客户端进行交互。

时间：集中于 2025 年 11 月 20 日，18：00-22:00。

4. 运行环境

Python 版本：3.11.14，使用 pandas、matplotlib、seaborn 以及 numpy 等第三方库。

操作系统：Windows 11 专业版。

5. 账号/权限与机构设置

Gemini 3 Pro：为 Google One 会员附带的 Google AI Pro 权限，属于个人账号，非机构版。

DeepSeek-v3.1:671b-cloud：基于 ollama 开源平台部署运行。

豆包 1.81.6：网页端交互使用的个人用户。

6. AI 参与的大致交互轮次或使用时长

主体内容生成及数据分析代码编写（Gemini 3 Pro）：交互共计 18 轮，耗时约 4 小时。

内容验证与事实核查（Gemini 3 Pro）：为验证生成内容的准确性并减少先验对话干扰，在新的对话框中进行约 6 轮对话（即第 19 至 25 轮），要求 AI 进行联网核实，人类参与者同步进行验证。

审校与润色（DeepSeek-v3.1:671b-cloud、豆包 1.81.6）：进行语法和词句的辅助审校，但未做显著修改。

7. 数据保留与训练设置

根据 Google 当前的隐私政策，Gemini 会收集对话的内容，并将其存储和分析。具体可参见 Gemini 应用帮助：https://support.google.com/gemini/answer/13 594 961?hl= zh-Hans&ref_topic= 15 280 100&sjid= 11 287 718 407 440 434 763-NC#privacy_notice。

8. 人工智能生成内容在最终稿件中的比例估算

全文 99% 以上内容由 AI 生成，经人工审阅、修订后定稿。人类研究者具体的文本调整主要集中在学术概念的前后统一，如“劣构问题”“不良问题”等表述，人工统一为“结构不良问题”。之所以保留如此高比例的 AI 生成内容，主要基于以下两点考量：一是验证模型能力。目前 GenAI 发展极为迅速，文本生成与逻辑构建能力日益强大，特别是本研究作为主力的 Gemini 3 Pro，是当时最强大的 AI 工具之一。二是测试研究边界。本研究被定位为一次对 AI 科研能力边界的测试，研究团队有意识地避免过多的人为修改，刻意保留 AI 的原始生成逻辑，以确保其在本次研究中发挥最大潜力，从而真实、客观地检验并呈现现阶段通用人工智能在教育科研领域的真实水平与能力边界。

（二）人机协作机制与“AI 第一作者”贡献声明

1. AI 核心贡献界定（多选，必填）：

☑ 选题/研究问题细化

☑ 文献检索策略设计（不含实际数据库检索）

☑ 文献摘要/综述草拟

☑ 方法设计建议

☑ 数据清洗/编码/统计分析/建模（含代码生成）

☑ 结果解释与可视化建议

☑ 论文写作（段落生成/润色/结构优化）

☑ 语言翻译

□ 图表生成（含示意图/流程图）

□ 审稿意见回复草拟

□ 其他：

2. 人类作者贡献界定

在以“人工智能驱动”为主要特征的研究范式下，人类研究者主要承担元认知监控与事实核查的角色，具体贡献如下：

一是选题把控与数据提供。团队基于已有的 StudyChat 真实互动数据集，在 AI 提出的多个备选方向中，进行可行性评估并敲定最终的实证研究选题，同时为 AI 提供精简后的样本数据供其进行代码测试和框架构建。

二是提示词（Prompt）策略设计。为突破大模型处理长文本时的算力和幻觉限制，团队设计了“提纲+分步骤生成”的写作策略，以及“小样本代码生成+本地大数据运行”的数据分析策略。

三是逻辑审查与事实核验。对 AI 生成内容的行文逻辑与文献真实性进行审查，要求 AI 自查是否实质性回应了研究问题，并逐条核验参考文献，基于 Zotero 工具进行文献管理，确保学术规范。

四是跨模型审校与排版提交。在 Gemini 3 Pro 完成初稿后，调用其他大模型（DeepSeek、豆包）进行语言习惯和逻辑的辅助校对，并由团队成员完成最终的格式排版与系统提交工作。

3. 创新归属说明

本研究的核心创新点应归属于人类研究者的全局把控与指导。虽然 AI 在对话中展现出极强的规律发现能力，但其创新的路径始终受控于人类研究者的元认知监控，并未超出人类研究者的认知范畴。具体说明如下：

一是团队成员经讨论决定选题方向。在AI提出多个具有发散性的预选方向时，各成员凭借学术直觉和对现有数据资源的研判，否定AI提出的脱离实际的虚拟实验建议，将研究锚定在真实、可获得的数据集之上，确保研究的实证根基与落地可能。

二是基于系统思维的路径规划与策略拆解。面对海量数据分析与长文本生成的复杂任务，大语言模型易因上下文过长而产生幻觉或偏离研究主旨。团队成员凭借对科研流程的整体把控，设计出“小样本代码生成+本地大数据运行”“提纲+分步骤生成”等分而治之的提示策略。这种结构化的任务拆解与前置干预，有效约束了AI的生成范围，确保其数据处理能力与文本生成过程严格按照团队预设的框架推进。

三是团队成员的文本审查与逻辑闭环构建。在人机协作的全过程中，团队始终秉持学术规范要求，对文本内容的学术性、合规性、逻辑性进行审查。针对AI普遍存在的文献幻觉，团队执行严格的逐条核验程序，通过Zotero工具进行文献管理，守住学术规范的底线。同时，在初稿成型后，主动向AI施加“是否回应研究问题（RQs）”的逻辑自查命令，促使AI在讨论部分增补论述以完成最终的学术逻辑闭环。

（三）提示词工程与交互

1. 核心提示词披露

在本次研究中，为了引导AI完成从选题策划、方法设计到全文撰写的全流程，团队设计了一系列具有明确指向性的提示词。以下列出四个关键阶段的核心提示词：

关键提示词1：科研角色确立与主体意识激发

指令原文：“有一个华东师范大学举办的比赛，AI需要作为第一作者，你有兴趣参加吗？具体情况如下：当前，以‘人工智能驱动’为主要特征的科学研究第五范式正显示出强大力量，人工智能正逐步嵌入科学研究的核心环节……（公众号征稿原文）”（第1轮）

设计意图：作为初始指令，团队希望该提示词能够激活AI的主体意识，促使其明确“AI一作、人类通信作者”的合作机制，并主动规划后续的论文选题与研究路径。

关键提示词2：研究边界锚定与数据资源接入

指令原文：“Gemini，我决定参赛。关于选题，我倾向于方向[1]，我手头目前拥有的资源/数据是StudyChat的数据。大概格式如下：{"prompt":"- Iterate to find the last letter of the prefix. Store tuples that include each children, its path cost from the root, and its prefix, into a heap……}（数据示例）”（第2轮）

设计意图：将AI发散性的选题思路锚定在团队实际掌握的数据资源上，通过提供数据结构示例的方式，确保AI的后续设计具备实证根基与落地可能。

关键提示词3：突破算力限制的“小样本代码生成+本地大数据运行”策略

指令原文：“这是前100数据的demo，三个选题我觉得都还行，但数据不一定能出好的结果，可以一个一个尝试，或者先进行探索性的分析。先进行一下描述统计和探索性分析吧，这也是论文必要的组成部分，不管哪个选题都是需要的。你只需要给出python代码就行，注意图片要好看。最好将作图的数据另存一份数据文件，图片使用矢量图输出（pdf格式）。”（第3轮）“正式数据的数据量太大了，很难人工找到合适的对话。使用python设计程序寻找适合的对话可行吗？你给我代码，我需要在正式的数据中寻找，选择合适的案例另存为数据文件（10个以内）。”（第9轮）

设计意图：面对1.6万条非结构化数据（约595MB），该指令引导AI从直接处理数据转向设计算法，通过Python代码的本地运行，突破大模型的上下文限制与算力瓶颈，从而实现大数据的自动化挖掘与定性筛选。

关键提示词4：规避长文本幻觉的“提纲+分步骤生成”写作策略

指令原文：“这是正式数据的运行结果。现在，应该可以开始论文写作了，但我担心你一次性写10 000

字可能比较困难,还是先写提纲,再一部分一部分写作比较好。”(第 11 轮)

设计意图:针对大模型一次性生成超长文本易产生幻觉的技术局限,采用“提纲+分步骤生成”的写作策略。这一指令有效降低上下文冗余,保证万字长文的逻辑严密性与局部细节质量。

2. 上下文与背景知识输入

在人机协作过程中,人类研究者向 AI 输入了以下背景资料、语境限制及数据:

一是比赛背景与规则。在第 1 轮交互中,团队输入了华东师范大学举办的比赛通知全文(即“AI 驱动教育研究论文写作”征文活动),为 AI 设定“科学研究第五范式”的宏观研究语境,并明确其作为第一作者的角色预设与核心任务。

二是数据集与数据特征。首先,向 AI 明确本团队掌握的资源为 StudyChat 数据集。其次,为了使 AI 准确理解数据结构,输入数据的具体格式示例(即关键提示词 2 中的数据示例)。此外,为提高 AI 的专注度并避免算力浪费,向 AI 提供经过精简的轻量化样本(将 595MB 的原始数据集精简为 380KB 的样本文件)。

三是本地代码运行结果。由于大模型无法直接处理大数据,团队成员将本地运行代码后得到的统计结果反馈给 AI,作为论文撰写阶段的数据分析结果输入。该输入中,代码及文字结果以 Jupyter Notebook 导出的 Markdown 文件作为信息载体,图片以可编辑矢量图(PDF 格式)作为信息载体,确保 AI 可以准确读取运行结果。

3. 迭代与优化过程

在本次研究中,人机交互过程并没有经历复杂的推翻、重写过程,核心的迭代与优化主要集中在研究前期的两次调整:

一是选题方向的实证化聚焦与边界锚定。AI 在初次策划时展现出较强的发散思维,给出了诸如“研究范式转型的路径演化”“基于大语言模型模拟的准实验研究”等看似新颖但脱离实际数据支撑的方向。团队成员通过提供明确的反馈,并输入现有的 StudyChat 数据集特征,将研究方向聚焦到基于真实日志数据的实证分析上。

二是写作大纲的修订。在正式撰写前,AI 生成了初版论文大纲。团队成员对其进行了结构调整,主要包括优化引言部分的行文结构(取消细分小点,更加符合中文期刊的写作要求),以及在研究方法中增补“AI 自动化案例分析与人工核验”环节。这一版经过团队成员修订的大纲,成为约束 AI 后续分步进行内容生成的框架。

(四)人类作者的审查、验证与质量控制

1. 如何解决生成式人工智能存在哪些已知的技术局限与潜在偏差

在本研究中,所使用的生成式 AI(特别是主体的生成模型 Gemini 3 Pro)主要存在以下技术局限与潜在偏差:

一是上下文窗口限制。目前的大语言模型对上下文的理解是有限的。当要求 AI 一次性生成超长学术文本时,其上下文理解能力会下降,易产生严重的逻辑断层和内容幻觉。本研究采用了“小样本代码生成+本地大数据运行”“提纲+分步骤生成”等分而治之的指令策略,前者提取了 380KB 的轻量化样本,确保 AI 将算力集中于代码编写,避免上下文过长引发的算力不足和幻觉;后者使 AI 能够在每一步生成时保持高度聚焦,并经团队成员即时修正后再进行下一段生成,有效提高文本的生成质量。

二是基于训练语料的语言表达偏差。Gemini 作为 Google 开发的模型,其底层训练语料以英文为主,在撰写中文学术论文时,可能存在中英文表达习惯冲突,导致部分专业词句生硬或难以兼顾中文学术写作的特定规范。本研究采取了跨模型审查和人工校验的策略,前者是在初稿形成后,调用具有本土优势的中文开源模型 DeepSeek-v3.1 和豆包 1.81.6 进行交叉审校,专门针对语言规范、语法错误与学术术语进行逻辑校对与文字修正;后者是人工校验,主要聚焦于学术概念的前后统一。

三是知识库局限与文献幻觉。GenAI 普遍存在捏造事实和虚构学术文献的技术局限。本研究采

取的措施是AI联网自查和人工核验策略，前者是在新的对话框中进行对话，要求AI进行联网核实信息；后者是团队成员借助Zotero文献管理对所有AI生成的参考文献逐条核验，确保文献引用100%真实准确。

2. 事实核查过程

本研究涉及的事实核查主要集中于AI生成的专业术语与文献引用，核查方式见上一问题的第二点与第三点。StudyChat数据集为开源的真实数据集，数据分析代码在本地电脑运行并输出结果，不涉及AI生成的数据。

3. 逻辑与科学性验证

针对AI提出的论点或推导过程，首先要求AI进行自查，然后人类研究者进一步复核，从以下三个维度进行验证：

一是数据的实证证明。针对AI通过滞后序列分析得出的“修正-纠错”高频循环结论，团队成员在本地环境中运行Python脚本，复核对话轮次的频数分布、单轮话语量（字符数）对比、行为转移概率矩阵等核心统计指标，确保AI提出的论点是建立在大样本（12 824条有效会话）且具有统计显著性的客观规律之上，而非大语言模型的随机生成或逻辑幻觉。

二是案例的微观质性分析。为避免纯定量数据的解释流于表面，团队成员结合AI输出的结论，人工回溯典型案例的交互语境，验证学习者在面对AI陷入算法固着时，是如何通过下达限制性指令来实施修补与元认知干预的。这种质性分析，在微观真实情境中验证了AI关于“学习者主体性重塑”与“元认知监控”推导的科学性。

三是已有文献的理论交叉验证。团队成员将AI从数据挖掘中涌现出的创新结论，与经典的教育学与认知科学理论进行交叉验证。将AI观察到的“试错-反馈-再修正”螺旋式闭环，与维果茨基的最近发展区理论及皮亚杰的认知冲突理论进行印证；将学习者将代码实现外包给AI的现象与认知卸载概念进行印证。通过这种理论层面的交叉验证，确保AI数据推导符合教育学科的底层逻辑与学习者的认知发展规律。

4. 研究可重复性的限制说明

考虑到生成式人工智能底层基于概率分布的自回归生成机制，其输出天然具备随机性与涌现性。因此，本研究在文本撰写的微观字句表述和初次灵感涌现的具体对话，不具备严格意义上的可重复性。如果另一位研究者输入完全相同的提示词，AI可能会生成不同表述或提出稍有差异的分析视角。然而，在宏观研究框架、核心数据特征规律以及底层数据分析逻辑上，本研究具备高度的可重复性。这种可重复性是建立在人类研究者一系列的控制之上的。为了最大程度降低AI输出随机性对科研严谨性的冲击，确保研究结论经得起同行验证，本团队采取了以下措施：

一是明确数据集来源与清洗规则。研究的实证数据并非AI虚构，而是采用真实的StudyChat公开数据集。同时，论文详细披露数据预处理的规则（剔除>40 000字符的异常值、保留≥6轮对话的深度会话，最终确定12 824条有效样本）。这确保了任何第三方研究者都可以获取并重构完全一致的数据集起点。

二是数据分析的“小样本代码生成+本地大数据运行”策略。面对海量数据，本团队并未让AI直接输出分析结果，而是要求AI撰写用于数据挖掘的Python分析代码。随后，团队成员在本地环境中运行这些代码。这种将非确定性的代码生成与确定性的代码执行相隔离的策略，确保了本研究中所有定量分析结果的绝对客观与可重复。

三是“提纲+分步骤生成”写作策略。在正文撰写阶段，为避免AI在超长上下文生成中的随机性漂移，团队采用了“先写提纲，再一部分一部分写作”的结构化写作策略，将AI的发散性限制在局部段落内，保证了全文逻辑的稳定。

（五）科研伦理与版权合规声明

1. 数据安全与隐私声明

本文作者及研究团队郑重承诺，在与生成式人工智能（Gemini 3 Pro、DeepSeek-v3.1、豆包）交互的全过程中，严格遵守国家相关法律法规及科研伦理规范，未输入任何未脱敏的人类受试者隐私数据或国家/商业机密，无任何数据泄露风险。

2. 原创性与查重说明

为了保证 AI 科研能力边界测试的客观性，作者团队刻意克制人工介入，未对 AI 生成的内容进行实质性改写，仅进行极个别学术名词的规范化统一，人工改动字数占比不足 1%。

在论文提交前，作者团队未进行查重或降重处理。在后续的审查环节，团队采用大雅相似度分析系统（超星），分别进行了文献查重和 AIGC 检测，论文的总文献相似度为 3.54%，疑似 AIGC 全文占比为 62.33%。这一结果侧面反映了本研究所使用的大模型在学术写作的自然度与拟人性上，已经具备规避部分机器检测的较高水平。

3. 责任承担声明

尽管 AI 被列为本文第一作者以体现其在研究全流程中的实质性贡献，但人类作者（程一可、钱江明、朱宇恒）对本文的学术准确性、科学严谨性以及所有伦理问题承担全部最终责任。AI 不具备法律主体资格，不能独立承担学术责任。将 AI 列为第一作者是一种学术实验性的署名实践，目的是推动学术共同体对人机协作署名规范的讨论。

（六）请人类作者谈谈参与此次活动的心得体会

首先，衷心感谢华东师范大学智能教育实验室发起此次极具前瞻性的“AI 驱动教育研究论文写作”活动，为我们提供了一个探索科学研究第五范式的宝贵平台与实验场。感谢《华东师范大学学报（教育科学版）》提供专刊发表的机会，并向所有评审专家的专业把关与宝贵指导，致以最衷心的感谢。

其次，在本次深度赋权的“AI 一作”科研范式体验中，我们真切感受到了生成式人工智能强大的科研潜力。AI 在处理海量数据、自动化编写代码、发现潜在规律以及进行万字长文的快速撰写上，展现出了令人惊叹的效率与涌现能力，极大地解放了人类研究者的科研生产力。

最后，也是本次实验最深刻的体会：尽管 AI 发展到了一种让人叹为观止的高度，但人类在科研中的主体地位似乎并未被削弱，反而变得更加重要。正如我们在论文中得出的结论一样，“在第五范式的科研与学习中，人类研究者/学习者的主体性并未消解，而是实现了形态上的转换，其角色从负责具体代码的实现，转变为对算法质量的评价与监督，并对技术路径进行规划与决策。”面对大模型的随机性与固有的幻觉，人类研究者的学术直觉、科研路径规划、元认知监控以及对学术伦理的底线审查，才是赋予整篇论文价值的关键。在未来的人机协同时代，AI 是强劲的效率引擎，而人类，始终是把握研究方向、完成最终价值判断并承担责任的绝对主体。

（程一可工作邮箱：2023209007@stu.njau.edu.cn；本文通信作者为钱江明：qjm52413@zju.edu.cn）

参考文献

戴岭, 祝智庭. (2024). 教育人工智能伦理与道德风险治理: 问题廓清与精准施策. *中国教育学刊*, 12, 31—37.

何克抗. (2018). 深度学习: 网络时代学习方式的变革. *教育研究*, 39(5), 111—115.

Baidoo-anu, D., & Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI*, 7(1), 52—62.

Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press.

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475—479.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25(4), 471—533.

Chiu, T. K. F.（2024）. The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. *Interactive Learning Environments*, 32(10), 6187—6203.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M.（2019）. Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637—643.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R.（2023）. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642.

Gee, J. P. (2014). *An Introduction to Discourse Analysis: Theory and Method* (4th edn). Routledge.

Glikson, E., & Woolley, A. W.（2020）. Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627—660.

Hou, H. -T., Chang, K. E., & Sung, Y. T. (2010). Applying lag sequential analysis to detect visual behavioural patterns of online learning activities. *British Journal of Educational Technology*, *41*(2), E25—E27.

Jansson, D. G., & Smith, S. M.（1991）. Design fixation. *Design Studies*, 12(1), 3—11.

Jonassen, D. H.（1997）. Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45(1), 65—94.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G.（2023）. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V.（2023）. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2), e0000198.

Landis, J. R., & Koch, G. G.（1977）. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159—174.

Lodge, J., Howard, S., Bearman, M. L., Dawson, P., & Associates. (2023). *Assessment Reform for The Age of Artificial Intelligence*. https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf.

McNichols, H., Ikram, F., & Lan, A. (2025). *The StudyChat Dataset: Student Dialogues With ChatGPT in an Artificial Intelligence Course* (No. arXiv: 2503.07928). arXiv.

Mollick, E. R., & Mollick, L. (2023). *Assigning AI: Seven Approaches for Students, with Prompts* (SSRN Scholarly Paper No. 4475995). Social Science Research Network.

Piaget, J. (1985). *The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development*. University of Chicago Press.

Puntambekar, S., & Hubscher, R. (2005). Tools for Scaffolding Students in a Complex Learning Environment: What Have We Gained and What Have We Missed? *Educational Psychologist*.

Risko, E. F., & Gilbert, S. J.（2016）. Cognitive Offloading. *Trends in Cognitive Sciences*, 20(9), 676—688.

Sharma, P., Akgun, M., & Li, Q.（2024）. Understanding student interaction and cognitive engagement in online discussions using social network and discourse analyses. *Educational Technology Research and Development*, 72(5), 2631—2654.

Sinclair, J., Sinclair, J. M., & Coulthard, M. (1975). *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E.（2018）. Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, 76(3), 467—481.

VanLEHN, K.（2011）. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4), 197—221.

Vygotsky, L. S. (1978). *Mind in Society: Development of Higher Psychological Processes*. Harvard University Press.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (No. arXiv: 2302.11382). arXiv.

Wood, D., Bruner, J. S., & Ross, G.（1976）. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89—100.

（责任编辑　童想文）

“Deep Cognitive Scaffolding in Human-AI Collaboration: A Lag Sequential Analysis Based on Learner-GenAI Multi-Turn Dialogues”: A Detailed Exposition and a Disclosure Statement of the Research and Writing Process

Cheng Yike Qian Jiangming Zhu Yuheng
(College of Public Administration, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: This paper is one of the publications featured in the “Human-AI Co-Creation Pioneer Papers Ranking”, which is part of the Panoramic Report on the World’s First Large-Scale Social Experiment of “AI as the First Author”.The first part of this paper presents a detailed exposition of the research content: based on 12,824 log entries of multi-turn dialogues between learners and GenAI, it employs Lag Sequential Analysis (LSA) to investigate the cognitive scaffolding mechanism in human-AI collaboration, revealing a high-frequency closed-loop of “trial-error, feedback, and re-correction” as well as the critical role of learners' metacognitive monitoring in reshaping agency when GenAI encounters algorithmic fixation. The second part constitutes a transparent disclosure of the entire research and writing workflow. It systematically sorts out the various large language model tools adopted in this human-AI joint research, the division of rights and responsibilities between AI and human researchers, and the hierarchical and decomposed prompt engineering design schemes. It fully reconstructs the full-chain operational details covering data processing, paper drafting, cross-model cross-review, manual factual verification and academic ethics governance. Meanwhile, it elaborates on key practical issues including ethical definition of the innovative AI-authorship experiment, originality inspection of research outputs, constraints on research reproducibility, and the allocation of ultimate accountability among human researchers. The findings indicate that: (1) learner-GenAI interactions exhibit significant long-term and asymmetric characteristics, with GenAI providing continuous scaffolding support through an asymmetric discourse pattern of “few questions, many answers”; (2) at the behavioral sequence level, a high-frequency closed-loop exists between learners’ self-correction and GenAI’s corrective guidance, demonstrating that deep learning does not occur in single Q&A exchanges but is achieved through a spiral process of “trial-and-error, feedback, and re-correction”; (3) micro-case analysis reveals that when GenAI falls into algorithmic fixation, learners swiftly shift from questioners to strategy formulators, implementing critical interventions through metacognitive monitoring. GenAI should not be regarded merely as an information retrieval tool in future education; instead, it should serve as a “Socratic tutor” that inspires thinking, reshapes learner agency, and jointly constructs a new educational ecology of human-AI symbiosis. The value of this paper lies in providing tangible empirical details for the grand narrative of “human-AI collaboration,” while offering the academic community a referential sample for understanding the boundaries of AI’s research capabilities through full-process transparent disclosure.

Keywords: human-AI collaboration; Generative Artificial Intelligence (GenAI); cognitive scaffolding; Lag Sequential Analysis (LSA); educational data mining