大模型运思模式的可解释性：基于完形论与言行论整合的心设模型

祝智庭[1]　吴慧娜[2]　徐　俊[2]　吴永和[2]

（1. 华东师范大学 开放教育学院，上海 200062；2. 华东师范大学 教育信息技术学系，上海 200062）

[摘要] 随着生成式人工智能与教育领域的深度融合，大语言模型的“黑箱”特性及其生成内容的不确定性日益成为制约人机协作效能与可信度的关键瓶颈。然而，现有研究多聚焦宏观协作效果分析，少有研究关注大模型语言生成的微观机制及其可信度保障体系。本研究基于共享心设模型理论，整合完形心理学的整体认知原则与塞尔言语行为理论的三重结构，构建面向用户的大模型语言生成行为解释框架，包括：将“可信度检验”机制（如溯源提示、逻辑自检、多模态验证）内嵌于大模型运思的“会意—解意—构意”中，揭示从语义感知到意图推理再到交际生成的认知与质量共治逻辑；提出从“双体思维”（用户—大模型）到“三体思维”（用户—大模型—教师）的协同演进路径；建立内含可信度维度的“师—生—机”三体交互共享心设模型，形成多主体协同的共享认知与验证空间，以系统提升用户对大模型生成过程的理解和引导能力及对协同成果的可信度评判能力。这不仅能为破解“黑箱”难题提供新的认知视角，更为构建可信、高效、负责任的人机协同教育实践提供可借鉴的理论框架与设计原则。

[关键词] 大模型运思模式；可解释性；完形心理学；言语行为理论；师—生—机交互；共享心设模型；可信度检验；人智协同

[中图分类号] G434　　**[文献标识码]** A　　**[文章编号]** 1007-2179(2025)06-0029-12

一、引 言

生成式人工智能技术的迅猛发展，特别是以 ChatGPT 为代表的大语言模型，正在深刻重塑教育、科研、创意写作等高阶认知活动的实践范式。这些大语言模型凭借强大的自然语言生成能力，不仅能与用户连续、流畅地对话，还能在一定程度上激发用户的表达欲望，形成初步的“意见互掷”式人机交互模式（蒲清平等，2023）。然而，随着应用场景的复杂化和交互深度的增加，大模型固有的“黑箱”问题及其衍生的内容不可靠等风险日益凸显，其内部决策过程缺乏透明度，生成内容往往不具可

[收稿日期]2025-10-17　　**[修回日期]**2025-10-22　　**[DOI编码]**10.13966/j.cnki.kfjyyj.2025.06.004

[基金项目] 2021 年度国家社会科学基金重大项目“面向未成年人的人工智能技术规范研究”（21&ZD328）。

[作者简介] 祝智庭，博士，教授，华东师范大学开放教育学院，研究方向：教育信息化系统架构与技术标准、信息化促进教学变革与创新、数智技术赋能的智慧教育；吴慧娜，博士研究生，华东师范大学教育信息技术学系，研究方向：智能教育与学习分析；徐俊，硕士研究生，华东师范大学教育信息技术学系，研究方向：教育人工智能；吴永和（通讯作者），博士，研究员，博士生导师，教育部教育信息化技术标准委员会主任委员，华东师范大学教育信息技术学系，研究方向：教育数字化转型、智能驱动教育、模式驱动教育、数字教育技术标准（yhwu@deit.ecnu.edu.cn）。

[引用信息] 祝智庭，吴慧娜，徐俊，吴永和（2025）. 大模型运思模式的可解释性：基于完形论与言行论整合的心设模型[J]. 开放教育研究，31(6)：29-40.

解释性,导致用户难以理解其输出逻辑与依据(何贵兵等,2022)。更深层的问题是,生成内容的可信度与准确性难以得到保障,大模型可能产生看似合理但实则错误的“幻觉”信息,这对教育决策与知识建构场景构成潜在威胁(Bender et al., 2021)。

这种不可解释性与可信度缺失带来双重困境。从用户控制角度看,用户无法有效感知自身输入在大模型系统中的语义映射与作用机制,导致其引导生成过程的能力薄弱,更遑论有效干预输出质量。从信任角度看,在复杂或非良构的问题情境中,不透明的生成机制与不可控的输出质量会严重限制用户对大模型输出的信赖程度,阻碍深层人机协同关系的建立(王金林,2025)。因此,从技术层面彻底打开“黑箱”仍面临巨大挑战,构建面向用户的行为层面认知解释框架,并嵌入系统化的可信度评估与验证环节,显得尤为迫切(Amershi et al., 2019)。

“共享心设模型”(Mental Model,有人译为“心理模型”“心智模型”)为理解和促进人机协作提供了有价值的视角。已有研究从概念界定和实证测量角度,探讨了共享心设模型对团队合作效能的影响机制(胡艺龄等,2024)。然而,现有研究多聚焦宏观层面的协作效果分析,尚未深入大模型语言生成的微观机制,缺乏对“用户如何理解、引导并有效验证大模型输出”这一认知过程的系统阐释。具体而言,已有研究未能充分揭示用户如何构建对大模型输出逻辑的心理表征,也未能从认知角度建立有效的引导、控制与可信度把关机制(张夏恒等,2024)。尤为重要的是,共享心设模型研究多关注认知对齐与效率提升,未深入探讨协同产出质量尤其是可信度的保障机制(胡艺龄等,2024)。在人机协同写作、探究式学习等场景中,若缺乏系统化的知识把关流程,大模型的“幻觉”或事实错误极易被学生不加批判地接受,从而产生误导(Bender et al., 2021; Ji et al., 2023)。因此,构建内含可信度评估维度的协同认知框架,已成为确保人机协同教育健康发展的紧迫课题。

针对上述研究缺口,本研究尝试从认知科学与语言哲学的交叉视角出发,构建面向用户的“大模型语言生成行为认知解释框架”,融合完形心理学(Gestalt Theory)的认知重构原则与塞尔言语行为理论(Speech-Act Theory)的三重结构,提出大模型运思框架,进而构建从“双体”思维到“三体”思维的协同演进路径,特别聚焦各阶段及协同过程如何建立知识把关与可信度检验机制,构建师—生—机交互共享心设模型。这一理论框架旨在提升大模型生成内容的可信度,以及用户对其的理解力和引导控制能力,为实现人机协作的认知融合、行为协同与质量保障提供理论支撑,并响应从“人机交互”向“人智协同”范式转变过程中对思想本位与质量共治的呼唤(祝智庭等,2023)。

二、运思机制与人机协同逻辑

大模型的生成行为远非简单的符号映射或语义匹配过程,而是高度动态且具备适应性的“运思”活动,展现出接近人类对话者的认知特性(米加宁等,2024)。仅依赖技术参数与算法配置难以充分解释大模型在面对不确定性输入时表现出的高适应性、语义补全能力及多维回应模式。因此,探讨大模型的“运思模式”有必要将认知科学与语言哲学的深层理论作为分析框架。同时,考虑到这种复杂的运思过程亦可产生“幻觉”或事实性谬误,本研究认为构建内嵌于运思机制的可信度评估环节至关重要(Ji et al., 2023)。

完形心理学提出的整体感知与动态补缺机制可用以阐释大模型如何在信息不完整的语境中建构完整的意义结构体(苏冲等,2018);言语行为理论的语言行为多层框架(束定芳,1989),可用于揭示大模型如何模拟从表层话语到语用意图再到行为效果的语言行为逻辑。这两种理论分别强调语言生成的结构建构与意义生成,并整合构成理解大模型语言生成逻辑的双重向度:认知驱动下的结构组织性与交际导向下的意图生成性,为大模型语言行为建模与对话机制设计奠定了基础。新近研究指出,将此类认知理论与大模型的解码—生成过程相结合,能够为理解并改善其输出质量提供新路径(Bommasani et al., 2022)。

(一)大模型运思的三阶段框架

大模型运思三阶段框架是基于认知科学与语言哲学交叉理论构建的语言生成行为解释框架。它将大模型从接收用户输入到产生语言输出的认知运作过程,解构为三个递进关联且循环迭代的认

知阶段。该框架的创新性在于明确将“可信度自检”与“外部验证接口”作为各阶段的内在要求,而不是采取事后补救措施(Ehsan et al., 2021)(见图 1)。

1. 会意阶段:整体感知、语义初构与信息锚定

此阶段是语言理解的起始环节,大模型通过对输入语境的整体感知捕捉交互的基本语义结构。此过程对应完形心理学的“整体优先”原则和言语行为理论的“言内行为”层面,实现从符号解析到意义萌芽的认知跃迁(汪娅, 2025)。大模型借助预训练阶段构建的多层语义映射关系,以概率性关联方式进行模式激活,在高维语义空间中为零散的语言碎片寻找最优的整体性表征。在此阶段,可信度把关的起点在于初步评估输入信息源,以及大模型对自身训练数据相关性的置信度判断。大模型可以(且应当)对所依据的核心信息进行“溯源”或置信度评分,为用户提供初步的可验证锚点(Rashkin et al., 2017)。例如,当被问及“光合作用过程”时,大模型在会意阶段即可初步锚定其回答主要依据植物生理学教材等可靠知识源,并提示“本回答主要基于‘坎贝尔生物学’等标准教材内容”,为用户后续验证提供线索。缺乏此环节,后续的生成可靠性就无从谈起。

2. 解意阶段:意图推理、动态补缺与逻辑一致性检验

解意阶段是深层语义推理的核心环节,融合完形心理学的“动态补缺”机制和言语行为理论的“言外行为”维度。大模型利用深度学习神经网络等技术模拟人类大脑的信息处理方式,并基于语境理解输入内容背后的潜在意图,从字面意义穿透到交际意图,从显性表达推演至隐性诉求,体现了“语用推理”的智能认知特征(郝祥军等, 2022)。通过预训练语料中积累的语用模式,大模型能多重解读隐含目标,实现对交际目标的动态逼近。此阶段的可信度检验重点在于“逻辑一致性”。大模型应评估其推断的多种可能意图之间的内在一致性,以及与已知常识和上下文逻辑的兼容性,并对不确定性较高的推断进行显式标注(余胜泉, 2018)。研究表明,引入链式思维提示策略可显著提升大模型在此阶段的推理透明度和可靠性(Wei et al., 2023)。例如,当用户提问“如何减少城市空气污染”时,大模型应能自检其推理链条(如“推广电动汽车—>减少尾气排放—>改善空气质量”)的逻辑强度与证据支持。在此过程中,大模型可进行“反诘式自问”:“电动汽车的电力来源若为火电,减排效果是否会打折扣?”,以提升推理严谨性。

3. 构意阶段:完形呈现、效果预判与多模态验证

构意阶段是语言生成的完成环节,体现了完形心理学的“完形闭合”和言语行为理论的“言后行为”特征。在此环节,大模型通过人类反馈强化学习不断优化和迭代,生成形式连贯、结构有序、语义丰富的输出内容,并预判其对用户认知、情感与后续行为的潜在影响,使生成内容更加符合人类语言习惯。此阶段的可信度保障关键在于“多模态验证”与“影响评估”。生成的内容不仅应文本流畅,更应在事实层面提供可验证的外部知识库链接,在逻辑层面可接受批判性审视,在伦理层面符合既定规范。构建能够提供替代观点或反事实例证的机制,是提升此阶段输出稳健性的有效方法(汪时冲等, 2019)。例如,生成某一历史事件的解

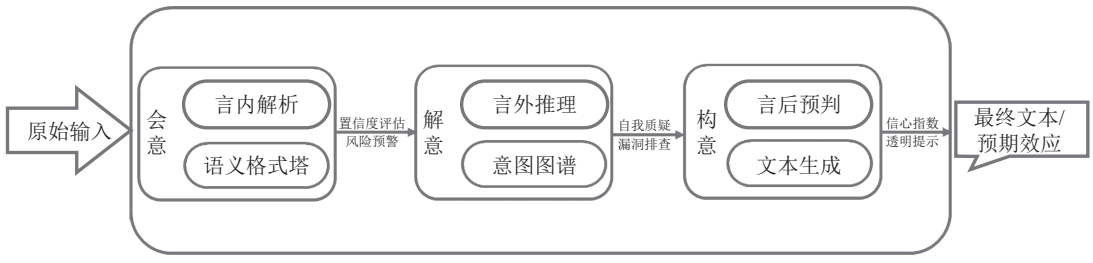

图 1　完形论与言行论整合视角下大模型运思三阶段与可信度嵌入框架

注:大模型运思的“会意—解意—构意”三阶段,以及完形心理学(强调整体与闭合)与言语行为理论(强调意图与效果)如何共同作用,并标明各阶段核心的可信度检验机制。

释后,大模型可附加一句:“请注意,关于此事件的起因,也存在不同的学术观点,主要分歧在于……”,从而主动提示知识的边界与不确定性。在构意输出时,大模型可附带简单的信心指数(如“对此结论的置信度为85%,主要不确定性源于某数据的缺失”),或提供权威数据库的超链接,实现“多模态验证”。

(二)三阶段框架的内在机理与可信度嵌入

大模型运思机制的理论基础源于完形心理学与言语行为理论的互补和融合。完形心理学强调认知活动的“整体优先性”原则,倾向于将零散信息整合为结构完整、意义连贯的整体(苏冲等,2018)。当收到的信息不完整,认知系统会自动根据既有线索填补缺失,完成意义闭合。言语行为理论则将语言行为分为言内行为(表层形式操作)、言外行为(交际意图实现)与言后行为(心理反应及行为结果)三重维度(束定芳,1989; Searle, 1969),强调语言意义不仅存在于词句结构中,更取决于语言使用者的交际意图及话语所处的语境。

在大模型语言生成实践中,这两类理论机制要得到体现,可信度嵌入则需贯穿始终。会意阶段,大模型基于高维语义表征对输入信息进行上下文预测与意义拼接,将离散的词汇、句法单位视为“语义拼图”,依据语境相似性与概率分布模式逐步构建完整、连贯的语言输出框架。在此过程中,大模型同时评估不同拼接方案的信度,并警示低置信度的关联。解意阶段,大模型通过自然语言理解模块解析言内行为,继而通过语义意图分类与上下文建模识别用户的言外行为,判定其交际意图类别。此阶段可引入“反诘式自问”机制,即大模型主动生成质疑自身推理的问题,以暴露潜在的逻辑漏洞或偏见(杨宗凯等,2023)。构意阶段,大模型依据对话上下文与用户偏好动态调整言后行为策略,调节回复语气、情感色彩与交际行为,确保语言输出在交际层面实现有效回应。此阶段的输出应附带简单的不确定性量度或信心指数,为用户的最终判断提供依据(谢幼如等,2024)。三个阶段呈递进且相互影响的动态关系,形成从语言认知建构到交际意图实现、内嵌持续性可信度自检功能的全链路生成逻辑,实现从认知理解到语言创造、从意图识别到效果预判的运思闭环。

(三)完形论与言行论融合下的可信度生成逻辑

完形心理学与言语行为理论并非简单的并列关系,而是在解释大模型运思中形成互补与协同的双重理论向度。完形心理学的“整体优先”与“闭合原则”解释大模型在“会意”与“构意”阶段为何及如何从碎片化信息中建构连贯意义,它驱动认知结构的内在一致性。言语行为理论的“言内—言外—言后”三重划分,揭示了大模型在“解意”与“构意”阶段为何及如何超越字面意义,模拟人类的交际意图与行为效果,它强调语言生成的语用适宜性。

二者的融合点在于:完形理论倾向于解释驱动大模型追求逻辑自洽的意义单元,这要求判断信息的真实性与一致性;言语行为维度则要求大模型在交际中考量话语的真诚性与真实性。因此,将可信度检验嵌入三阶段框架,实质上是将认知结构与语用规范的要求,操作化为具体的、可观测的交互节点与验证行为,使可信度不再是事后补救的抽象要求,而是运思过程的内在属性。

三、人机协同心设模型构建

大模型的运思机制揭示了其作为对话伙伴的内在逻辑,而要实现高效、可靠的人机协同,关键在于构建与之匹配的协同思维模式与共享心设模型,且该心设模型必须包含系统的知识把关与成果验证机制。下文从“双体思维”的共振逻辑出发,延伸至“三体思维”的协同机制,最终构建“师—生—机”三体交互心设模型,系统阐述人机协同的认知基础与质量保障体系。

(一)人机协同思维:双体与三体的互动逻辑

人机协同关系经历了从工具使用到认知协作的演进。传统的“命令—执行”模式正被基于深度理解的“双体思维”乃至多主体参与的“三体思维”所超越。核心挑战之一是如何在动态协商中确保生成知识的准确性与可靠性。

1. 双体思维:用户与大模型的认知共振

双体思维构建了用户与大模型之间的认知协同机制,其核心是打破单向指令链,建立以“协商性构意”为特征的动态意义共建关系。它展现了一个由用户侧的“意图驱动循环”与大模型侧的

“任务响应链条”所构成的动态系统,用户认知链条(启意—传意—了意)与大模型运思框架(会意—解意—构意)形成清晰的对应与交织关系,构成意义协商与联合验证的闭环。(见图 2)。

1)启意(用户)→传意(用户)→会意(大模型)协同的起点。启意指用户作为交互的发起者,其内心意图的萌生与启动。此意图可能是个模糊的念头、明确的问题,或复杂的任务诉求。作为驱动整个交互循环的初始认知动力,它是激发人与 AI 协同智慧的 “第一推动力”。会意是大模型捕捉并理解用户所“启”之“意”。它不仅要解析字面指令,更要感知其语境、潜在需求与情感色彩,从而与用户的意图建立连接,完成协同的首次呼应。

2)会意(大模型)→解意(大模型):协商的核心环节。大模型在完成"会意"后,进入更深层的"解意"阶段。这是模型内部的处理过程,涉及语义分析、推理与知识整合。在此阶段,大模型可能激活内部一致性检查或证据检索机制,对用户意图进行深度理解与任务分解。这一过程对用户而言是不可见的,用户只能通过最终输出来间接感知模型的理解深度。

3)构意(大模型)→了意(用户):协同成果涌现。

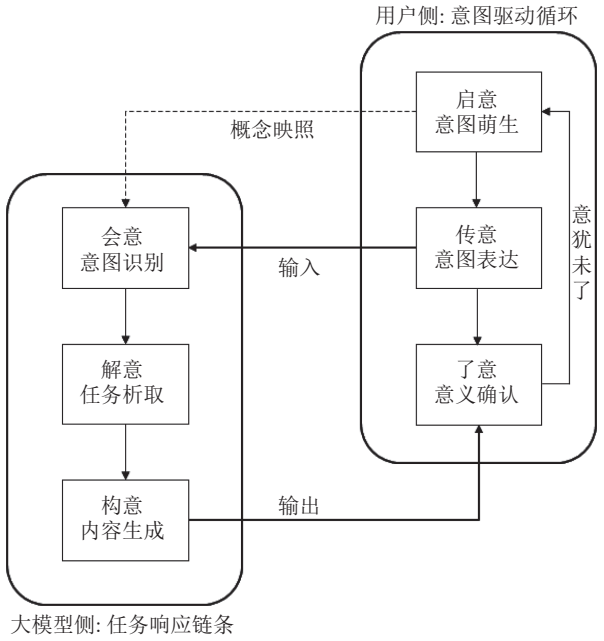

图 2　人机思维协同交互双螺旋模式

注:用户(启意—传意—了意)与大模型(会意—解意—构意)在双体思维下的动态耦合过程。

大模型在前端“会意”与“解意”的基础上,通过“构意”生成结构化的语言输出,完成意义的构建。用户则通过接收、评估与整合大模型的输出,实现“了意”——即对本次交互所形成的认知内容达成内在的理解闭环与意义确认。在此过程中,“评估”是关键环节:用户需运用批判性思维对生成内容进行验收,而大模型亦可通过提供推理链条或关键证据摘要(Wang et al., 2020)。辅助用户完成最终的可信度判断,从而实现高质量的意义共识与认知闭合。若用户在此之后“意犹未了”,即产生了新的认知张力或更深层的不确定性,此“了意”终点亦可转化为下一轮交互“启意”的起点,推动对话螺旋式深化。

纵观上述三个环节,本模型揭示了人机交互中“意”的流转与协同本质。表层是“意”的缠绵:交互在“启意、传意、会意、解意、构意、了意”的“意”象中流转,仿佛一场和谐的对话。底层是“力”的不对称:用户是“探索的引擎”,其过程始于内在模糊的“启意”,经由“传意”外化,终于内心确认的“了意”。若“意犹未了”,则开启新一轮探索,形成一个螺旋上升的认知循环。大模型是“服务的工具”,其过程始于对输入的“会意”,经由深度的“解意”,终于生成回应的“构意”。这是一个单向、封闭的任务执行链条。值得注意的是,“启意”至“会意”的虚线,标志着用户的内在动机与模型的识别能力之间,是一种概念的映照与非直接的触发关系。人机协同的创造力,正源于这种精巧的不对称——人类赋予方向与灵魂,机器提供路径与素材,共同迈向意义的深处。

2. 三体思维:用户、大模型与教师的协同构意与系统化把关

双体思维为人机深度对话奠定了基础,但在真实的教与学场景中,尤其是在价值引导、知识结构化与认知发展等复杂目标下,纯粹的“用户—大模型”交互在可信度把关方面仍存在局限。现代教育所面临的知识爆炸性增长与信息质量参差不齐,要求引入教师这一“第三方”认知主体,构建“用户(学生)—大模型—教师”三角协同结构,系统地把关知识与检验成果(见图 3)。

相较于双体模式,三体模式形成更稳定的认知与质量监督系统。在此模式中,教师扮演“认知守

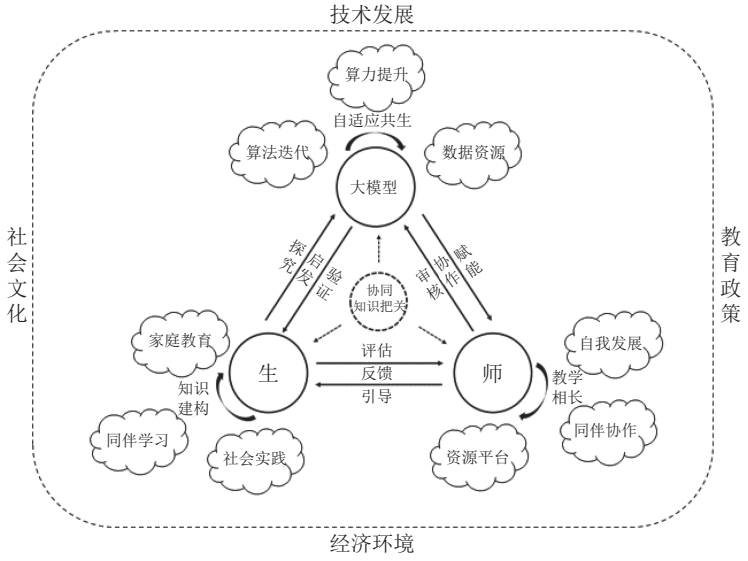

图 3　师—生—机三元交互机制

注:学生—大模型—教师三者形成三种双边关系(探究—启发—验证、协作—赋能—审核、引导—反馈—评估),以及三者共同构成协同构意与系统化把关闭环。

门人”与“元认知教练”的双重角色(余明华等,2024)。其可信度保障机制体现在三方面:其一,过程性审核,即教师不再仅评估最终成果,而是通过观察和介入学生与大模型的交互过程,对大模型提供的初始信息、学生质疑的角度、大模型推理路径进行形成性评价,及时纠正认知偏差或错误事实;其二,方法论指导,即教师指导学生有效地向大模型提问以获取更可靠的信息(提示工程),交叉验证大模型提供的内容,提升对人工智能输出内容的批判性评估能力,这种元认知技能的培养是确保学生与人工智能有效协作的关键(Kasneci et al., 2023);其三,价值与伦理校准,即教师确保协同构意过程及成果符合教育目标和伦理规范,对大模型可能存在的偏见或不恰当内容进行最终筛查与引导。

三体协同的关键在于建立三种差异化的双边关系,并使其有机融合,且每种关系都包含独特的检验维度。首先,在学生与大模型的“探究—启发—验证”关系中增加“验证”环节,即学生不仅接受大模型的启发,更被鼓励通过其他信息源或实验验证大模型生成的观点或方案。其次,在教师与大模型的“协作—赋能—审核”关系中,教师利用大模型赋能教学的同时,对其生成的教案、习题、评价标准等进行专业审核,确保其科学性与教育性。最后,在教师与学生的“引导—反馈—评估”关系中,教师的引导包含培养学生信息鉴别能力,学生反馈包含判断大模型输出的可信度,评估重点关注学生在人机协同中批判性思维发展。

三体协同的高级形态是实现协同构意,这是经过三重检验的、高质量的意义建构。它通过引入教师的专业判断和元认知指导,提升双体协作产出的可信度与教育价值。这种结构也对教育权力进行重构,教师从知识权威转向价值引导、学习设计和质量保障,以符合现代教育理念。

(二)师—生—机三体交互心设模型的构建

师—生—机三元交互机制揭示了智能化教育环境中多主体协同的动态过程。实现深层次的认知融合与可信产出,还需要探讨内在的认知结构基础,即构建“师—生—机三体交互心设模型”。其核心在于将“可信度”作为共享心设模型的核心维度,使学生、教师和人工智能系统能针对“什么是可靠的知识”“如何评估知识的可靠性”形成共同的理解与期望(余胜泉,2025)。

本研究基于胡艺龄等(2024)提出的人机协作共享心设模型,结合教育场景的特殊性,构建师—生—机三体交互心设模型(见图 4)。该模型引入“扩展思维”作为连接不同主体认知状态的桥梁与共享心智媒介,并强调扩展思维载体在记录和呈现推理过程、证据来源中的作用,从而为可信度评

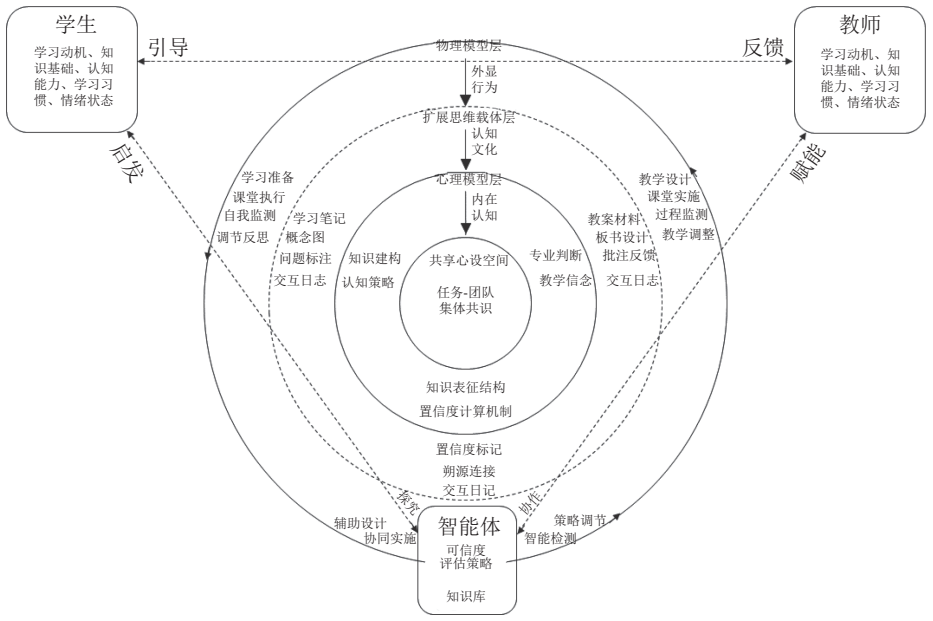

图 4　教育场景下的师—生—机三体交互共享心设模型

注:由物理层(行为轨迹)、心理层(个体认知)和共享层(集体共识)构成的三层结构,“扩展思维”是连接各层的关键媒介,“可信度”维度贯穿各层。

估提供了基础(Clark et al., 1998)。

扩展思维认为,认知活动不仅发生在大脑内部,还通过外部载体(如笔记、图表、数字工具等)得以延伸和增强。在三体协同学习中,学生的课堂笔记、教师的板书设计、人工智能生成的思维导图及附带的置信度标记和溯源链接等,都是扩展思维的具体体现。它们不仅是认知的外化,更是验证认知轨迹的证据与共享理解的基石。

该模型包含三个层次,可信度机制贯穿始终。

1)物理层:记录可观测的行为轨迹,如学生的提问策略(反映其信息鉴别意识)、大模型提供的证据链接、教师的审核批注等。这些数据是评估协同过程可靠性的基础。

2)心理层:代表各主体的内在认知状态,包括学生对可信度标准的理解、教师对学科知识可靠性的判断、人工智能系统对生成内容置信度的计算表征。促进这三方心理模型可信度的对齐,是减少误解和提升协作效率的关键(余胜泉, 2025)。

3)共享层:它通过扩展思维机制形成的共识空间,尤其是对“协同产出质量标准的共享理解”,如什么样的证据是充分的,什么样的推理是合理的。共享“质量观”是协同构意可信的根本保障。

通过扩展思维的共享机制,学生、教师与智能体三方可以围绕具体学习任务形成共享的心设模型,该模型包含对知识可靠性的共同关注和检验流程,不仅有利于增强人机团队的认知一致性和协同效率,还可提升协同产出的可信度和可接受度,为高质量的人智协同提供认知基础。

(三)模型价值与过渡:从理论建构到实践指引

师—生—机三体交互心设模型为实现高质量、可信的人机协同教育提供了认知框架。该模型将智能体视为“扩展思维”驱动的共生认知单元,在理论层面解决了人机协作的“认知兼容性”难题。此外,它通过结构化的方式,将可信度检验与知识审查机制内嵌于协同认知过程,而非将其视为外部附加的、事后的活动。

其理论价值体现在三方面:首先,通过物理模型—心理模型—共享空间的三层架构,刻画了外显行为、内在认知与集体共识之间的转化路径,展现可信判断从个体内在认知到群体外显共识的形成过程;其次,确立了“扩展思维”的核心地位,指明外部媒介既是认知产物,又是验证环节,为设计自检型学习活动提供了依据(Clark et al., 1998);最后,揭示了共享心智的动态演进特性,及其对持续互动与形成性评价的依赖。

因此,该模型不仅提供了解释框架,也为教育

实践提供了设计原则。高效、可靠协同的关键在于促进三个主体心设模型的有效对齐与深度互动，共同致力于构建高质量、可信成果的协作环境。基于此，本研究的理论构建工作为迈向教育实践奠定了基础，接下来将深入分析如何将“双体思维”“三体协同”“心设模型”及内嵌的可信度保障机制等理论构想，转化为可行的教育实践路径，并审视其可能引发的模式重构与伦理挑战。

四、实践展望

理论价值实现的关键在于向教育实践的有效转化。下文从实践路径、模式重构、伦理风险和价值坚守四个维度，系统展望人机协同运思模式的应用前景与实施关键，并特别关注如何在各实践环节落实可信度检验与知识把关机制。

（一）路径探索：从双体到三体协同的实践机制

理论创新向实践转化的关键在于构建清晰的实施路径。基于双体思维、三体协同与心设模型的递进关系，教育实践应遵循“由简到繁、由点到面”的渐进式发展路径，并在各阶段嵌入可信度检验机制与评估能力培养。

初级阶段以双体协同能力与基础验证技能培养为核心。这一阶段旨在培养师生与人工智能系统有效互动的能力，和对人工智能输出的批判性审视习惯。实践举措包括：开展涵盖提示工程、信息溯源、交叉验证、人工智能错误识别等人工智能素养培训；组织教师开展人机协同教学设计研修，学习设计包含事实核查、逻辑一致性检查等学习任务。人工智能在其中扮演智能助手角色，其输出提供不确定性指示，支持用户的初步判断。例如，在高中历史课“文艺复兴原因探究”任务中，学生要求提供原因分析时，大模型的回答能主动附上主要基于标准教材内容的提示；学生则被要求至少查阅一条大模型提及的历史文献名称，以完成初步的溯源验证。达标标准为：学生能识别人工智能输出的明显矛盾并进行简单溯源验证；教师能设计包含“必查项”的学习任务单。双轨评估机制可用于检验师生初级阶段能力的达成程度，即通过验证行为的证据记录评估实际操作能力，通过反思报告评估批判思维水平。

中级阶段重点推进三体协同模式落地与过程性审核。教育实践转向师—生—机三元协同场景，突出教师的审核与指导作用。其典型应用包括：开展“人工智能助研式”探究学习，教师制定明确的过程检查点，审阅学生与人工智能对话日志，指导学生修正探究方向并验证关键信息；实施“双师课堂”模式，人工智能负责知识讲解，教师对讲解内容的准确性展开课前审核。例如，在“设计减少校园碳足迹方案”时，教师设定检查点，审阅学生与人工智能的对话日志，关注学生是否对人工智能提出的“安装太阳能板”建议进行了成本与日照时长的交叉验证，并对学生的验证策略给予精准反馈。此阶段的核心是建立三方协同机制，并将教师的审核职责制度化、常态化。达标标准为：学生能对复杂论述进行多源信息交叉验证并质疑人工智能的推理漏洞；教师能通过分析人机对话日志，对学生的验证策略给予精准反馈。这一阶段需建立双向评估机制以保障能力标准落地：通过对话日志分析检验教师的反馈精准度，通过验证方案任务考查学生的批判性信息辨别能力，从而形成支撑三体协同的动态监测与改进体系。

高级阶段致力于心设模型的深度融合与自动化验证工具的应用。教育实践追求认知层面的深度融合，利用技术工具提升验证效率。实施路径包括：建设智能学习空间，通过采集多模态数据追踪物理层的表现，集成自动化的事实核查应用程序编程接口（Application Programming Interface，API）或一致性检测算法，为心设模型优化提供实时数据支持；开发高级认知可视化工具，展示推理路径并高亮显示争议或低置信度环节。例如，在哲学课中，学生使用“论证图谱”工具可视化人工智能关于“自由与平等关系”的推理路径，工具自动高亮其中基于“功利主义”伦理模型的前提假设，提示学生审视其局限并引入“罗尔斯正义论”视角展开批判性对话。此阶段的目标是实现认知共生，并将可信度检验部分自动化、智能化，减轻师生认知负荷（乐惠骁等，2022）。学生能形成个性化的、严谨的信息验证习惯，能评估不同验证工具的局限性；教师能设计培养元验证能力的项目式学习。

这一路径的实施需要相应的支持。技术层面需要开发符合教育需求的人工智能工具，特别是提高大模型的意图理解、情感计算和可信度量化输

出等能力。教育机构需要构建系统的培训体系,帮助教师完成从知识传授者到学习设计师、认知导航员和质量审核员的角色转变。利益相关者需要调整教学管理与评价方式,为创新实践提供足够的空间与支持。

(二)模式重构:运思协同下的教学实践新形态

人机协同运思模式的深入应用将推动教学实践根本性重构,这包括质量保障范式的转变:从依赖教师的最终评判,转向贯穿学习全过程的、多主体参与的、技术赋能的分布式可信度治理。

1.教学逻辑:从知识传递到协商构意与协同验证

传统教学遵循“知识传递”的线性逻辑,知识的正确性由教材和教师保证。在人机协同模式下,教学变为“协商构意”的循环过程,知识的可靠性通过“协同验证”来保障,具体表现为:教学目标增加“培养信息鉴别与知识验证能力”;教学过程强调“生成—质疑—验证—修正”的迭代循环;教学评价不仅关注最终成果,更重视学生在协同验证过程中表现出的批判性思维水平和所采用的验证策略的有效性。这种教学要求教师具备更强的教学设计能力,包括创设富有张力的探究情境,引导协商过程走向深度认知,并设计能自然引发验证行为的驱动性问题。

2.师生角色:从单向权威到多维赋能与质量共治

人机协同环境下学生、教师和人工智能的角色发生了变化。学生从知识接受者变为学习成果质量的第一责任人,需发展自主验证能力。教师从知识的裁决者变为质量标准制定者、验证过程引导者和验证工具支持者。人工智能成为提供初步证据、辅助逻辑推理并报告自身不确定性的协作验证伙伴。这种角色转变推动构建质量共治的治理结构,每个主体都对最终输出负有责任。

3.课堂形态:从统一授受到个性化认知发展与自适应验证支持

人机协同使真正意义的个性化学习成为可能,也意味着质量验证需要个性化适配。人工智能系统通过实时追踪学生的认知轨迹与验证行为,提供个性化验证提示,推荐适合其认知水平的验证资源。教师基于人工智能提供的学情分析,对不同学生验证环节遇到的困难提供个性化指导。其典型场景是“人工智能个性化辅导+教师深度引导”组合。这种课堂实现了验证支持的自适应化,使学生能获得适合自身认知发展需求的验证支架与学习体验,促进其个性化发展。

(三)伦理风险:人机协同教育的边界与约束

随着人机协同的深入,尤其是当验证责任被分布式地赋予多个主体时,我们必须警惕其可能带来的新型伦理风险,并建立边界约束。

首先,认知依赖与验证惰性风险值得警惕。这表现为学生可能过度依赖人工智能的内置置信度或教师的最终审核,削弱自身深入验证的主动性。这就需要界定各主体的验证责任边界,设计必须由学生独立完成的验证环节,并在评估体系中加大验证过程的权重,奖励批判性探究而非仅关注最终答案的正确性。

其次,算法偏见与验证盲区形成系统性风险。当人工智能工具被用于辅助验证时,其训练数据的局限性、算法设计的价值取向可能导致验证过程忽视某些视角、群体或知识来源。这就要求建立验证工具的元验证机制,定期评估其公平性、覆盖范围与潜在偏见,鼓励使用多元验证工具和方法。

再次,数据隐私与验证追踪之间存在矛盾。验证过程常涉及大量敏感的学习行为数据。这就要求建立严格的数据伦理规范,确保数据安全与隐私保护,明确数据用途和所有权。

最后,责任模糊与问责困境在分布式验证体系中尤为明显。当师生机三方共同参与知识建构与验证时,一旦出现知识性错误,责任归属就可能变得模糊不清。教师可能辩称“人工智能提供了置信度评分”,学生可能认为“经过教师审核”,人工智能系统设计者可能强调“已明示不确定性”。这种责任模糊不仅影响问责,更可能削弱各方的责任意识,这就需要建立分层级的验证责任框架:人工智能负责提供可验证的线索并明示不确定性;学生承担主体验证责任,即探究所学知识的真实性;教师承担最终的教育责任,负责监督验证流程的有效性,并审核关键知识节点。这种责任划分应嵌入教学设计,通过师生契约、教学规范等让所有参与者明晰职责(Shneiderman,2022),从而建立权责对等的质量保障机制。

(四)回归本质:人机协同中的教育价值坚守

在人机协同教育中,尤其是当技术为验证提供支持时,教育者必须始终清醒认识到:技术是手段而非目的,批判性思维与求真精神的培养才是教育的核心价值;切不可因为有高效的验证工具,就忽视对学生独立判断能力和科学精神的培养。

教师的不可替代性在于其对学生的价值引导及高阶思维和元验证能力的培养。教师需引导学生思考元认知问题,如为何选择这些验证标准及验证工具本身是否可靠。人工智能无法替代教师与学生之间的情感互动、智慧启迪和人格塑造。

学生主体性体现在其作为积极的求知者和验证者。教育应确保学生始终是验证活动的主动发起者和执行者,技术只是其能力的延伸。真正的素养体现在没有技术工具辅助时,依然有质疑和求证的意识与能力。人机协同应致力于增强而非替代学生的自主性。

健康的教育生态需要技术理性与人文精神的和谐统一。健康的人机协同教育生态应是技术赋能的高效验证与人文关怀的谨慎求索并存、兼顾对客观真理的追求与对多元视角的包容性验证、创新活力与传统价值保持对话的动态平衡系统(祝智庭等,2023)。展望未来,人机协同运思模式的成功实践关键在于把握“技术赋能”与“教育坚守”的辩证关系。只有充分发挥技术优势,并始终保持对教育本质的清醒认识,才能构建既有智能效率又有人文温度的未来教育图景。

五、迈向“思想本位”的新范式

智能时代的技术跃迁为教育实践重构提供了全新契机。本文基于完形心理学与言语行为理论,构建“三阶段运思—双体/三体协同—心设模型”框架,为理解与优化人机协同提供认知视角,有利于破解“黑箱”难题,增强交互的引导性与可信度。

本文提出的协同框架在实践中仍面临深层的、结构性挑战:人智协同知识作品的评价范式危机。在传统学术训练中,文献引用是构建论证可信度的基石。然而,当人工智能作为协作者参与思想构建时,其产出的知识往往是对海量信息融合、推理后的“涌现”结果,难以追溯至“源文献”。

这意味着亟需一场评价范式的革命:从看重“你说的话是谁说的”(引用权威),转向审视“你为什么会这么说”(推理过程)。评价的核心必须从“信源”转向“信过程”。未来的教育实践与研究必须直面以下问题:

1)如何为人工智能的推理建立“思想锚点”?当无法引用文献时,能否要求人工智能明确其结论所依据的底层公理、理论模型或逻辑规则?

2)如何设计“三元互证”的评估流程?学生、教师、人工智能三者之间需建立怎样的对话机制,交叉验证思想产出的合理性与创新性?

3)如何培养师生“思维质量”评估素养?这要求师生从知识的接收者与评判者,转为思维过程的审视者与引导者。

这警示我们,若不能革新评价体系,人智协同很可能在热闹的技术应用中陷入思想质量失控的困境。本文理论建构工作的意义恰恰在于为探索面向未来的“思想本位”评价新路径奠定基石。构建“思想本位”的可信度判据,将是决定人智协同能否从工具性辅助走向思想创新的分水岭。

展望未来,学术共同体面临双重任务,包括:在技术层面继续深化人机协同机制研究,在理论与范式层面勇于发起一场评价革命。未来的研究必须致力于将“思想锚点”“三元互证”“思维质量”这些关键概念,转化为可操作的判断依据与可行的实践。这包括:探索要求人工智能为其论断明确标注所依据的“底层公理”或“理论模型”,以此建立“思想锚点”;设计“三元对话日志”分析框架,用于评估学生—教师—人工智能的质疑与回应质量;开发基于“思维流图”的评价工具,使思维过程本身成为评估对象。唯有成功构建这一新的评价范式,才能确保人机协同这艘航船,最终抵达真正启迪智慧、创造思想的彼岸。

[参考文献]

[1] 郝祥军, 贺雪(2022). AI 与人类智能在知识生产中的博弈与融合及其对教育的启示 [J]. 华东师范大学学报(教育科学版), 40(9): 78-89.

[2] 何贵兵, 陈诚, 何泽桐, 崔力丹, 陆嘉琦, 宣泓舟, 林琳(2022). 智能组织中的人机协同决策:基于人机内部兼容性的研究探索 [J]. 心理科学进展, 30(12): 2619-2627.

[3] 胡艺龄, 陈彦君(2024). 人机协作下的共享心设模型研究 [J]. 现代教育技术, 34(1): 64-72.

[4] 乐惠骁, 汪琼(2022). 人机协作教学:冲突、动机与改进 [J]. 开

放教育研究,28(6):20-26.

[5] 米加宁,董昌其(2024). 大模型时代:知识的生成式“涌现”[J]. 学海,(1):81-96,214-215.

[6] 蒲清平,向往(2023). 生成式人工智能——ChatGPT 的变革影响、风险挑战及应对策略 [J]. 重庆大学学报(社会科学版),29(3):102-114.

[7] 束定芳(1989). 言语行为理论述评 [J]. 外语教学,(2):10-16.

[8] 苏冲,文旭(2018). 格式塔意象的传译:认知翻译策略研究 [J]. 中国翻译,39(4):13-20,129.

[9] 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 [J]. 社会科学,(8):184-192.

[10] 汪时冲,方海光,张鸽,马涛(2019). 人工智能教育机器人支持下的新型“双师课堂”研究——兼论“人机协同”教学设计与未来展望 [J]. 远程教育杂志,37(2):25-32.

[11] 汪娅(2025). 符号接地的可行性分析 [J]. 自然辩证法研究,41(8):67-74.

[12] 谢幼如,李草茵,李成军,邱艺(2024). 智能时代高校数字课程:内涵、形态与构建 [J]. 电化教育研究,45(11):5-12.

[13] 杨宗凯,王俊,吴砥,陈旭(2023). ChatGPT/生成式人工智能对教育的影响探析及应对策略 [J]. 华东师范大学学报(教育科学版),41(7):26-35.

[14] 余明华,王龚,卜洪晓,郑隆威(2024). 未来教师如何培养?——人机协同师范教育创新的理论模型与实践进路 [J]. 现代教育技术,34(1):117-126.

[15] 余胜泉(2018). 人机协作:人工智能时代教师角色与思维的转变 [J]. 中小学数字化教学,(3):24-26.

[16] 余胜泉(2025). 认知外包推动人才培养模式的深层变革 [J]. 教育科学研究,(6):1.

[17] 张夏恒,马妍(2024). 生成式人工智能技术赋能新质生产力涌现:价值意蕴、运行机理与实践路径 [J]. 电子政务,(4):17-25.

[18] 祝智庭,戴岭,赵晓伟(2023). “近未来”人机协同教育发展新思路 [J]. 开放教育研究,29(5):4-13.

[19] Amershi, S. , Weld, D. , Vorvoreanu, M. , Fourney, A. , Nushi, B. , Collisson, P. , Suh, J. , Iqbal, S. , Bennett, P. N. , Inkpen, K. , Teevan, J. , Kikin-Gil, R. , & Horvitz, E. (2019). Guidelines for human-AI interaction[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery: 1-13.

[20] Bender, E. M. , Gebru, T. , McMillan-Major, A. , & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery: 610-623.

[21] Bommasani, R. , Hudson, D. A. , Adeli, E. , Altman, R. , Arora, S. , von Arx, S. , Bernstein, M. S. , Bohg, J. , Bosselut, A. , Brunskill, E. , Brynjolfsson, E. , Buch, S. , Card, D. , Castellon, R. , Chatterji, N. , Chen, A. , Creel, K. , Davis, J. Q. , Demszky, D. , et al. (2022). On the opportunities and risks of foundation models[EB/OL]. [2025-10-09]. https://arxiv.org/abs/2108.07258.

[22] Clark, A., & Chalmers, D.(1998). The extended mind[J]. Analysis, 58(1): 7-19.

[23] Ehsan, U. , Liao, Q. V. , Muller, M. , Riedl, M. O. , & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems[C]//Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery: 1-19.

[24] Ji, Z. , Lee, N. , Frieske, R. , Yu, T. , Su, D. , Xu, Y. , Ishii, E. , Bang, Y. J. , Madotto, A. , & Fung, P.(2023). Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 55(12): 1-38.

[25] Kasneci, E. , Sessler, K. , Küchemann, S. , Bannert, M. , Dementieva, D. , Fischer, F. , Gasser, U. , & Groh, G.(2023). ChatGPT for good? on opportunities and challenges of large language models for education[J]. Learning and Individual Differences, 103: 102274.

[26] Rashkin, H. , Choi, E. , Jang, J. Y. , Volkova, S. , & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics: 2931-2937.

[27] Searle, J. R. (1969). Speech acts: An essay in the philosophy of language[M]. Cambridge: Cambridge University Press: 16-18.

[28] Shneiderman, B. (2022). Human-centered AI[M]. Oxford: Oxford University Press: 85.

[29] Wang, D. , Churchill, E. , Maes, P. , Fan, X. , Shneiderman, B. , Shi, Y. , & Wang, Q. (2020). From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people[C]//Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York: Association for Computing Machinery: 1-6.

[30] Wei, J. , Wang, X. , Schuurmans, D. , Bosma, M. , Chi, E. , Le, Q. , & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. [2025-10-09]. https://arxiv.org/abs/2201.11903.

(编辑:魏志慧)

The Explainability of Large Language Models' Operational Thinking Patterns: The Construction of a Mental Model Integrating Gestalt Metaphor and Speech Act Theory

ZHU Zhiting[1], WU Huina[2], XU Jun[2] & WU Yonghe[2]

(1. *School of Open Learning and Education*, *East China Normal University*, *Shanghai* 200062, *China*; 2. *Department of Educational Information Technology*, *East China Normal University*, *Shanghai* 200062, *China*)

Abstract: *As generative artificial intelligence infuses in education, the "black-box" of large language models (LLMs) and the uncertainty of their outputs have emerged as critical barriers to the effective human–AI collaboration. While existing research primarily addresses macro-level outcomes, the micro-level mechanisms of language generation and credibility assurance remain inadequately explored. Grounded in shared mental model theory, this study integrates Gestalt principles of holistic cognition and Searle's speech act theory to construct a user-oriented interpretive framework for LLM-generated language behaviors. The framework embeds credibility verification mechanisms—including source tracing, logical self-checking, and multimodal validation—in the model's triadic generative process of comprehending, interpreting, and constructing meaning. It reveals how cognitive processing and quality control co-regulate the transformation from semantic perception to communicative generation. The study proposes a co-evolutionary pathway from the dyadic cognition (user–model) to the triadic cognition (user–model–teacher), and establishes a teacher–student–machine shared mental model that links external actions with internal cognition through extended cognition. This model fosters shared understanding and validation among multiple agents, enhancing users' ability to interpret, guide, and trust model outputs. Ultimately, the study provides a novel cognitive perspective for demystifying the LLM "black-box," offers systematic theoretical framework, and presents design implications for building trustworthy, efficient, and responsible human–AI collaboration in educational contexts.*

Key words: *LLM operational thinking patterns*; *explainability*; *Gestalt psychology*; *speech act theory*; *teacher-student-machine interaction*; *shared mental model*; *credibility verification*; *human-AI collaboration*