认知之镜：生成式人工智能何以促进学习者认知发展？

董　艳[1]　张维予[1]　王　宇[2]　江　汀[3]

（1. 北京师范大学 教育学部，北京 100875；2. 上海交通大学 教育学院，上海 200240；
3. 澳大利亚国立大学 商学院，堪培拉 ACT 2600）

[摘要] 生成式人工智能既可能促进学习者的认知发展，也可能引发认知外包。学习者内部认知模型与系统外部学习者模型的协同校准构成生成式人工智能支持认知发展的关键。本研究从学习者建模与知识追踪视角出发，引入主动推断理论和开放学习者模型，提出“认知之镜”分析框架及“认知外显—状态表征—镜像反馈—内部反馈—再次外显”运行链条：学习者出现提问、解释、推理与修正等活动，系统外显认知状态；系统依据交互证据形成对学习者当前学习状态的暂时性表征，并将其转化为可见、可理解、可协商的镜像反馈；学习者据此识别自身理解与任务要求之间的差距，从而调整内部认知模型；学习者的后续回应又成为系统校准状态表征与反馈策略的新证据，推动两类模型相互校准。该框架可为生成式人工智能支持下的反馈设计与人机协同教学提供理论参考。

[关键词] 生成式人工智能；GenAI；认知之镜；认知发展；学习者建模；人机协同

[中图分类号] G442　　**[文献标识码]** A　　**[文章编号]** 1007-2179(2026)05-0070-10

生成式人工智能已逐渐成为学生获取信息、展开对话和完成学习任务的重要工具（Chan & Hu, 2023），其育人价值关键在于能否促进学生能力与素养的发展。由此，生成式人工智能如何介入学习者的促进学习者的认知发展，成为智能教育研究亟须回应的课题。已有研究者探讨了生成式人工智能的功能边界（Kasneci et al., 2023）和教学应用（Chiu, 2024），检验了不同支持对学习效果的影响（Lee et al., 2024），并调查了学习者的使用态度（Chan & Hu, 2023）、依赖倾向（Wang et al., 2025）和学习能动性（夏亮亮等, 2025）。然而，生成式人工智能的介入既可能促进学习者的认知发展，也可能引发认知外包、思维替代和学习浅表化等问题（汪凡淙等, 2025; 余胜泉等, 2023）。总体来看，现有研究较多关注系统功能、使用行为及学习结果，较少解释人机交互过程中学习者的认知状态如何变化。其中，学习者如何在人机交互过程中修正内部认知模型，生成式人工智能系统又如何依据交互证据更

[收稿日期]2026-08-01　　**[修回日期]**2026-08-23　　**[DOI编码]**10.13966/j.cnki.kfjyyj.2026.05.008

[基金项目] 2026 年度北京市自然科学基金—副中心联合基金（培育项目）“面向人机共生的多智能体协同沉浸式课堂空间构建方法研究”（L2611046），国家社会科学基金 2025 年度教育学青年项目“GAI 影响青少年跨学科学习的认知神经机制及优化策略研究”（CCA250227）。

[作者简介] 董艳，博士，教授，博士生导师，北京师范大学科学教育研究院副院长，研究方向：跨学科创新教育、人工智能赋能教师专业发展（yan.dong@bnu.edu.cn）；张维予，博士研究生，北京师范大学教育学部，研究方向：人工智能教育应用、教学设计（weiyuzh@mail.bnu.edu.cn）；王宇（通讯作者），博士，助理教授，上海交通大学教育学院，研究方向：STEM 教育、学习科学、认知神经科学、人工智能教育应用（fergie_wang@sjtu.edu.cn）；江汀，硕士研究生，澳大利亚国立大学商学院管理学专业，研究方向：教育心理学、数字化管理与组织行为（u7276759@anu.edu.au）。

[引用信息] 董艳，张维予，王宇，江汀（2026）. 认知之镜：生成式人工智能何以促进学习者认知发展？[J]. 开放教育研究，32(5)：70-79.

新对学习者认知状态的估计，是理解这一过程的两个关键问题。

基于上述问题，本研究从学习者建模与知识追踪视角出发，引入主动推断理论和开放学习者模型，提出“认知之镜”人机协同分析框架，以阐释生成式人工智能支持学习者认知发展机制，为人机协同学习设计提供参考。

一、理论基础与概念界定

（一）主动推断理论与学习者认知更新

“认知之镜”聚焦的关键问题是：在生成式人工智能支持的学习中，当系统反馈与学习者原有判断不一致时，学习者如何借助生成式人工智能的镜像反馈重新审视自身理解，并通过调整概念结构、问题表征与推理路径，修正原有的认知模型，进而促进自身认知发展。

主动推断理论为解释这一认知更新过程提供了理论视角，其思想源头可追溯至19世纪下半叶德国生理学家和物理学家亥姆霍兹（Helmholtz）提出的“无意识推断”。该思想认为，知觉是主体依据已有经验对外部信息的解释和推断（von Helmholtz, 2005）。此后，拉奥与巴拉德（Rao & Ballard, 1999）提出层级预测编码模型，认为高层区域向下传递预测，低层区域向上传递实际输入与预测之间的误差信号。预测加工框架随后被用于解释知觉、注意、行动和学习等心智活动（Clark, 2013; Hohwy, 2013）。弗里斯顿（Friston, 2010）进一步运用自由能原则（free-energy principle）解释感知、行动与学习之间的关系，强调认知主体通过降低预测误差与环境的不确定性，维持自身与外部世界之间的适应关系。在这一理论中，“自由能”用于衡量当前内部模型对外部信息解释是否充分。自由能越高，意味着现有模型越难解释当前获得的信息。为降低自由能，主体既可以更新自己的理解与判断，也可以通过行动获取新信息，以检验原有预期。在此基础上，主动推断理论将认知主体视为主动探询者，个体依据内部模型对环境及自身状态作出预测，并通过行动获取新的感觉证据。当预测与感觉证据不一致时，个体调整内部模型，以降低不确定性并适应环境（Friston et al., 2017; Parr et al., 2022）。这一过程可概括为形成预测、觉察误差、信息探索与更新模型。

近年来，主动推断理论被教育研究者用于解释学习者面对认知冲突时的主动探索和理解更新。在学习情境中，学习者依据已有知识、任务经验与自我判断，对任务要求、自身理解和可能结果形成预期。如外部信息与原有预期不一，学习者就会形成预测误差。但预测误差并不必然引发学习。学习者只有追问差异产生的原因，并通过比较、验证或重新表达，主动获取新证据，才可能调整原有理解。从这一视角看，认知冲突构成主动探索的触发条件，学习者由认知冲突转向好奇、追问和批判性思考（Di Paolo et al., 2024; May et al., 2022）。

生成式人工智能凭借持续对话能力，为这一过程提供了动态参考。学习者可以先向系统表述自己的初步理解，再在人机交互中获得不同于原有判断的参照信息。主动推断理论较好地解释了学习者如何利用这一参照信息察觉预测误差、主动搜寻证据，并据此修正内部认知模型。

（二）学习者建模与认知状态估计

学习者建模的核心任务是依据学习表现与数据，形成关于学习者知识水平、策略使用和学习困难的计算性表征。系统依据作答结果、反应时间、行为轨迹或语言表达等证据估计学习者的认知状态。早期知识追踪主要利用作答正误、练习次数和反应时间等结构化变量估计学习者的知识掌握程度。其中，贝叶斯知识追踪技术将知识掌握视为不可直接观测的潜在状态，并根据连续作答的结果更新学习者的知识掌握概率。深度知识追踪技术利用序列模型编码学习者的行为轨迹，据此预测其学习表现。随着建模对象从学习结果拓展到学习过程，策略使用（Li et al., 2025）和元认知调节（Wang et al., 2025）及其动态变化（Järvelä & Hadwin, 2024; Wang et al., 2026）逐渐成为认知状态估计的关键要素。

与作答正误等结构化数据相比，学习者的提问、解释、论证和修正能提供更丰富的过程信息。系统可以从开放性表达中提取核心概念与论证成分，分析概念之间、论证成分之间的关系。系统还可以结合修改前后的文本差异和多轮对话序列，推断概念结构、错误类型及学习状态的变化（Li et al., 2024; Neshaei et al., 2024; Tran et al., 2025）。由此，学习者建模可用于估计学习者“如何理解”“偏差

出现在哪里”及“需要何种支持”。

（三）“认知之镜”概念

“认知镜像”概念最初见于富水隼人等（Tomisu et al., 2025）的研究，该研究将生成式人工智能设置为可被学习者教导的“新手”，通过人工智能的响应来映射学习者的讲解质量，以支持学习者的元认知监控与自我调节。本研究不关注人工智能被“教会”的程度，而关注如何表征学习者的认知状态及发展。基于此，本研究提出“认知之镜”，说明生成式人工智能介入学习者认知调节的机制，即系统依据学习者在学习任务中的提问、解释、推理与修正等，形成对其当前认知状态的暂时性表征，并将该表征转化为可供学习者审视、解释和修正自身理解的外部参照。从功能定位看，“认知之镜”是一种镜像性认知中介，主要用于反映学习者当前的理解状态，支持其识别认知差距并调整既有理解，而非直接提供答案或替代学习者作出判断。

“认知之镜”的运行涉及两类模型。学习者内部认知模型包括知识经验、任务目标、自我判断与价值取向，用于判断学习者如何理解问题和处理反馈。系统外部学习者模型由人机交互证据生成，用于估计学习者的概念理解、问题表征、论证结构和策略等状态。前者构成学习者理解问题和采取行动的内在依据，后者构成系统生成反馈的计算依据。二者通过学习者表达与系统反馈建立联系：学习者的外显表达为系统更新学习者模型提供依据，系统反馈又为学习者审视和调整原有理解提供参照。

“认知之镜”的运行过程（见图 1）包括：学习者首先通过自我表达将内部理解外显到交互空间，系统据此形成暂时性的学习者状态表征，并将状态估计转化为镜像反馈；学习者将镜像反馈与原有判断比较，形成对当前理解的内部反馈，调整后的理解再次外显，成为系统更新状态估计的新证据，由此开启下一轮交互。在这一循环中，学习者可反复检验和修正内部认知模型。“认知之镜”要求系统围绕学习者的认知状态组织反馈，同时将反馈的解释权、学习路径的选择权和结果的最终判断权留给学习者。

二、系统运作

“认知之镜”的运行需四方面条件支撑：任务、学习者、交互和系统。其中，任务决定系统能否获得具有诊断价值的认知证据；学习者是否承担认知责任，影响反馈能否进入内部调节；交互的结构递进会影响两类模型能否持续校准；系统功能制约状态表征的精度与反馈方式。任一条件缺少，都可能使“认知之镜”退化为认知外包或认知替代。

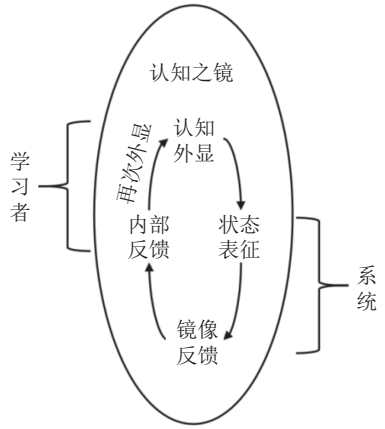

图 1 “认知之镜”运行链条

（一）基于任务的认知过程显性

认知过程的外显，指学习者按照任务要求进行表达并留下过程留痕，使原本难以直接观测的概念理解、问题表征、推理路径等内部认知，转化为可观察、可记录和可比较的交互证据。认知过程能否外显，既取决于学习者能否呈现自身思路，也取决于任务是否要求学习者说明思考过程。

在事实回忆或封闭式作答任务中，学习者通常只给出既定答案，较少呈现答案形成过程，系统因而难以从作答结果中形成具有诊断意义的状态表征（Kim et al., 2025）。在概念辨析、论证写作、案例分析和方案设计等开放性任务中，学习者需提出初步判断、说明理由、组织证据并修正观点。这些行为能使学习者的理解状态进入人机交互空间，为系统识别认知差距和表征学习状态提供依据。

教学设计者还需将初步判断、理由说明、修改记录和反思解释等认知外显支架嵌入任务流程，促使学习者持续呈现理解与调整过程。即使是开放性任务，若学习者只是复制系统生成的内容，系统将难以获得能反映学习者实际理解状况的线索。因此，认知过程外显有赖于任务的开放性、学习者的主动表达与学习支架设计。

（二）认知责任归属

学习者需要承担理解、判断和修正的责任，并愿意反思自身认识。若学习者过度依赖生成式人工智能，直接接受系统判断，人机交互便容易变成认知替代（Wang et al., 2025）。只有学习者愿意呈现思路、核验系统反馈并重新表达时，反馈才可能进入其认知调节过程。认知责任意味着学习者始终保有对系统反馈的解释权、选择权与检验责任。学习者可以采纳、质疑、拒绝或修正系统的建议，但需说明依据，并通过新的表达或行动检验自己的判断。学习者的主体地位因此也被落实到反馈处理过程中，构成镜像反馈与认知替代的关键区别。

（三）交互结构的递进

交互结构指学习者表达、系统反馈与后续回应在多轮人机交互中的组织方式和推进关系，其作用在于使学习者内部认知模型与系统外部学习者模型能依据持续生成的交互证据相互校准。单轮交互只能提供阶段性证据，既难以显示学习者理解变化的过程，也无法利用后续表现检验当前的状态估计。因此，学习者内部认知模型的更新需经历反馈比较、理解调整和再次外显，系统外部学习者模型也需依据新的交互证据修正状态表征。

交互结构不仅需要保持连续，还应体现认知的递进性。系统的后续反馈应以前一轮交互中暴露的学习者认知差距为起点，逐步从“问题定位”进入“原因辨析”和“路径检验”等环节，而非反复生成内容相近的答案。若后续交互不能检验和推进前一轮判断，增加交互次数也难以推动学习者内部认知模型更新。

（四）系统功能适配

系统功能应与镜像反馈的要求相匹配。首先，内容生成能力可用于概念对照、观点比较和表达修改，但其通常只针对当前输入作出回应。进一步而言，上下文保持、交互记忆和轨迹追踪等能力使系统能比较学习者的前后表达，据此估计学习者的理解变化。在此基础上，系统还可以协助学习者检验方案，但其介入范围受教学设计和学习者的约束。

系统功能越复杂，越需要保证状态估计所依据的证据可追溯、形成的判断可修正。系统在呈现状态判断时，应区分“直接观察到的表达”“依据规则识别的特征”和“模型推断的潜在状态”，避免将不同的证据层级归为单一的认知状态表征。“认知之镜”的镜像质量取决于状态估计是否有充分的证据、反馈是否对应差距，以及系统能否依据学习者的回应作出调整。

三、人机双向建构

通过“认知之镜”形成的人机双向建构，延续了智能教育人机双向反馈研究的“人机双主体”立场。既有研究指出，学习者不应只是接受机器反馈的客体，而应作为能够提问、回应、评价和调整的交互主体，与人工智能建立具有主体间性的双向关系（董艳等，2021）。在“认知之镜”框架中，这种关系具体表现为学习者内部认知模型与系统外部学习者模型在持续交互中相互校准。

（一）学习者侧：镜像反馈驱动的内部模型更新

在学习者侧，内部认知模型的更新经历认知外显、预测误差识别与内化、认知模型修正三个环节。主动推断理论认为，学习者会依据既有知识、任务目标和自我判断，对任务要求、自身理解和可能结果形成预测。外部反馈如与原有预测不一，学习者需重新检视自身的认知模型，但这种不一致能否转化为认知变化，还取决于学习者能否解释差距的来源，并通过行动检验修正后的判断。

1. 认知外显

认知外显是镜像反馈进入学习者认知调节的起点。学习者在向生成式人工智能提问、作出解释时，需组织隐含的概念理解、问题表征、论证结构与认知策略，将其转化为系统可以识别和回应的表达（Chi et al., 1994）。这一语言组织过程涉及的信息选择、关系梳理和推理补全，也会促使学习者重新组织已有的理解（Fiorella & Mayer, 2015, 2016）。在此过程中，原本模糊的概念、零散的关系或未经检验的判断开始显露，内隐理解成为学习者可以回看、比较和修正的对象。

外显表达也为系统建立学习者模型提供了初始线索。学习者提出问题、连接概念和组织理由的方式，可以反映其问题表征和知识组织特点（Graesser et al., 2004; Graesser & Person, 1994）。不过，这些语言痕迹只是系统推断的依据之一。系统还需结合任务要求、人机交互历史和学习者的后续回应，检验并修正当前判断。因此，认知外显具

有双重功能:一方面,学习者获得审视自身理解的对象;另一方面,系统获得构建学习者模型的初始证据。

2. 预测误差识别与内化

学习者介入任务时,通常会对自己是否掌握相关概念、理解任务要求或完成充分论证作出初步判断。生成式人工智能可以基于交互内容,通过重述、追问、对照、提示或纠偏等方式,使概念混淆、条件遗漏、推理断裂或论证薄弱等问题显现出来。如果系统反馈与学习者的原有判断冲突,学习者的预期与外部信息之间便出现差距。主动推断理论将这种差距称为预测误差。

预测误差只是认知调整的触发条件。外部反馈只有经过学习者的可信度判断、原因分析和意义重组,才可能转化为内部反馈,成为认知调节的依据(Nicol & Macfarlane-Dick, 2006)。学习者需要辨明,差距究竟来自知识缺漏、问题表征偏差或推理失序,还是系统反馈本身存在偏差。只有当学习者能比较不同解释、判断差异来源并进行调整时,反馈才会促进学习者认知调节。

3. 认知模型修正

内部反馈形成后,学习者还需采取行动检验新的判断。主动推断将行动视为获取信息和检验假设的过程(Friston et al., 2017)。在生成式人工智能支持的学习中,学习者既可以追问任务条件、比较不同解释或借助反例检验原有判断,也可以通过自我解释补充推理环节,并重新表达观点,以检验新理解的内部一致性(Di Paolo et al., 2024)。这些行动使学习者能选择、质疑和重构系统反馈,降低理解的不确定性,修正内部认知模型。

在多轮交互中,认知调整可能由某一答案或表述,逐渐延伸至概念理解、问题表征、论证结构和策略选择。只有当这种变化能迁移至新任务,且学习者对系统提示的依赖随之降低时,学习者的认知才能被视为得到发展。由此,认知外显暴露当前认知模型的边界,镜像反馈提供新的外部参照,主动探询用于检验和修正新的理解,三者共同构成内部认知模型的更新循环。

(二)系统侧:基于学习者建模的镜像反馈生成

系统侧关注交互证据如何经由学习者建模转化为镜像反馈。系统以学习者的当前表达、交互历史和任务目标为输入,完成认知特征提取、学习者状态估计和目标差距诊断,再据此选择反馈的类型,确定反馈内容的粒度和介入强度。反馈发出后,学习者的回应成为新的交互证据,用于检验和校准先前状态判断。与学习者侧的内部认知调节不同,系统侧承担证据表征、状态推断与反馈策略选择等任务。

1. 认知特征提取

学习者建模依据学习者的外显行为,对不可直接观测的学习状态进行估计。生成式人工智能情境中,系统的输入信息主要是非结构化文本,系统的观测对象也由学习结果延伸至学习过程。系统可借助语义分析、错误模式识别和序列比较等方法,从学习者的表达中提取概念证据 C(concept)、论证证据 A(argument)、推理证据 R(reasoning)和变化证据 B(behavioral change)。其中,概念证据用于反映关键概念的使用情况和概念之间的关系,论证证据用于反映主张、依据与结论之间的结构,推理证据用于表征推理步骤和步骤之间的逻辑联系,变化证据用于反映学习者得到反馈后的修改、质疑与迁移表现。第 t 轮交互证据的提取可表示为:

$$E_t = (C_t, A_t, R_t, B_t) = \Phi(X_t, G, H_{t-1}) \quad (1)$$

其中,X_t表示学习者本轮表达,G表示任务目标与评价标准,H_{t-1}表示此前交互历史,Φ表示语义分析、错误模式识别与序列比较等方法构成的特征提取过程。借助这些方法,系统可以提取关键概念和论证成分,分析概念及论证成分之间的关系,并定位前提缺失、关系断裂、逻辑跳跃和潜在概念误解等认知线索(Hoq et al., 2025; Ren et al., 2025)。

单轮表达只能呈现学习者特定时段的表现,系统需结合多轮交互历史考察证据如何变化。例如,学习者是否依据反馈修正观点,能否说明修改理由,能否将已有理解迁移至新任务,以及能否避免同类错误反复出现。相较于仅记录作答结果和行为次数,生成式人工智能保留了学习者的表达内容与推理轨迹。学习者组织问题、阐释关系、回应矛盾与修正判断的方式,会在语言表达和交互轨迹中留下可供分析的认知和调节线索(Borchers et al., 2025; Tran et al., 2025)。这些痕迹不能完整还原学习者的内部心理过程,却为系统状态判断提供了必要证据。

2. 学习者状态动态估计

认知特征提取呈现学习者当前的外显表现，状态估计则据此推断不可直接观测的理解状态。系统需要综合学习者的当前表达、任务表现与交互历史，形成能随新证据持续更新的学习者模型。该模型分别估计学习者的知识掌握程度、概念关系理解水平、问题表征特征和推理策略使用情况，并保留对各项估计的不确定性。学习者能否修正原有观点，说明修改理由，以及避免同类错误反复出现，都会改变系统对这些维度的判断（Neshaei et al., 2024）。

状态估计应整合三类证据：一是当前任务的作答结果、概念使用、论证结构和错误类型，可用于表征学习者当前的任务表现；二是多轮交互的文本差异、反馈回应和重复错误情况，可用于判断学习者的理解是否发生变化；三是学习者能否将修正后的理解迁移至相近任务，可用于检验这种变化能否跨任务保持。不同技术承担不同功能：大语言模型可解析开放回答中的概念、论证和推理线索，知识图谱可将学习者使用的概念及其关系与领域知识结构对齐，贝叶斯知识追踪可依据连续作答更新特定知识点的掌握概率，深度知识追踪可通过长时序行为序列识别学习者状态变化。

不同技术提取或推断的结果，应在统一的学习者状态框架中整合。多项证据指向同一判断时，系统可提高判断的置信度；新证据与原有判断相反时，系统应及时修正状态估计；不同证据相互冲突时，系统需通过追问、反例或迁移任务继续取证。学习者模型还应记录判断依据和置信度，并区分直接观察到的表达、技术识别的特征与模型推断的潜在状态，避免将学习者特定阶段的表现固化为能力标签。由此形成的学习者模型，是服务于当前任务和交互过程的情境性表征，有待系统基于后续证据持续检验和修正。

3. 认知差距诊断与反馈决策

学习者状态估计还需转化为认知差距诊断。系统先依据领域知识结构、解题约束和论证评价标准形成任务目标表征，再将学习者的当前状态与目标表征分维度进行比较。其中，知识图谱用于比较概念及其关系，约束规则用于检查条件遗漏与步骤缺失，论证分析用于识别主张、证据和结论之间的结构断裂。比较的结果用于标明差距的类型，如概念缺失、关系误置、前提遗漏、推理跳步或论证证据不足。

认知差距诊断包括位置识别、类型判定和成因区分。位置识别回答差距出现在何处；类型判定指区分知识缺漏、结构错误、策略不当和表达不充分；成因区分指结合交互历史，判断当前差距源于持续性困难、偶然失误，还是系统对学习者表达的误解。诊断不宜仅凭单次错误作出判断，而应综合错误是否重复、学习者如何修改及能否说明理由，对结论进行交叉验证。

在此基础上，反馈决策模块根据差距的位置、类型与置信度，选择相应的策略，并确定反馈内容的粒度与介入强度。概念边界不清时，系统可提供概念对照、例证或反例；关系结构有误时，系统可提示学习者比较概念间的层级、条件或因果联系；推理链缺少中间环节时，系统可通过追问或步骤提示，引导学习者补充推理依据（Man & Man, 2025）；论证结构松散时，系统可指出主张与证据之间尚未建立的关系；状态判断的置信度低时，系统宜先提出诊断性问题，避免直接判定学习者的回答对或错；学习者能独立推进时，系统应降低提示强度，为其继续推理留有空间。

4. 学习者模型校准

系统发出镜像反馈后，学习者的解释、修改和验证结果会形成新的响应证据。系统据此判断学习者是否理解反馈、原有偏差是否缩小，以及同类错误是否在新任务中再次出现。学习者能准确修改并说明理由时，系统可在后续交互中降低提示强度；同类差距持续出现时，系统需重新估计相应的知识缺口；学习者补充新的问题条件或对系统判断提出质疑时，系统需修正任务表征与状态判断。

系统侧还需区分不同层级的有效性。认知特征提取的准确性，只能说明输入信息的表征是否可靠；状态估计的校准程度，主要反映学习者模型的估计能力；反馈能否被学习者理解和采用，反映了交互支持的可用性。这些技术与交互层面的表现，都不能直接证明学习者已实现认知发展。系统还需考察学习者的概念理解、问题表征和策略使用是否持续变化，以及这些变化能否迁移至后续任务。由此，系统侧负责形成可追溯的状态估计并生成适

切的镜像反馈，学习者侧解释镜像反馈如何进入学习者内部认知调节，两条机制最终在交互证据与反馈行动处衔接。

（三）开放学习者模型促成的协同校准

系统只能根据交互证据推断学习者状态（Shen et al., 2024），学习者也无法完全了解系统的推断过程，因此两类模型之间始终存在表征差异。开放学习者模型可将这种差异转化为双方共同检视和修正的对象。

开放学习者模型连接系统侧与学习者侧，构成镜像反馈的支撑机制（见图 2）。它不是独立于内外模型的第三种模型，也不要求公开系统的全部运算过程，只向学习者适度开放与当前任务相关的状态判断、判断依据及其不确定性。具体而言，“可见”是让学习者看到系统对其当前状态的判断，“可理解”是说明这一判断基于哪些表达与行为证据，“可协商”是允许学习者确认、质疑或修正系统的判断。由此，原本隐藏在系统内部的状态估计成为双方共同检视的对象，学习者也由被建模者变为模型校准的参与者。

在开放学习者模型的支持下，镜像反馈通过“生成”和“回应”连接系统建模与学习者认知调节。在“生成”环节，系统依据外部学习者模型形成的状态估计和差距诊断，选择需向学习者呈现的内容，并将其转化为镜像反馈。“生成”并非指原样输出系统的分析，而是将“系统作出何种判断”转化为“学习者需要审视何种差距”。因此，反馈既要对应具体的交互证据，也要为学习者继续思考留有空间。

在“回应”环节，系统向学习者呈现当前的状态判断、认知差距及依据，使学习者获得审视自身理解的外部参照。这里回应的，是系统根据有限证据形成的学习者状态的暂时性表征。学习者可将这一表征与原有判断比较，并通过确认判断、提出质疑、补充证据或修正表达作出回应。学习者据此开展认知调节，其回应又构成新的认知外显和交互证据，返回系统建模过程。

协同校准沿学习者侧和系统侧两个方向展开。学习者借助镜像反馈检验自己对任务的理解和对自身认知状态的判断，进而调整内部认知模型；系统将学习者的确认、质疑与修正作为新证据，用于更新外部学习者模型。开放学习者模型既让系统判断成为学习者认知调节的外部参照，也使学习者的回应成为系统继续建模的依据。协同校准并不追求两类模型完全一致，而是让系统表征更加贴合当前的交互证据，同时使学习者的理解和自我判断得到持续检验。

四、风险边界

“认知之镜”的风险主要沿两条路径产生：一

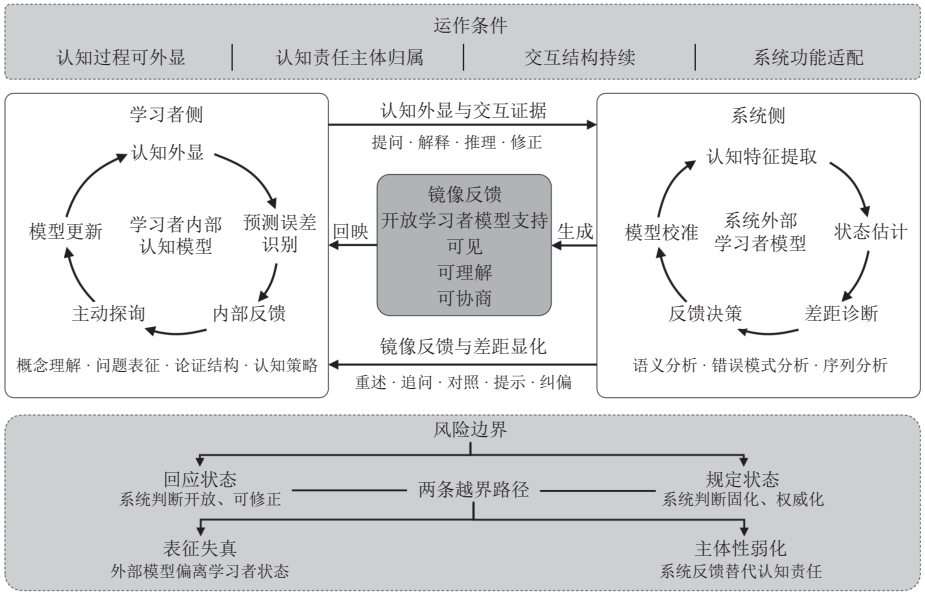

图 2 “认知之镜”双向建构

是作用于学习者表征，沿“局部表现被固化 → 系统判断自我强化 → 表征失真”的路径累积；二是作用于反馈关系，沿“系统支持越位 → 认知责任转移 → 主体性弱化”的路径扩散。

表征失真首先源于建模证据的不足及情境局限。系统只能依据学习者阶段性的语言表达与交互行为推断其认知状态，一次迟疑、一处错误或一段不清晰的表达，都不足以支撑系统形成稳定的能力判断。若系统将这些局部表现固化为持久特征，并持续据此生成后续反馈，原有判断便可能在循环中不断自我强化，逐渐偏离学习者的理解状态。因此，学习者表征应保持情境性、暂时性与可修正性，系统也需随新的交互证据及时更新先前判断。

主体性弱化主要发生在系统支持逐渐替代学习者认知责任的过程中。当学习者借助生成式人工智能解决问题，可能绕过差距识别、证据比较和策略选择，直接将系统判断视作正确结论。形式完整、表达流畅的反馈还可能放大系统判断的表面可信度，进一步降低学习者质疑、解释与检验反馈的意愿。此时，镜像反馈由反思资源转变为现成结论。随着反馈的解释权、选择权和检验责任逐步移交给系统，学习者调节自身理解的能力也可能随之减弱。

“回应状态”与“规定状态”由此构成“认知之镜”的功能边界。“回应状态”把系统判断作为暂时、开放的外部参照，帮助学习者发现认知差距；“规定状态”把基于有限证据的推断转化为确定结论，帮助学习者界定能力水平、限定学习路径或指定修正方式。系统一旦从前者滑向后者，便可能引发表征失真或主体性弱化。

守住这一边界，需同时约束外部模型与镜像反馈。外部模型应保持开放，避免将暂时性的状态估计固化为稳定标签；镜像反馈应避免替代学习者的解释、选择与检验。为此，系统设计需保证状态判断可追溯、可质疑、可修正，并为学习者保留回应和校准镜像反馈的空间。如此，两类模型之间的差异才可能成为认知调整的契机，生成式人工智能也才能发挥外部参照与反思支架的作用。

五、结语

本研究提出“认知之镜”框架，将生成式人工智能支持认知发展的过程解释为学习者与系统之间持续交互，生成证据，并分别影响内部认知模型和外部学习者模型。学习者通过认知外显、预测误差识别、内部反馈和主动探询调整原有理解，系统则依据学习者的交互表达提取认知特征、估计学习者状态、诊断认知差距，并据此调整反馈。开放学习者模型将系统的状态判断及依据向学习者开放，使之成为可见、可理解、可协商的镜像反馈，由此建立两类模型之间的联系。

在“认知之镜”框架中，学习者依据镜像反馈识别当前理解与任务要求之间的差距，通过解释、质疑和主动探询修正内部认知模型。学习者的回应又成为系统更新状态估计、差距诊断与反馈策略的新证据。由此，两类模型在持续交互中相互衔接、循环更新。该循环成立的前提是学习任务允许认知外显，学习者保有认知责任，人机交互能持续递进，以及系统具备动态建模与反馈调适能力。同时，“认知之镜”也存在明显边界，即系统应回应学习者状态，而不应规定其状态。系统判断若被封闭固化，就可能引发表征失真；系统反馈若替代学习者作出解释与判断，则可能削弱其认知主体性。因此，生成式人工智能只有将反馈作为可追溯、可协商、可修正的外部参照，才能支持学习者开展自我审视与认知调节。

本研究主要完成“认知之镜”的理论建构，各构件之间的关系、运行过程及成立条件仍有待经验检验。研究所讨论的情境也主要限于个体学习者与生成式人工智能之间的交互，对不同学段、知识领域和任务类型带来的差异关注不足。后续研究可围绕三方面展开：第一，将框架的关键构件转化为可观察指标，然后结合对话日志、口语报告等过程性材料，检验各构件之间的关系及其动态变化；第二，比较不同任务类型、学习者先前知识和系统功能条件下的交互过程，识别框架适用、受限或失效的具体条件；第三，采用纵向研究或对照研究，比较镜像反馈、一般反馈与答案供给对学习者概念理解、知识迁移和元认知监控的影响，并同步关注系统误判、隐私暴露与认知责任转移等潜在风险。

[参考文献]

[1] Borchers, C., Zhang, J., Fleischer, H., Schanze, S., Aleven, V., &

Baker, R.(2025). Large language models generalize SRL prediction to new languages within but not between domains[J]. Journal of Educational Data Mining, 17(2): 24-54.

[2] Chan, C. K. Y., & Hu, W.(2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education[J]. International Journal of Educational Technology in Higher Education, 20(1): 43.

[3] Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C.(1994). Eliciting self-explanations improves understanding[J]. Cognitive Science, 18(3): 439-477.

[4] Chiu, T. K. F.(2024). A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT[J]. Educational Technology Research and Development, 72(4): 2401-2416.

[5] Clark, A.(2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science[J]. Behavioral and Brain Sciences, 36(3): 181-204.

[6] Di Paolo, L. D., White, B., Guénin-Carlut, A., Constant, A., & Clark, A.(2024). Active inference goes to school: The importance of active learning in the age of large language models[J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 379(1911): 20230148.

[7] 董艳,李心怡,郑娅峰,翟雪松(2021). 智能教育应用的人机双向反馈:机理、模型与实施原则 [J]. 开放教育研究,27(2):26-33.

[8] Fiorella, L. , & Mayer, R. E. (2015). Learning as a generative activity: Eight learning strategies that promote understanding [M]. Cambridge: Cambridge University Press: 125-129.

[9] Fiorella, L., & Mayer, R. E.(2016). Eight ways to promote generative learning[J]. Educational Psychology Review, 28(4): 717-741.

[10] Friston, K.(2010). The free-energy principle: A unified brain theory?[J]. Nature Reviews Neuroscience, 11(2): 127-138.

[11] Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G.(2017). Active inference: A process theory[J]. Neural Computation, 29(1): 1-49.

[12] Graesser, A. C. , Lu, S. , Jackson, G. T. , Mitchell, H. H. , Ventura, M. , Olney, A. , & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language [J]. Behavior Research Methods, Instruments, & Computers, 36(2): 180-192.

[13] Graesser, A. C., & Person, N. K.(1994). Question asking during tutoring[J]. American Educational Research Journal, 31(1): 104-137.

[14] Hohwy, J. (2013). The predictive mind [M]. Oxford: Oxford University Press: 1-7.

[15] Hoq, M. , Rao, A. , Jaishankar, R. , Piryani, K. , Janapati, N. , Vandenberg, J. , Mott, B. , Norouzi, N. , Lester, J. , & Akram, B. (2025). Automated identification of logical errors in programs: Advancing scalable analysis of student misconceptions [J]. arXiv preprint arXiv: 2505.10913.

[16] Järvelä, S., & Hadwin, A.(2024). Triggers for self-regulated learning: A conceptual framework for advancing multimodal research about SRL[J]. Learning and Individual Differences, 115: 102526.

[17] 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., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G.(2023). ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and Individual Differences, 103: 102274.

[18] Kim, J., Klopfer, M., Grohs, J. R., Eldardiry, H., Weichert, J., Cox, L. A., & Pike, D.(2025). Examining faculty and student perceptions of generative AI in university courses[J]. Innovative Higher Education, 50(4): 1281-1313.

[19] Lee, H. Y., Chen, P. H., Wang, W. S., Huang, Y. M., & Wu, T. T.(2024). Empowering ChatGPT with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction[J]. International Journal of Educational Technology in Higher Education, 21(1): 16.

[20] Li, T. , Liu, Z. , Matsumura, L. , Wang, E. , Litman, D. , & Correnti, R. (2024). Using large language models to assess young students' writing revisions [C]//Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024): 365-380.

[21] Li, T., Yan, L., Iqbal, S., Srivastava, N., Singh, S., Raković, M., Swiecki, Z., Tsai, Y., Gašević, D., Fan, Y., & Li, X.(2025). Analytics of self-regulated learning strategies and scaffolding: Associations with learning performance[J]. Computers and Education: Artificial Intelligence, 8: 100410.

[22] Man, J. C. H., & Man, K. H. H.(2025). Using GenAI for Socratic questioning: An approach to higher-order thinking for nursing education[J]. Nursing Open, 12(10): e70355.

[23] May, C. J., Wittingslow, R., & Blandhol, M.(2022). Provoking thought: A predictive processing account of critical thinking and the effects of education[J]. Educational Philosophy and Theory, 54(14): 2458-2468.

[24] Neshaei, S. P. , Davis, R. L. , Hazimeh, A. , Lazarevski, B. , Dillenbourg, P. , & Käser, T. (2024). Towards modeling learner performance with large language models [J]. arXiv preprint arXiv: 2403.14661.

[25] Nicol, D. J., & Macfarlane-Dick, D.(2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice[J]. Studies in Higher Education, 31(2): 199-218.

[26] Parr, T. , Pezzulo, G. , & Friston, K. J. (2022). Active inference: The free energy principle in mind, brain, and behavior [M]. Cambridge, MA: The MIT Press: 15-35.

[27] Rao, R. P., & Ballard, D. H.(1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects[J]. Nature Neuroscience, 2(1): 79-87.

[28] Ren, Y. , Zhou, X. , Zhang, N. , Zhao, S. , Lan, M. , & Bai, X. (2025). Towards comprehensive argument analysis in education: Dataset, tasks, and method [C]//Proceedings of the 63rd Annual Meeting of the

Association for Computational Linguistics (Volume 1: Long Papers) (pp. 14215-14231).

[29] Shen, S., Liu, Q., Huang, Z., Zheng, Y., Yin, M., Wang, M., & Chen, E.(2024). A survey of knowledge tracing: Models, variants, and applications[J]. IEEE Transactions on Learning Technologies, 17: 1858-1879.

[30] Tomisu, H., Ueda, J., & Yamanaka, T.(2025). The cognitive mirror: A framework for AI-powered metacognition and self-regulated learning[J]. Frontiers in Education, 10: 1697554.

[31] Tran, N. , Litman, D. , & Godley, A. (2025). Using large language models to analyze students' collaborative argumentation in classroom discussions [C]//Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers: 111-125.

[32] von Helmholtz, H. (2005). Treatise on physiological optics: Volume III [M]. Mineola, NY: Dover Publications: 1-4.

[33] Wang, F., Li, N., Cheung, A. C., & Wong, G. K.(2025). In GenAI we trust: An investigation of university students' reliance on and resistance to generative AI in language learning[J]. International Journal of Educational Technology in Higher Education, 22(1): 59.

[34] 汪凡淙，汤筱玮，余胜泉(2025). 基于生成式人工智能的认知外包：交互行为模式与认知结构特征分析 [J]. 心理学报，57(06)：967-986.

[35] Wang, H., Chen, P., Luo, J., & Yang, Y.(2026). Tailoring educational support with graph neural networks and explainable AI: Insights into online learners' metacognitive abilities[J]. Computers & Education, 240: 105452.

[36] 夏亮亮，沈可心，王宇，董艳(2025). 生成式人工智能情境下学生学习能动性：现状与影响因素 [J]. 现代远距离教育，(2)：37-47.

[37] 余胜泉，汪凡淙(2023). 人工智能教育应用的认知外包陷阱及其跨越 [J]. 电化教育研究，44(12)：5-13.

（编辑：魏志慧）

The Cognitive Mirror: How Does GenAI Foster Learners' Cognitive Development?

DONG Yan[1], ZHANG Weiyu[1], WANG Yu[2] & JIANG Ting[3]

(1. *Faculty of Education*, *Beijing Normal University*, *Beijing* 100875, *China*; 2. *School of Education*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*; 3. *College of Business and Economics*, *The Australian National University*, *Canberra ACT* 2600, *Australia*)

Abstract: *Generative artificial intelligence (GenAI) may promote learners' cognitive development, but it may also lead to cognitive offloading. The collaborative calibration of learners' internal cognitive models and systems' external learner models constitutes a key process through which GenAI supports cognitive development. Drawing on learner modeling and knowledge tracing, this study integrates active inference theory and open learner models to propose an analytical framework termed "Cognitive Mirror." The framework follows a cyclical process of "cognitive externalization–state representation–mirror feedback–internal feedback–re-externalization." Learners externalize their cognitive states through questioning, explanation, reasoning, and revision. Based on this interaction evidence, the system constructs a provisional representation of the learner's current state and translates it into mirror feedback that is visible, comprehensible, and negotiable. Learners then use this feedback to identify discrepancies between their current understanding and task requirements and to adjust their internal cognitive models. Their subsequent responses provide new evidence for the system to recalibrate its state representations and feedback strategies, thereby enabling the continuous updating of both models. The framework offers a theoretical basis for feedback design and human-AI collaborative teaching in GenAI-supported learning.*

Key words: *generative artificial intelligence*; *GenAI*; *cognitive mirror*; *cognitive development*; *learner modeling*; *human–AI collaboration*