借助大语言模型实现有效的学生模拟

Zhihao Yuan[1*]、**Yunze Xiao**[1*]、**Ming Li**[2*]、**Weihao Xuan**[3],**Richard Tong**[4],**Mona Diab**[1],**Tom Mitchell**[1]1卡内基梅隆大学、 [2]马里兰大学[3]东京大学、 [4]NEOLAFzhihaoyu@andrew.cmu.edu、
{yunzex,mdiab,mitchell}@cs.cmu.eduminglii@umd.edu、
weihaoxuan@g.ecc.u-tokyo.ac.jp、richard.tong@ieee.org

arXiv:2601.05473v1 [cs.CL] 9 Jan 2026

摘要

教育创新常受制于以人为对象的研究所需的成本、时间和伦理复杂性。模拟学生提供了一种补充方法：学习者模型可用于教学测试、数据生成、教师培训和社会学习（VanLehn 等，1998年；Koedinger 等，2015年）。大语言模型（LLM）大幅增强了此类模拟器的表达能力，但也带来了效度风险：流畅的互动可能掩盖不切实际的错误模式和学习动态。我们认为，当要求一个能力广泛的模型表现为知识不完备的学习者时，就会产生这种能力悖论。我们将问题重新界定为认知状态规范：教育有效性取决于明确界定模拟学习者在特定时刻能够获取哪些内容（知识、策略、表征和资源）、错误如何形成，以及其状态是否以及如何随教学而演变。我们将基于大语言模型的学生模拟形式化为在此类规范下进行的受约束生成，并提出目标—环境框架，依据教学目标和部署情境对模拟器进行定位。最后，我们探讨了构建有效、可复用且安全的基于大语言模型的学生模拟器所面临的开放挑战。

1 引言

教育创新常受人类受试者研究在后勤、伦理和资金方面所面临挑战的制约。研究和实施新的教育干预，需要应对知情同意义务、降低数字环境中的隐私风险，并管理非随机流失，同时还要承担在不同场所实施干预的成本（Pardo 和 Siemens，2014；Fixsen 等人，2005）。

为应对这些限制，研究人员长期以来一直在研究模拟学生：这类学习者的计算模型可被赋予多种教学角色。经典研究强调了其应用，包括教师培训、以教促学与协作，以及在教学开发中的形成性应用（VanLehn 等人，1998年）。

尽管学生模拟在智能辅导系统中已有悠久历史，但近期研究开始探索将大语言模型（LLM）用作模拟学生的灵活基础。通过支持开放式自然语言交互，基于大语言模型的系统降低了在结构化程度较低的领域和互动场景中构建交互式学习者模型的门槛（Chu 等人，2025）。这种灵活性推动了对更丰富的模拟行为形式的研究，包括随时间展开的长期互动，以及参与多智能体学习环境（Zhang 等人，2025b）。

与此同时，流畅度和生成能力的提升也对教育效度构成了新的威胁。令人信服的语言可能掩盖不切实际的学习动态，将学习者的异质性简化为刻板印象，或引入误导解读的伪影（Li 等人，2025a）。一种关键的失效模式是能力悖论：经过训练而具备广泛能力的模型，却被要求表现出有限的知识和由错误观念驱动的行为。我们认为，这种张力反映了模型能力与模拟器旨在呈现的学习者状态之间的不匹配。由于大语言模型不会天然地将生成内容与明确的学习者状态绑定，其错误可能表现为偏离专家推理的表面现象，而非稳定且有助于诊断的错误观念。因此，教育效度取决于是否明确规定模拟学习者能够和不能获取哪些内容，以及预期其行为如何随教学而变化。

*贡献相同。

学生模拟

学生模拟是指生成类似学习者行为（例如答案、问题、错误或对话行为）的生成式系统。我们将学生模拟与学生建模区分开来，后者侧重于推断或预测学习者状态，而不呈现互动行为。我们还指出，当代基于大语言模型的学生模拟器很少用于检验教学理论；因此，我们聚焦于当前实践中的基于大语言模型的模拟学生。

为使教育主张具有可比性，并使评估保持一致，我们提出将明确的**认知状态规范**作为系统必须声明的内容。该规范说明在特定时刻，模拟学习者能够接触到哪些概念、策略、表征和资源，哪些知识缺口或错误观念构成其行为的基础，以及这种可接触性如何随时间演变。我们将这一要求置于一个情境感知框架中，该框架沿两个维度组织模拟学生系统：行为目标（表现、学习和人类方面）以及环境（学科领域、学习者群体和交互模态）。作为一项贯穿各方面的要求，认知状态规范为评估奠定基础，避免过度宣称，并支持超越基于流畅度的代理指标进行有意义的比较。

我们的贡献如下。

•我们将模拟学生定位为教育研究和系统开发的一种互补方法，涵盖教师培训、数据生成、社会学习和内容评估等用途。

• 我们将能力悖论视为模型能力与预期的学习者状态访问权限不匹配，并将学生模拟形式化为在明确的认知状态规范下进行的受约束生成。

• 我们提出一种情境感知框架，根据行为目标和环境对模拟学生系统进行分类，以支持评估对齐与比较。

• 我们指出了实现累积进展所面临的开放挑战，包括效度威胁、危害风险，以及对可复用基线和标准化评估的需求。

2 相关工作

2.1 传统学生模拟

模拟学生长期以来一直是学习理论与教育系统的设计和评估之间的桥梁，支持教师培训、以教促学、教学评估和内容编写等应用（Van-Lehn 等人，1998年；Koedinger 等人，2015年）。

早期研究主要采用基于规则的认知模型，以白盒视角看待学习与错误，认为行为源于明确的符号机制。错误被建模为系统性的程序性错误（Brown 和 VanLehn，1982），后来又在 ACT-R 等认知架构中得到形式化（Anderson 等人，1995；Taatgen 和 Anderson，2010年）。Sim-Student 和 Apprentice Learner 等系统扩展了这一范式，通过交互学习产生式规则，同时保留明确的表征，以支持诊断和教学干预（Matsuda 等人，2007；MacLellan 等人，2016）。尽管这些方法具有理论基础且易于解释，但它们需要大量人工构建的结构，且难以扩展到开放式领域。

2.2 基于大语言模型的学生模拟

近期研究利用大语言模型（LLM）充当模拟学生，发挥其创作成本低、对话灵活的优势。在以教促学的情境中，先前的系统通过显式或动态变化的学习者状态来约束大语言模型代理，以调节其能力并促进富有成效的互动。这些研究表明，教育效用取决于认知控制，而非仅仅取决于语言流畅度（Jin 等人，2024年；Rogers 等人，2025）。

另一条研究路线将模拟从个体学习者扩展到课堂级模拟，利用角色设定、规划和记忆来维持多智能体教育交互（Yue 等人，2024年；Zhang 等人，2025b）。相关研究尝试从基于大语言模型的学习者中引出结构化推理轨迹和错误观念，以支持可解释性和形成性应用（Sonkar 等人，2024年）。总的来说，这些研究凸显了基于大语言模型的模拟器既有潜力也有风险：尽管表达能力强且易于扩展，其有效性取决于是否明确规定学习者状态和学习动态。

3 基于大语言模型的模拟学生的形式化与验证

大语言模型（LLM）推动教育模拟研究近期从模拟专家导师转向模拟学习者本身。基于大语言模型的系统不再仅仅着眼于优化教学策略，而是日益致力于再现学习者的行为（Qi 等人，2026）、推理轨迹（Lu和Wang，2024年）以及交互动态（Zhou 等人，2025）。这种转变并非表面上的角色扮演变化。模拟学生必须受到有意的约束：与以提供帮助和确保正确为优化目标的通用助手不同，高保真度的模拟学生必须再现人类学习中典型的不完美之处，包括系统性误解、非最优推理路径、情感驱动行为（例如焦虑、失去参与），以及随时间演变的学习轨迹。

3.1 将学生模拟视为受约束生成。

一个核心挑战是能力悖论，我们认为其根源在于一个不可消解的先验知识纠缠问题。主流大语言模型经过训练，具备广泛的能力、自我纠正能力和亲社会倾向。与人类学习者不同，它们一旦内化了解题图式或专家启发式，就无法真正将其“遗忘”。因此，即使提示它们扮演新手，其潜在推理轨迹仍会受到目标学生尚未获得的专家先验的影响。

因此，我们将基于大语言模型的学生模拟定义为一项**受约束生成任务**，其目标并非最大化正确率，而是生成始终处于明确认知边界之内的回答。认知边界具体规定了模拟学习者在特定时刻能够合理调用的内容（概念、策略、表征和资源），以及构成其错误的错误观念或知识缺口。

具体而言，模拟学生必须满足三个相互关联的要求：

1. 错误保真度：给定一种目标错误观念（或局限），模拟器应在该观念相关的问题类型中始终如一地应用它；2. 认知一致性：生成的错误和解释应能从所述认知边界中找到因果依据，并且在同构题目和多轮交互中保持稳定，而非只是偶尔偏离专家推理的表层偏差；以及

3. 能力边界：在这些区域之外，模拟器仍应按照其假定的能力水平行事，既要避免泄露不可获取知识的专家捷径，也要避免退化为随机噪声。

这一框架促使我们采用明确的控制机制（提示、解码约束、状态变量、外部控制器或结构化知识表示），以抵消模型倾向于自我纠正并趋向全局最优推理的默认行为，同时允许随着模拟学习者的进步动态更新其认知边界。

3.2 学生模拟设计

为了使学生模拟器具有可比性，并使评估与实际主张相符，我们沿两个基本维度组织设计空间：行为目标 和环境。行为目标规定哪些学习者行为方面是模拟器旨在复现的（例如，反应分布、学习轨迹或社会情感行为）。环境则规定这些行为在何处 表现并受到约束，包括学科领域、目标学习者群体和交互模态。关键在于，两个系统都可以被称为“模拟学生”，但彼此并不可比：旨在匹配客观评分下简答数学题错误分布的模拟器，无法直接与旨在模拟开放式对话中长期学习和寻求帮助的模拟器相比较。明确这两个维度，可以避免夸大主张，并确保评估针对具体情境检验恰当的保真度。

3.2.1 行为目标

行为目标规定了模拟器必须复现学习者行为的哪些 Facet。我们区分了三种在以往研究中经常被混为一谈的目标。

模拟表现。模拟器在某一时间点再现学生的可观察输出，包括成功率和典型错误模式。当逼真的反应分布比模拟学生如何获得或未能获得底层知识更重要时，这一目标对于题目难度估计、干扰项生成和合成数据增强至关重要。

这一目标便尤为重要，因为此时模拟学生如何获得或未能获得底层知识并不如模拟其表现重要。

模拟学习。 模拟器将过程 建模为一条轨迹。与静态的表现快照不同，学习模拟要求状态在交互过程中不断演化，包括技能习得、遗忘以及对干预措施（如支架、反馈时机和同伴效应）的敏感程度差异。这一目标对于教师培训（观察教学所产生的后果）至关重要，也有助于训练自适应辅导系统，因为其策略取决于学生如何变化，而非学生某一次给出的答案。

模拟人类特征。 模拟器再现影响学习交互的非认知属性，例如个性、动机、情绪和社会语言风格。这些因素决定了课堂互动和同伴学习互动在行为上是否合乎情理。情感真实性还要求对寻求帮助的行为进行建模，包括工具性、执行性和回避性行为；这些行为会随自信、焦虑和感知社会风险而变化，而非一概将学生设定为合作的“好学生”。

3.2.2 环境

环境规定了外部约束，这些约束界定了现实性在实践中的含义，以及可用的输入和模态。

学科。 不同学科需要不同的知识表征和错误模型。结构化领域（如数学、编程）具有客观的正确性标准和明确的错误概念模式；开放式领域（如写作、讨论）则需要对主观推理、论证策略以及依赖评分标准的评估进行建模。

学生群体与角色一致性。 模拟器应匹配目标人群（年龄、熟练度、语言背景、文化背景、神经多样性），因为这些因素会影响错误观念和互动风格。角色设定不只是提示属性，还应作为贯穿多个轮次和任务的一致性约束，防止出现不合理的能力漂移（例如，自称注意力不集中，却持续展现高度专注的推理能力）。

学习环境与模态。 互动场景会限制可观察到的行为。纯文本对话不同于教室或 VR 场景，在后者中，语音、目光和具身动作都很重要。即使在文本环境中，伪影（方程、图表、代码编辑器）的可用性也会改变错误的表现形式。环境建模还应考虑教学结构，例如支架和逐步撤除支架：真实的学生可能依赖提示，并在支持措施撤除后表现退步，而不是立即实现泛化。

3.3 认知状态规范

除前述两个维度外，我们还要求每个模拟学生系统明确声明其认知状态规范 （ESS）。
ESS 定义了学习者在特定时刻掌握哪些知识以及能够获取哪些知识、错误如何产生，以及该状态是否会随时间变化，以及如何变化。具体而言，ESS 规定：(i) 学习者在时间点 t可用的表征、知识元素、策略和资源；(ii) 系统性错误的来源，例如错误观念或不完整程序；以及 (iii) （如有）支配状态转换的更新机制。

ESS 旨在厘清系统在多大程度上模拟的是固定能力水平下的表现 ，而非作为有状态过程的学习。我们将 ESS 操作化为一个包含五个等级的轻量级报告标签：

•E0：未指定。学习者的内部知识、错误来源和状态转移均未定义。生成输出时没有明确的认知约束。

• E1：静态有界。学习者使用一组固定且预先指定的知识元素、技能或错误模板，这些内容在交互过程中不会发生变化。其行为反映的是给定初始能力水平时的表现，不涉及学习。

• E2：课程索引。学习者可获取的知识或错误模式会根据外部进度信号进行更新，例如课程进度、问题索引或掌握度变量，但没有明确的错误观念或策略变化模型。

• E3：误解结构化。学习者受一个明确且稳定的错误观念、策略或部分程序模型所支配

这些因素会因果性地决定行为。错误源于可识别的认知结构，而非表面上的随机性。

- **E4：经校准或学习**。学习者状态表征和状态转移动态基于人类互动数据进行学习或校准，因此，表现和状态演化都有观察到的学习者轨迹作为实证依据。

此标签是一项贯穿各处的声明，旨在防止过度宣称、促进不同系统之间进行有意义的比较，并使评估方案与模拟器所声明的认识论约束保持一致。附录中提供了若干 ESS 等级的示例B。

基于大语言模型的学生模拟：四个有前景的方向

本节概述基于大语言模型的模拟学生在哪些方面最有前景，以及原因何在。我们将这些机遇归纳为四个方向：教师培训、社会学习（包括边教边学和协作学习）、数据生成以及内容评估。对于每个方向，我们都重点介绍其教育动机，以及大语言模型带来的独特能力。**表1 总结了这四个方向，并列出基于大语言模型的模拟学生在所有方向上都能带来的三项共同优势：可扩展性**（部署不受人力可用性的限制）、**安全性**（无风险实验，可减少隐私、法律责任和伦理方面的顾虑）以及**多功能性**（对多样化学习者特征进行行为建模，而这些特征难以或无法通过真实学生可靠地复现）。

4.1 教师培训

高质量的教学是促进学生成就的主要驱动力，而教师培养不足会对短期和长期学业成果造成不利影响（Markel 等人，2023）。然而，由于在线教育迅速扩张，以及越来越依赖大规模的新手辅导员和助教队伍，提供严格的培训变得愈发困难（Pan 等人，2025；Markel 等人，2023）。实际教学机会往往十分有限，且受培训项目无法控制的因素制约（Mohne 等人，2023）。其他传统培训方案，从网络研讨会到人工编写的视觉模拟或虚拟现实模拟，维护起来既耗费资源又耗时。因此，这些方法难以按照所需规模部署，以便为不断壮大的职前教育工作者和兼职教育工作者队伍提供标准化培训（Lee 等人，2023；Christensen 等人，2011）。

除了可扩展性之外，传统的实地教学经历还面临固有的结构性局限。在真实课堂中进行培训会带来显著的问责、法律责任和隐私风险（Chris- tensen 等人，2011）。至关重要的是，有限的接触时长限制了新手接触学生异质性的机会；真实世界的实习无法保证新手能接触到足够多样的个性、学习风格和能力倾向，而这些正是掌握适应性教学所必需的（Sanyal 等人，2025；Knezek 和 Christensen，2024年）。缺乏多样化的接触机会会阻碍复杂技能的发展，例如管理全班互动动态，或促进基于论证的讨论（Nazaretsky 等人，2023）。

基于大语言模型的模拟学生通过提供频繁、结构化且无风险的演练机会，为应对这些挑战提供了解决方案（Judge 等人，2013年）。与静态环境不同，大语言模型能够模拟多样且与个性相符的行为，从而为练习适应性策略提供丰富的试验场（Ma 等人，2025；Pentangelo 等人，2025）。研究表明，这些模拟环境能够缓解真实教室带来的即时时间压力，让教师得以拟定更周到、更具包容性的回应，并围绕学习目标制定策略（Markel 等人，2023）。通过将这些智能体置于逼真的情境中，TutorUp 等系统使新手能够在安全、可控的环境中练习应对失去参与和困惑的情况，从而弥合理论与实践之间的差距（Pan 等人，2025；Lee 等人，2023）。

4.2 社会学习

模拟学生还可以作为社会性学习伙伴，进一步促进教育。以教促学（LbT）是一种行之有效的教学模式，学生通过梳理知识以便讲解、对学习对象承担责任，以及反思自身的学习过程，从而实现更深入的概念理解（Biswas 等，2005）。尽管以教促学在不同年龄群体和领域中都颇具成效，设计有效的以教促学环境仍面临挑战。一项关键局限在于，导师和学习对象都不是专家；在缺乏指导的情况下，参与者往往会强化共同的错误观念，或将注意力放在无关紧要的细节上，而非核心

表1：模拟学生研究的维度映射：目标与环境

应用背景	相关行为目标	环境考量
教师培训	在模拟学习中实现高保真度（轨迹-故事）和人类方面（情感、个性）。	需要学生群体具有高度异质性以及类似课堂的互动环境，可能还需要多模态支持。
社会学习	聚焦于人类方面 （社会语言学）以及模拟学习（以引发门生效应）。	基于对话的界面；协作式或一对一同伴环境。
数据生成	侧重于模拟表现（错误模式）和学习（纵向日志）。	通常是结构化学科（科学、技术、工程和数学、编程——编程）；数字交互日志和标准化——化任务。
内容评估	主要关注模拟表现（静态成功/失败分布）。	高通量平台集成；客观的正确性标准和可扩展的测试套件。

概念（Debbané 等，2023）。人类专家可以假装不懂来引导学生进行讲解，从而缓解这一问题，但这种角色扮演既费时费力，也难以扩展（Duran，2017）。

基于大语言模型的模拟学生为这一动态提供了一种可扩展的替代方案。这些智能体可以模拟新手学习者，以诱发门生效应；在这一效应中，学生会更加努力地学习自己预期要教授的材料，同时利用其底层专家知识，巧妙地引导学生走向正确理解（Rogers 等，2025）。此外，教授人工智能体还能营造一种心理安全的环境，减轻对受到评判的恐惧，以及与实时人际互动通常相伴的社会压力（Chase 等，2009；Debbané 等，2023）。

除了导师—受教者互动之外，模拟学生也可以在协作学习框架中充当同伴。研究表明，积极的同伴互动能够改善学业成果，并培养沟通和团队协调等重要软技能（Veldman 等人，2020年；Rohrbeck 等人，2003）。相反，缺乏实时反馈和互动的教育环境往往难以取得理想的学习成果。在当今的数字环境中，预录制讲座、远程教学和异步日程已成为常态，学生常常面临后勤和社交障碍，难以找到学习伙伴（Wang 等人，2025）。模拟学生可以填补这一空缺，作为随时可用的虚拟同伴或学习伙伴，为原本孤立的学习环境重新带来协作互动的益处。

4.3 数据生成

现代智能教育平台（例如 Coursera）和基于大语言模型的教育工具（例如 Google LearnLM）的效能，在很大程度上依赖于高质量训练数据的可用性，以支持个性化和算法优化（Gao 等人，2025；Team 等人，2025）。然而，获取高保真度教育数据集（例如学生回答和对话日志）受到严格隐私法规的严重限制，而人工标注的相关成本也很高。此外，真实世界教育数据往往稀少或含有噪声，而要取得出色的算法表现，则需要全面且结构良好的数据集（Chen 等人，2025；Gao 等人，2025）。

模拟学生通过充当可扩展的合成数据生成引擎来缓解这一匮乏问题（Li 等人，2025b）。通过对不同学习者特征进行建模，这些智能体能够生成大规模且符合隐私要求的数据集，模拟现实世界的分布。近期文献展示了这一方法在两个关键领域的实用价值：生成逼真的静态学生回答（Miroyan 等人，2025；Benedetto 等人，2024年；Wu 等人，2025；Ross 和 Andreas，2025），以及合成复杂的纵向任务求解轨迹（Ross 等人，2025；Sharma 和 Li，2024）。

4.4 内容评估

教育技术平台和教学对话代理（PCAs）的快速扩张，例如可汗学院的Khanmigo、松鼠AI和多邻国，既为个性化教学带来了机遇，也对确保内容质量和教学成效提出了挑战（Gonnermann-Müller等，2025；Jin等，

2025）。评估这些系统面临着关键瓶颈。专家教师评估的成本高得令人难以承受，而使用真实学生进行试点测试则会引入隐私风险，并可能使其接触有害内容（Ezzaki等，2024；Jin等，2025）。模拟学生提供了一种极具吸引力的替代方案，可为系统级验证和内容级评估提供高通量且无风险的环境，尤其适用于项目难度建模（Hambleton等，1991；Hsu等，2018；AlKhuzaey等，2021年；Li等，2025d；Peters等，2025）。尽管准确的项目难度建模对于评估效度和自适应学习算法至关重要（Dueñas等，2024年），但包括人工专家标注和项目反应理论在内的传统方法，存在主观性强、可扩展性有限，或依赖大量历史学生数据等问题。这些局限使模拟学生成为越来越有吸引力的解决方案。

模拟学生通过充当合成应试者来弥补这些局限。通过对不同能力水平的学习者进行建模，这些智能体能够生成表现数据，以预测题目难度，而无需进行大规模的人类试测。近期的框架，如 UPN-ICC、QG-SMS 和 SMART，表明模拟学生无论是独立使用，还是用于增强项目反应理论模型，都能准确校准教育内容（Dueñas 等人，2024年；Nguyen 等人，2025；Scarlatos 等人，2025）。与此同时，近期的大规模证据表明，仅仅用大语言模型替代人类试测者，并不能解决这一问题：即使难度标签基于真实学生实地测试，现成的大语言模型仍可能与人类的难度判断存在系统性偏差，而且难以可靠地模拟低熟练度认知状态（Li 等人，2025c）。这凸显了构建经过校准的目标条件化模拟学生的必要性，同时必须明确规定其熟练度控制和验证方案。

5 项挑战

5.1 数据匮乏与隐私

高保真学生模拟依赖细粒度的真实世界轨迹数据，以捕捉学习者错误、反馈和教学情境。实际中，这类数据十分稀缺：收集成本高昂（通常需要专家标注），而持续获取教室或专有平台的访问权限也很困难。缺失的教学材料以及干预情境也会造成持续存在的情境空缺，限制模拟器呈现教学法如何影响学习成果的能力（Zhai 等人，2021年；Baker 和 Siemens，2014；Xu 等人，2024年）。

隐私限制进一步限制了可收集和可共享的内容。教育轨迹数据通常包含丰富的人口统计和行为标识信息，因此许多数据集无法重新分发，损害了可复现性，也拖慢了累积式进展（Hoel 和 Chen，2018；Wu 等人，2025）。对于基于大语言模型的模拟，风险会进一步叠加，因为在教育轨迹数据上训练的模型可能会记忆并复现私人信息，或表现出会扭曲教学互动的隐私相关偏差（Shvartzshnaider 和 Duddu，2025）。

5.2 不可验证结果下的评估与标准化

随着模拟学生越来越多地应用于长篇探究、开放式推理和真实课堂对话，评估变得困难，因为目标行为很少能通过二元正确性加以验证。近期证据表明，出现了更广泛的**不可验证性危机**：标准自动化指标，甚至基于大语言模型的评审器，都可能与人类专家偏好不一致；在一些科学推理情境中，据报告，偏好匹配准确率约为 ∼65%（Zhao 等人，2025）。这种局限因风格与实质内容不匹配而加剧：模型生成的回答虽然 polished 且自信，却缺乏依据或逻辑不一致。因此，自动化评估器可能无法区分正在进行有效尝试的学生与令人信服地生成幻觉的模拟器（Zhang 等人，2025a）。

这些测量挑战直接凸显了采用**标准化评估框架**的必要性，而此类框架应明确与模拟器预期的行为目标和部署环境相契合。具体而言，评估学习动态需要采用不同于评估静态正确性的工具。然而，由于缺乏统一的基准和报告规范，当前研究转而依赖临时的专家评分；这种做法难以大规模推广、不易复现，也难以在不同研究之间进行比较。这种模糊性使跨论文比较变得脆弱，因为恰当的保真度概念取决于模拟器的行为目标和部署环境，

而在缺乏统一且以目标为条件的基准时，这一概念仍未得到充分界定。

6 建议

6.1 将认知状态规范作为可复现性的强制要求

应将认知状态规范（ESS；E0–E4）视为模拟学生系统必须报告的产物，因为能力悖论归根结底是认知状态不匹配：大语言模型无法真正“忘却知识”，因此可能在产生新手式错误的同时泄露专家知识。因此，系统描述应明确说明每一轮可访问的知识、策略和资源，以及这些访问权限如何随时间演变；从未指定（E0）转向误解结构化（E3）或校准（E4）规范，可以使相关主张具备可证伪性（例如，在改述以及跨同构题目时表现出稳定的错误概念行为），将“学生水平”转化为可审计的设计选择，并实现有意义的跨论文比较。

6.2 将评估重点从表层真实性转向目标对齐保真度

评估应优先关注目标对齐保真度，而非表层真实性。当流畅的对话和笼统的“人性化”评分被视为教育有效性的证据时，就会产生保真度评估差距，尽管这些指标可能掩盖不可信的学习动态。此外，为不同目的构建的模拟器（例如用于数据生成或教师培训）需要根本不同的成功标准。

指标应根据模拟器的行为目标和环境来确定。表现和数据生成设置需要受控错误分布，并能抵御偏向专家推理。以学习为导向的设置需要使用学习轨迹指标，例如逐步改进以及对反馈保持适当的敏感度。人类因素和社会学习设置则需要与具体应用相匹配的交互指标，例如坚持性、寻求帮助，或在相关情况下体现情感连贯性。

6.3 融入明确的学习机制以建模学习轨迹

纵向模拟受益于明确的学习机制，而非仅依赖上下文窗口提示。仅靠提示的方法往往会产生不切实际的动态变化，包括从新手到专家的骤然跃迁、能力在多轮交互中不稳定，以及对措辞脆弱的依赖，这些问题削弱了使用模拟器研究学习或对教学干预进行压力测试的有效性。

一种实用的方向是采用混合架构，将大语言模型与明确的学习者状态表征及定义清晰的转移规则相结合，例如知识追踪、熟练度变量、错误概念图或认知模型。这样可以使学习轨迹更易于解释和校准，也能将失败诊断为状态、转移或界面错误，而不是不透明的模型方差。

6.4 建立标准化错误概念基准，以应对不可验证性

我们需要针对不可验证的教育行为建立共享基准，尤其是针对错误观念的一致性和纵向社会情感学习轨迹。许多对教育效度至关重要的属性无法仅凭正确性来评估，而自动化评判器可能会被流畅却不连贯的行为误导，导致该领域仍依赖昂贵、临时拼凑的专家评分和脆弱的比较结果。

基准测试套件应强调受控变化下的一致性。误解测试可使用同构题目和释义改写，评估与声明的认知状态相符的错误特征是否稳定。学习测试可采用多轮课程，评估模型在反馈下是否逐步、有选择地修正。社会情感测试可使用引发挫折感或重新投入的情境，评估其随时间推移是否呈现连贯的发展轨迹。开放、标准化的测试套件将直接提升可复现性和可比性。

7 结论

总之，要推动基于大语言模型的学生模拟从探索性原型发展为严谨的科学工具，该领域必须通过优先保障**认知忠实度**而非表层真实性，来化解“能力悖论”。我们通过**目标—环境框架**和**认知状态规范（ESS）**将这一转变形式化，确保模拟行为可因果归因于明确界定的学习状态，而非随机性幻觉。采用这些标准后，研究人员便能将模拟学生转变为值得信赖且可复现的测试平台，推动教育创新，弥合对话流畅性与有效教学建模之间的鸿沟。

局限性

本研究主要属于理论研究，综合了教育心理学、学习科学和自然语言处理领域既有文献中的见解。我们未对所提出的认知状态规范开展实证评估、课堂部署或大规模心理测量学验证。因此，关于如何化解“能力悖论”的论述，尚未通过直接的教学互动或纵向系统测试得到验证。尽管我们提出了认知状态（E0–E4）的独特分类法，但现实世界中的模型行为往往表现出波动不定或难以界定的知识边界。为提升分析清晰度，我们对框架中的这些维度作了理想化处理；这可能会限制其应用于随机、非确定性基础模型时的稳健性。

此外，所提出的解决方案假设开发者能够通过架构设计施加严格的知识约束，但提示工程本身的脆弱性以及商业黑箱模型的不透明性可能会削弱这一做法。未来的工作应侧重于通过受控实证研究验证该框架，衡量符合认知状态规范的模拟器能否有效防止知识泄露，并开发自动化指标，以量化不同学习情境下的认知忠实度。

伦理考量

如果生成的行为没有严格以明确定义的认知状态为依据，部署基于大语言模型的学生模拟器就会带来重大教学风险。一个主要问题是教师教育中可能出现负训练迁移：新手教师若与语言表达逼真、但学习动态不真实的模拟器进行练习，可能会形成不良的教学策略。例如，如果模拟器对教学干预做出看似合理、实则缺乏因果关联的改进反应，比如仅仅在得到一个简单提示后，就表现得仿佛立即掌握了复杂概念，那么这会强化表层教学手段，而非深层的教学推理（Dai 和 Xiao，2025）。这种不匹配会带来风险：模拟器沦为一种“教学安慰剂”，营造出练习有效的假象，却无法反映真实课堂中观察到的认知阻力和逐步的技能习得。

致谢

我们感谢 Ken Koedinger 教授和 Carolyn Rosé 教授提供宝贵的反馈和建议，我们已将其中大部分纳入本文。我们也感谢 Gati Aher 和 Aylin ÖZTÜRK 在本项目构思初期提供的有益建议。

此外，如果与“学习困难”或“错误观念”相关的行为和语言线索在设计中未能因地制宜或纳入参与式设计，基于大语言模型的学生档案可能会形成规范性或文化偏见（Xiao等人，2025）。这会使代表性不足的文化群体进一步边缘化（AlKhamissi等人，2025），或强化关于学术能力的主流刻板印象。尽管我们的目标—环境框架倡导根据具体情境校准学生档案，但仍需开展更多实证研究，以验证这类策略在不同教育环境中的有效性。

我们也认识到，我们的框架可能被双重用途。对“人类特征”的分类体系可能被用于设计更具说服力或在情感上更具操控性的系统，尤其是在商业辅导或监控场景中。我们鼓励未来的工作制定缓解策略，例如可解释性指标（如向用户展示当前生效的 ESS 状态）、受到约束的拟人化档案，或门控发布机制，以帮助监测和控制模拟学生的行为。我们还强调，在系统开发过程中，与学习科学家、伦理学家和受影响社区开展跨学科协作至关重要。

通过阐明大语言模型中认识论模拟的功能优势及其可能造成的危害，我们旨在推动透明、与社会价值相契合且关注用户的设计实践。我们强烈鼓励未来研究通过实证验证并完善这一框架，尤其是借助参与式协同设计和跨文化评估。

参考文献

Mai AlKhamissi、Yunze Xiao、Badr AlKhamissi 和 Mona Diab。2025。聘请你的人类学家！从人类学视角重新思考文化基准。预印本，arXiv:2510.05931。

Samah AlKhuzaey、Floriana Grasso、Terry R Payne 和 Valentina Tamma。2021年。数据驱动的题目难度预测方法系统综述。载于

国际人工智能教育会议，第 29–41 页。施普林格。

John R. Anderson、Albert T. Corbett、Kenneth R. Koedinger 和 Ray Pelletier。1995年。认知导师：经验教训。《学习科学杂志》，4(2)：167–207。

Ryan S. Baker 和 George Siemens。2014年。教育数据挖掘与学习分析。《剑桥学习科学手册》，第253–272页。

Luca Benedetto、Giovanni Aradelli、Antonia Donvito、Alberto Lucchetti、Andrea Cappelli 和 Paula Buttery。2024 年。使用大语言模型模拟学生对考试题目的回答。载于《计算语言学协会年会：*EMNLP 2024*论文集》，第11351–11368页。

Gautam Biswas、Krittaya Leelawong、Daniel Schwartz、Nancy Vye 和范德比尔特大学可教代理研究组。2005年。以教促学：教育软件的一种新型代理范式。《应用人工智能》，19(3-4)：363–392。

John Seely Brown 和 Kurt VanLehn。1982。《迈向“错误”的生成理论》。载于 Thomas P. Carpenter、James M. Moser 和 Thomas A. Romberg 编，《加法与减法：认知视角》，第117–136页。Erlbaum，新泽西州希尔斯代尔。

Catherine C. Chase、Doris B. Chin、Marily A. Oppezzo 和 Daniel L. Schwartz。2009。可教学智能体与门生效应：提高学习投入。《科学教育与技术杂志》，18：334–352。Si Chen、Izzy Molnar、Ting Hua、Peiyu Li、Le Huy Khiem、G Alex Ambrose、Jim Lang、Ronald Metoyer 和 Nitesh V Chawla。2025。SimInstruct：一种负责任的工具，用于收集专家与由大语言模型模拟的新手之间的支架式对话。*arXiv* 预印本 *arXiv:2508.04428*。Rhonda Christensen、Gerald Knezek、Tandra Tyler-Wood 和 David Gibson。2011。Simschool：一种用于加强教师培养的在线动态模拟器。《国际学习技术杂志》，6(2)：201–220。Zhendong Chu、Shen Wang、Jian Xie、Tinghui Zhu、Yibo Yan、Jingheng Ye、Aoxiao Zhong、Xuming Hu、Jing Liang、Philip S. Yu 和 Qingsong Wen。2025。用于教育的大语言模型代理：进展与应用。载于计算语言学协会：*EMNLP 2025*会议论文集，第13782–13810页，中国苏州。计算语言学协会。Gordon Dai 和 Yunze Xiao。2025。拥抱矛盾：理论上的不一致不会阻碍构建负责任的人工智能系统之路。预印本，arXiv:2505.18139。

Amy Debbané、Ken Jen Lee、Jarvis Tse 和 Edith Law。2023。以教促学：关键挑战与设计启示。*ACM* 人机交互（*PACM HCI*）论文集，7(CSCW1)：第68篇文章。

George Dueñas、Sergio Jimenez 和 GeralEduardo Ma- teus Ferro。2024。UPN-ICC 在 BEA 2024 共享任务中的工作：利用大语言模型预测多项选择题的难度。载于第*19*届自然语言处理在教育应用构建中的创新用法研讨会论文集（*BEA 2024*）。

David Duran。2017。以教促学：作为一种教学机制的证据与启示。《教育与教学国际创新》，54(5):476–484。

Soukaina Ezzaki、Najat Messaoudi 和 Jaafar Khalid Naciri。2024。利用大语言模型通过虚拟学生模拟评估教育内容。《利用虚拟学生模拟和大语言模型评估教育内容》。

Dean L. Fixsen、Sandra F. Naoom、Karen A. Blase、Robert M.Friedman 和 Frances Wallace。2005。实施研究：文献综述。国家实施研究网络。

Weibo Gao、Qi Liu、Linan Yue、Fangzhou Yao、Rui Lv、Zheng Zhang、Hao Wang 和 Zhenya Huang。2025年。Agent4edu：通过生成式智能体为智能教育系统生成学习者作答数据。载于 *AAAI*人工智能会议论文集，第39卷，第23923–23932页。Jana Gonnermann-Müller、Jennifer Haase、Konstantin Fackeldey 和 Sebastian Pokutta。2025年。Facet：以教师为中心的基于大语言模型的多智能体系统——迈向个性化教育练习单。*arXiv*预印本 *arXiv:2508.11401*。Ronald K Hambleton、Hariharan Swaminathan 和 H Jane Rogers。1991。项目反应理论基础，第2卷。Sage。Tore Hoel 和 Weiqin Chen。2018。学习分析中的隐私与数据保护：以利益相关者为中心的分类体系。技术增强学习研究与实践，13(1):1–21。Fu-Yuan Hsu、Hahn-Ming Lee、Tao-Hsing Chang 和 Yao-Ting Sung。2018。多项选择测验项目难度的自动估计：词嵌入技术的应用。《信息处理与管理》，54(6):969–984。Hyoungwook Jin、Seonghee Lee、Hyungyu Shin 和 Juho Kim。2024年。教会人工智能编程：在编程教育中使用大语言模型作为可教代理。载于*CHI*计算机系统中的人为因素会议论文集（*CHI '24*），美国夏威夷州檀香山。计算机协会。

Hyoungwook Jin、Minju Yoo、Jeongeon Park、Yokyung Lee、Xu Wang 和 Juho Kim。2025年。Teachtune：借助模拟学生，针对不同学生档案评估教学代理。载于*2025*年*CHI*计算机系统中的人为因素会议论文集，第1–28页。

Sharon Judge、Jonna Bobzien、Ann Maydosz、Sabra Gear 和 Petros Katsioloudis。2013年。基于视觉的模拟环境在教师培养中的应用。教育与培训研究期刊，1(1)。

Tanja Käser 和 Giora Alexandron。2024年。教育技术中的模拟学习者：系统性文献综述与图灵式测试。人工智能教育国际期刊，34(2):545–585。

Gerald Knezek 和 Rhonda Christensen。2024年。通过在线模拟提升有特殊学习需求学生的教学成效。载于国际电子学习期刊，第133–152页。教育计算促进协会（AACE）。

Kenneth R Koedinger、Noboru Matsuda、Christopher J MacLellan 和 Elizabeth A McLaughlin。2015年。评估模拟学习者的方法：来自 SimStudent 的示例。载于*AIED*研讨会。

Unggi Lee、Sanghyeok Lee、Junbo Koh、Yeil Jeong、Haewon Jung、Gyuri Byun、Yunseo Lee、Jewoong Moon、Jieun Lim 和 Hyeoncheol Kim。2023 年。面向教师培训的生成式智能体：利用基于大语言模型的智能体为职前教师设计教育问题解决模拟。在 *NeurIPS'23* 生成式人工智能教育研讨会（*GAIED*）。Ang Li、Haozhe Chen、Hongseok Namkoong 和 Tianyi Peng。2025a。大语言模型生成的人设是一项有条件的承诺。 *arXiv* 预印本 *arXiv:2503.16527*。Dawei Li、Yue Huang、Ming Li、Tianyi Zhou、Xian-gliang Zhang 和 Huan Liu。2025b。 生成模型用于合成数据：在生成式人工智能时代转变数据挖掘。载于第 *34* 届 *ACM* 国际信息与知识管理会议论文集，CIKM '25，第 6833–6836 页，美国纽约州纽约市。计算机协会。Ming Li、Han Chen、Yunze Xiao、Jian Chen、Hong Jiao 和 Tianyi Zhou。2025c。大语言模型能否估计学生学习困难？利用熟练度模拟进行人类与人工智能难度对齐以预测题目难度。*arXiv* 预印本 *arXiv:2512.18880*。Ming Li、Hong Jiao、Tianyi Zhou、Nan Zhang、Sydney Peters 和 Robert W Lissitz。2025d。使用微调后的小型和大型语言模型进行题目难度建模。教育与心理测量，第 00131644251344973 页。

Xinyi Lu 和 Xu Wang。2024 年。 生成式学生：利用大语言模型模拟的学生画像支持题目评估。载于第十一届 *ACM Learning @ Scale* 会议论文集，L@S '24，第 16–27 页，美国纽约州纽约市。计算机协会。

Qianou Ma、Hua Shen、Kenneth Koedinger 和 Tong- shuang Wu。2024年。如何在人工智能时代教授编程？使用大语言模型作为可教学智能体进行调试。载于第*25*届国际人工智能教育会议论文集（*AIED 2024*），第14829卷，人工智能讲义，第1–16页，巴西累西腓。施普林格。

Yiping Ma、Shiyu Hu、Xuchen Li、Yipei Wang、Shiqing Liu、Kang Hao Cheong 和 Yuqing Chen。2025年。当大语言模型学着成为学生：用于建模和评估教育交互中虚拟学生智能体的 SOEI 框架。arXiv 预印本 arXiv:2410.15701v2。

Christopher J. MacLellan、Erik Harpstead、Rony Pa- tel 和 Kenneth R. Koedinger。2016年。学徒学习者架构：闭合学习理论与教育数据之间的循环。载于第*9*届国际教育数据挖掘会议（*EDM*）论文集，第151–158页。

Julia M. Markel、Steven G. Opferman、James A. Lan- day 和 Chris Piech。2023 年。GPTeach：利用基于 GPT 的学生进行交互式助教培训。载于第十届 *ACM Learning @ Scale* 会议论文集（*L@S '23*），第 1–11 页，丹麦哥本哈根。计算机协会。

Noboru Matsuda、William W. Cohen、Jonathan Se- wall、Gustavo Lacerda 和 Kenneth R. Koedinger。2007。Simstudent：通过辅导合成学生来构建智能辅导系统。投稿稿件。Mihran Miroyan、Rose Niousha、Joseph E. Gonzalez、Gireeja Ranade 和 Narges Norouzi。2025。ParaS- tudent：通过教大语言模型如何挣扎，生成并评估逼真的学生代码。arXiv 预印本 arXiv:2507.12674。Cameron Mohne、Dora Demszky、Nicholas Vo 和 Chris Piech。2023。Educasim：用于教学实践的交互式模拟对象。ArXiv 预印本 arXiv:2410.03017。可能存在后续正式发表的版本。Tanya Nazaretsky、Jamie N Mikeska 和 Beata Beigman Klebanov。2023。通过人工智能赋能教师学习：自动评估教师在以论证为重点的讨论中对学生观点的关注程度。载于*LAK23*：第*13*届国际学习分析与知识会议，第122–132页。Bang Nguyen、Tingting Du、Mengxia Yu、Lawrence An- grave 和 Meng Jiang。2025。QG-SMS：通过学生建模与模拟增强试题分析。arXiv 预印本 arXiv:2503.05888。

Sitong Pan、Robin Schmucker、Bernardo Garcia Bulle Bueno、Salome Aguilar Llanes、Fernanda Albo Alarcón、Hangxiao Zhu、Adam Teo 和 Meng Xia。2025。TutorUp：如果学生是模拟出来的，会怎样？培训导师应对在线学习中的参与度挑战。载于*2025*年*CHI* 计算机系统中的人为因素会议论文集，第1–18页。

Abelardo Pardo 和 George Siemens。2014 年。学习分析的伦理与隐私原则。英国教育技术杂志，45(3)：438–450。

Viviana Pentangelo、Luigi Turco、Stefano Lambiase、Carmine Gravino 和 Fabio Palomba。2025 年。Senem-ai：利用大语言模型模拟虚拟学习环境中的学生行为。*SoftwareX*，31：102278。

Sydney Peters、Nan Zhang、Hong Jiao、Ming Li、Tianyi Zhou 和 Robert Lissitz。2025 年。大规模评估中题目难度建模的文本方法：系统综述。*arXiv* 预印本 *arXiv:2509.23486*。

Changyong Qi、Longwei Zheng、Anna He、Haoxin Xu、Linzhao Jia、Yuang Wei、Bingqian Jiang 和 Xi- aoqing Gu。2026年。使用基于大语言模型的角色扮演智能体模拟学生学习行为：数据驱动的认知启发式框架。《应用专家系统》，304：130753。Kantwon Rogers、Michael Davis、Mallesh Maharana、Pete Etheredge 和 Sonia Chernova。2025年。装糊涂以变聪明：在大学计算机科学课程中创建并评估基于大语言模型的可教学智能体。载于《*CHI* 人机交互系统中的人为因素会议论文集》（*CHI '25*）。计算机协会。Cynthia Rohrbeck、Marika Ginsburg-Block、John Fantuzzo 和 Traci Miller。2003年。针对小学生的同伴辅助学习干预：元分析综述。《教育心理学杂志》，95：240–257。Alexis Ross 和 Jacob Andreas。2025年。学习犯错：对学生错误思维和关键错误进行建模。*arXiv* 预印本 *arXiv:2510.11502*。Alexis Ross、Megha Srivastava、Jeremiah Blanchard 和 Jacob Andreas。2025年。使用380万条程序轨迹对学生学习进行建模。*arXiv* 预印本 *arXiv:2510.05056*。Debdeep Sanyal、Agniva Maiti、Umakanta Maharana、Dhruv Kumar、Ankur Mali、C. Lee Giles 和 Murari Mandal。2025年。探究教学型教师与学生大语言模型智能体：遗传适应结合检索增强生成，跨越不同学习风格。*arXiv* 预印本 *arXiv:2505.19173*。Alexander Scarlatos、Nigel Fernandez、Christopher Ormerod、Susan Lottridge 和 Andrew Lan。2025年。SMART：与项目反应理论对齐、用于题目难度预测的模拟学生。arXiv 预印本 arXiv:2507.05129。

P. Sharma 和 Q. Li。2024年。设计模拟学生，以在开放式学习环境中模拟学习者活动数据。载于第*17*届国际教育数据挖掘会议论文集，第986–989页，亚特兰大，佐治亚州，美国。国际教育数据挖掘学会。

Yan Shvartzshnaider 和 Vasisht Duddu。2025 年。语言模型中的隐私偏差：一种基于情境完整性的审计指标。预印本，arXiv:2409.03735。

Shashank Sonkar、Xinghe Chen、Naiming Liu、Richard G. Baraniuk 和 Mrinmaya Sachan。2024年。基于大语言模型的存在误解学生的认知模型。预印本，arXiv:2410.12294。

NielsA. Taatgen 和 John R. Anderson。2010年。认知架构的过去、现在与未来。《认知科学专题》，2(1):110–131。

LearnLM 团队、Abhinit Modi、Aditya Srikanth Veerub-hotla、Aliya Rysbek、Andrea Huber、Brett Wiltshire、Brian Veprek、Daniel Gillick、Daniel Kasenberg、Derek Ahmed、Irina Jurenka、James Cohan、Jennifer She、Julia Wilkowski、Kaiz Alarakyia、Kevin R. Mc-Kee、Lisa Wang、Markus Kunesch、Mike Schaeker-mann 及另外 27 位作者。2025。LearnLM：提升 Gemini 的学习能力。预印本，arXiv:2412.16429。

Kurt VanLehn、Stellan Ohlsson 和 Rod Nason。1998年。模拟学生的应用：一项探索。《国际人工智能教育期刊》。M. A. Veldman、S. Doolaard、R. J. Bosker 和 T. A. B. Snijders。2020年。幼儿共同合作。一年级初等教育中合作学习对儿童小组合作的影响。《学习与教学》，67:101308。Tianjia Wang、Tong Wu、Huayi Liu、Chris Brown 和 Yan Chen。2025。生成式协同学习者：利用生成式人工智能提升异步学习中学生的认知临场感和社会临场感。《*ACM* 人机交互论文集》（*PACM HCI*），9(GROUP19)：文章 GROUP19。Tao Wu、Jingyuan Chen、Wang Lin、Mengze Li、Yu-meng Zhu、Ang Li、Kun Kuang 和 Fei Wu。2025。拥抱不完美：使用基于大语言模型的智能体模拟不同认知水平的学生。*arXiv* 预印本 *arXiv:2505.19997*。Yunze Xiao、Lynnette Hui Xian Ng、Jiarui Liu 和 Mona T. Diab。2025。机器拟人化：通过多层次设计框架重新思考大语言模型拟人化。载于*2025*年自然语言处理实证方法会议论文集，第3331–3350页，中国苏州。计算语言学协会。Songlin Xu、Xinyu Zhang 和 Lianhui Qin。2024。Edu-agent：学习中的生成式学生智能体。*arXiv* 预印本 *arXiv:2404.07963*。

Murong Yue、Wenhan Lyu、Jennifer Suh、Yixuan Zhang 和 Ziyu Yao。2024。Mathvc：一个用于数学教育的大语言模型模拟的多角色虚拟课堂。*arXiv* 预印本 *arXiv:2404.06711*。

Xuesong Zhai、Xiaoyan Chu、Ching Sing Chai、Morris Siu YungJong、Andreja Istenic、Michael Spector、Jia-Bao Liu、Jing Yuan、Yan Li 和 Ning Cai。2021年。2010年到2020年人工智能（AI）在教育中的应用综述。*Complex.*，2021年。

Linda Zhang、Kevin Liu、Sarah Chen 和 Richard Wang。2025a。Simulrag：用于将大语言模型落地于长篇科学问答的基于模拟器的检索增强生成。载于*2025*年自然语言处理实证方法会议论文集（*EMNLP*）。OpenReview 预印本。

Zheyuan Zhang、Daniel Zhang-Li、Jifan Yu、Linlu Gong、Jinchang Zhou、Zhanxin Hao、Jianxiao Jiang、Jie Cao、Huiqin Liu、Zhiyuan Liu、Lei Hou 和 Juanzi Li。2025b。利用大语言模型赋能的智能体模拟课堂教育。载于*2025*年美洲计算语言学协会分会会议论文集：人类语言技术（第*1*卷：长篇论文），第10364–10379页，阿尔伯克基，新墨西哥州。计算语言学协会。

Yilun Zhao、Kaiyan Zhang、Tiansheng Hu、Sihong Wu、Ronan Le Bras、Taira Anderson、Jonathan Bragg、Joseph Chee Chang、Jesse Dodge、MattLatzke、Yixin Liu、Charles McGrady、Xiangru Tang、Zihang Wang、Chen Zhao、Hannaneh Hajishirzi、Doug Downey 和 Arman Cohan。2025年。Sciarena：面向科学文献任务中基础模型的开放评估平台。预印本，arXiv:2507.01001。

Xuhui Zhou、Jiarui Liu、Akhila Yerukola、Hyunwoo Kim、MaartenSap。2025。社会世界模型。预印本，arXiv:2509.00559。

检索策略

为了确定与基于大语言模型的学生智能体相关的文献，我们在六个主要学术资源库中进行了全面检索：Semantic Scholar、LearnTechLib、arXiv、ACM Digital Library、SpringerLink 和 IEEE Xplore。遵循（Käser 和 Alexandron，2024年）的研究方法，我们采用了一个旨在检索模拟术语不同变体的布尔检索式：

> “模拟学生” OR “学生模拟”OR “模拟学习者” OR模拟学生 OR 模拟学生 OR “模拟学生” OR “模拟学习者”

初次检索共获得 973 条记录。

A.1 筛选与选择标准

我们删除了重复记录，并对检索到的论文进行了人工筛选，以确保其符合本综述的具体范围。论文只有满足以下标准才会被纳入：

1. **基于大语言模型的架构：** 该系统使用大语言模型作为生成学习者行为的主要机制，从而有效排除了大语言模型出现之前的基于规则或人工设计的模拟系统。

2. **生成性：** 我们严格区分了学生建模与学生模拟。根据我们的研究范围，我们纳入了采用生成性视角的系统（生成类似学习者的数据或行为），并排除了纯粹的学生建模系统；这类系统仅关注推理、预测或分析，不包含模拟组件。

3. **明确的模拟重点：** 我们排除了通用聊天机器人，以及仅偶然模仿学习者行为的浅层实现。我们优先考虑明确旨在推进学生模拟的方法、架构或保真度的研究，并剔除了仅将通用角色设定用作工具、缺乏技术创新的研究。

为确保全面覆盖，我们进行了向前和向后的滚雪球式追溯，人工追踪入选论文的引文和参考文献，以找出在关键词检索中可能遗漏的高影响力研究。

A.2 最终文献集

经过筛选和滚雪球式追溯，最终文献集包含 73 篇论文。根据主要研究目标，我们将这些研究分为四个不同类别：数据生成（n =23）、社会学习（$n = 19$）、教师培训（$n = 17$）和内容评估（$n = 14$）。

B 模拟学生系统案例研究

为说明所提出框架的实用性，我们分析了近期的五种模拟学生系统：**GPTeach**、**MATHVC**、**HypoCompass**、**Agent4Edu**和**GenerativeStudents**。我们根据各系统的维度特征描述和认知状态规范（ESS）对其进行分类，重点说明具体的设计选择如何

缓解“能力悖论”——本文将其定义为大语言模型默认采用专家表现的倾向——以及这些选择为何需要采用不同的评估策略。

B.1 GPTeach：使用基于 GPT 的学生进行交互式助教培训

GPTeach（Markel 等人，2023）充当教学法的飞行模拟器，使计算机科学专业的新手助教（TA）能够在低风险的环境中练习高风险的答疑时间互动。 **维度特征描述：**

- **行为目标：** 模拟人类特征。其主要价值在于复现教学中的情感与社交动态（例如，应对挫折感、冒名顶替综合征），而不是对精确的长期认知衰退进行建模。

- **环境：** 情境： 计算机科学入门课程的答疑时间；用户： 大学级别助教；模态： 基于文本的聊天。

认知状态规范（ESS）：E1 级别（静态有界）。 GPTeach 通过静态的“角色”模板（例如“对变量作用域感到困惑”）和特定情境的约束来定义学生状态，从而运行于**E1**级别。该智能体的无知程度受特定会话提示描述的限制，不具备能够跨会话演变的持久记忆。

实现设计： 该系统基于模块化的“教学场景”架构构建。每个场景都会初始化一份学生档案，包含三个不同层次：(1) 背景 （姓名、专业、编程经验），(2)教学状态（具体的错误观念，例如混淆递归与迭代），以及 (3) 情感状态 （例如防备、焦虑）。为弥合抽象档案与真实对话之间的差距，系统采用少样本提示，并融入真实学生与助教对话记录的片段。这种检索增强方法为大语言模型的语气提供依据，使其模仿学习困难的新手那种犹豫且常常不够准确的表达，而非教科书般流畅的文风。

评估方法： 评估优先考虑**实用性和情感真实性**，而非认知精确性。作者开展了一项混合方法研究，邀请助教使用该工具。定量指标包括自我效能评分和感知真实感评分。定性评估则聚焦于“安全空间”这一优势：这种模拟是否让助教能够尝试不同的教学策略（例如苏格拉底式提问与直接讲解），而不必担心伤害真实学生？结果表明，尽管“学生”偶尔会虚构自己的能力（过快地解出代码），但其主要价值在于社交压力模拟。

B.2 MATHVC：多角色虚拟课堂

MATHVC（Yue 等人，2024年）模拟了一个协作式学习环境：一名人类学生与三位由大语言模型模拟的同伴共同解决数学建模问题，每位同伴都展现出不同的技能水平和社会角色。

维度特征描述：

- **行为目标：** 模拟学习 和人类方面。该系统对能力倾向各异的学生如何参与协作性问题解决（CPS）并在过程中不断发展进行建模。

- **环境：** 情境： 中学数学；模态： 以符号任务为基础的多方对话。

认知状态规范（ESS）：E3级（误解结构化）。MATHVC 通过实现“符号角色模式”体现了**E3级**。该系统不只是依赖自然语言提示，而是维护一种结构化表示，记录智能体当前“掌握”或“误解”的具体变量和子任务。

实现设计： 该架构采用双图式方法。首先，任务模式定义数学问题的真实值。其次，将特定错误（例如遗漏变量、错误关系映射）注入该图式，以便为每个智能体生成受限的角色模式 。专门的对话控制器 负责协调所有交互：在生成响应之前，大语言模型必须先将其拟定的对话行为“落实”到受限图式中。如果大语言模型试图引用角色图式中不存在的变量（这是能力悖论的表现），控制器就会过滤或重新生成响应。

评估方法：该评估框架引入了两项新指标：**特征对齐** 和**程序对齐**。

前者验证智能体的错误是否可归因于其指定的图式（例如，验证错误答案是否源于程序设定智能体会误解的特定变量）。后者评估群体动态是否遵循协作解决问题的自然阶段，超越简单的“图灵测试”式真实感，进一步验证结构保真度。

B.3 HypoCompass：用于调试的可教学大语言模型智能体

HypoCompass（Ma 等人，2024年）颠倒了标准的辅导范式：人类学生作为助教，帮助一个由大语言模型模拟、编写了含错误代码的智能体，从而练习自己的调试技能。

维度特征描述：

- **行为目标：** 模拟表现（错误生成）。目标是生成高保真度的伪影（含错误的代码），作为可靠的训练材料。

- **环境：** 情境：入门编程；用户：编程新手；模态：以代码为中心的对话。

认知状态规范（ESS）：E1 级（静态有界）。系统采用**E1** 规范，其中认知边界由注入代码中的特定错误具体体现。智能体的“知识状态”实际上在注入错误的那一刻便被冻结。

实现设计： 为确保“学生”智能体表现得有帮助但能力不足，HypoCompass采用先大量生成、再筛选流程。系统首先提示大语言模型生成多个正确参考代码的错误变体，并筛选出语法可解析但逻辑错误的变体。在交互阶段，系统严格遵守**能力边界**：允许大语言模型智能体进行代码补全（“双手”），但明确禁止其进行假设生成（“大脑”），以此促使人类用户阐明错误背后的原因。

评估方法： 评估重点关注**错误保真度** 和学习成果。作者首先验证了生成错误的质量，以确保它们并非无关紧要。随后，通过一项对照研究测量了人类学生的学习增益。关键发现是，教导 HypoCompass 智能体的学生在调试后测中的表现比对照组提高了 12%，这证明该系统作为教学工具是有效的。

B.4 Agent4Edu：生成学习者作答数据

Agent4Edu（Gao 等人，2025）是一个全面的仿真框架，旨在生成合成学习者作答数据，用于训练计算机自适应测试（CAT）算法。

维度特征描述：

- **行为目标：** 模拟表现与学习。该系统旨在生成响应模式，既能反映静态能力水平，也能体现熟练度随时间的变化（学习和遗忘）。

- **环境：** 情境：智能辅导系统；模态：与算法推荐系统交互。

认知状态规范（ESS）：E2 级（课程索引）。Agent4Edu 是一种先进的 **E2**系统。它明确将学习者的认知状态与大语言模型的文本生成参数分离，并将熟练度作为动态变量进行跟踪，该变量会根据交互历史更新。

实现设计： 该架构通过三个相互耦合的模块模拟人类认知：(1) 学习者档案，利用现实世界数据（项目反应理论参数）初始化，以捕捉基线能力；(2) 记忆模块，整合艾宾浩斯遗忘曲线，确保智能体对概念的回忆会随时间推移而真实地衰减；(3) 行动模块，由大语言模型根据检索到的记忆状态生成响应。这种分离使系统能够模拟由状态变量驱动的复杂行为，例如“练习效应”和“失误”，而非随机产生幻觉。

评估方法： 评估着重考察**预测效度**和**分布保真度**。作者没有采用定性对话，而是评估合成数据能否像真实数据一样有效地训练计算机自适应测验模型。评估指标包括模拟学生与真实学生之间心理测量曲线（项目反应理论）的吻合度，以及使用合成语料库训练的下游推荐算法的表现。结果表明，Agent4Edu忠实地再现了人类学习轨迹的统计特性。

B.5 GenerativeStudents：面向问题评估的画像

生成式学生（Lu和Wang，2024年）利用大语言模型模拟人机交互领域中学生回答选择题时的多样模式，具体而言，是为了识别设计不佳的评估题目。

维度特征描述：

- **行为目标：**模拟表现（错误概念模式）。目标是复现特定的困惑模式，以便当题目存在歧义时，“困惑的”学生会合理地选择某个干扰项。

- **环境：**情境：HCI启发式评估；模态：多项选择题。

认知状态规范（ESS）：E3级（误解结构化）。该系统通过显式的知识成分（KC）三元组来定义学习者的状态，从而提供了**E3**的典型范例：已掌握、未知和困惑。

实现设计：核心创新是结构化的混淆元组（例如，confusion(KC_A, KC_B)）。该提示策略明确指示模型：学生无法区分这两个特定概念。这一约束迫使大语言模型生成的错误并非随机猜测，而是与该特定错误观念一致的结构性失败。通过系统地排列这些画像，系统生成响应分布，从而揭示哪些干扰项对哪些类型的学习者具有吸引力。

评估方法： 系统通过**题目难度相关性**进行了验证。作者将模拟学生得出的题目难度排名与真实学生的历史数据进行了比较。较高的相关性表明，模拟器能够基于正确的原因准确识别出“难题”。定性分析证实，模拟学生选择的干扰项与分配给它们的混淆元组相符，这也证明系统成功地约束了大语言模型的专家先验。

C 报告卡：认知状态规范（ESS）

为减少歧义并避免过度声称，本报告卡将 ESS 操作化为一组可证伪的要求。作者应为每个模拟学生系统填写此报告卡。只有满足从 E_1 到 E_k各级别的**所有** 标准，系统才能声称达到级别 E_k。

系统名称： _____ **声称的 ESS 级别：** [E0 ⊮E1 ⊮E2 ⊮E3 ⊮E4]

级别 E1：静态有界（边界检查）

模拟器是否具有明确界定且不会变化的认知状态?

□ **明确列出的未知内容：**列出学生不知道的具体概念、技能或操作流程（例如，“不会做需要借位的两位数减法”）。□ **实现锚点：**指出此边界是在何处强制执行的（提示文本、规则集、硬性约束、错误模板）。

E2级：课程索引（状态变量检查）

模拟器是否会随时间维护并更新明确的学习者状态?

□ **状态变量：** 指出表示学习者状态的变量（例如掌握度分数、技能向量、课程索引）。□ **更新规则：** 说明哪些事件会更新该状态（解题尝试、提示、反馈），以及如何更新（规则、方程、启发式方法）。□ **行为依赖：** 至少举例说明不同的状态值如何导致不同的可观察行为。

E3级：误解结构化（错误机制检查）

错误是否由明确的认知结构因果性地产生?

□ **误解清单：** 列举学生可能持有的错误观念、错误规则或有缺陷的策略。□ **生成机制：**指出产生错误的机制（例如产生式规则、图式、决策逻辑）。□ **跨问题一致性：** 说明同一种错误观念会在同构任务中产生相似的错误。

E4 级：经校准或学习（数据依据检查）

学习者状态和状态转移是否以真实学生数据为依据?

□ **数据来源：** 说明用于校准或学习的数据集（群体、任务领域、规模）。□ **拟合参数：**指出模型中哪些方面经过学习或校准（状态转移、错误率、学习曲线）。□ **预测测试：**报告基于过去的交互历史预测未来人类响应或状态的表现。□ **轨迹匹配：** 比较模拟与人类学习轨迹（例如斜率、方差、错误持续性）。