arXiv:2603.00994v1 [cs.HC] 1 Mar 2026

VizQStudio：使用模拟学生进行可视化素养选择题的迭代式设计

Zixin Chen, Yuhang Zeng, Sicheng Song*, Yanna Lin, Xian Xu, Huamin Qu, Mong Xia*

摘要——选择题是一种广泛使用的教育工具，尤其适用于可视化素养等需要广泛覆盖概念并支持多样化现实应用的领域。然而，设计高质量的可视化素养选择题仍然颇具挑战，因为教师必须协调多模态要素（如图表、题干和干扰项），涵盖多种可视化任务，并兼顾背景各异的学习者。现有的可视化素养评估主要依赖标准化的固定题库，对能够适应学习者能力、背景和推理策略差异的迭代式问题设计支持有限。为应对这些挑战，我们提出了*VizQStudio*，这是一种可视分析系统，利用多模态大语言模型驱动的模拟学生，支持教师迭代设计和完善可视化素养选择题。教师可以设定不同的学生画像，涵盖人口统计特征、知识水平和学习相关特质。随后，系统会将模拟学生对不同问题组成部分的推理过程和作答情况可视化，帮助教师在课堂部署前探索潜在的错误观念、难度校准和设计权衡。我们通过混合方法评估考察*VizQStudio*，包括专家访谈、案例研究、课堂部署和大规模在线研究。研究结果表明，借助*VizQStudio*设计的选择题能够取得与成熟基准题目相当的学习成果，同时在设计过程中提供更高的灵活性和可扩展性。总体而言，本研究将测评编写中的基于多模态大语言模型的学生模拟重新定位为一种设计阶段的探索性辅助工具。通过在真实教学情境中考察其价值与局限，我们总结出设计方面的启示，为未来系统如何支持以教师为中心、迭代且负责任地运用人工智能进行可视化素养及相关领域的多模态评估设计提供参考。

索引词—**学生模拟、可视化素养教育、选择题设计、多模态大语言模型智能体**

I. 引言

随着数据可视化图表通过社交媒体和工作场所日益融入日常生活，可视化素养——即准确解读可视化图表并从中提取见解的能力——已成为一项基本技能 [1]–[5]。随着这一需求不断增长，可视化素养领域的研究和教育实践迅速发展 [6]。

然而，由于可视化素养涉及的概念范围广泛、推理任务多样且学习者背景各异，如何有效地开展教学和评估仍然颇具挑战。选择题（MCQ）已成为一种广泛使用的评估形式，能够很好地满足这些需求 [7], [8]。其灵活的结构能够适应不同语境和认知要求，因此特别适用于评估可视化素养——从简单的信息检索到更高层次的比较分析 [6], [9]–[11]。

尽管取得了这些进展，现有研究仍主要聚焦于标准化可视化素养评估，这类评估使用固定的问题集来衡量熟练程度，而非支持教学 [9], [12], [13]。此类测试通常在每个知识领域仅包含一到两个问题，且旨在进行总结性基准评估，因此难以根据不同的学习者背景或教学需求调整问题。因此，教育工作者和学生缺乏足够优质、多样且有针对性的教学资源来支持有效的教学与练习，导致可视化素养教育领域缺少经过严格审核的选择题题库。

此外，设计高质量的可视化素养选择题本身就极具挑战且十分耗时 [14], [15]。即使是经验丰富的测验开发者，每天也只能编写少数几道设计完善的问题，而且由于学生反馈有限，可能会无意中引入个人偏差 [16], [17]。可视化问题的多模态特性进一步增加了设计难度，要求教师生成合适的视觉数据表示、编写连贯的题干和干扰项，并在每道题中纳入多项可视化任务。要让选择题适应不同领域学习者多样的技能组合——例如计算机科学、商业和设计领域——则又增加了一层复杂性，使教育工作者承受沉重负担，因为他们通常需要投入大量人工和专业知识 [18]–[21]。

多模态大语言模型（MLLM）的最新进展为生成选择题等教育材料带来了可喜的机遇。然而，其在可视化素养教育中的潜力仍未得到充分挖掘。目前的研究主要聚焦于基于文本的选择题 [22]–[27] 或从图示中提取问题 [28]。因此，可视化素养教育中至关重要的大多数教学考量，例如多模态对齐、多样化的推理任务以及针对学习者的调整，仍未得到充分解决。

为应对这些挑战，我们推出*VizQStudio*，这是一个可视分析系统，支持教师迭代创建和完善用于可视化素养教育的高质量多模态选择题。*VizQStudio* 融入教学工作流程，使教育工作者能够指定示例题目、定义可视化任务、调整图表特征并配置其他题目设置。该系统还集成了一个由多模态大语言模型驱动的学生模拟模块，教师可在其中定义多样化的学生画像——包括人口统计属性、知识水平和认知特征——以获得针对性反馈并进行全面的题目评估。教育工作者

*通讯作者。稿件收到日期：2021年4月19日；修订日期：2021年8月16日。

可以通过直观的面板交互式调整问题特征，通过自然语言提示完善细节，或直接修改特定元素。同时，教师会实时收到模拟学生反馈，并借助可视分析了解其背后的推理过程。这种以学生为中心的迭代方法有助于教师发现误解、调整问题以满足不同学习者的需求，并在课堂部署前提升整体质量。

通过对学生模拟和问题生成模块进行全面的定量评估，并结合说明性用例研究、深入的专家访谈、真实的课堂实验以及在线用户研究，我们展示了*VizQStudio*能够有效支持设计稳健、以学习者为中心的可视化素养教育选择题。总而言之，我们的主要贡献包括：

- 一项形成性研究，探讨授课教师在设计可视化素养选择题时的需求，重点关注他们对模拟学生反馈和迭代优化流程的期望；
- *VizQStudio*，一个可视分析系统，探索教师与多模态大语言模型如何协作，在多种可视化素养教育场景中迭代设计多模态选择题；
- 一项多方面评估，通过在真实教学情境中部署 *VizQStudio* ，总结设计启示与实践经验，并重点阐述对可视化素养教育及更广泛教学实践的关键考量。

II. 相关工作

本节回顾相关研究，包括可视化素养教育、面向教育内容创作的可视分析、AI 辅助选择题生成，以及用于学生模拟的大语言模型。

A. 可视化素养教育

可视化素养长期以来一直被视为一个重要议题，尤其是考虑到它在日常生活中的相关性日益增强 [29]–[30]。这是一种多方面的能力，涵盖阅读和解读可视化图表、从数据中提取信息、批判性推理，以及创作有效可视化图表等技能。其涵盖范围之广，使可视化素养教育成为一项广泛且充满挑战的工作 [31], [32]。

现有文献大多聚焦于总结有效视觉传播的指导原则 [33]–[37] ，并识别妨碍理解的常见陷阱 [38]–[40]。另一些研究则提出了标准化可视化素养评估，这些评估主要关注阅读与批判性思维，并提供带有总结性评分的固定题库 [9], [12], [41]。尽管这些工具对于基准测试和总结性评估颇具价值，但其静态、题目固定的性质限制了对多样化教学语境和学习者群体的适应能力 [10]。这一局限在可视化素养教育中尤为突出：相关技能已被证明是多维的，且高度依赖语境，而非统一或固定不变的 [42]。因此，授课教师往往需要获得形成性评估设计方面的支持，以便根据具体的教学目标和学习者背景来定制问题。

受这一需求的启发，*VizQStudio* 探讨如何支持教师迭代生成并完善自适应多模态选择题。该系统并非要取代现有评估实践，而是借鉴成熟的可视化素养框架和以推理为导向的评估视角，支持教师开展形成性问题设计。

B. 面向教育内容创作的可视分析

可视化分析已被广泛用于支持教育内容创作，帮助授课教师检视并反思学习者行为。既有系统通常侧重于分析学生的解题过程，从而为内容的迭代优化提供依据。例如，QLens [43] 与 StuGPTViz [44] 借助桑基图或树图对学生推理进行可视化，使授课教师能够识别常见策略与错误观念。其他方法则利用汇总的学习者回答来辅助内容创作，例如通过总结众包解答来生成面向同伴的提示 [45]。

近期研究已开始将人工智能辅助融入由教师主导的创作工作流。例如，SPROUT [46], 将交互式可视化图表与大语言模型相结合，以支持教程创作。然而，这些系统在很大程度上依赖授课教师的专业知识或历史学习者数据来指导优化。它们对探索不同学习者可能如何回应新创作内容的支持仍然有限。

这种对先前收集数据的依赖带来了一个根本性局限：对于全新的问题或教学材料，此类数据并不存在。因此，授课教师缺乏在设计阶段主动推演潜在学习者回答的工具。*VizQStudio* 通过使用模拟学生画像来可视化合理的推理路径，从而弥补这一空白，使授课教师能够预判错误观念，并在课堂使用前迭代优化问题。

*C. AI*辅助选择题生成

选择题在教育中被广泛使用 [47],，但设计高质量的选择题仍然出了名地具有挑战性 [14], [15]。即便是经验丰富的测验开发者，往往每天也只能产出少数几道精心编写的问题 [17],，而当涉及图像和图示时，创建带有多样化干扰项的同构问题就变得尤为困难。生成式人工智能的最新进展，尤其是大语言模型，为协助教师设计选择题开辟了新途径。先前的研究主要聚焦于从文本段落中自动生成信息抽取题 [22], [24], [48], [49],、改进干扰项设计，或支持不同的题目难度 [25], [26]。

然而，这些方法往往缺乏现实世界可视化素养教育所需的多样化教学覆盖，包括全面的知识覆盖面、学生多样性，以及可视化无障碍性等领域特定要求。虽然一些工具利用人机协作来改进带有教学

考量的问题设计，它们通常仅限于开放式形式 [50], [51], 诸如编程等学科 [52], 或用于学习的同伴生成题目 [53]。相比之下，我们的工作利用多模态大语言模型，专门支持面向可视化素养的自适应、多模态选择题设计，确保覆盖多样的知识点，同时兼顾不同背景与技能水平的学习者。

D. 用于学生模拟的大语言模型

近期研究探索了使用大语言模型在各种教育场景中模拟学生行为。研究者提出了多智能体系统，用于模拟课堂互动或虚拟学习环境 [54]–[56], ；另一些工作则侧重于通过模拟辅导、问题解决场景或基于对话的互动来支持教师培训 [57]–[60]。总体而言，这些研究证明了将大语言模型用作学生替代模型的可行性。然而，它们主要关注教学过程中的教学互动，而非为评估编写提供设计阶段反馈 —— 在这一过程中，由于缺乏实证学生数据，教师必须在新问题投入使用前，推断不同学习者可能会如何作答。

与此同时，研究者通过纳入认知因素（例如知识水平）和非认知特征（例如个性或思维模式），进一步丰富了学生模拟 [61]–[64]。尽管这类领域无关模型可应用于各种任务，但在可视化密集型领域的评估设计中，它们往往只能提供有限的诊断价值。可视化素养依赖于特定领域的推理瓶颈，例如对视觉编码的解读、比例比较和知觉捷径，而这些因素并未在通用模拟中得到明确体现。因此，这些方法难以深入揭示不同学习者在解答具体可视化问题时的推理过程，限制了其在迭代评估优化中的实用性。

我们的工作立足于这一交集，但旨在实现不同的设计目标。我们并非将学生模拟用于教学或教师培训，而是探索其作为测评编写设计阶段工具的作用。通过让基于大语言模型的智能体以领域特定的可视化素养因素（例如视觉处理能力、图表解读策略）为依据，并将其模拟推理可视化，*VizQStudio* 支持授课教师预判错误观念，并在部署前迭代改进选择题。在这一框架下，当无法获得真实作答时，学生模拟可通过提供探索性反馈来补充实证学习者数据。

III. 形成性研究

为了更深入地了解教师在设计可视化素养问题时面临的挑战，以及他们对涉及模拟学生的工具的需求，我们对六位领域专家（E1–E6）进行了半结构化访谈。受访者包括三位助理教授（E1–E3）、一位讲师（E4）和两位博士后研究人员（E5 和 E6）；每位受访者在本科和硕士数据可视化课程的教学或助教工作方面均至少有三年经验，并且在作业和考试设计方面经验丰富。

访谈期间，我们探讨了专家的问题设计工作流程、使用人工智能工具（如大语言模型）的过往经历、对学生模拟的期望，以及其他功能需求。下面总结了这些讨论中提炼出的关键设计要求。

R1：灵活的样例问题与需求规范。 授课教师开始设计问题时，准备程度往往各不相同。有些教师带着完整开发好的样例，有些只有课程幻灯片中的原始截图或预制的可视化图表，还有些则仅有一个宽泛的想法。因此，系统必须适应这些不同的起点，使教师既能基于现有材料构建内容，也能从头生成新内容。此外，系统还应提供灵活的配置选项，例如调整问题难度、指定情境细节（如领域场景）、选择图表类型，以及确定核心知识点，从而使授课教师能够根据具体教育目标和学生的多样化需求定制问题。

为进一步明确这些配置需求，我们对每位专家进行了后续访谈。我们展示了来自 VLAT [9] 和 CALVI [12],的选择题样例，询问哪些可调设置和学生档案属性最为有益。为减少主观偏差并加强理论依据，我们将专家反馈与既有的教育及可视化素养框架进行三角互证。这一过程形成了四类选择题特征（共 14 项），如图 1-A 所示。具体而言，我们将认知复杂度 与修订版布鲁姆分类法 [65] 对齐，以刻画从低阶任务（如数值提取）到高阶分析推理的递进。与此同时，内容相关特征（如视觉编码和图表解读）以经过验证的可视化素养框架 [9], [12], 为依据，确保问题设计反映可视化教育中的既有构念。

R2：多样化学生画像考量。参与者强调，在设计与评估问题时，使用大语言模型驱动的模拟来覆盖学生多样性的广度十分重要。他们指出，需要表征广泛的年龄、背景、知识水平和个人特质。为以理论为依据落实这些考量，我们将专家意见与学习科学及可视化认知领域的既有研究进行三角互证。由此形成了三类共 15 项学生档案特征——人口统计属性、学习特征以及可视化相关知识点——如图 1-B 所示。

具体而言，参与者强调了视觉化素养特有的因素，例如视觉加工能力和工作记忆。这一重点以认知负荷理论 [66]和图形理解模型 [67], 为依据，这些理论和模型指出，当学习者在推理数据关系的同时解读视觉编码时，这些能力是关键制约因素。通过将这些特征纳入学生画像，系统旨在更好地反映学习者在先前接触经验和认知需求方面的差异，为授课教师在设计问题时分析新手与进阶学习者之间的差异提供结构化视角。

此外，参与者指出，模拟学生应

为其答案提供推理过程，使授课教师能够识别错误观念并完善问题设计。最后，对于授课教师缺乏对学生群体详细了解的情形，他们建议提供一组默认的学生画像。这些画像有助于评估问题在无需定制输入的情况下对一般学习者的适用程度，从而促进评估设计的公平性和广泛适用性。

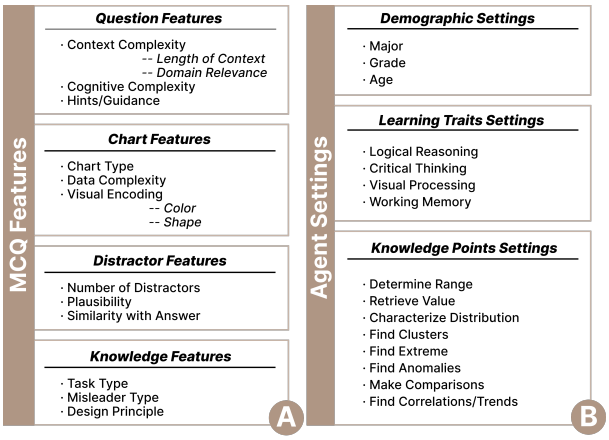


图1. 所提出的选择题特征和学生智能体画像设置概览，详细定义见补充材料。（**A**）选择题特征分为四类：问题、图表、干扰项和知识。（**B**）智能体设置涵盖人口统计属性、学习特征，以及与可视化素养相关的知识点。

R3：简要总结模拟结果，并按需提供详细信息。使用多模态大语言模型进行问题生成和学生模拟的一项关键优势，是能够生成多样化的题目设置，并预测学习者可能会如何作答。然而，生成的问题和模拟数据数量庞大，可能会让教师应接不暇。考虑到设计每道高质量选择题的复杂性，参与者建议一次聚焦一道题，并提供简明的学生反馈总结。他们还建议将相似的学生分组，呈现汇总结果，同时突出预测难度、错误率和学生感知等关键指标，帮助教师快速评估问题质量并确定改进方向。

此外，参与者认为按需获取学生推理过程的细粒度分析很有价值。对于特定答案，能够深入探究个体的思维过程，有助于揭示错误观念，并确认其与预期设计的一致性。他们建议采用分层反馈方法：先提供高层次的见解，再根据需要深入探究详细的推理过程。

R4：迭代式问题设计与完善。要创建能够有效纳入模拟学生反馈的高质量选择题，核心要求之一是支持教师持续主导对所有问题组成部分（例如题干、选项及相关可视化图表）进行完善。访谈中，参与者强调需要提供简单直接的调整方式：既可以向多模态大语言模型描述期望的修改，也可以通过简单按钮直接调整设置（例如图表类型、数据集复杂度、配色方案或问题情境）。每次更新后，系统应重新模拟学生回答，使授课教师能够比较修改前后的表现。这一反馈循环使教师能够迭代式地完善选择题，确保教学严谨性。

R5：平衡问题集中的难度。教师强调，设计时不应只关注单个问题，还应打造一套难度各异且平衡良好的问题集。多样且经过校准的问题集更能支持学习者的进步，并确保全面涵盖可视化素养技能。

IV. 问题生成与学生模拟

本节详细介绍基于多模态大语言模型的问题生成模块和学生模拟模块，它们构成了本系统迭代设计工作流的核心。

A. 教师与多模态大语言模型协作生成问题

为可视化素养设计高质量选择题，需要生成准确、多样且有时会刻意误导的图表——这是一项重大的技术挑战。我们首先概述所选用的技术栈，以及用于提升多模态大语言模型生成图表质量的检索增强生成（RAG）方法，随后介绍用于问题生成和迭代优化的后端流程。

1）可视化生成方法选择：系统应将多样化的用户输入——从粗略描述和讲义幻灯片到完整样本——转化为新的选择题（**R1**）。一个关键挑战是生成精确的可视化图表，因为当前的多模态大语言模型仍然不稳定 [49]。尽管先前工作常使用 Vega-Lite，但其高层语法缺乏处理涉及误导因素的任务所需的灵活性（例如截断坐标轴、非线性刻度）。因此，我们利用多模态大语言模型生成 D3.js 代码，以实现细粒度控制并支持复杂的可视化设计 [68]。

2）基于模板的生成与迭代优化：直接从头生成 D3.js 往往会得到不稳定的脚本 [68]。为提高可靠性，我们采用受检索增强生成（RAG）启发的方法：系统并非完全从头创建可视化图表，而是查询、编辑并变换现有模板。为此，我们整理了 80 个标准化 D3.js 模板，涵盖十种常见图表类型以及现有可视化素养测试中的全部误导因素 [9], [12], [13]。

如图 2 所示，我们开发了一个包含迭代优化流程的两阶段问题生成流程。首先，需求分析模块解析用户的文本或图像输入，提取四类选择题特征（图 1-A）。系统根据这些特征检索最相关的 D3.js 模板和 CSV 数据。随后，问题生成模块根据用户配置对其进行修改，引入变化（例如随机颜色或数据分布），以丰富生成的图表（图 1-B）。该流程能够可靠地生成满足可视化素养选择题多样化需求的图表可视化效果；第 VI 节中的定量实验对此进行了进一步评估。

此外，图 3 展示了迭代优化流程：教师可以通过修订提示和特征选择，更新选择题的任意部分（图表、代码、

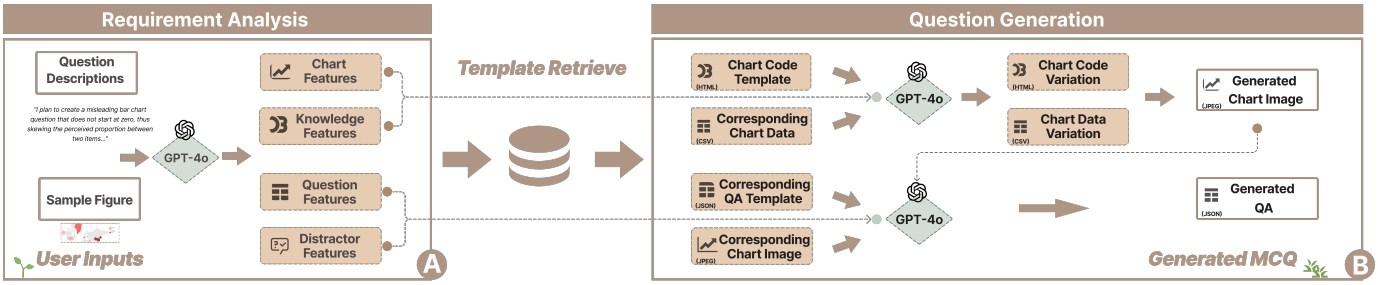


图2. 问题生成流程。**(A)** 系统使用 GPT-4o 模块分析教师的输入，例如文本描述、示例图或讲义幻灯片。随后，它依据图1中所述的特征集提取特征需求（例如，图表特征、知识点、干扰项特征）。**(B)** 这些提取出的特征用于从我们收集的外部数据库中检索最相似的选择题模板，以及相应的图表数据和代码。**(C)** 第一步，系统使用检索到的 D3.js 代码、数据和问答模板，并结合相应的图表图像和用户指定的特征需求，修改模板代码并生成定制图表。第二步，额外的多模态大语言模型模块处理生成的图表图像、模板问答以及教师对问题内容和干扰项的要求，以生成最终的问答集。这一两步流程产生的输出选择题完全符合教师的设计意图。

数据或问答）。系统利用多模态大语言模型整合这些更改并生成更新版本，使教师能够逐步提升问题质量和教学严谨性。

B. 基于多模态大语言模型的学生模拟

为了对生成的选择题提供精准反馈和可操作的修订建议，我们实现了一个基于多模态大语言模型的学生模拟模块。该模块根据需求**(R2)**设计，整合了图 1 中的可控特征，并支持三个核心流程：生成学生画像、对学生进行聚类，以及模拟学生作答。不同于依赖历史学习者作答数据的项目反应理论 (IRT) [69] 或贝叶斯知识追踪 (BKT) [70]，我们的方法基于用户建模框架 [71], [72]构建合成学习者角色。通过这种设计，无需预先收集学习者交互数据，即可为新编写的问题提供形成性设计阶段反馈。

生成与更新学生画像：一项核心要求是呈现多样化的学生群体（**R2**）。我们开发了一个画像生成模块，利用 GPT-o3-mini 动态合成学生画像，并可精细控制人口统计、认知和知识属性

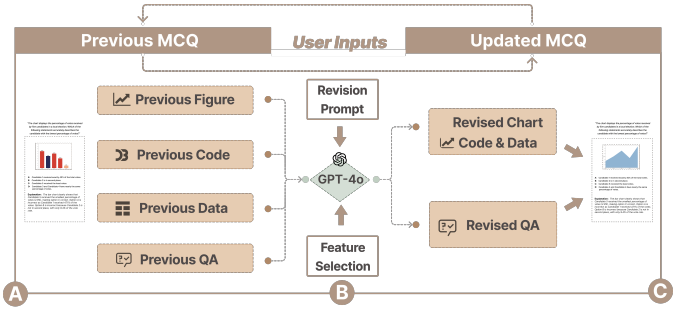


图 3. 选择题迭代更新流程概览。(A) 系统以先前的选择题（图表图像、代码、数据和问答内容）作为参考。(B) 除这些参考内容外，多模态大语言模型模块还会接收用户的修订提示，以及从**特征视图**中选取的特征。(C) 随后，多模态大语言模型会更新图表代码和 CSV 数据，渲染修订后的图表，并修改问答内容，从而生成符合用户指示的更新版选择题。

（图 1-B）。系统遵循教师设定的约束，否则采用合理的默认值。例如，默认模拟群体规模设为 20，反映了大学辅导课或讨论小组的典型规模 [73]。这一选择既能较好地展现学习者多样性，又能维持实时交互所需的计算效率。授课教师可以通过指定任意群体规模来覆盖此默认值。我们经过多轮迭代完善提示模板，以准确构建画像；完整模板见补充材料。

用于汇总分析的学生聚类：大规模分析个体回答可能会令人不堪重负，因此我们的系统使用 K 均值（默认 $K = 4$），根据三个层次的特征对学生进行聚类（图 1-B）；这是一种教育场景中常用的方法 [74], [75]。系统使用雷达图呈现结果，以总结群体特征并支持跨群体比较（**R3**）。这使授课教师能够高效评估群体层面的表现，并识别在特定题型上遇到困难的群体。

模拟学生回答：为评估选择题的有效性，我们采用基于多模态大语言模型的推理框架来模拟学生回答，并以每个智能体的人口统计属性、认知特征和领域知识为条件。这样便能探索不同学生如何回答问题，从而支持对答案分布、潜在错误观念以及后续题目优化的分析。

模拟流程会检索学生档案、目标选择题（题干和选项）以及经过 Base64 编码的图表图像，以构建结构化推理提示。系统将每位学生的认知能力（例如逻辑推理、视觉处理）映射为不同推理路径的选择概率。对于每道题，模拟学生会给出：（i）所选答案；（ii）逐步推理过程记录；以及（iii）从六个维度对问题进行评分——情境清晰度、图表复杂度、数据难度、视觉编码复杂度、总体认知挑战和提示依赖度。

需要特别指出的是，所有这些输出均由多模态大语言模型生成式推断得出。由于新颖选择题不存在任何先前的学生数据，因此我们提示模型在遵循学生档案和知识配置中所规定约束的同时，逐步进行“思维朗读”。由此产生的

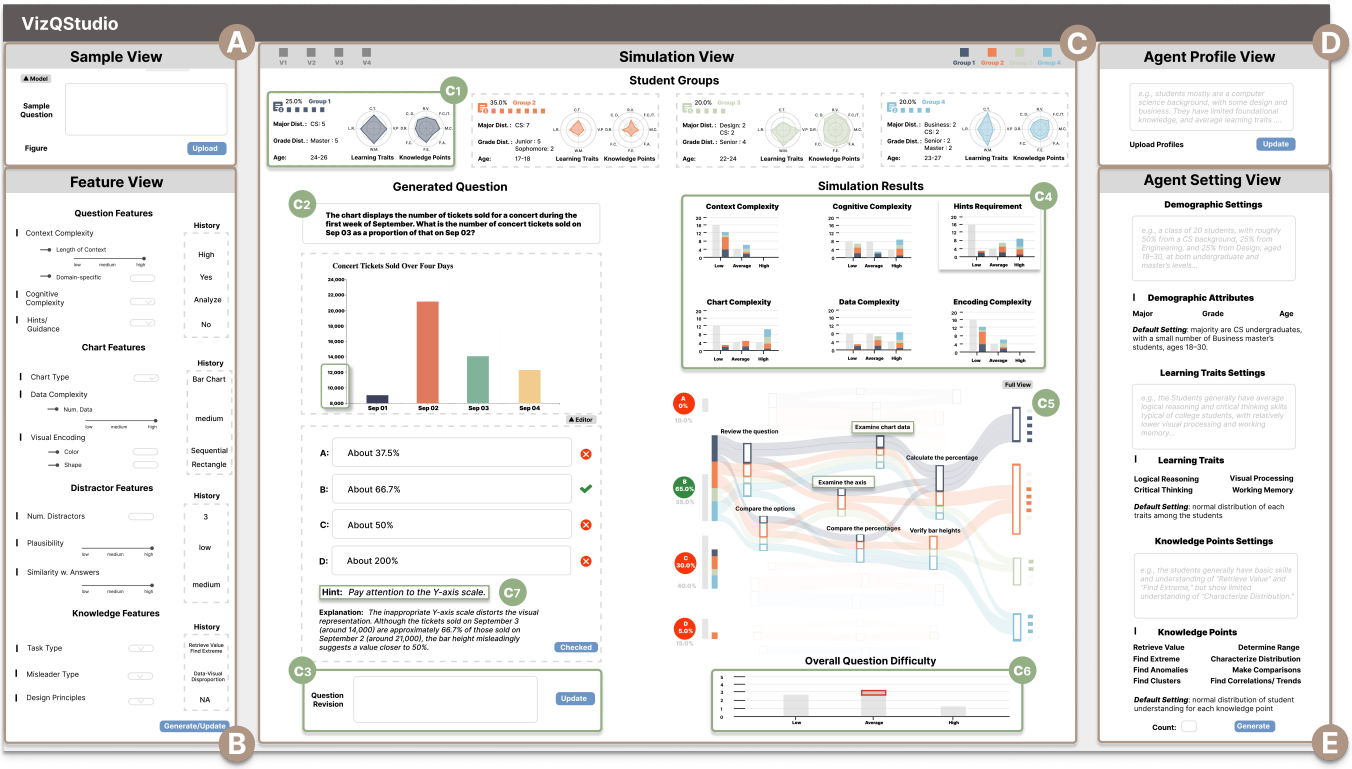


图4. *VizQStudio*的系统界面。界面由五个主要面板组成：(A) **样例视图**，用户可在此提供样例问题和/或参考图；(B)**特征视图**，显示四类特征及用户历史选择，用于控制和调整问题生成；(C) **模拟视图**，对模拟学生进行分类，并展示生成的问题和模拟结果；(D) **智能体档案视图**，用户可在此概括定义学生智能体的特征；(E) **智能体设置视图**，支持对多模态大语言模型模拟的学生的档案进行精细控制和调整。四个侧边面板 (A、B、D、E) 用作用户控制区域，而中央的 (C) **模拟视图** 则在模拟结果 区域中，将生成的问题 的各个组成部分与相应的模拟反馈对应起来。该布局使用户能够即时查看每轮修改或特征调整对学生回答的影响。

推理轨迹因此不被视为对个体学习者认知的真实解释，而是被视为合理的推理路径；我们通过在大规模在线研究（第VI节）中将其与真实学生自我报告的推理进行对齐，实证考察这些路径的教学相关性。

这些输出既支持汇总，也支持对问题质量进行细粒度诊断，使授课教师能够检查不同学习者画像可能如何推理问题，以及可能出现哪些错误观念或意料之外的捷径。为提高效率，我们通过多线程 API 调用并行运行模拟，以降低大规模运行时的延迟。系统会解析输出，提取答案、推理轨迹、论证和诊断指标，从而支持迭代式选择题优化，同时确保与多样化的学习者画像保持一致。

V. 系统设计

基于第 III 节中确定的设计要求和第 IV 节中的模块，我们开发了*VizQStudio*，这是一个可视分析系统，使教师能够借助基于多模态大语言模型的模拟学生反馈，迭代设计和完善可视化素养选择题。本节首先概述系统工作流程，然后介绍四个主要输入面板（**A**、**B**、**D**、**E**），最后展示模拟视图（**C**），其中呈现生成的问题和模拟结果。

A. 系统概述

VizQStudio 采用迭代工作流程，帮助授课教师借助基于多模态大语言模型的模拟学生反馈，设计、测试和完善可视化素养选择题。教师首先起草问题或导入示例幻灯片，然后调整问题特征（例如图表类型、知识点），并配置学生画像（例如背景、学习特征）。随后，系统模拟学生回答及其推理过程，及时揭示潜在的错误观念和问题质量。授课教师根据这些反馈完善问题并重新运行模拟，直至选择题符合教学目标并满足不同学习者的需求。

B. 用户界面

VizQStudio 的界面分为五个主要面板（图 4），分别对应问题设计流程中的一个阶段。

1) 样例视图与特征视图：教师通过模型按钮选择偏好的模型后，便可在**样例视图**（图 4-A）中起草或导入问题；他们既可以描述要点，也可以粘贴示例幻灯片和问题。这种灵活的设置使授课教师能够针对不同场景设计问题，同时确保其符合真实的课堂需求（**R1**）。

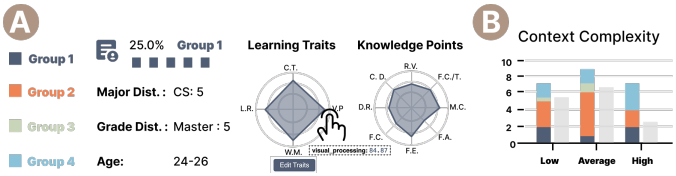


图5. (A) 每张彩色编码群组卡片显示簇大小和人口统计详情。雷达图展示每个群组的平均特质和知识水平。用户可以交互式拖动雷达图以调整群组水平。(B) 每个堆叠条形图展示每个学生群组对题干及相关特征的反馈。彩色条表示当前版本的反馈，而灰色条表示先前版本的回答，以便快速比较。

起草好问题后，教师会进入**特征视图**（图 4-B），对选择题的具体方面进行微调。系统根据此前收集的教师需求（**图 1-A**），提供四类核心特征——问题特征、图表特征、干扰项特征，以及知识特征。每类特征中的具体选项均以滑块或下拉菜单的形式呈现，便于教师轻松调整相关数值（**R4**）。每个特征控件旁边的历史记录区域会显示此前选定的数值，方便用户比较不同的问题版本。最后，点击底部的生成/更新按钮，即可重新生成问题以及相邻**模拟视图**（图 4-C）中的学生模拟结果。

2) 智能体档案与设置视图：为使用基于多模态大语言模型的模型模拟潜在学生，教师依赖于**智能体档案视图**（图4-D）和**智能体设置视图**（图4-E）。与样本与特征视图类似，授课教师既可以上传现有学生信息（例如包含人口统计属性和学习特征的电子表格），也可以先在**智能体档案视图**（图4-D）中描述学生特征的总体分布，然后在**智能体设置视图**（图4-E）中细化和更新具体的智能体设置。基于先前研究以及收集到的教师需求**(R2)**，我们将智能体配置分为三个主要特征类别（人口统计设置、学习特质设置和知识点设置），每一类都包含多个可调参数（图1-B）。通过结合这两个视图，教师可以建模多样化的学习风格和能力。由此产生的学生群组会自动分类并显示在**模拟视图**（图4-C）中，授课教师可在此进一步查看每个群组的表现和反馈。

3）模拟视图：**模拟视图**（图 4-C）是分析生成的问题和模拟结果的核心工作区。在顶部，系统会根据用户定义的智能体设置，将模拟学生聚类为带有颜色标记的学生群组（图 4-c1）。默认情况下，系统通过 k 均值聚类生成四个群组。教师可以指定不同的群组数量或聚类标准（图 4-c1）。具体而言，每个群组面板（图 5-A）都会显示重要摘要，例如簇大小、平均专业或背景信息，以及两个用于展示平均认知水平和知识点水平的雷达图。

在**模拟视图**的下半部分，系统将生成的问题（图4-c2）中的各个组成部分——包括题干、图表、答案选项和相关知识点的解释——排列在左侧，并在右侧显示相应的学生反馈和回答。这种并排布局使授课教师能够清楚地看到**特征视图**（图4-B）中的设置如何影响问题更新的各个部分，同时也便于将其与先前版本进行比较，先前版本以灰色条标示（**R4**）。

模拟学生的回答汇总在模拟结果区域中。该区域一目了然地展示了两行堆叠条形图（图 4-c4、图 5-B）：上行显示各学生群组如何理解题干，下行反映他们对图表设置的评价，为学生反馈提供简明摘要，便于快速完善问题（**R3**）。再往下，桑基图（图 4-c5、图 6-A）能深入呈现学生的选项选择和思维过程（**R3**）。授课教师可以点击某个选项（图 6-a1），仅突出显示选择该选项的学生，从而减少杂乱，并揭示共通的错误观念或设计缺陷。例如，32% 的学生选择了错误选项 B，主要来自深蓝色和橙色学生群组（图 6-a1）。此外，桑基图式可视化还会追踪每位学生的推理过程，并将采用相同逐步策略的学生归入统一区块；水平间距则表示根据各智能体多模态大语言模型的推理令牌长度推断出的平均“思考时间”（图 6-a2）。例如，橙色群组中有三名选择了选项 B 的学生，他们首先都“比较各个选项”。其中两人接着“比较坐标轴”，第三人则直接转而“比较百分比”，并花了更长时间推理这一步骤。为解决步骤长度和视觉杂乱方面的潜在问题，后端算法在推理分析过程中加入了对步骤长度的额外控制。同时，用户可以点击桑基图上方的“完整视图”按钮，在弹出窗口中查看放大后的版本。更新问题特征后，教师可以直接编辑图 4-c2 或在图 4-c3 中输入自然语言指令，然后重新运行模拟，观察这些修改如何影响学生回答。底部的摘要图表（图 4-c6）会比较当前版本与之前各轮迭代的正确率，帮助教师快速评估每次完善对问题复杂度和清晰度的影响（**R5**）。通过将左侧的问题内容与右侧的学生反馈和推理过程对应起来，**模拟视图**可以帮助教师轻松发现含糊的措辞、令人困惑的视觉元素或难度过高的知识点。教师还可以直接修改问题，或在图 4-(c2 和 c3) 中与多模态大语言模型交互。如有需要，他们还可以展开图 7-(A–C) 中的图表编辑器面板，以编辑原始图表代码或 CSV 数据。此外，教师可以调整图 4-B 中的特征，或更新图 4-(D 和 E) 中的学生画像。完成任何更新后，他们都可以重新运行模拟，直至选择题符合教学目标并满足不同学习者群体的需求。

设计替代方案。在设计过程中，我们与专家协商，探索了不同的可视化方法。例如，一个想法是采用节点链接布局，展示各组的推理轨迹（图 6-B）。

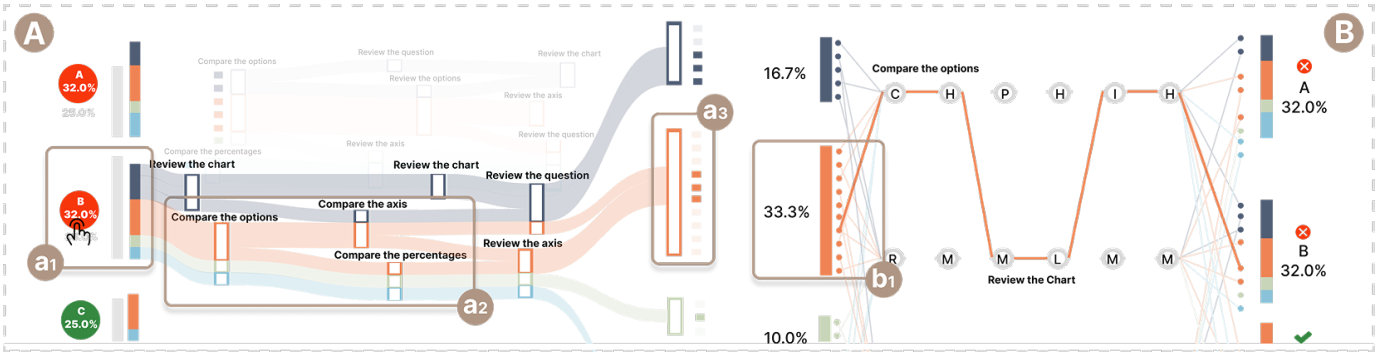


图 6。（A）用于可视化学生回答和推理过程的桑基式布局。带有选项名称和编号标签的彩色圆点表示选择某个特定（通常是错误）选项的学生比例。教师点击某个答案选项（a1）后，可以突出显示相关学生子集并隐藏其他学生，以减少视觉杂乱，从而更深入地了解学习者为何选择某些选项，以及如何改进问题。堆叠条形图（a1）展示了各学生群组的构成；桑基图则将采用相同策略的学生合并为一个区块，并通过水平间距编码平均“思考时间”，该时间以多模态大语言模型的推理令牌长度（a2）近似表示。每条策略路径最终都会重新连接到学生原属的群组（a3）。（B）另一种节点链接设计，用于可视化不同群组的推理轨迹。专家放弃了这种方案，因为它不便于比较不同学生的问题解决过程。

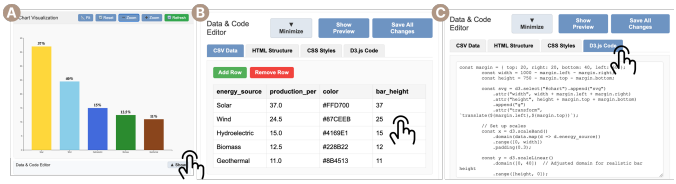


图 7. 用户可以展开 (A) 中的图表编辑器面板，查看原始 CSV 数据 (B) 和 D3.js/HTML 代码 (C)，并直接进行编辑。

垂直对齐的节点表示学生在每一步所做的行动选项，悬停在某个圆点上（图 6-b1）会突出显示对应的问题解决路径。尽管这种布局能更直接地揭示个体推理，但专家最终更倾向于类似桑基图的设计，因为它强调汇总后的学生表现，并便于比较各组的问题解决过程。

VI. 评估

我们通过以下方式评估*VizQStudio* ：(1) 对模拟与问题生成模块的定量评估；(2) 展示系统功能的案例研究；(3) 面向学生设计高质量可视化素养选择题的真实课堂实验；(4) 衡量学习增益与教学效果的大规模在线研究；以及 (5) 对六位领域专家的访谈；

A. 评估学生模拟与问题生成

本评估的首要目标是考察模拟学生是否表现出与其预期学习者画像一致的推理行为。正如关于合成用户的既有文献所强调的，评估此类智能体需要考察其对行为特征和推理过程的遵循程度，而不能仅看任务成功与否 [76]。因此，通用大语言模型基准测试（如 MMLU）侧重于整体智能水平 [77]，，并不适合评估学生模拟中的角色忠实度。因此，我们的评估聚焦于两个目标：（1）使用基于对齐的指标衡量角色忠实度；（2）在交互式设计约束下为模型选择提供依据。

基于对齐的模拟忠实度评估。遵循基于角色的大语言模型评估框架 [78]，，我们通过三个互补的对齐维度对角色忠实度进行了操作化：**认知对齐**，用于刻画所生成的推理是否反映预定义的认知优势与劣势；**推理步骤对齐**，用于评估推理是否遵循学习者画像所隐含的预期问题解决策略（例如，视觉优先与逻辑优先的方法）；以及**语义对齐**，使用语义文本相似度（语义文本相似度） [79]衡量学习者画像描述与所生成推理过程记录之间的整体主题连贯性。这些对齐指标的形式化定义与实现细节见补充材料。

所有对齐得分均归一化至 $[0, 1]$，并通过加权和（认知占 40%、推理步骤占 40%、语义占 20%）合并为总体一致性得分。该权重有意优先考虑过程层面的保真度，而非表层语义相似度。这一权重分配体现了我们的设计目标：模拟学生如何完成可视化任务中的推理，而不是模拟他们如何流畅地描述自己。

既有研究并未为解读对齐得分设定绝对阈值；相反，这类指标用于相对衡量 角色遵循度，以便比较和选择模型 [78]。在此语境下，总体一致性得分约为 0.6，表明模型在多次生成中能够始终纳入档案约束，而非生成与角色无关或高度不稳定的推理轨迹。因此，我们认为这一对齐水平足以支持对常见推理模式和潜在错误观念进行设计阶段探索，但不足以证明该模拟能够准确预测个体学习者行为。

确认模型选择。为了在所提出的基于对齐的评估下验证我们的设计选择，我们使用相同的聚类学习者档案和可视化素养问题（20名模拟学生× 50 轮）将GPT-4o与o1进行了比较。GPT-4o取得了更高的总体对齐



图 8。（A）经专家 E1 和 E5 调整后的学生模拟与聚类结果。两个聚类主要由能力水平各异的计算机科学专业学生组成，另外两个聚类则由设计或商科学生组成，每个群体分别专攻视觉处理或知识点。（B）调整前后针对生成的问题的模拟学生反馈。上半部分中，几乎所有学生都从六个角度将原始问题（C 上图）评为简单。相比之下，专家修订图表类型并增加更多干扰项后（下半部分），更多模拟学生认为更新后的问题（C 下图）更加复杂。（C）专家生成的问题的原始版本（上）和修订版本（下）。

一致性高于o1（0.6611 vs. 0.6237），同时每次响应所需的延迟也大幅更低（35.05秒 vs. 146.43秒）。鉴于我们的目标是在交互约束下实现合理稳定的人设遵循，这一结果确认了我们将GPT-4o用作默认模拟模型。

问题生成与修订的可靠性。我们还考察了问题生成工作流的运行可靠性。在 50 个随机生成与修订场景中，初始问题生成阶段每个实例平均耗时 12.67 秒，人工成功率为 76%；修订阶段平均耗时 9.82 秒，成功率为 82%。这里的成功是指大语言模型的输出是否正确遵循了指定的生成或修订指令。这些结果表明，该系统能够在交互式创作语境中支持及时生成和迭代修订，为授课教师迭代开发和完善问题设计提供了切实可行的基础。

B. 案例研究

我们邀请了形成性研究中的六位专家（E1–E6）评估 *VizQStudio*。在简要介绍系统的背景和工作流程后，专家使用*VizQStudio*迭代设计可视化素养选择题。根据这些环节，我们归纳出两个案例研究，展示系统在不同情境和学生背景下的实用性。

1) 基于初稿的迭代式问题设计：专家 E1 和 E5 使用 *VizQStudio* 从一个粗略想法出发创建选择题，通过反馈和模拟反复完善每道题，直到达到他们的标准。

学生画像设置与调整。专家首先在**智能体档案视图**（图 4-D）中描述了他们的班级：“我的学生大多具有计算机科学背景，也有一些设计专业和商科学生。他们刚开始接触数据可视化，基础知识有限，学习特征处于平均水平。”初步模拟显示出四个聚类，其中大多是计算机科学专业学生，他们的学习特征和可视化知识各不相同。为了更好地反映班级的人口统计特征，专家调整了智能体设置视图中的**人口统计设置**（图 4-E），提高了设计专业和商科学生的比例，并减少了计算机科学专业学生的人数。调整后，形成了四个均衡的聚类：其中两个主要由能力各异的计算机科学专业学生组成；一个主要由视觉处理能力较强、且掌握某些知识点的设计专业学生组成；另一个则主要由批判性思维能力较强的商科学生组成（图 8-A）。

粗略的问题描述与迭代调整。设置好学生画像后，专家转到**样例视图**（图 4-A），概述他们对问题的要求。他们首先给出一个宽泛的描述：“我需要一道条形图题，用于提取并比较不同组别的数值。题目不应过于直白，并应包含有趣的情境背景。”他们基本保留**特征视图**（图 4-B）的默认设置，只修改了图表类型。

在完成初始问题生成和学生作答模拟后，专家审阅了生成结果（图 8-B、C 上半部分）。他们确认了问题的正确性和整体质量；模拟学生反馈（图 8-b1）表明，大多数方面（如图表解读）都很简单，而语境的复杂程度适中——这与专家的预期相符。然而，初始正确率高达 95%（图 8-c1），说明难度不够。只有第2组的一名大三学生作答错误，其各项特质明显偏低（图 8-a1）。为提高挑战性，专家重新查看了**特征视图**（图 4-B），同时增加干扰项数量和似真性，并将图表类型改为折线图，以增加复杂度。重新生成问题并进行模拟后，更新后的问题包含四个似真的干扰项，正确率降至更理想的 75%（图 8-c2）。新加入的干扰项采用了视觉上相似的折线图，根据更新后的学生反馈，作答者需要更加仔细地检查（图 8-b2）。专家对新的难度水平感到满意，遂将问题标记为“已检查”，以供课堂使用。

2）为教育目的设计误导性选择题：作为本科数据可视化课程授课教师的专家 E2 和 E3，使用*VizQStudio* 基于讲课材料反复设计问题，并根据学生的具体需求进行调整。

幻灯片上传与需求说明。为了突出误导因素不恰当的刻度范围——即非零起点的 y 轴会扭曲条形图中比例的感知——他们上传了相关幻灯片截图（见补充材料），并提供了明确的说明："我计划设计一道误导性条形图问题，让坐标轴不从零开始，从而扭曲对两个项目之间比例的感知。"确认模拟学生档案聚类与其班级构成相符后，专家便开始模拟学生对初稿的作答。

识别常见错误推理。生成的问题（图 4-c2）符合专家的预期，没有明显错误。不过，模拟学生既不觉得问题过于复杂，也不觉得它特别刁钻，但总体模拟正确率仅为 35%（图 9）。为了解原因，专家查看了详细的推理轨迹（图 9-A）。他们发现，答错的学生通常在理解问题或理解选项之后，便直接核查条形高度，忽略了坐标轴刻度。相比之下，答对的学生则会检查图表数据，或明确检查图表坐标轴，呈现出截然不同的推理策略（图 9-B）。尽管只有少数学生表示需要提示（图 9-C），专家还是决定加入一个含蓄的提示，引导学生注意坐标轴刻度，以强化这一关键的可视化素养概念，而不是单纯地给学生设陷阱。

问题修订与重新模拟。回到问题修订面板（Fig.4-c3）后，专家添加了一条文字指示："在题干中添加一个不易察觉的提示。"应用这一微小改动重新生成问题后（图 4-c7），重新模拟显示情况有了明显改善：现在，超过 65% 的学生会先检查坐标轴刻度，再比较柱形高度，从而正确识别出误导性元素

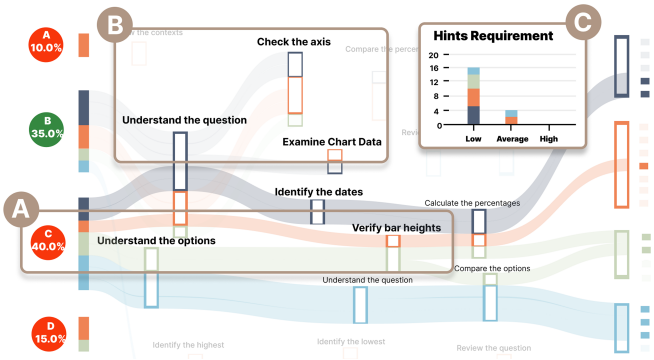


图 9. 第 6.2.2 节中误导性条形图问题第一版的模拟结果，模拟学生的答对率为 35%。(A) 落入"陷阱"并选择选项 C 的学生，在理解问题和选项后，往往会直接比较每个条形的高度。(B) 答对的学生通常会检查图表坐标轴或数据。(C) 大多数模拟学生忽略了截断坐标轴，认为问题很简单；他们通常认为提示要求较低。

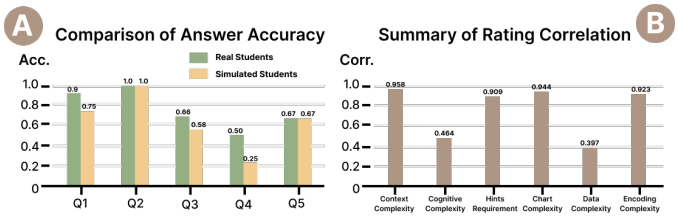


图 10. (A) 真实学生与模拟学生代理（由专家 E3 设计）在各问题上的答题准确率比较。有关问题的详情，请参阅补充材料。(B) 汇总皮尔逊相关系数，以展示模拟智能体与真实学生在所有问题上的五项反馈指标评分之间的一致性。

（图4- c5）。值得注意的是，模拟学生反馈（图4-c4）明确提到了识别率的提高，以及这一提示的实用性。原始问题和修订后的问题均已保存，这进一步体现了引导提示在这道颇具挑战性的问题中的价值。

C. 真实课堂实验

专家访谈结束后，我们开展了一项真实课堂实验（经机构审查委员会批准），从学生的角度评估使用*VizQStudio*设计的选择题质量，并考察其模拟学生表现与真实学生结果的吻合程度。我们与 E3 合作，为一场关于数据叙事的有效可视化设计的讲座设计了一份课堂测验。

测验设计与设置。 E3旨在设计选择题，评估讲义中的四个关键知识点：(1) 热力图的有效配色方案；(2) 用于可视化变量关系的散点图；(3) 减少折线图中的视觉杂乱；(4) 当条形图的y轴不从零开始时可能产生的误读。最初，E3纳入了两道此前设计的误导性条形图问题——一道包含明确提示，另一道不包含提示——以探究提示是否能有效引导学生识别误导性元素。随后，E3使用*VizQStudio* 上传相关讲义幻灯片和初步问题描述，并根据模拟学生反馈反复完善问题。这个过程兼顾了整体难度，最终形成一份包含五道问题的测验。学生在下一堂课上完成了这份测验，参与完全自愿、匿名，并经机构审查委员会批准。共有12名学生同意分享数据用于研究。他们根据与模拟指标一致的六个维度（第IV-B节：模拟学生回答）对每道问题进行评分，并对测验的问题设计提出意见，为评估问题质量以及模拟智能体在多大程度上反映了真实学生的体验提供了参考。

学生反馈与智能体比较。总体而言，学生反馈非常积极。许多人指出，题目与课程材料高度契合，例如："这些题目与课程材料相符。我能看出图表设计中运用了某些启发式和偏差，以操纵和误导数据。" 其他人则称赞其视觉清晰度，称其"赏心悦目"，并形容图表"有趣且展示精美"。我们比较了真实学生与 E3 设计的模拟智能体在每道题平均正确率上的表现，

并计算了真实学生评分与模拟学生反馈之间的皮尔逊相关系数（图 10）。结果表明，模拟智能体通常能够准确预测学生表现，但往往会略微低估真实学生的成功率（图 10-A）。至于智能体对学生评分的建模，其表现因指标而异（图 10-B）：智能体在捕捉语境和视觉理解方面更为准确，而在建模数据复杂性或整体认知挑战方面则相对较弱。

总之，我们的实验证明，模拟学生模块能够有效、可靠地捕捉真实的学生感知，并为选择题优化提供有价值的指导。

D. 关于学习成果与模拟有效性的大规模在线研究

我们开展了一项大规模在线研究，以考察*VizQStudio*的教育影响，以及其学生模拟在支持教学设计方面的有效性。该研究探讨了两个研究问题：（1）在 *VizQStudio* 的支持下设计的选择题能否作为有效的教学材料，并在可视化素养方面带来可衡量的学习成果；（2）问题设计过程中使用的大语言模型模拟推理策略，能否可信地近似常见的学习者思维模式。

研究设计概述。本研究采用两个阶段的设计，模拟真实的教学工作流程：学习者画像、问题设计、提供反馈的教学，以及教学后的评估。在所有阶段中，我们通过聚焦于一组固定的五种可视化误导因素来控制学习目标，这些误导因素构成了核心知识构念。

第一阶段：学习者画像、基线评估与分组。在第一阶段，我们通过 Prolific 招募了 100 名参与者，其中 99 人提供了有效回答。参与者报告了背景信息，并完成了自我报告项目，以评估其认知和可视化相关能力；这些项目与*VizQStudio*学生模拟中使用的学习者属性相一致。参与者还完成了一项包含五道题的可视化素养前测（分数范围：0–5）；题目由教师从 VLAT 中选取，并与教程中针对的可视化误导因素相一致。

根据前测表现和人口统计特征，参与者被分为两组，两组的基线分数相当，人口统计分布也相似（第1组：$N = 50$，均值 $= 2.10$，标准差 $= 1.15$；第2组：$N = 49$，均值 $= 2.12$，标准差 $= 1.17$）。为减少记忆效应，前测题目使用了与后续阶段不同的图表类型来呈现目标误导因素。

教师指导的问题设计（使用 *VizQStudio*）。曾教授过可视化素养课程的教师受邀使用*VizQStudio*设计教学选择题。他们获得了第1组的学习者画像，并设计了十道选择题，针对五种误导因素（每种误导因素两道题）。在此过程中，*VizQStudio*为每道题生成了全部50名学习者的模拟推理轨迹，记录了教师的编辑日志，并对模拟推理进行聚类，以提取每道题中出现频率最高的五种推理模式。

第二阶段：概念教学、基于选择题的学习和推理对齐任务。第二阶段在第一阶段三天后进行，参与者有一周时间完成所有任务。首先，所有参与者都学习了一份教程，了解与五种误导因素相关的可视化素养概念，以确保他们具备共同的教学基础。

随后，参与者完成了一项包含十道选择题的学习活动，这些题目作为教学材料，而非仅仅用于评估。每回答一道题后，参与者都会看到正确答案及其解释，以帮助理解概念。

第1组的参与者完成了使用 *VizQStudio* 设计的十道选择题，第2组的参与者则完成了十道涵盖相同知识点的 VLAT 选择题。前测、教学选择题和后测中的误导因素保持一致，同时改变图表类型和视觉编码，以促进概念迁移。对于第1组的参与者，每回答一道选择题后，还需完成一项额外的推理匹配任务。具体而言，参与者会看到针对该问题、从大语言模型模拟的学生推理轨迹聚类结果中提取的五种最常见推理策略，并被要求选出与自身思考过程最相符的一种。如果这些策略都无法充分反映其推理过程，参与者可以选择“其他”选项，并用自由文本描述自己的推理。完成学习活动后，所有参与者都完成了包含五道题的后测；这些 VLAT 题目经过挑选，与前测评估的知识点相对应。

在完成第一阶段的 99 名参与者中，80 人在取得有效同意并完成教程后进入第二阶段，最终第1组有 39 份有效回答（11 份未完成的提交），第2组有 41 份有效回答（8 份未完成的提交）。

1）学习成果与学习增益：在两种实验条件下，参与者接受教学后均表现出可视化素养的提升。两种实验条件下的后测表现相当（第1组：均值 $= 2.95$，标准差 $= 1.30$，$N = 39$；第2组：均值 $= 2.88$，标准差 $= 1.31$，$N = 41$）（图 11-A），这表明学习者完成任一组教学选择题后，都能展现出相近水平的可视化素养。

基于匹配的前后测数据（$N = 79$），两组相较于各自的基线表现均取得了具有统计显著性的学习增益。第1组的平均增益为 0.74（标准差 $= 1.67$，$N = 38$；配对t检验，$p = 0.010$），而第2组的平均增益为 0.63（标准差 $= 1.37$，$N = 41$；配对t检验，$p = 0.005$）。所有参与者的平均学习增益为 0.68（标准差 $= 1.52$），95% 置信区间为 $[0.34, 1.02]$，表明参与教学活动后，可视化素养整体得到稳定提升。

为比较不同教学条件，同时控制基线差异，我们以后测分数为因变量、前测分数为协变量进行了方差分析。分析发现，实验条件没有显著影响（组别系数 $= -0.04$，$p = 0.887$）。本分析并非旨在超越 VLAT 等标准化工具，而是评估使用*VizQStudio*编写的选择题是否

能在灵活得多的设计条件下取得相当的学习成果。未发现显著差异，这表明，尽管*VizQStudio*依赖教师驱动的生成与模拟，而非固定的精选题库，但其设计的问题并不会损害学习效果。这一结果凸显了*VizQStudio*在设计阶段的潜在作用：在因课程不匹配、覆盖范围有限或需要频繁迭代而不适合采用标准化测试的教学情境中，支持可扩展、低成本且灵活适应的题目编写。

2）大语言模型模拟的推理策略的有效性： *VizQStudio* 的一个核心目标是通过模拟合理的学生推理模式来支持教学设计。因此，我们的评估聚焦于大语言模型模拟的推理策略能否充分代表问题解决过程中观察到的常见学生思维模式。

学生推理模式的覆盖率。在第1组的39名参与者与10个问题中（即390个推理对齐实例），94.62%的回答选择了通过对模拟推理轨迹进行聚类得到的五种代表性推理策略之一。在问题层面，所有题目的覆盖率均超过92%，表明大多数参与者能够识别出与自身思维过程密切匹配的模拟推理策略。

与模拟推理策略的感知对齐度。 除选择匹配的策略外，参与者还使用五点李克特量表评价所提供策略与自身推理过程的契合程度。在第1组的全部问题与参与者中，参与者报告模拟推理策略与自身思维之间总体具有较高的感知对齐度（均值 $= 3.62$，标准差 $= 1.14$，$N = 39$），表明大语言模型模拟的策略常常被认为捕捉到了学生推理过程的关键方面。该量化评估通过刻画模拟推理与学生自我报告的推理之间的感知相似程度，对覆盖率结果形成了补充。

不对齐案例分析。在参与者选择“其他”选项或报告对齐度较低的情况下，开放式回答显示，不对齐通常源于启发式估算（例如，仅凭视觉进行大致比较，而没有明确的逐步推理）、个人视觉可及性问题，或个性化的推理捷径。例如，一位参与者表示，他们“粗略地目测比较了条形的长度，然后选择最接近的选项”，并未明确遵循模拟策略中描述的多步比较。这些回答表明，不匹配往往反映的是推理粒度或表达方式的差异，而非截然不同的问题解决方法。

模拟与学习成果的契合度与学习成果之间的关系。为考察与模拟推理策略的契合度是否与学习成果相关，我们在第1组中开展了探索性分析。参与者的平均契合度评分以及选择“其他”的频率，均未显示与后测表现存在显著的皮尔逊相关（$|r| < 0.25$，$p > 0.20$） [80]。这一结果表明，与模拟推理策略的契合度与个体学习成果并无密切关联，而模拟的主要作用在于支持教学设计，而非预测学习者的表现。

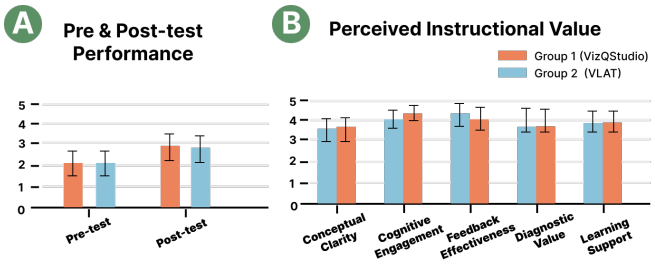


图11. (A) 两种教学条件下参与者的前测与后测表现。(B) 参与者对*VizQStudio*设计的问题（第1组）和基准测试 VLAT 问题（第2组）的多项教学价值维度的感知。

3）感知教学价值：教程结束后，参与者对作为学习材料的选择题给予了积极评价（图11-B）。对于*VizQStudio*设计的选择题（第1组），参与者认为问题清晰（概念清晰度：均值 $= 3.54$，标准差 $= 1.19$），且具有认知参与度（认知挑战／参与度：均值 $= 3.97$，标准差 $= 0.99$），这表明使用*VizQStudio* 制作的教师编写的问题易于理解，同时仍能促使参与者进行主动推理。参与者还表示，所提供的答案与解释有效地帮助他们理解内容（反馈有效性：均值 $= 4.15$，标准差 $= 1.04$）。

除了表层的清晰度和参与度外，参与者还认为由*VizQStudio*设计的选择题在诊断和支持学习方面具有教学意义。具体而言，参与者表示，这些问题有助于揭示误解（诊断价值：均值 $= 3.64$，标准差 $=$1。20），并有助于为回答类似的可视化素养问题做好准备（学习支持／准备度：均值 $= 3.79$，标准差 $= 1.08$）。这些看法表明，这些选择题不仅是练习题，也发挥了反思性学习提示的作用，促使学习者审视自己的推理过程。

为了解这些看法的背景，我们将第1组与 VLAT 基准测试实验条件（第2组）进行了比较。后者在相同的教学价值维度上总体评分相近（图 10-B）（概念清晰度：均值 $= 3.59$，标准差$= 1.26$；认知参与度：均值 $= 4.17$，标准差 $= 0.77$；诊断价值：均值 $= 3.68$，标准差 $= 1.11$；反馈有效性：均值 $= 4.05$，标准差 $= 0.97$；学习支持：均值$= 3.83$，标准差 $= 1.14$）。综合来看，这些结果表明，使用*VizQStudio*的教师能够设计出学习者认为有效的学习材料，其总体教学价值与教程式学习情境中的 VLAT 问题相当。

4）教师修订模式与人在环路中的安全护栏：在教师指导的问题设计过程中，*VizQStudio*记录了详细的修订日志，记录授课教师在部署前如何反复完善每道选择题。针对第二阶段设计的十道教学选择题，授课教师对每道问题都进行了多次修订，平均每道选择题修订 8.3 次

，平均每道问题的编写时间为 9.6 分钟（标准差 $= 3.1$）。

根据修订日志和教师反馈，约一半的修订侧重于问题的结构方面，例如修改图表数据或视觉编码，以确保正确性和清晰度；其余修订则是通过检查模拟的学生推理，并相应调整难度、干扰项或解释来完成的。值得注意的是，尽管与图表相关的修订占修订事件的相当大一部分，但所耗费的总创作时间占比却较小（约三分之一），这得益于灵活的编辑功能，例如直接操作图表，以及通过自然语言与大语言模型交互。

总的来说，这些结果表明，*VizQStudio* 支持一种由专家主导的迭代设计流程，模型生成的草稿可通过基于模拟的探索得到高效修正和教学层面的完善。第七节将讨论更广泛的影响和局限性。

E. 专家访谈

在案例研究之后，我们对六位专家（E5–E10）进行了一对一访谈，其中两位参与过案例研究，另外四位是新邀请的助理教授，来自两所大学，研究方向为可视化。每次访谈持续 60–80 分钟，并提供 70 美元酬金。对于新邀请的专家，访谈包括背景介绍、系统演示、动手试用，以及围绕易用性、工作流程和模拟可靠性的反馈。

系统工作流程。所有专家都称赞了*VizQStudio*工作流程的清晰度、灵活性和功能性。他们尤其认可系统多样化的问题生成方法，例如将粗略输入（如讲义幻灯片截图）转化为与特定知识点精准匹配的问题。专家还强调了系统直观的问题和学生智能体编辑功能、支持在创建问题时自由探索、通过清晰的并排界面无缝整合实时模拟更新，以及可靠的版本保存功能。这些特性显著简化了迭代设计流程。

问题生成和学生模拟模块。所有专家都认为问题生成和学生模拟模块高效、稳定且富有启发性。他们表示运行速度很快——问题生成约需20秒，每次由大语言模型驱动的更新约需10秒，而模拟一个典型的30–40人班级则需30秒。专家尤其赞赏生成的问题在质量和灵活性方面的表现。例如，E8表示：“生成的问题令人印象深刻。即使只有粗略的想法或一张幻灯片，系统也能准确理解我的意图，并生成带有清晰解释的图表和问答。编辑面板进一步确保了问题的可靠性，同时让我能够进行细致的控制。”

专家高度赞扬了学生模拟模块，认为它有潜力洞察不同背景学生的推理过程，尤其能帮助他们设计并测试针对学生需求的问题。正如E10所说，“我有意调整了两组学生——一组具有较高的一般学习特征但可视化知识较低，另一组则相反——以观察他们在答案和推理上的差异。结果令人印象深刻，能够帮助我为不同群体设计更有针对性的问题，预判他们可能的思考方式，并评估题集是否有效支持其预期的学习成果。”

视觉设计与交互。专家们一致认为，该系统的视觉设计与交互清晰直观。堆叠条形图尤其有助于快速了解学生对问题的总体反馈，而桑基图则有效展示了各学生群组最终答案和推理路径的分布。正如 E9 所指出的，“我的首要关注点是学生的看法和总体问题难度，而该系统能够有效地对此进行总结，同时仍允许我查看每位学生的推理过程。”丰富的交互方式也提供了极大的灵活性，使专家能够通过自然语言输入、按钮和滑块交互以及直接编辑来设计和修改问题与学生档案。

建议。专家们提出了若干改进*VizQStudio*的想法。E9 建议减少问题特征中的默认选项数量，仅在需要时提供完整选项，以简化系统。E10 和 E7 建议更加突出总体问题难度图表，因为他们通常会先查看整个问题集的难度。E8 进一步提出，应加强桑基图中的推理步骤与学习特征和知识点之间的联系，从而更深入地洞察不同学生群组。

七、讨论

本节综合了我们在设计与评估中的见解，将我们的方法置于更广泛的教育与学习者建模研究背景下，并讨论其局限性、伦理考量以及对未来工作的启示。

A. 评估设计中的大语言模型：多层级人工监督是设计的必要条件。

尽管大语言模型能够有效生成多样的问题草稿、干扰项和合理的推理路径，但在实际应用中，它们最适合定位为生成式与探索性工具。正确性、教学一致性和最终验证必须由教师掌控。

我们的研究得出的一个核心见解是，要有效利用大语言模型进行评估设计，需要教师进行多层级监督。在**较高层级**，教师受益于这样一种交互范式：它允许他们以贴近真实教学实践的方式向大语言模型传达期望的修改，例如用自然语言描述问题、请求替代表述，或指出（例如通过草图绘制、注释或高亮标记）问题或可视化中需要修改的部分。这种交互支持流畅的探索与改进，使教师无需直接接触技术表示，便能调整大语言模型生成的问题框架、干扰项或解释。

与此同时，我们的观察表明，对于高风险教育材料，仅有高层级交互并不足够。教师需要对最终学习产物进行精确、确定性的控制，以确保事实与概念的正确性；对于基于可视化的问题尤其如此，因为数据值、尺度或编码的细微变化都可能改变所评估的核心概念。因此，需要**较低层级的控制**，使教师能够直接核验和修改评估材料的具体组成部分。

此外，将大语言模型生成建立在领域专用参考资料的基础上，已成为对高层和低层监督的重要补充。纳入教师提供的示例、以往的评估或公开可用的教育资源，并结合检索增强生成等技术，有助于使生成内容立足于课程背景，并弥补通用模型知识的局限。

综上，这些见解揭示了更广泛的设计启示：在教育环境中，有效地借助大语言模型进行创作，关键不在于尽量减少人的参与，而在于让授课教师能够在表达性强的高层交互、精确的低层控制以及结合语境的内容 grounding 之间灵活切换。优先支持这种多层次监督的系统，更有利于促进生成式模型得到负责任且符合教学原则的应用。

B. 评估设计中的学生模拟：价值、局限与设计启示。

虽然我们展示了模拟能够捕捉整体推理模式，但其主要教学价值在于支持授课教师探索评估设计过程中可能出现的多种推理模式。通过明确呈现这些模式，模拟使授课教师能够审视自己对学生如何完成可视化素养任务进行推理的假设，既看到其优势，也认识到其局限。

大规模在线研究的结果表明，在总体层面上，大语言模型模拟的高频推理策略往往能够涵盖学生自我报告的思维过程。许多参与者至少找到了一个与自身推理高度吻合的模拟策略，这表明模拟能够有效呈现许多学习者共有的主导性且与教学相关的推理模式。这种总体层面的覆盖对于评估设计尤为重要，因为教师通常需要分析常见误解、典型解题路径和普遍存在的困难，而非个体的认知轨迹。

与此同时，我们的研究结果表明，在模拟真实的学生回答中可能出现的机会主义推理策略和启发式推理策略方面，模型存在盲点。模拟学生在呈现结构化或“理想化”的推理模式方面相对有效，例如逐步比较或仔细解读，但往往无法捕捉付出最少努力或采用捷径的行为。例如，一些参与者表示，他们依靠粗略的视觉估算（如“我只是大致看了看柱子的长度，然后选了最接近的答案”），而非模拟策略所体现的多步骤推理。这些案例凸显了一个模型推理与学生在实践中采用的务实、有时较为粗略的策略之间存在差距。

这些局限揭示了未来学生模拟系统设计中的重要启示。尽管我们目前的学生画像主要依据人口统计属性、认知能力水平和目标知识点构建，已足以呈现许多常见推理模式，但仍不够全面。学习科学领域的既有研究表明，学生的任务表现还受到元认知调节、动机、情感状态和参与度等因素的影响，而仅以能力或知识为核心的画像难以充分体现这些因素。将这些维度纳入未来的模拟方法，或许能够更丰富、更真实地呈现学生推理，但这需要更紧密地以学习理论和有关学习者认知的实证研究为基础。

总体而言，我们的研究结果表明，基于大语言模型的学生模拟是一种反思性设计工具，而非学习者行为的真实模型。通过呈现可能的推理路径及其失效之处，模拟有助于授课教师探索并比较评估设计可能如何与学习者多样性相互作用，同时将教学判断与验证牢牢交由人类负责。

C. 对可视化素养教育的启示

我们的研究提出了若干观察结果，可为可视化素养教育领域的广泛讨论提供参考。在我们的样本中，参与者自评其可视化相关能力时信心较高，但其基线表现表明，他们对核心概念的掌握程度较为有限。在第一阶段，参与者的自评可视化能力平均分超过五分制中的4分，而前测分数平均为5分制中的2.11分。尽管应谨慎解读这一结果，但它显示，学习者感知到的自身能力与他们在可视化素养任务中展现出的能力之间可能存在错位。

这种错位反映了可视化教育面临的一项更广泛挑战：困难往往不在于表层解读，而在于对误导性编码、比例关系和语境线索进行推理。静态评估或仅以正确性为依据的评估不太可能揭示这些差距。从这一角度看，着重考察推理过程而非仅关注答案的形成性评估，对可视化素养尤为重要。基于模拟的问题设计提供了一种方式，促使授课教师思考不同的推理策略、误解或捷径可能如何出现，并设计出更有效地揭示概念性薄弱环节的问题。

D. 局限性与伦理考量

我们的研究存在若干值得指出的局限。首先，问题生成和学生模拟都依赖现成的大语言模型，其行为受到训练数据、提示策略以及我们无法控制的持续模型更新的影响。因此，不能假定生成的问题、答案和模拟推理轨迹在不同时间和部署情境下始终准确、公正或可复现。这些特点意味着

必须将大语言模型的输出视为暂定产物，而非正确性或学习者行为的权威表征。

其次，尽管模拟能够有效揭示总体层面常见的推理模式，但它仍无法完整呈现真实的学生行为。机会主义、启发式或投入极少努力的策略未能得到持续捕捉，而依据人口统计属性、认知能力水平和知识点定义的学习者画像，必然简化了真实学习情境的丰富性。这些限制降低了对模拟推理的可解释性，并进一步表明，模拟应被定位为设计辅助工具，而非预测模型。

在本研究中，我们尝试通过去标识化、可选的学习者画像，以及让教师明确掌控模拟参数和评估内容，来缓解这些局限。尽管如此，这些措施只能提供部分保障。未来仍需制定更稳健的实践，以提升模拟支持的评估设计中的偏见审计、透明度和可复现性。

E. 普遍性与未来方向

除了本研究考察的可视化素养语境外，我们的发现还指出了模拟驱动的评估设计可进一步探索的若干方向。其中一个反复出现的见解是，将模拟推理用于题目编写之外。代表性的推理路径——无论正确与否——不应仅仅作为设计阶段的工具，还可有选择地用于为学习者提供反思性反馈，支持他们进行比较和元认知反思，而非仅仅验证答案。

授课教师还讨论了利用模拟为具有不同先前知识或学科背景的学习者调整教学材料的价值。通过探索不同学习者画像下合理推理模式的变化，授课教师可以在准备面向不同受众或不同教学语境的材料时，利用模拟来调整问题框架、解释或教学支架。同样，模拟也被认为有助于思考群体层面动态，例如背景知识的差异可能如何影响协作学习。总体而言，这些方向表明，在学习依赖复杂推理过程和教师判断的教育环境中，模拟驱动的设计具有更广泛的应用前景。

VIII. 结论

我们介绍了*VizQStudio*，这是一个可视分析系统，可支持教育工作者为可视化素养教育设计和完善选择题。通过将基于多模态大语言模型的题目生成与模拟学生推理相结合，我们使授课教师能够在设计阶段针对不同的学习者画像探索并迭代完善问题设计。通过多方法评估，我们考察了这一方法的价值与局限，并提出了设计方面的启示，说明如何结合可视分析与多模态大语言模型，以支持以教师为中心、灵活且负责任的可视化素养及相关教育场景中的评估设计。

参考文献

[1] Z. Pousman、J. Stasko 和 M. Mateas，“休闲信息可视化：日常生活中的数据描绘”，IEEE Transactions on Visualization and Computer Graphics，第 13 卷，2007 年，第 6 期，页码 1145–1152。[2] E. Segel 和 J. Heer，“叙事可视化：用数据讲故事”，IEEE Transactions on Visualization and Computer Graphics，第 16 卷，第 6 期，页码 1139–1148，2010 年。[3] J. Heer 和 D. Boyd，“Vizster：在线社交网络可视化”，载于 IEEE 信息可视化研讨会，2005 年。INFOVIS 2005。IEEE，2005 年，页码 32–39。[4] M. Ruckenstein，“被可视化与互动的生活：个人分析以及与数据分身的互动”，Societies，第 4 卷，2014 年，第 1 期，页码 68–84。[5] D. Huang、M. Tory、B. A. Aseniero、L. Bartram、S. Bateman、S. Carpendale、A. Tang 和 R. Woodbury，“个人可视化与个人可视分析”，IEEE Transactions on Visualization and Computer Graphics，第 21 卷，第 3 期，页码 420–433，2014 年。[6] S. Stokes，“教学与学习中的视觉素养：文献视角”，Electronic Journal for the Integration of Technology in Education，第 1 卷，第 1 期，页码 10–19，2002年。[7] A.-M. Brady，“使用选择题评估学习”，Nurse Education in Practice，第 5 卷，第 4 期，页码 238–242，2005 年。[8] A. C. Butler，“教育中的选择题测验：评估的最佳实践是否也有利于学习？”，Journal of Applied Research in Memory and Cognition，第 7 卷，第 3 期，页码 323–331，2018 年。[9] S. Lee、S.-H. Kim 和 B. C. Kwon，“VLAT：视觉素养评估测试的开发”，IEEE Transactions on Visualization and Computer Graphics，第 23 卷，第 1 期，页码 551–560，2016 年。[10] E. E. Firat、A. Joshi 和 R. S. Laramee，“交互式可视化素养：最新进展”，Information Visualization，第 21 卷，第 3 期，页码 285–310，2022 年。[11] B. Alper、N. H. Riche、F. Chevalier、J. Boy 和 M. Sezgin，“小学阶段的可视化素养”，载于《2017年 CHI计算机系统中的人为因素会议论文集》，2017 年，页码 5485–5497。[12] L. W. Ge、Y. Cui 和 M. Kay，“CALVI：面向可视化素养的批判性思维评估”，载于《2023年 CHI计算机系统中的人为因素会议论文集》，2023 年，页码 1–18。[13] S. Pandey 和 A. Ottley，“Mini-VLAT：一种简短而有效的可视化素养测量工具”，载于 Computer Graphics Forum，第 42 卷，第 3 期。Wiley Online Library，2023 年，页码 1–11。[14] J. Collins，“为继续医学教育活动和自我评估模块编写选择题”，Radiographics，第 26 卷，第 2 期，页码 543–551，2006 年。[15] M. H. Towns，“开发高质量、可靠且有效的选择题评估指南”，《化学教育杂志》，第 91 卷，第 9 期，页码 1426–1431，2014 年。[16] J. Considine、M. Botti 和 S. Thomas，“用于护理研究与教育的选择题的设计、格式、效度与信度”，Collegian，第 12 卷，第 1 期，页码 19–24，2005 年。[17] M.-K. Kim、R. A. Patel、J. A. Uchizono 和 L. Beck，“将布鲁姆分类法融入药物治疗学课程的选择题考试题目中”，American Journal of Pharmaceutical Education，第 76 卷，第 6 期，页码 114，2012 年。[18] J. K. Gilbert，“可视化：科学与科学教育中的一种元认知技能”，载于《科学教育中的可视化》。Springer，2005 年，页码 9–27。[19] J. Sorva、V. Karavirta 和 L. Malmi，“面向编程入门教育的通用程序可视化系统综述”，ACM Transactions on Computing Education (TOCE)，第 13 卷，第 4 期，页码 1–64，2013 年。[20] N. Presmeg，“数学学习与教学中的可视化研究”，《数学教育心理学研究手册》，页码 205–235，2006 年。[21] C. N. Knaflic，《用数据讲故事：面向商业专业人士的数据可视化指南》。John Wiley & Sons，2015 年。[22] W. Fu、B. Wei、J. Hu、Z. Cai 和 J. Liu，“QGEval：题目生成评估基准”，arXiv 预印本，页码 arXiv–2406，2024 年。[23] S. Lamsiyah、A. El Mahdaouy、A. Nourbakhsh 和 C. Schommer，“通过强化学习微调大语言模型以生成教育题目”，载于《国际人工智能教育会议》。Springer，2024 年，页码 424–438。[24] K. Li 和 Y. Zhang，“先规划，再提问：一种用于可控题目生成的大语言模型引导的方法”，载于《计算语言学协会 ACL 2024 论文集》，2024 年，页码 4715–4729。

[25] Y. Chen、C. Wu、S. Yan、P. Liu 和 Y. Xiao，“Dr. academy：评估大语言模型教育提问能力的基准测试”，载于《计算语言学协会第 62 届年会论文集（第 1 卷：长篇论文）》，2024 年，页码 3138–3167。[26] H. Luo、Y. Deng、Y. Shen、S. K. Ng 和 T.-S. Chua，“示例链：增强多模态教育题目生成中的干扰项生成”，载于《计算语言学协会第 62 届年会论文集（第 1 卷：长篇论文）》，2024 年，页码 7978–7993。[27] W. Xing、T. Zhu、J. Wang 和 B. Liu，“教育领域多模态大语言模型综述：应用与未来方向”，Future Internet，2024 年。[28] W. Ding、Y. Zhang、J. Wang、A. Jatowt 和 Z. Yang，“探索视觉区域的并集与交集，以生成问题、答案和干扰项”，载于《2024年自然语言处理实证方法会议论文集》，2024 年，页码 1479–1489。[29] K. B¨orner、A. Bueckle 和 M. Ginda，“数据可视化素养：定义、概念框架、练习与评估”，《美国国家科学院院刊》，第 116 卷，第 6 期，页码 1857–1864，2019 年。[30] M. Avgerinou 和 J. Ericson，“视觉素养概念综述”，《英国教育技术学杂志》，第 28 卷，1997 年，第 4 期，页码 280–291。[31] P. Ruchikachorn 和 K. Mueller，“通过类比学习可视化：利用可视化变形促进视觉素养”，IEEE Transactions on Visualization and Computer Graphics，第 21 卷，第 9 期，页码 1028–1044，2015 年。[32] L. Y.-H. Lo、Y. Ming 和 H. Qu，“学习可视化工具：教授数据可视化教程”，载于《2019年 IEEE 可视化会议（VIS）》。IEEE，2019 年，页码 11–15。[33] L. Yang、C. Xiong、J. K. Wong、A. Wu 和 H. Qu，“通过示例进行讲解：从众包信息可视化入门描述中汲取的经验”，IEEE Transactions on Visualization and Computer Graphics，第 29 卷，第 3 期，页码 1638–1650，2021 年。[34] G. Kindlmann 和 C. Scheidegger，“一种用于可视化设计的代数过程”，IEEE Transactions on Visualization and Computer Graphics，第 20 卷，第 12 期，页码 2181–2190，2014 年。[35] C. X. Bearfield、L. Van Weelden、A. Waytz 和 S. Franconeri，“相同数据，不同视角：可视化引发相互竞争解读的力量”，IEEE Transactions on Visualization and Computer Graphics，2024 年。[36] C. Nobre、K. Zhu、E. M¨orth、H. Pfister 和 J. Beyer，“读懂像素之间的信息：探究可视化素养障碍”，载于《2024年 CHI计算机系统中的人为因素会议论文集》，2024 年，页码 1–17。[37] Q. Zhu、L. Y.-H. Lo、M. Xia、Z. Chen 和 X. Ma，“偏差感知设计助力知情决策：提升对用户评分与评论中自我选择偏差的认识”，ACM 人机交互，第 6 卷，CSCW2 期，页码 1–31，2022 年。[38] L. Y.-H. Lo、Y. Cao、L. Yang 和 H. Qu，“为何要修改我的设计：通过可探索的解释说明构造不佳的可视化设计”，IEEE Transactions on Visualization and Computer Graphics，第 30 卷，第 1 期，页码 955–964，2023 年。[39] L. Y.-H. Lo、A. Gupta、K. Shigyo、A. Wu、E. Bertini 和 H. Qu，“受可视化误导：我们能从误导性可视化中学到什么？”，载于 Computer Graphics Forum，第 41 卷，第 3 期。Wiley Online Library，2022 年，页码 515–525。[40] J. Hong、C. Seto、A. Fan 和 R. Maciejewski，“大语言模型具备可视化素养吗？通过修改后的可视化评估其数据解读泛化能力”，IEEE Transactions on Visualization and Computer Graphics，2025 年。[41] Y. Cui、W. G. Lily、Y. Ding、F. Yang、L. Harrison 和 M. Kay，“可视化素养的自适应评估”，IEEE Transactions on Visualization and Computer Graphics，第 30 卷，第 1 期，页码 628–637，2023 年。[42] S. Beschi、D. Falessi、S. Golia 和 A. Locoro，“数据可视化素养的特征：系统性文献综述”，arXiv 预印本 arXiv:2503.14468，2025 年。[43] M. Xia、R. P. Velumani、Y. Wang、H. Qu 和 X. Ma，“QLens：用于分析多步问题解决行为、改进问题设计的可视分析”，IEEE Transactions on Visualization and Computer Graphics，第 27 卷，第 2 期，页码 870–880，2020 年。[44] Z. Chen、J. Wang、M. Xia、K. Shigyo、D. Liu、R. Zhang 和 H. Qu，“StuGPTViz：一种理解学生与 ChatGPT 互动的可视分析方法”，IEEE Transactions on Visualization and Computer Graphics，2024 年。[45] E. L. Glassman、A. Lin、C. J. Cai 和 R. C. Miller，“学习者众包个性化提示”，载于第 19 届 ACM 会议论文集，会议名称为

计算机支持的协同工作与社会计算，2016年，第1626–1636页。[46] Y. Liu、Z. Wen、L. Weng、O. Woodman、Y. Yang和W. Chen，《Sprout：一款借助大语言模型可视化来生成编程教程的交互式创作工具》，IEEE Transactions on Visualization and Computer Graphics，2024年。[47] J. L. Little 和 E. L. Bjork，《优化选择题测试，使其成为学习工具》，Memory & cognition，第43卷，第14–26页，2015年。[48] X. Yuan、T. Wang、Y.-H. Wang、E. Fine、R. Abdelghani、P. Lucas、H. Sauz´eon和P.-Y. Oudeyer，《从预训练大语言模型中筛选更好的样本：一项关于问题生成的案例研究》，arXiv预印本 arXiv:2209.11000，2022年。[49] X. Zeng、H. Lin、Y. Ye和W. Zeng，《通过可视化参考指令微调，推动多模态大语言模型在图表问答中的应用》，IEEE Transactions on Visualization and Computer Graphics，2024年。[50] X. Wang、S. T. Talluri、C. Rose和K. Koedinger，《Upgrade：收集学生开放式解答，创造可扩展的学习机会》，载于第六届（2019年）ACM Learning@Scale会议论文集，2019年，第1–10页。[51] X. Lu、S. Fan、J. Houghton、L. Wang和X. Wang，《Readingquizmaker：支持授课教师设计高质量阅读测验题的人机自然语言处理协作系统》，载于2023年CHI计算机系统中的人为因素会议论文集，2023年，第1–18页。[52] S. Sarsa、P. Denny、A. Hellas和J. Leinonen，《使用大语言模型自动生成编程练习和代码解释》，载于2022年ACM国际计算教育研究会议论文集第一卷，2022年，第27–43页。[53] I. Yeckehzaare、T. Barghi和P. Resnick，《Qmaps：通过自主问题生成与关联来提升学生参与度》，载于2020年CHI计算机系统中的人为因素会议论文集，2020年，第1–14页。[54] D. R. Thomas、S. Gupta、E. Gatz、C. Tipper和K. R. Koedinger，《想成为辅导员吗？成人辅导员的专业发展与基于情境的培训》，载于Learning Ideas会议。Springer，2023年，第537–547页。[55] Z. Zhang、D. Zhang-Li、J. Yu、L. Gong、J. Zhou、Z. Hao、J. Jiang、J. Cao、H. Liu、Z. Liu等，《借助大语言模型赋能的智能体模拟课堂教育》，arXiv预印本 arXiv:2406.19226，2024年。[56] S. Xu、H.-N. Wen、H. Pan、D. Dominguez、D. Hu和X. Zhang，《Classroom simulacra：在在线教育中构建情境化学生生成式智能体，以模拟学习行为》，arXiv预印本 arXiv:2502.02780，2025年。[57] U. Lee、S. Lee、J. Koh、Y. Jeong、H. Jung、G. Byun、Y. Lee、J. Moon、J. Lim和H. Kim，《用于教师培训的生成式智能体：为职前教师设计基于大语言模型智能体的教育问题解决模拟》，载于NeurIPS’23教育生成式人工智能研讨会（GAIED），2023年。[58] S. Pan、R. Schmucker、B. G. B. Bueno、S. A. Llanes、F. A. Alarc´on、H. Zhu、A. Teo和M. Xia，《Tutorup：如果学生是模拟出来的，会怎样？培训辅导员应对在线学习中的参与度挑战》，arXiv预印本 arXiv:2502.16178，2025年。[59] J. M. Markel、S. G. Opferman、J. A. Landay和C. Piech，《GPTeach：与基于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，2023年，第226–236页。[60] H. Jin、M. Yoo、J. Park、Y. Lee、X. Wang和J. Kim，《TeachTune：利用模拟学生评估针对不同学生画像的教学智能体》，arXiv预印本 arXiv:2410.04078，2024年。[61] Y. Choi、E. J. Kang、S. Choi、M. K. Lee和J. Kim，《Proxona：利用大语言模型驱动的人设，加深创作者对其受众的了解》，arXiv预印本 arXiv:2408.10937，2024年。[62] Y.-S. Chuang、A. Goyal、N. Harlalka、S. Suresh、R. Hawkins、S. Yang、D. Shah、J. Hu和T. T. Rogers，《利用基于大语言模型的智能体网络模拟意见动态》，arXiv预印本 arXiv:2311.09618，2023年。[63] J. He-Yueya、N. D. Goodman和E. Brunskill，《利用大语言模型判断评估并优化教育内容》，arXiv预印本 arXiv:2403.02795，2024年。[64] X. Lu和X. Wang，《生成式学生：利用大语言模型模拟的学生画像支持试题评估》，载于第十一届ACM Learning@Scale会议论文集，2024年，第16–27页。[65] D. R. Krathwohl，《布鲁姆分类法的修订：概述》，Theory into practice，第41卷，第4期，第212–218页，2002年。[66] J. Sweller，《问题解决过程中的认知负荷：对学习的影响》，Cognitive science，第12卷，第2期，第257–285页，1988年。[67] P. A. Carpenter和P. Shah，《图形理解中的感知与概念过程模型》，Journal of experimental psychology. Applied，第4卷，第2期，第75–100页，1998年。

[68] Z. Chen、S. Song、K. Shum、Y. Lin、R. Sheng和H. Qu，《揭开欺骗性视觉内容的面纱：对多模态大语言模型在误导性图表问答方面的基准测试》，*arXiv*预印本 *arXiv:2503 18172*，2025年。.[69] T. Kinkner和J. McKillip，《项目反应理论基础》，评估实践，第14卷，第2期，第191–193页，1993年。[70] A. T. Corbett和J. R. Anderson，《知识追踪：程序性知识习得的建模》，用户建模与用户自适应交互，第4卷，第4期，第253–278页，1995年。[71] E. Rich，《通过刻板印象进行用户建模》，认知科学，第3卷，第4期，第329–354页，1979年。[72] A. Kobsa，《通用用户建模系统》，用户建模与用户自适应交互，第11卷，第1-2期，第49–63页，2001年。[73] K. Exley和R. Dennick，小组教学：辅导课、研讨课及更多形式。Routledge，2004年。[74] M. S. Sharif、M. T. Tamang、D. Mani和W. Elmedany，《运用聚类算法，通过推荐模型提升个性化学习》，载于*2024*年信息学、计算与技术创新与智能国际会议（*3ICT*）。IEEE，2024年，第547–552页。[75] E. Arslan、A. ¨Ozkaymak和N. ¨O. D¨onmez，《基于游戏化用户类型和学习风格对学生进行聚类》，*arXiv*预印本 *arXiv:2310.14430*，2023年。[76] P. H¨am¨al¨ainen、M. Tavast和A. Kunnari，《评估大语言模型生成合成的人机交互研究数据：一项案例研究》，载于*2023*年*CHI*计算机系统中的人为因素会议论文集，2023年，第1–19页。[77] Y. Chang、X. Wang、J. Wang、Y. Wu、L. Yang、K. Zhu、H. Chen、X. Yi、C. Wang、Y. Wang等，《大语言模型评估综述》，*ACM*智能系统与技术汇刊，第15卷，第3期，第1–45页，2024年。[78] N. Wang、Z. Peng、H. Que、J. Liu、W. Zhou、Y. Wu、H. Guo、R. Gan、Z. Ni、J. Yang等，《RoleLLM：大语言模型角色扮演能力的基准测试、引导与增强》，载于计算语言学协会年会论文集：*ACL 2024*，2024年，第14 743–14 777页。[79] N. Reimers和I. Gurevych，《Sentence-BERT：使用孪生BERT网络生成句子嵌入》，*arXiv*预印本 *arXiv:1908.10084*，2019年。[80] J. Benesty、J. Chen、Y. Huang和I. Cohen，《皮尔逊相关系数》，载于语音处理中的噪声降低。Springer，2009年，第1–4页。



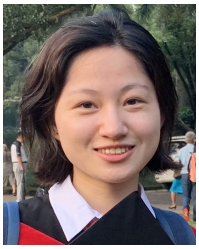
Zixin Chen 目前是香港科技大学的博士候选人。他的研究聚焦于数据可视化与人机协作，尤其侧重于将其应用于教育大语言模型。他设计交互式工具，以支持教学与教学决策、个性化学习体验，并使大语言模型行为对学习者和教育工作者都更加可信。与此同时，他还探索大语言模型驱动的模型与系统在其他重要领域中的更广泛适用性（AI4Health 与 AI4Science）。更多信息：https://cinderd.github.io/。



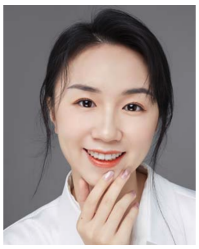
Yuhang Zeng 目前是卡内基梅隆大学计算数据科学专业的硕士研究生。他获得了香港科技大学数据科学与技术及计算机科学学士学位。他的研究兴趣包括数据可视化、人机交互和交互式人工智能系统，尤其关注将大语言模型融入可视分析与教育工具。



Sicheng Song 于2024年获得华东师范大学计算机科学与技术学院博士学位。他现为香港科技大学计算机科学与工程方向的博士后研究员。他的主要研究兴趣包括信息可视化和可视分析。更多详情请参见 https://ByShawn.github.io/。



Yanna Lin 目前是滑铁卢大学的博士后研究员。她于香港科技大学获得博士学位，并于中山大学获得工学学士学位。她的研究兴趣包括数据可视化与人机交互（HCI），重点是设计直观的界面并开发算法，以增强数据的传达与分析，最终支持数据驱动的决策。更多信息请访问：https://yannahhh.github.io/。



Xian Xu 在中国中央戏剧学院获得导演学学士学位，在戏剧研究领域获得艺术硕士学位，并在香港科技大学可视化实验室获得计算媒体与艺术博士学位。她是香港岭南大学数字艺术与创意产业系的长聘轨助理教授，也是牛津大学和剑桥大学的访问学者。除研究工作外，她还是影视编剧。



Huamin Qu 现任香港科技大学跨学科学院院长、新兴跨学科领域学部主任，以及计算机科学与工程系讲座教授。他在中国西安交通大学获得数学学士学位，并在石溪大学获得计算机科学硕士和博士学位。他的主要研究兴趣包括可视化和人机交互，重点研究城市信息学、社交网络分析、电子学习、文本可视化和可解释人工智能 更多信息：http://huamin.org/。



Meng Xia 是得克萨斯农工大学计算机科学与工程系的助理教授。她在香港科技大学获得博士学位，并先后在卡内基梅隆大学和韩国科学技术院担任博士后。她的研究兴趣主要集中于人与人工智能交互、数据可视化和教育技术。更多详情请访问 https://www.xiameng.org。