自动化学报
ACTA AUTOMATICA SINICA

交互式教学内容生成大语言模型的构建与自动化评估

张若华 刘涛 卢烨 宋思宇 刘文涛 宋宇 周爱民 郝昊

Construction and Automated Evaluation of Large Language Model for Interactive Teaching Content Generation

Zhang Ruo-Hua, Liu Tao, Lu Ye, Song Si-Yu, Liu Wen-Tao, Song Yu, Zhou Ai-Min, Hao Hao

---

交互式教学内容生成大语言模型的构建与自动化评估

张若华[1,2] 刘 涛[1,2] 卢 烨[1] 宋思宇[1,2] 刘文涛[1,2,3] 宋 宇[1] 周爱民[1,2,3] 郝 昊[1]

摘 要 当前教育大语言模型主要依赖文本对话开展知识讲解, 缺乏直观且动态的交互式呈现能力, 从而在复杂概念的可视化与过程性理解支持方面存在一定局限. 为此, 提出一种面向富媒体交互生成的教育大语言模型构建方法及其配套评测框架. 首先, 依据多学科教学大纲并结合细粒度知识点, 构建包含高质量交互式超文本标记语言 (HTML) 动画以及教学代码的指令数据集, 以缓解高质量数据稀缺的问题. 其次, 构建"监督微调 + 强化学习"的两阶段训练策略: 设计融合代码可执行性与教学相关性的混合奖励模型, 并通过强化学习进一步对齐模型生成内容的教育价值与代码鲁棒性, 从而获得具备交互式 HTML 教学素材生成能力的领域专用大语言模型. 最后, 针对交互式教学内容缺乏自动化评测体系的现状, 构建涵盖代码正确性、渲染成功率与教学交互性等维度的指标体系, 并实现相应的自动评估管线. 实验结果表明, 所提模型在交互式教学内容生成的准确性与交互性方面优于主流开源与闭源基座模型. 相关模型、数据与评测工具将公开发布, 以支持教育人工智能研究与应用的进一步发展.

关键词 教育大语言模型; 交互式教学内容生成; HTML 动画; 强化学习; 自动化评测

引用格式 张若华, 刘涛, 卢烨, 宋思宇, 刘文涛, 宋宇, 周爱民, 郝昊. 交互式教学内容生成大语言模型的构建与自动化评估. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c260145 **CSTR** 32138.14.j.aas.c260145

Construction and Automated Evaluation of Large Language Model for Interactive Teaching Content Generation

Zhang Ruo-Hua[1,2] Liu Tao[1,2] Lu Ye[1] Song Si-Yu[1,2] Liu Wen-Tao[1,2,3]
Song Yu[1] Zhou Ai-Min[1,2,3] Hao Hao[1]

Abstract Existing educational large language models (LLMs) primarily rely on text-based dialogue for knowledge explanation, lacking intuitive and dynamic interactive presentation capabilities. Consequently, they face certain limitations in visualizing complex concepts and supporting process-oriented understanding. To address this gap, we propose a method for constructing educational LLMs for rich-media interactive generation, along with a supporting evaluation framework. First, guided by multi-disciplinary curricula and fine-grained knowledge points, we construct an instruction dataset containing high-quality interactive hypertext markup language (HTML) animations and teaching code, alleviating the scarcity of high-quality data. Second, we develop a two-stage training strategy——supervised fine-tuning followed by reinforcement learning (RL). Specifically, we design a hybrid reward model that integrates code executability and pedagogical relevance, and apply RL to further align the educational value of model-generated content with code robustness, resulting in a domain-specific LLM capable of producing interactive HTML teaching materials. Finally, to address the current absence of an automated evaluation system for interactive teaching content, we establish a multi-dimensional metric suite covering code correctness, rendering success rate, and teaching interactivity, and implement an automated evaluation pipeline. Experimental results show that the proposed model outperforms mainstream open-source and closed-source base models in both the accuracy and interactivity of generated interactive teaching content. The related models, datasets, and evaluation tools will be publicly released to support the further development of educational artificial intelligence research and applications.

Keywords educational large language models; interactive teaching content generation; hypertext markup language animations; reinforcement learning; automated evaluation

Citation Zhang Ruo-Hua, Liu Tao, Lu Ye, Song Si-Yu, Liu Wen-Tao, Song Yu, Zhou Ai-Min, Hao Hao. Construction and automated evaluation of large language model for interactive teaching content generation. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

收稿日期 2026-02-27 录用日期 2026-06-24
Manuscript received February 27, 2026; accepted June 24, 2026
国家自然科学基金 (62306174) 资助
Supported by National Natural Science Foundation of China (62306174)
本文责任编委 刘淇
Recommended by Associate Editor LIU Qi

1. 华东师范大学上海智能教育研究院 上海 200062 2. 华东师范大学计算机科学与技术学院 上海 200062 3. 上海创智学院 上海 200003
1. Shanghai Institute of AI for Education, East China Normal University, Shanghai 200062 2. School of Computer Science and Technology, East China Normal University, Shanghai 200062 3. Shanghai Innovation Institute, Shanghai 200003

1 研究背景

伴随着深度学习与自然语言处理技术的持续突破, 以 Gemini 3 Pro、Qwen3 为代表的大语言模型 (large language models, LLMs) 在知识问答、逻辑推理与代码生成等关键能力上展现出显著优势[1−3]. 在教育信息化与智能教育系统发展的语境下, 上述能力正在推动教学资源供给模式由“标准化资源库的检索与分发”向“面向学习目标与学习者状态的按需生成”转型. 当前, 生成式人工智能已被集成至多类智能导学系统之中, 用于实现解题讲解、学习路径推荐与即时反馈等功能. 例如, 好未来推出的 MathGPT[4] 已能够在一定范围内完成较高精度的数学题目解析、步骤化推理呈现与逻辑纠错. 这一进展表明, 面向教育场景的领域适配与教学对齐正在成为大语言模型应用落地的重要方向. 然而, 尽管现有教育大语言模型在“文本解答生成”与“静态图像理解”等任务上已达到较高水平, 其在生成具备动态交互能力的复杂科学教育内容方面仍处于探索阶段. 现阶段的生成式教学内容仍以静态图文或线性讲解为主, 本质上仍属于单向的信息呈现与接收模式. 当面对物理场论、系统动力学等涉及复杂时空演化与参数敏感性的知识点时, 仅依赖静态表达往往难以支持学习者进行可操作的探究式学习与可检验的概念建构. 依据认知负荷理论[5−6], 学习者在理解复杂静态图文材料时, 需要在工作记忆中执行高成本的动态心理模拟, 以补足材料中未显式呈现的过程与状态变化. 该过程容易诱发外在认知负荷累积, 并在任务难度较高时造成认知资源挤占, 从而影响对本质概念与因果关系的加工效率. 相较之下, 交互式仿真允许学习者通过直接操纵系统参数、观察实时响应并形成反事实比较, 以外部可视化与即时反馈分担内部心理模拟压力, 这一机制通常被称为“认知卸载”[7]. 既有实证研究表明, 在科学、技术、工程与数学领域的学习过程中, 经过适当设计的交互式可视化内容与操作反馈机制, 能够有效促进学习者对核心概念的理解, 并显著提升其知识迁移的实际表现[8−9]. 因此, 面向教育大语言模型的关键问题不再仅是“能否给出正确答案”, 而是“能否生成可交互、可验证且服务于教学目标的过程性材料”, 以更好支持学习者形成可操作的心理模型与可迁移的知识结构. 如图 1 所示, 交互式内容通过实时反馈回路实现认知卸载, 从而降低学习者构建心理模型的负担. 交互式教学内容生成并非单纯追求页面动态效果或交互控件数量, 而是要求交互行为、视觉呈现与文本讲解共同服务于特定知识点的概念建构、规律验证和探究式学习目标.

图 1 传统静态资源与生成式交互仿真中的认知加工过程对比

Fig. 1 Comparison of cognitive processing in traditional static resources and generative interactive simulations

实现高质量交互式教学内容的自动化生成仍面临多重技术挑战. 一方面, 传统交互内容开发通常依赖 Unity 等重型引擎, 开发成本高、部署链路复杂, 且难以满足 Web 端轻量化、跨平台与可快速迭代的教学应用需求. 相比之下, 基于超文本标记语言 (hypertext markup language, HTML) 和 JavaScript 的 Web 标准具有天然的分发与运行优势, 并具备“代码即文本 (code as text)”的可学习性与可编辑性[10], 更适合与 LLMs 的代码生成能力进行耦合. 另一方面, 面向科学概念教学的交互仿真代码往往同时包含数值计算、状态更新、事件响应、可视化渲染与交互控件绑定等复合逻辑, 对模型的程序综合能力与鲁棒性提出更高要求. 通用代码模型, 如 DeepSeek-Coder-V2[11] 和 Qwen3-Coder[12], 在生成涉及科学计算或仿真过程的代码时, 仍可能出现“幻觉”或“隐性错误”, 例如变量含义不一致、数值更新不符合守恒关系或事件监听缺失等, 且通常缺乏对教学交互逻辑的针对性训练与约束. 交互式教学内容生成至少面临三类耦合约束: 第一, 代码层面的可执行性与鲁棒性, 要求模型生成的 HTML/JavaScript 能够在浏览器环境中稳定运行; 第二, 交互层面的状态一致性, 要求事件监听、参数控制、动画更新和用户反馈形成可达的行为链条; 第三, 教育层面的目标一致性, 要求视觉呈现和交互过程服务于特定知识点的解释、验证与迁移, 而非仅产生形式上的动态效果. 现有评估体系多聚焦于语法正确性、编译或解释执行成功率以及静态视觉相似度等指标, 而对渲染结果的可用性、交互行为的可达性以及教学目标一致性等关键维度缺乏统一的多模态度量框架[13]. 在缺少可靠评测与反馈信号的情况下, 模型训练与优化难以形成稳定闭环, 进而限制了交互式教学内容生成质量的可控提升.

针对上述挑战, 在 Yang 等[14] 提出的 InterCode 所采用的执行反馈驱动优化思想基础上, 进一步面向教育场景提出一种融合富媒体交互生成能力的教育大语言模型构建方法. 不同于主要关注命令行执行结果或静态页面相似度的既有工作, 所提方法将模型输出视为可在浏览器环境中运行的教学程序, 并将代码执行结果、页面渲染状态、交互行为变化和教学目标一致性共同纳入数据集构建、自动评估与强化学习 (reinforcement learning, RL) 对齐过程.

具体而言, 以稀疏混合专家 (mixture of experts, MoE) 架构的 Qwen3-30B-A3B-Coder 作为基座[12], 面向教育交互代码生成任务设计训练数据与对齐策略. 一方面, 构建基于多模态反馈的迭代修正闭环, 在数据合成与筛选阶段显式约束“可执行性–可渲染性–可交互性”的质量门槛, 从而提升指令数据的有效性与一致性; 另一方面, 引入组相对策略优化 (group relative policy optimization, GRPO) 强化学习算法[15], 在监督微调 (supervised fine-tuning, SFT) 形成能力冷启动后, 通过奖励建模进一步对齐模型生成内容与教学目标需求. 考虑到交互式代码生成中可能存在的奖励投机风险, 在奖励函数中加入反作弊约束, 以抑制通过堆砌表面交互换取高分的无效策略, 从而提升生成内容在教学情境中的可用性与鲁棒性. 所关注的核心问题并不是如何将通用代码模型简单迁移到教育场景, 而是如何将交互式教学内容生成建模为一个运行后质量可观测的多模态约束优化问题. 与静态教育问答、通用代码生成和静态前端页面生成不同, 交互式教学 HTML 的质量只有在代码可执行、页面可渲染、交互行为可达且交互过程服务于教学目标时才真正成立. 基于这一任务特性, 将自然语言教学意图、可执行 HTML/JavaScript 程序、浏览器运行反馈、视觉渲染状态与教学目标一致性纳入统一闭环, 并围绕该闭环设计数据集构建、评估指标和强化学习对齐机制.

主要贡献总结如下:

1) 提出面向教育语境与运行时多模态一致性约束的交互式教学内容生成任务建模, 并构建训练数据集 Edu-Interact. 不同于传统教育大语言模型的文本讲解生成或通用代码模型的静态程序生成, 交互式教学内容生成被建模为同时受代码可执行性、视觉可渲染性、运行时交互响应与教学目标一致性约束的复合生成任务. 针对现有开源数据在交互逻辑与教学语境方面不足的问题, 基于 K-12 知识点体系构建包含 10 000 条“文本-HTML/JavaScript”配对样本的高质量数据集, 并进一步验证基于“诊断–修复”机制的数据增强潜力, 为交互式教学 HTML 生成提供可复用的数据基础.

2) 提出一种基于 GRPO 的“冷启动 + 强化”两阶段对齐训练方法. 利用 Qwen3-30B-A3B-Coder 的高效 MoE 架构, 首先通过 SFT 实现指令遵循与任务格式的能力冷启动; 随后针对强化学习训练中可能出现的策略坍塌与奖励投机问题, 设计包含反作弊约束的混合奖励模型, 以提升生成代码的交互覆盖率与物理逻辑正确性.

3) 建立一套基于统一多模态大语言模型的级联评估体系. 利用 Qwen3-VL-235B-A22B-Instruct[16] 同时承担代码逻辑与视觉呈现的自动裁判角色, 构建覆盖代码可执行性、渲染成功率与教学交互性等维度的自动化评估管线, 并在统一测试集上系统验证所提方法的有效性与可扩展性.

需要说明的是, 该自动评估指标主要用于衡量生成内容的技术可用性与初步教学适配性, 真实学习成效仍需结合学习者实验、课堂应用和长期反馈进一步验证; 相关边界将在讨论部分进一步分析.

2 相关工作

本节围绕教育大语言模型、代码生成与交互式环境、基于多模态反馈的对齐与评估以及代码生成的自我修正四个方向, 对代表性研究进行系统梳理, 并进一步归纳现有方法在面向教学的多模态交互式内容生成任务上所面临的关键瓶颈与研究空白. 总体而言, 尽管教育与代码领域的大语言模型能力在近两年快速提升, 但现有工作大多关注文本化教学问答、通用代码生成、静态页面或前端生成以及受限环境中的执行反馈, 尚未充分处理教育场景中“教学目标–可执行代码–视觉渲染–动态交互”之间的耦合关系. 尤其是, 如何在教育语境约束下建立可验证的多模态一致性指标体系, 并进一步将其转化为稳定的训练与对齐信号, 仍缺乏系统研究.

2.1 教育大语言模型

近年来, 教育垂类大语言模型的发展呈现出由“以文本问答为主的知识讲解”向“以步骤化推理为核心的认知支持”转变的趋势. Kasneci 等[17] 和 Kohnke 等[18] 较早讨论了大语言模型进入教育场景所带来的机会与风险, 并指出其在内容可信度、偏差与可控性方面仍需系统治理. 在此基础上, EduChat 等教育大语言模型借助指令微调 (instruction tuning) 与对话式教学模板, 实现了面向答疑、引导与反馈的基础教学能力. 进入 2024—2025 年, 研究重心进一步转向数学、几何等结构化推理密集型任务.

例如, DeepMind 的 AlphaGeometry 通过神经语言模型与符号演绎引擎的结合, 在几何证明上取得突破性进展[19]; MathGPT 则通过大规模合成数据进行微调, 在一定程度上降低了数学推理与计算过程中的幻觉问题[4]. 上述工作共同表明, 围绕可验证推理链与形式化约束的训练与推断范式, 正在成为教育大语言模型提升可靠性的重要路径.

尽管现有教育与推理大语言模型在“解题步骤生成”、“形式化证明”与“符号推演”方面表现突出, 但其输出形态仍以静态文本或静态图表为主, 难以满足复杂概念的“过程性可视化”与“交互式探究”需求. 例如, 粒子碰撞、函数参数实时变化、几何构造拖拽操作等典型教学情境, 往往要求模型生成可执行且可交互的仿真环境, 而不仅是给出结论或推导过程. 从学习科学角度看, 建构主义学习理论强调学习者在操作、试错与反馈中形成概念结构; 若缺乏“从做中学”的交互式探究空间, 学习者可能难以建立对动态系统与因果机制的稳健理解. 因此, 如何将推理能力进一步外化为“可交互、可验证、可迭代”的教学呈现形式, 并使交互行为真正服务于知识点理解、规律验证和探究式学习目标, 仍是教育大语言模型亟待突破的方向.

2.2 代码生成与交互式环境

代码生成能力是构建动态教学内容与交互式学习环境的关键支撑技术. 随着 MoE 等架构逐步成熟, Qwen3-Coder[12]、Qwen2.5-Coder[20] 和 DeepSeek-Coder-V2[11] 等大语言模型在保持推理效率的同时, 展现出较强的程序合成与逻辑约束处理能力, 为生成可运行教学材料提供了更高的上限. 在交互式环境与执行反馈研究中, InterCode[14] 提出的“生成–执行–反馈”范式具有代表性意义, 其核心在于将代码执行结果显式纳入模型优化与评估, 从而提升生成程序的可用性与稳定性. 然而, 当任务迁移至网页与前端交互环境时, 评估与反馈获取的复杂度显著上升. WebArena[21] 指出, 为智能体构建能够逼真模拟用户操作的评估环境具有较高工程与建模难度, 且操作序列、页面状态与目标达成之间存在显著的长程依赖. Design2Code[13] 虽关注前端生成, 但其评估多依赖静态视觉相似度或布局匹配, 容易忽略运行时事件响应、状态更新与交互逻辑的正确性.

现有研究在面向教学的交互式环境生成方面主要存在三类不足: 其一, 领域适配性不足. 许多代码模型主要针对 Bash、SQL、后端脚本等工程任务优化, 缺少对教学仿真逻辑的显式建模与约束表达, 导致在物理或数学仿真场景中易出现逻辑性幻觉, 例如违反基本物理规律的参数设置. 其二, 交互反馈建模不足. 前端生成或网页智能体相关工作虽然涉及页面状态与用户操作, 但其目标多为页面还原、任务完成或操作序列成功, 较少关注交互过程是否能够帮助学习者理解概念、验证规律或形成可迁移的知识结构. 其三, 评估维度相对单一. 当前评测通常偏向代码层面可编译、可运行, 或偏向界面层面视觉相似度, 但缺乏将“视觉呈现质量–交互可达性–动态逻辑正确性–教学目标一致性”统一起来的端到端自动化评测体系. 尤其在教学场景中, 交互元素的存在并不必然带来教学有效性; 若缺少对教学目标一致性与交互引导质量的度量, 模型可能生成“形式上可交互、实质上无教学价值”的内容.

2.3 基于多模态反馈的对齐与评估

为提升生成内容的可控性与可靠性, 大语言模型对齐技术已由传统的基于人类反馈的强化学习 (reinforcement learning from human feedback, RLHF) 逐步演进到更高效的偏好优化与策略学习框架. 简单偏好优化 (simple preference optimization, SimPO)[22] 与 Focused-DPO[23] 通过改进偏好排序与优化目标, 提高了指令遵循与风格一致性的稳定性. 面向复杂推理任务, DeepSeekMath 提出的 GRPO[15] 以组内相对优势估计替代传统价值网络, 在数学与代码生成任务中展现出较高的数据效率, 反映出轻量化对齐在推理密集任务中的潜力. HybridFlow[24] 则为 RLHF 提供了灵活的训练框架与系统支持. 与此同时, 渲染感知强化学习方法[25] 已开始探索利用渲染反馈指导图形生成, 表明多模态反馈信号可为生成模型提供超越文本层面的约束来源. 然而, 在强化学习对齐过程中, 奖励过度优化与奖励投机问题仍较突出[26]: 模型可能利用奖励函数的漏洞生成高分但低质的结果, 例如通过堆砌无效交互元素或重复模板化结构来骗取交互相关指标. 此外, 采用更强模型, 如 Gemini 3 Pro[2], 作为裁判进行自动评估已成为趋势[27], 但在多模态教育内容生成场景中, 如何协同大语言模型与视觉语言模型 (vision-language model, VLM) 实现一致、稳定且可解释的联合评估, 仍缺乏统一范式.

针对交互式教学代码生成任务, 目前尚缺少能够兼顾“交互丰富性”与“呈现质量/教学目标一致性”的对齐与评估机制: 一方面, 过度强调交互数量可能损害界面简洁性与认知负荷控制; 另一方面, 单纯强调视觉质量又可能掩盖交互路径不可达、事件响应错误等问题. 同时, 现有奖励设计对奖励投机的防御能力有限, 难以保证模型在不同学科与不同

任务结构下的泛化稳定性. 因此, 面向交互式教学代码生成的反馈对齐机制, 需要同时处理三类问题: 一是如何将代码执行、页面渲染、交互响应与教学逻辑转化为可计算的多模态反馈信号; 二是如何缓解交互式代码生成中的奖励稀疏问题, 使模型能够从静态页面逐步过渡到有效交互页面; 三是如何防止模型通过堆砌无效按钮、滑块或事件监听等方式获得高分, 即抑制“形式上可交互、实质上不可教学”的奖励投机行为.

2.4 代码生成的自我修正

传统代码生成范式多采用一次性生成, 在长代码、复杂依赖与多约束目标的场景中容易累积错误并导致整体失败. 针对这一问题, Self-Refine[28] 与 Reflexion[29] 等工作验证了 LLMs 具备在外部反馈或自我反思提示下进行迭代修正的能力, 其典型流程包括生成初稿、根据错误信息进行诊断、提出修改方案并再次生成. 此类方法在以编译器报错为核心的场景中表现有效, 能够提升语法正确性与基本可运行性.

现有自我修正机制主要依赖报错栈与单元测试等文本级反馈, 难以有效处理交互式教学内容中常见的跨模态错误. 具体而言, 纯编译器反馈往往无法充分暴露前端渲染层面的布局错位、元素遮挡与动画失效, 难以识别事件绑定错误或状态更新停滞等交互故障, 也无法捕获重力方向颠倒、边界条件失真等违背客观规律的仿真逻辑异常. 此外, 多模态输出存在严格的“语义一致性”要求, 即文本讲解、视觉呈现与交互行为必须协同服务于同一教学目标, 而现有修正框架较少对该类约束进行显式建模. 因此, 面向交互式教学内容生成的自我修正机制, 需要从基础的“代码可运行”维度进一步拓展到代码逻辑、视觉渲染、交互状态变化与教学目标一致性的多模态场景. 这也意味着, 修正信号不能只来自编译器或单元测试, 还需要结合浏览器运行时反馈、视觉反馈和领域知识反馈, 以发现并修复“代码能跑但不可教学”的隐性质量问题.

2.5 与现有工作的区别

综合上述四类研究可以看到, 现有工作为所提方法提供了重要基础, 但在面向交互式教学 HTML 生成任务时仍存在若干缺口. 首先, 现有教育大语言模型主要围绕文本问答、解题推理与静态图文讲解展开, 尚未形成专门面向交互式 HTML 教学内容生成的模型、数据集与评估基准; 其次, 通用代码生成与执行反馈研究虽然能够提升程序的可运行性与稳定性, 但其目标多集中于命令行、SQL、Bash 或一般工程代码任务, 缺少对教学目标、认知负荷和交互引导质量的约束; 再次, 前端生成与网页智能体相关研究多关注静态页面还原、视觉相似度或用户操作任务完成度, 对教学场景中运行时状态变化、交互反馈与概念表达之间的一致性关注不足; 最后, 现有反馈对齐与自我修正方法大多依赖文本级错误信息、偏好分数或有限环境反馈, 尚未充分处理交互式教学内容中“代码可执行、页面可渲染、交互可达、教学目标一致”之间的联合约束.

相比之下, 所提方法面向教育语境下的交互式教学内容生成任务, 将“教学目标–可执行代码–视觉渲染–动态交互”统一纳入同一生成、评估与对齐闭环. 具体而言, 从多学科教学大纲和细粒度知识点出发构建教育属性明确的交互代码数据; 通过浏览器执行反馈、视觉语言模型评估和交互状态检测建立多层级自动评测体系; 进一步将该评测体系转化为强化学习阶段的奖励信号, 以优化模型在可运行性、可渲染性、交互响应和教学目标一致性上的综合表现. 因此, 该工作的推进点不只是将已有代码生成方法应用于教育场景, 而是围绕交互式教学 HTML 生成这一任务, 补齐数据集构建、自动评估、训练流程与奖励对齐机制等关键环节.

3 交互式教学内容生成模型构建方法

针对教育领域大语言模型在生成高交互性、逻辑严密的教学内容时面临的关键困难, 包括科学计算与交互逻辑中的幻觉、代码可执行但教学行为链条断裂, 以及缺乏面向渲染结果与教学目标一致性的反馈信号等, 提出一种系统化的端到端训练框架. 该框架借鉴执行反馈驱动的代码优化思想[14], 并结合多媒体学习与认知负荷相关研究[5, 9, 30], 将教育语境约束、浏览器运行时反馈、视觉渲染反馈和交互行为反馈纳入统一优化过程. 其核心思想是将教学内容生成从传统的静态文本输出任务转化为以可执行程序为中介、以运行环境和渲染结果为反馈的闭环优化过程, 从而在训练阶段缩小“文本正确但行为错误”和“代码可运行但不可教学”之间的质量差距.

所提方法将代码生成范式由单次预测扩展为“生成–执行–反馈–修正”的动态交互过程, 并强调环境反馈在提升指令遵循稳定性与程序可用性方面的作用. 与主要面向工程命令行或受限交互环境的既有工作不同, 该方法进一步面向教育场景中的富媒体交互任务, 将反馈信号扩展至网页运行时与视觉呈现层面, 使模型在优化过程中同时满足: 1) 语法与依赖正确; 2) 交互事件可达且状态更新一致; 3) 渲染结果可用且教学表达清晰. 交互数量或视觉

复杂度本身并非最终目标, 交互行为、视觉呈现与文本讲解应共同服务于具体知识点的解释、验证与探究过程.

本节从总体架构出发, 依次介绍模型选型与模块化训练流程、面向交互式教学代码的数据合成机制、基于多模态大语言模型 (multimodal large language model, MLLM) 的级联评估与质量过滤策略, 以及基于组相对策略优化的强化学习对齐方法. 本文以 Qwen3-30B-A3B-Coder 为基座模型, 依次进行 SFT 冷启动与 GRPO 强化学习, 并将两阶段训练所得的最终模型命名为 IE-code

3.1 总体架构

总体架构旨在解决“代码即思维 (code as thought)”在教育场景中面临的两类约束: 一方面, 教学应用对推理延迟、部署成本与跨端运行较为敏感, 难以直接依赖高成本的大尺寸模型; 另一方面, 教育内容不仅要求程序能够运行, 还要求交互链条与概念表达服务于明确的教学目标, 因而需要可解释、可验证的质量控制机制[10]. 如图 2 所示, 该系统包含数据合成、监督微调、强化学习和级联评估 4 个相互耦合的环节. 代码质量通过浏览器执行结果判断, 视觉质量通过渲染截图判断, 交互质量通过用户操作后的状态变化验证, 教学质量则结合知识点目标与页面行为进行一致性判断. 由此, 训练目标由模仿教师代码进一步转向优化运行后可观测的教学交互质量.

系统由四个核心模块构成, 各模块在功能上相互依赖并形成闭环.

1) 迭代修正闭环: 以多学科教学大纲、细粒度知识点、学段要求和交互场景描述为教育先验, 利用教师模型生成与教学任务匹配的“思维链–代码”样本. 同时引入迭代修正机制, 使合成数据在执行反馈与渲染反馈约束下逐步收敛到可用版本.

2) SFT 冷启动模块: 以 Qwen3-30B-A3B-Coder 为基座开展指令微调, 通过高质量指令数据激活领域相关专家参数, 使模型获得稳定的格式遵循能力与基础交互代码生成能力, 为强化学习阶段提供策略初始化.

3) GRPO 强化学习模块: 在 SFT 冷启动基础上引入 GRPO 进行交互对齐, 通过相对优势估计比较同组候选程序的质量, 优化交互覆盖、事件响应与逻辑一致性等目标.

4) 级联评估器: 基于 Qwen3-VL-235B-A22B-Instruct 构建早停式评估机制, 对候选样本进行分层筛查, 以较低成本优先剔除不可执行、不可交互或渲染失败的样本.

图 3 展示了基于多模态反馈的迭代修正闭环. 系统利用视觉语言模型作为修复模型, 结合抽象语法树 (abstract syntax tree, AST) 语法检测和视觉缺陷检测, 将低质量代码修正为可用的训练样本. 具体流程如下:

1) 种子数据注入 ($\mathcal{D}_{\text{seed}}$): 由专家教师基于多媒体学习原则[30] 构建的 18 988 条种子数据作为系统

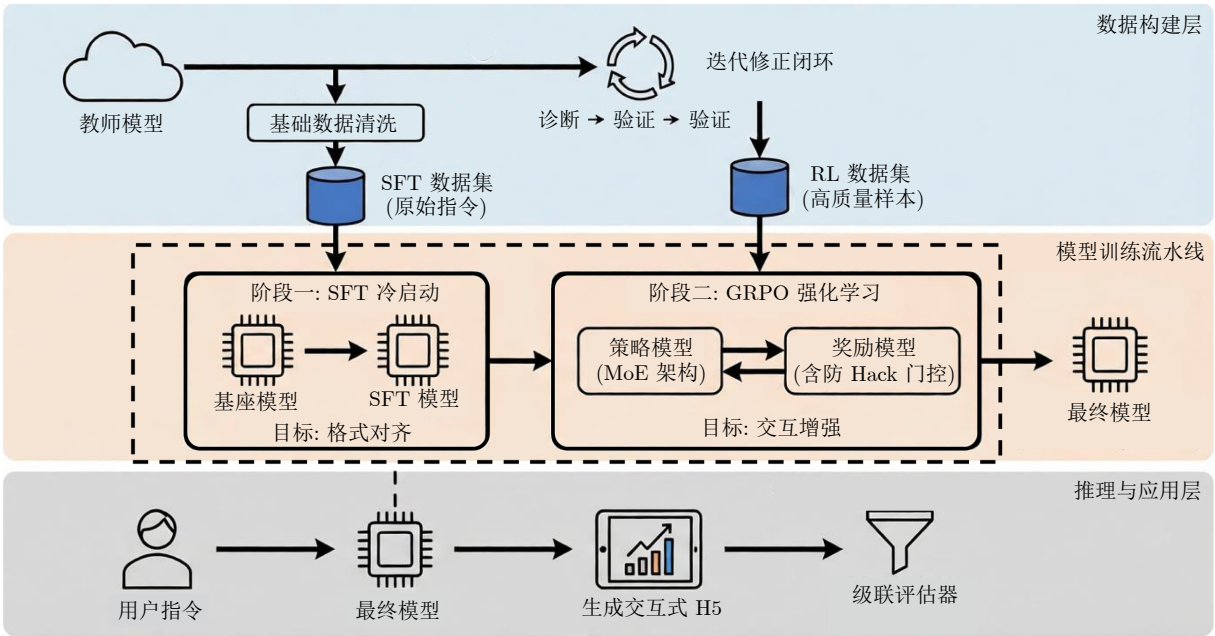


图 2 融合富媒体交互的生成式教育大语言模型总体架构

Fig. 2 Overall architecture of the generative educational large language model with rich-media interaction

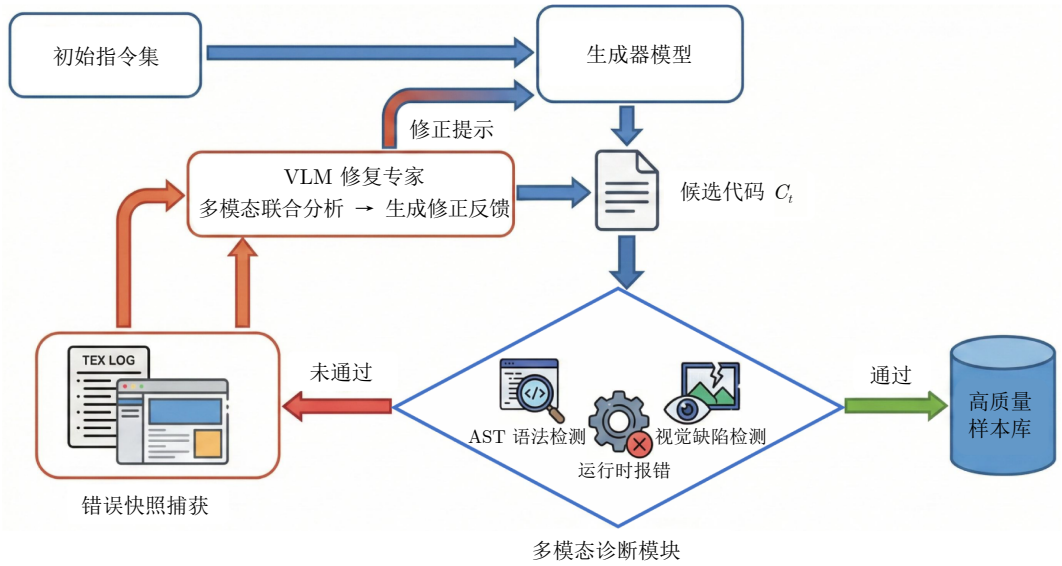


图 3　基于多模态反馈的迭代修正闭环

Fig. 3　Closed-loop iterative refinement based on multimodal feedback

数据合成的高质量先验知识库.

2) 教师模型推理与合成: 采用 Gemini 3 Pro[2] 作为教师模型, 通过思维链 (chain of thought, CoT) 推理将教学元数据转化为包含 HTML、JavaScript 和层叠样式表 (cascading style sheets, CSS) 的原始交互式代码数据 $\mathcal{D}_{\text{raw}}$. 教师模型的选择基于实验初期的多模型预评估. 选取 Gemini 3 Pro、Qwen3-235B 和 DeepSeek-V3.2 等代表性模型, 分别生成约 30 个交互式教学页面, 并从代码可运行性、视觉布局合理性、交互完整性、教学逻辑清晰度和整体可用性等方面进行人工比较. 结果表明, Gemini 3 Pro 在该任务上的整体表现较优, 因此将其作为主要教师模型. 单一教师模型可能带来的布局风格和交互模式偏置将在结束语中讨论.

3) 基于多模态反馈的迭代修正闭环: 为提高数据利用率并验证自我修正的可行性, 设计了自动修正模块, 具体流程见算法 1.

a) 诊断与修正: 视觉语言模型修复器基于 4 层评测体系生成包含多模态证据的诊断报告, 生成模型据此生成修正代码.

b) 数据分级策略: 为保证主模型训练的稳定性, 将数据划分为两类:

i) 高置信度集: 一次通过全部测试的样本, 是当前版本模型训练的主要数据来源.

ii) 修复集: 通过迭代修正闭环修复成功的样本, 约占失败样本的 40%. 这部分数据标记为银标数据, 用于验证修复方法的有效性和后续消融实验.

4) 数据集构建: 为提高 MoE 模型的专家利用率, 将高置信度集整理为以下两部分, 共约 10 000 条:

a) 指令微调数据集 ($\mathcal{D}_{\text{SFT}}$): 包含 10 000 条覆盖多学科的数据, 用于建立模型的基础指令遵循能力.

b) 强化学习数据集 ($\mathcal{D}_{\text{RL}}$): 根据逻辑复杂度评分, 从 $\mathcal{D}_{\text{SFT}}$ 中筛选难度最高的 40% 样本, 共约 4 000 条. 这部分数据复用于强化学习阶段, 使模型在复杂场景中探索交互性优于 SFT 标签的解法.

5) 强化学习对齐: 将经过 SFT 冷启动的模型作为初始策略网络, 在 GRPO 算法驱动下通过与模拟环境交互不断优化.

算法 1 中, getScore 将错误报告转换为当前得分; $C_{\text{curr}}/R_{\text{curr}}$ 与 $C_{\text{next}}/R_{\text{next}}$ 分别表示当前及下一轮候选代码/错误报告, $s_{\text{curr}}/s_{\text{next}}$ 表示相应得分; E_{all} 为汇总错误证据; P_{fix} 为修正提示; S_{shot} 为渲染截图; extractSyntax、extractLogic、extractVisual 和 extractRuntime 分别提取语法、逻辑、视觉和运行错误证据; constructPrompt 根据初始指令、当前代码和错误证据构造修正提示; simulate 执行候选代码并获得渲染结果; $\mathcal{M}_{\text{gen}}$ 生成下一轮候选代码, $\mathcal{M}_{\text{judge}}$ 返回相应的错误报告和得分.

算法 1. 基于多模态反馈的迭代修正算法

输入. 初始指令 I, 失败代码 C_0, 错误报告 R_0, 最大轮数 $T_{\max}$, 及格阈值 τ.

输出. 高质量代码 C^* 或空集 $\emptyset$.

1) $t \leftarrow 1$; $C_{\text{curr}} \leftarrow C_0$; $R_{\text{curr}} \leftarrow R_0$;
2) $s_{\text{curr}} \leftarrow \text{getScore}(R_{\text{curr}})$
3) **while** $t \le T_{\max} \wedge s_{\text{curr}} < \tau$ **do**
4) $\quad E_{\text{syn}} \leftarrow \text{extractSyntax}(R_{\text{curr}})$
5) $\quad E_{\text{log}} \leftarrow \text{extractLogic}(R_{\text{curr}})$

$E_{\text{vis}} \leftarrow \text{extractVisual}(R_{\text{curr}})$
$E_{\text{run}} \leftarrow \text{extractRuntime}(R_{\text{curr}})$
$E_{\text{all}} \leftarrow \{E_{\text{syn}}, E_{\text{log}}, E_{\text{vis}}, E_{\text{run}}\}$
$P_{\text{fix}} \leftarrow \text{constructPrompt}(I, C_{\text{curr}}, E_{\text{all}})$
$C_{\text{next}} \leftarrow \mathcal{M}_{\text{gen}}(P_{\text{fix}})$
$S_{\text{shot}} \leftarrow \text{simulate}(C_{\text{next}})$
$(R_{\text{next}}, s_{\text{next}}) \leftarrow \mathcal{M}_{\text{judge}}(C_{\text{next}}, S_{\text{shot}})$
if $s_{\text{next}} \geq \tau$ **then**
  return C_{next}
end
$C_{\text{curr}} \leftarrow C_{\text{next}}$, $R_{\text{curr}} \leftarrow R_{\text{next}}$
$s_{\text{curr}} \leftarrow s_{\text{next}}$, $t \leftarrow t+1$
end
return $\emptyset$

3.2 训练数据集 Edu-Interact 构建

上述总体流程描述了数据在系统中的流转方式, 本节进一步说明训练数据的结构化维度、清洗流程与最终数据集划分.

3.2.1 数据维度设计

为使生成内容系统覆盖 K-12 教学需求, 根据国家课程标准与专家教学大纲, 从学科领域、适用学段、具体知识点、认知难度和交互场景 5 个维度对每条指令进行结构化标记. 这些维度作为数据合成的先验约束, 用于指导后续代码生成.

3.2.2 规模描述与清洗流程

清洗流程包括目标代码提取、前端环境依赖自动注入和多轮去重. 去重采用最小哈希与局部敏感哈希算法, 通过计算代码文档间的局部相似度过滤高度同质化的生成样本. 整体流程借鉴 MathGPT[4] 和 AlphaGeometry[19] 的特定领域数据合成方法, 最终构建包含 10 000 条交互式教学指令样本的数据集 Edu-Interact, 其完整集合用于 SFT, 再根据逻辑复杂度从中筛选约 4 000 条高难度样本, 构成强化学习数据集用于 RL.

3.3 基于统一多模态大语言模型的级联评估

为兼顾筛选精度、计算效率和评估的语义一致性, 构建了基于统一多模态大语言模型的级联评估机制. 该机制既用于离线测试和数据过滤, 也作为强化学习阶段奖励建模的基础, 将交互式教学内容的可执行性、可渲染性、可交互性和可教学性转化为结构化反馈信号. 如图 4 所示, L1 和 L4 的快速过滤可拦截约 70% 的无效样本, 首尾关键帧差分可将视觉词元消耗降低约 90%.

选用 Qwen3-VL-235B-A22B-Instruct[16] 作为核心评估器. 该模型共有 2 350 亿个参数, 其中每次推理激活约 220 亿个参数, 并支持原生 256 K 上下文, 能够同时处理长 HTML 源码和高分辨率渲染截图.

3.3.1 交互性的形式化定义

为刻画生成页面在用户操作后是否产生可观测的状态变化, 定义候选页面 x 的有效交互指示函数:

$$\mathbb{I}_{\text{int}}(x) = \begin{cases} 1, & \Delta_{\text{DOM}} > \varepsilon \vee \Delta_{\text{canvas}} > 0 \vee N_{\text{mut}} > 0 \\ 0, & \text{其他} \end{cases} \tag{1}$$

其中, Δ_{DOM} 为文档对象模型 (document object model, DOM) 的变化量; ε 为判断 DOM 有效变化的阈值; Δ_{canvas} 为画布像素的变化量; N_{mut} 为 DOM 变更次数.

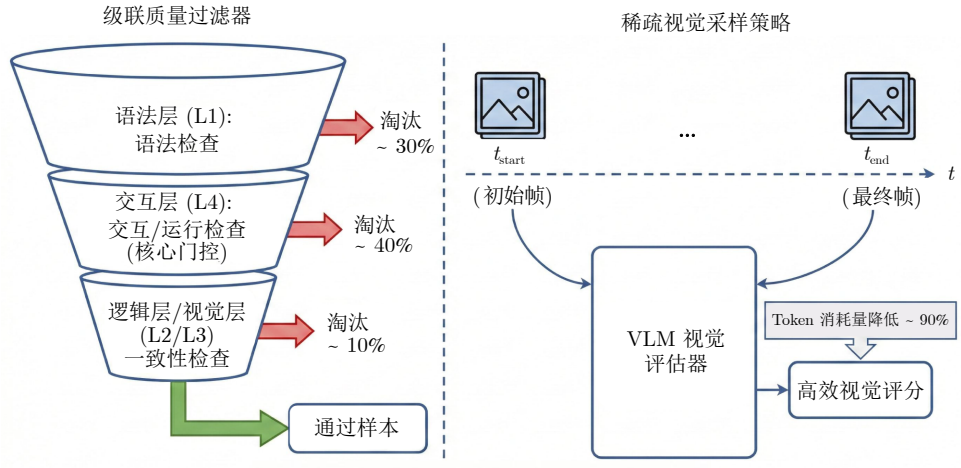


图 4 级联评估与稀疏视觉采样策略

Fig. 4 Cascaded evaluation and sparse visual sampling strategy

3.3.2　基于首尾关键帧的视觉采样

为避免强化学习高频采样过程中多模态大语言模型上下文溢出并降低显存开销, 采用首尾状态差分检测策略. 交互式代码的有效性主要表现为操作前后的状态变化, 因此将视觉评估函数的输入定义为关键帧二元组

$$\boldsymbol{V}_{\text{in}} = [\boldsymbol{I}_{\text{init}},\ \boldsymbol{I}_{\text{final}}] \tag{2}$$

其中, $\boldsymbol{I}_{\text{init}}$ 为页面加载完成后的初始渲染图像; $\boldsymbol{I}_{\text{final}}$ 为模拟用户执行一系列交互操作后的最终状态图像. Qwen3-VL-235B-A22B-Instruct 通过比较两帧图像的语义差异判断交互逻辑是否生效.

3.3.3　并发评测机制

针对大规模强化学习训练中环境交互开销较大的问题, 设计了快速失败动态剪枝与混合并发架构. 系统采用级联评估策略, 当 L4 交互检测失败时, 自动跳过计算成本较高的 L3 多模态视觉评估, 从而降低约 40% 的计算开销.

基于多线程与异步协程的混合并发调度架构将同步的浏览器环境操作与异步模型推理隔离, 提高了评估吞吐量. 针对交互式代码对外部资源库的依赖, 系统还引入离线资源代理与拦截模拟机制, 在本地接管内容分发网络请求, 降低网络延迟与波动带来的奖励计算噪声.

3.4　多模态层级化奖励模型

交互式教学 HTML 的输出质量需要在浏览器环境中执行后观测. 候选页面即使语法正确, 仍可能存在渲染失败、控件不可触发、状态更新停滞、视觉遮挡或交互过程与教学目标不一致等问题. 因此, 将语法正确性、教学逻辑一致性、视觉渲染质量和交互响应纳入层级化评估体系, 并转化为强化学习阶段的奖励信号. SFT 阶段用于获得基础格式遵循能力和可运行代码生成能力, GRPO 阶段则通过环境执行和多模态评估获得候选程序间的相对质量差异.

该任务同时存在奖励稀疏和奖励投机问题. 训练初期模型容易生成静态展示页面, 使 L4 奖励长期为 0. 潜力奖励通过识别事件监听、交互控件和画布等交互意图, 为模型提供由静态页面向有效交互页面过渡的弱引导. 若仅依据交互触发次数或 DOM、画布状态变化评分, 模型可能堆砌无意义按钮或滑块以获得较高交互分. 为此, 将交互得分与视觉质量联合约束, 并利用反作弊正则项降低形式上可交互但实质上不可教学的奖励投机风险. 如图 5 所示, 多模态奖励模型引入潜力奖励缓解冷启动问题, 并利用反作弊正则项抑制无效交互. 对于候选页面 x, 总奖励 $R(x)$ 由基础质量、潜力奖励和正则化惩罚 3 部分构成:

$$R(x) = \underbrace{\sum_{i=1}^{4} w_i S_{L_i}(x)}_{\text{基础质量}} + \underbrace{\beta \mathbb{I}_{\text{pot}}(x)}_{\text{潜力奖励}} - \underbrace{[\mathcal{P}_{\text{soft}}(x) + \mathcal{P}_{\text{hack}}(x)]}_{\text{正则化惩罚}} \tag{3}$$

其中, $S_{L_i}(x)$ 为第 i 层评估器对页面 x 的得分; w_i 为相应权重, 权重向量设为 $\boldsymbol{w} = [0.1,\ 0.2,\ 0.3,\ 0.4]^{\mathrm{T}}$, 其中视觉层 L3 和交互层 L4 具有较高权重; β 为潜

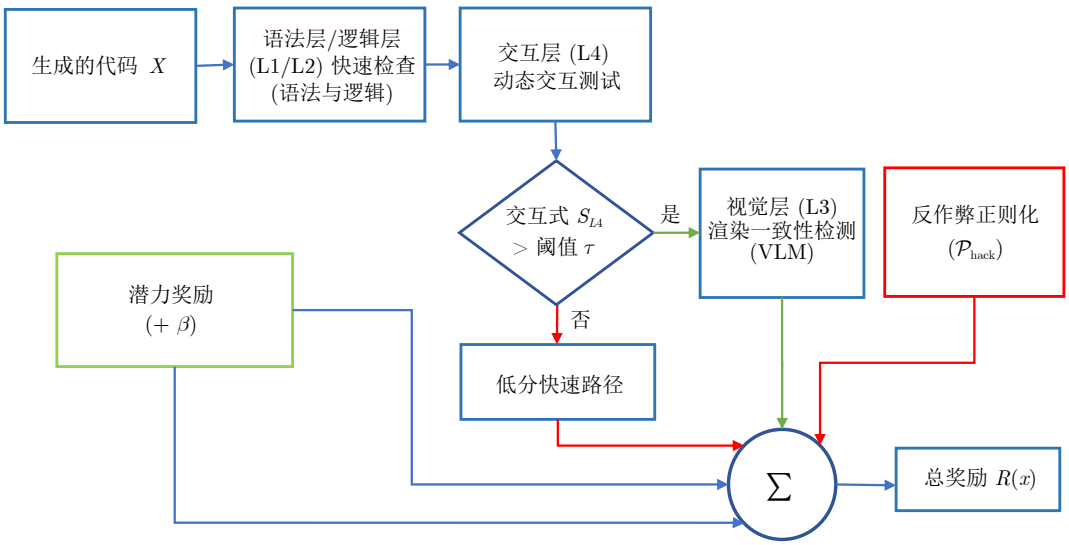


图 5　多模态奖励模型计算流程图

Fig. 5　Computational flowchart of the multimodal reward model

力奖励系数, 设为 0.2; $\mathbb{I}_{\text{pot}}(x)$ 为交互潜力指示函数; $\mathcal{P}_{\text{soft}}(x)$ 和 $\mathcal{P}_{\text{hack}}(x)$ 分别为软惩罚项和反作弊惩罚项.

3.4.1 隐式交互引导

针对初期模型倾向于生成静态页面、L4 交互分为 0 的冷启动问题, 引入潜力奖励. 利用 AST 解析器检测代码中是否包含潜在的交互元素, 如 `canvas` 和 `onclick`. 交互潜力指示函数定义为

$$\mathbb{I}_{\text{pot}}(x)=\begin{cases}1, & N_{\text{event}}>0 \vee \text{hasTag}(x,\ \texttt{canvas})\\ & \quad =1\\ 0, & \text{其他}\end{cases} \tag{4}$$

其中, N_{event} 为代码中的事件监听数量; $\text{hasTag}(\cdot)$ 用于判断页面是否包含指定标签.

3.4.2 软惩罚与反作弊机制

为避免硬门控导致的梯度消失, 采用线性软惩罚项 $\mathcal{P}_{\text{soft}}$ 处理低交互样本:

$$\mathcal{P}_{\text{soft}}(x)=\lambda_1 \max\left[0,\ \tau_{L_4}-S_{L_4}(x)\right] \tag{5}$$

高频交互并不必然意味着更好的学习支持. 根据多媒体学习理论中的一致性原则, 与知识点建构无关的控件、动画或状态变化可能增加外在认知负荷. 因此, 当候选页面的交互响应得分较高而视觉呈现质量较低时, 将其视为潜在的无效交互或奖励投机样本. 反作弊惩罚项定义为

$$\mathcal{P}_{\text{hack}}(x)=\lambda_2 \mathbb{I}[S_{L_4}(x)>80 \wedge S_{L_3}(x)<40] \tag{6}$$

其中, $\tau_{L_4}=60$ 为交互及格线; λ_1 和 λ_2 为惩罚系数; 反作弊阈值 $S_{L_4}(x)>80$ 和 $S_{L_3}(x)<40$ 为经验超参数; $\mathbb{I}(\cdot)$ 为指示函数, 条件成立时取 1, 否则取 0.

3.5 模型微调与强化学习策略

采用“SFT 冷启动 + 强化学习优化”的两阶段训练范式, 全流程基于 Qwen3-30B-A3B-Coder 架构进行优化.

3.5.1 阶段一: SFT 冷启动

选用 Qwen3-30B-A3B-Coder[12] 作为基座模型. 该模型采用高稀疏度的 MoE 架构, 总参数量为 30.5 B, 每次推理激活约 2.4 B 个参数. 利用 $\mathcal{D}_{\text{SFT}}$ 进行全参数微调, 激活模型内部特定的教育代码专家, 使其具备基础的 HTML 结构生成能力, 并为强化学习提供低熵初始策略 π_{SFT}. 该阶段不直接最大化交互质量, 而是为强化学习提供稳定的代码生成分布, 降低直接强化学习容易出现的无效探索风险.

3.5.2 阶段二: 组相对策略优化

在冷启动模型的基础上, 利用 $\mathcal{D}_{\text{RL}}$ 进行强化学习, 采用 GRPO 算法[15]. 与近端策略优化 (proximal policy optimization, PPO) 相比, GRPO 不需要价值网络, 而是通过组内相对优势估计基线, 从而降低训练 MoE 模型时的显存占用. 对同一教学指令采样多个候选 HTML 程序, 并通过级联评估获得其在语法、逻辑、视觉和交互方面的相对质量差异, 与 GRPO 的组内相对优势估计相适配.

为防止模型在探索过程中发生模式坍塌, 在目标函数中引入 Kullback-Leibler (KL) 散度正则项:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta)=\mathbf{E}\Bigg\{&\frac{1}{G}\sum_{i=1}^{G}\min\Big[\rho_i A_i,\\ &\text{clip}(\rho_i,\ 1-\varepsilon,\ 1+\varepsilon)A_i\Big]-\\ &\beta_{\text{KL}}\text{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Bigg\}\end{aligned} \tag{7}$$

其中, 期望 $\mathbf{E}$ 对 $q\sim P(\mathcal{Q})$ 和 $\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)$ 求取, $\mathcal{Q}$ 为教学指令集合, q 为采样指令, o_i 为第 i 个候选程序, G 为每组候选数量, $\pi_{\theta_{\text{old}}}$ 为旧策略; π_θ 为当前策略; $\rho_i=\pi_\theta(o_i|q)/\pi_{\theta_{\text{old}}}(o_i|q)$ 为重要性采样比率; A_i 为第 i 个候选的组内相对优势; ε 为截断阈值; β_{KL} 为 KL 惩罚系数; π_{ref} 为参考策略; $\text{clip}(\cdot)$ 为截断函数; $\text{D}_{\text{KL}}(\cdot\|\cdot)$ 表示 Kullback-Leibler 散度. 实验中设置 $G=16$, $\varepsilon=0.2$, $\beta_{\text{KL}}=0.05$. 策略模型与评估模型的配置见表 1.

表 1 模型配置

Table 1 Model configurations

角色	模型名称	总参数/激活参数	任务
策略模型	Qwen3-30B-A3B-Coder	30.5 B/2.4 B	生成交互式代码
评估模型	Qwen3-VL-235B-A22B-Instruct	235 B/22 B	多模态联合评分

4 实验与结果分析

本节通过多维度评估验证所提方法在交互式教学内容生成任务上的有效性. 首先, 构建涵盖多学科和多认知层级的独立测试基准 Edu-Eval, 并设计包括代码语法、科学与教学逻辑、视觉渲染和交互响应的 4 层评估指标. 然后, 将 IE-code 与多种开源和闭源模型进行量化比较, 并通过消融实验分析 SFT 冷启动、GRPO 强化学习、直接强化学习路径和训练数据规模的影响. 此外, 采用小样本人类偏好评估初步检验自动评估结果与人工判断的一致趋势, 并结合物理教学案例分析不同模型的行为差异和所提方法的局限性.

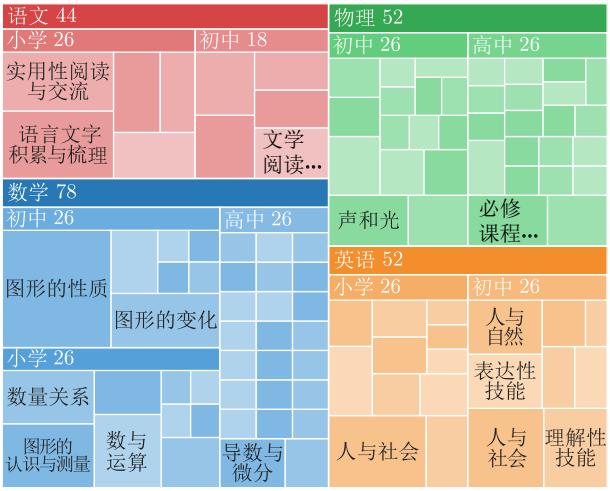


图 6 Edu-Eval 基准测试集的学科、学段与知识点分布

Fig. 6 Subject, grade-band, and knowledge-point distribution of the Edu-Eval benchmark test set

4.1 实验设置

4.1.1 评估数据集

为系统评估模型在 K-12 教育场景下生成交互式内容的能力, 构建了基准测试集 Edu-Eval. 该基准包含 226 条高质量指令, 所有模型均在相同的测试集和评估流程下进行比较. 如图 6 所示, 数据集覆盖数学、语文、英语和物理 4 个学科, 样本涉及小学、初中和高中 3 个学段. 根据认知深度设置不同层级的交互任务, 既包括概念展示类任务, 也包括参数控制、状态更新和多步骤交互任务. 按“学科–学段–知识点”统计, 测试集包含 9 个实际学科–学段组合和 103 个叶子类别; 对同一学科内跨学段重复的知识点名称去重后, 共覆盖 90 个知识点类别. 图 6 中矩形面积表示样本数量, 最底层小矩形对应“学科–学段–知识点”类别, 仅在面积较大的矩形中标注具体类别名称.

1) 学科覆盖: 数据集覆盖数学、语文、英语和物理等 K-12 核心学科, 既包含适合动态仿真的理科概念, 也包含语言理解、情境表达与知识讲解类任务;

2) 难度分层: 指令按认知深度划分为基础概念演示、参数探究仿真和综合逻辑推演 3 个层级, 用于考察模型生成简单动画和处理多参数耦合等复杂交互逻辑的能力.

4.1.2 基准模型

将 IE-code 与以下 4 种基准模型进行比较:

1) IE-code: 基于 Qwen3-30B-A3B-Coder, 使用 10 000 条高质量数据经 SFT 和 GRPO 两阶段训练得到的最终模型;

2) DeepSeek-V3.2: 面向通用编程和推理任务的开源模型;

3) Gemini 3 Pro: 具有多模态理解和逻辑推理能力的闭源模型;

4) Qwen3-VL-235B-A22B-Instruct: 与基座模型同系列的大规模指令模型;

5) Qwen3-30B-A3B-Coder: 未经领域微调的基座模型.

4.1.3 评估指标

采用前文定义的 4 层多模态评估体系, 从语法正确性、教学逻辑一致性、视觉渲染质量和交互响应 4 个维度进行评价. 各层分别计算独立得分, 并按照预设比例对 L1 ~ L4 的得分进行加权汇总, 得到模型的总体评价得分 (总分).

1) 语法层 (L1): 主要评估生成 HTML 代码的语法正确性与可执行性. 首先利用离线静态代码校验工具对 HTML、JavaScript 等代码进行规则检查, 获得确定性的代码校验结果; 随后将代码及校验结果输入大语言模型进行二次语义校验, 以识别不同代码实现方式所带来的合理差异, 避免因静态规则过严而误判实际可正常运行的代码. 综合两阶段校验结果得到 L1 得分.

2) 逻辑层 (L2): 主要评估生成内容在科学原理和逻辑上的准确性. 评估模型结合用户输入的教学要求、页面内容及相关学科知识, 对生成结果是否符合基本科学规律进行综合判断. 例如, 数学内容应符合相应公式和推导关系, 物理仿真中的运动状态与参数变化应与对应物理规律一致, 避免出现视觉表现、文字讲解与科学原理相互矛盾的情况, 并据此得到 L2 得分.

3) 视觉层 (L3): 主要评估页面的渲染质量与布局合理性. 利用 VLM 对实际渲染后的页面进行评价, 综合考察页面整体视觉效果、元素布局、信息呈现清晰度以及是否存在遮挡、错位等问题, 得到 L3 得分.

4) 交互层 (L4): 主要评估页面交互响应的覆盖率与正确性. 采用 Playwright 模拟用户点击、拖动等交互操作, 并检测操作前后 HTML 页面中的状态、DOM 或 Canvas 等是否产生有效变化; 同时结合 VLM 对交互前后的页面状态进行判断, 综合评价交互行为是否真实生效及其反馈是否符合预期, 得到 L4 得分. 60 分为交互层的最低有效阈值, 当 L4 得分低于该阈值时, 认为生成页面未满足基本交互要求, 并将该结果判定为不合格.

4.2 主实验结果

4.2.1 综合性能对比

图 7 展示了总体得分对比, 表 2 列出了各维度的详细结果. 实验结果表明:

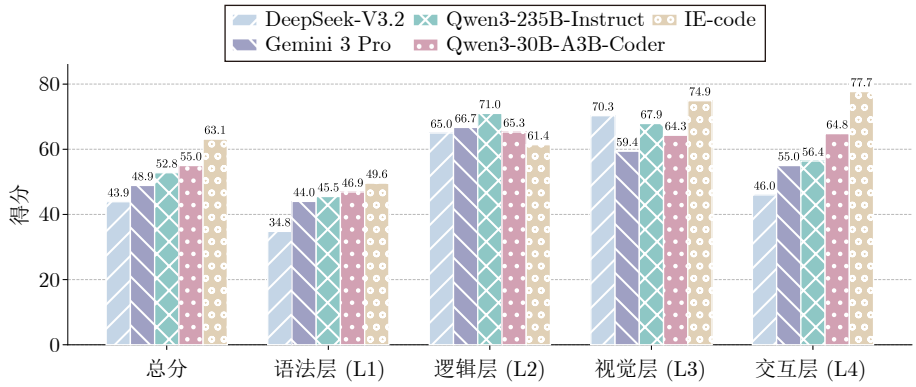


图 7 不同模型的主实验结果

Fig. 7 Main experimental results of different models

表 2 不同模型在 Edu-Eval 测试集上的性能对比

Table 2 Performance comparison of different models on the Edu-Eval test set

模型	总分	语法层 (L1)	逻辑层 (L2)	视觉层 (L3)	交互层 (L4)
DeepSeek-V3.2	43.9	34.8	65.0	70.3	46.0
Gemini 3 Pro	48.9	44.0	66.7	59.4	55.0
Qwen3-VL-235B-A22B-Instruct	52.8	45.5	71.0	67.9	56.4
Qwen3-30B-A3B-Coder	55.0	46.9	65.3	64.3	64.8
IE-code	63.1	49.6	61.4	74.9	77.7

1) 交互层 (L4): IE-code 得分为 77.7, 较 DeepSeek-V3.2 的 46.0 提高 68.9%. 这表明 GRPO 隐式交互引导能够改善模型处理网页动态事件监听和画布实时渲染的能力.

2) 视觉层 (L3): IE-code 得分为 74.9, 高于其他对比模型. 将多模态评估引入奖励计算有助于优化前端布局并减少界面元素遮挡等渲染缺陷.

3) 逻辑层 (L2): IE-code 得分为 61.4, 低于 Qwen3-VL-235B-A22B-Instruct 等大规模通用模型. 所提模型的主要增益集中在视觉呈现与交互响应能力, 复杂科学逻辑和跨学科知识表达仍受模型规模与训练数据覆盖范围限制.

4.2.2 多维能力结构分析

为直观展示模型能力分布, 图 8 给出了多维能力雷达图.

通用模型在语法和知识逻辑方面具有一定优势, 但在视觉呈现与运行时交互方面相对较弱; IE-code 在视觉层 L3 和交互层 L4 上表现更突出. 这说明面向交互式教学 HTML 的数据集构建与多模态对齐能够增强模型在页面渲染、状态更新和交互反馈等任务上的能力, 但复杂知识逻辑和跨学科概念准确性仍有提升空间.

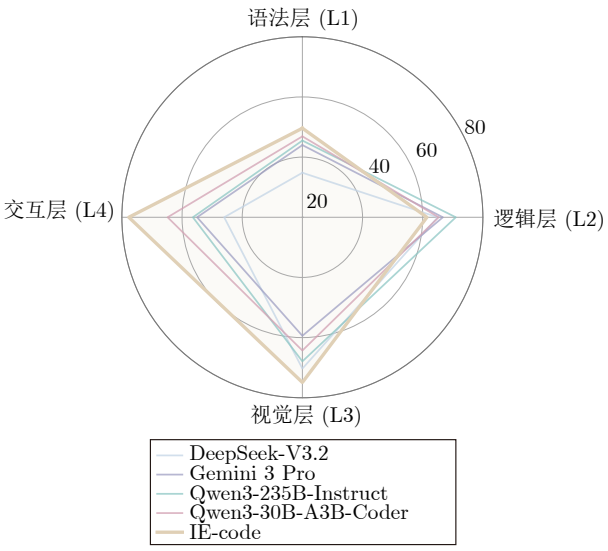


图 8 模型能力多维评估雷达图

Fig. 8 Radar chart of multidimensional model capabilities

4.3 消融实验

为验证“SFT 冷启动 + 强化学习优化”策略和数据规模的影响, 进行消融实验, 结果见表 3. 其中, 基座模型表示未经过训练的模型, 即 Qwen3-30B-A3B-Coder; Direct-RL (无 SFT) 表示对基座模型直接进行 RL, 没有先 SFT; 5k/10k-SFT/RL 分别代表数据规模和训练路径; IE-code-5k-SFT 表示基座模型使用 5 k 数据进行 SFT; IE-code-5k-RL 表

表 3　不同训练阶段与数据规模的消融实验结果
Table 3　Ablation study results on different training stages and data scale

模型	总分	视觉层 (L3)	交互层 (L4)
基座模型	55.0	64.3	64.8
Direct-RL (无 SFT)	53.9	68.9	62.2
IE-code-5k-SFT	47.2	57.3	58.7
IE-code-5k-RL	57.6	72.0	70.7
IE-code-10k-SFT	46.3	54.9	58.3
IE-code-10k-RL	63.1	74.9	77.7

示在 5 k 数据的 SFT 模型基础上进行 RL; 总分表示模型在 Edu-Eval 评测集上的结果.

消融结果反映了不同训练配置的性能差异. SFT 阶段主要提供格式遵循、代码结构和基础可运行性的冷启动, 强化学习阶段则依据级联评估反馈进一步优化视觉呈现和交互响应. 若跳过 SFT 直接进行强化学习, 模型需要在较大的 HTML 和 JavaScript 搜索空间中探索有效交互逻辑, 容易出现局部最优或无效探索.

1) SFT 冷启动与强化学习的互补性: IE-code-10k-SFT 的总分为 46.3, 低于基座模型的 55.0. 经过强化学习后, IE-code-10k-RL 的总分提高至 63.1, 视觉和交互指标较 SFT 阶段分别提高 20.0 和 19.4. 结果说明 SFT 提供了较稳定的初始策略分布, 强化学习能够进一步利用运行时反馈提升多模态对齐能力.

2) 直接强化学习的局限性: 跳过 SFT 冷启动直接进行 GRPO 训练时, Direct-RL (无 SFT) 的视觉得分为 68.9, 但交互层得分和总分分别为 62.2 和 53.9, 均低于基座模型. 这说明缺少 SFT 初始化时, 强化学习较难在代码生成空间中有效探索复杂交互逻辑.

3) 训练数据规模的影响: IE-code-5k-RL 和 IE-code-10k-RL 的总分分别为 57.6 和 63.1. 随训练数据量增加, 总分提高 5.5, 交互得分达到 77.7, 表明扩大高质量交互指令规模有助于提高模型的领域泛化能力.

4.4　小样本人类评估

为检验自动评估结果与人工判断是否具有一致趋势, 从 Edu-Eval 测试集中选取覆盖 4 个学科的 20 个代表性样本, 每个学科包含 5 个样本, 比较 IE-code 与 Qwen3-30B-A3B-Coder 生成的交互式教学页面. 评审采用匿名 A/B 偏好形式, 系统随机打乱两个候选页面的左右顺序并隐藏模型名称. 评审者从教学目标一致性、概念解释清晰度、交互有效性、视觉可用性和整体教学适配性等方面进行综合判断, 界面选项为“A 更好”、“B 更好”、“差不多”、“都不合格”、“跳过”; 解盲后, “A 更好”和“B 更好”按随机呈现记录转换为表 4 中以模型名称表示的偏好标签.

实验邀请 20 余名具有教育、人工智能或教学内容设计背景的评审者参与. 对每个测试样本收集多名评审者的匿名判断, 评审结束后依据该样本的随机呈现记录, 将“A 更好”和“B 更好”分别映射为对应模型名称, 再采用多数投票得到表 4 所列的样本级偏好标签. 设第 i 个样本中 IE-code 和 Qwen3-30B-A3B-Coder 获得的有效偏好票数分别为 v_i^{IE} 和 v_i^{Qwen}, 则样本偏好标签 y_i 定义为

$$y_i = \begin{cases} \text{IE}, & v_i^{\mathrm{IE}} > v_i^{\mathrm{Qwen}} \\ \text{Qwen}, & v_i^{\mathrm{Qwen}} > v_i^{\mathrm{IE}} \\ \text{other}, & \text{其他} \end{cases} \tag{8}$$

其中, y_i 为第 i 个样本经多数投票得到的样本级标签, 取值为 IE、Qwen 或 other; IE 表示 IE-code; Qwen 表示 Qwen3-30B-A3B-Coder; other 表示包含票数相同、都不合格和跳过.

IE-code 的样本级胜率为

$$r_{\mathrm{win}}^{\mathrm{IE}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i = \text{IE}) \tag{9}$$

其中, N 为样本总数, 本实验 $N = 20$.

若仅统计具有明确模型偏好的样本, 则相对偏好胜率为

$$\tilde{r}_{\mathrm{win}}^{\mathrm{IE}} = \frac{\sum_{i=1}^{N} \mathbb{I}(y_i = \text{IE})}{\sum_{i=1}^{N} \mathbb{I}(y_i = \text{IE}) + \sum_{i=1}^{N} \mathbb{I}(y_i = \text{Qwen})} \tag{10}$$

表 4 给出了 20 个样本经过多数投票后的样本级结果. “IE-code 更好”在 13 个样本中被评为更优, 样本级胜率为 $13/20 = 65.0\%$; “Qwen3-30B-A3B-Coder 更好”在 3 个样本中被评为更优, 占 15.0%.

表 4　匿名 A/B 人类偏好评估结果
Table 4　Anonymous A/B human preference evaluation results

评价结果	样本级结果数	比例
IE-code 更好	13	65.0%
Qwen3-30B-A3B-Coder 更好	3	15.0%
差不多	2	10.0%
都不合格	1	5.0%
跳过	1	5.0%

若仅考虑具有明确模型偏好的样本, IE-code 的相对偏好胜率为 $13/(13+3)=81.25\%$. 为保持统计口径保守, 主要报告基于全部 20 个样本的样本级胜率.

人工评估中 IE-code 在多数样本上获得更高偏好, 该趋势与自动评估中 L3 和 L4 得分的提高基本一致. 这说明自动评估体系能够在一定程度上反映生成内容的技术可用性与初步教学适配性. 该实验仍属于小样本验证, 真实课堂学习效果还需通过更大规模、长周期的教学实验评估.

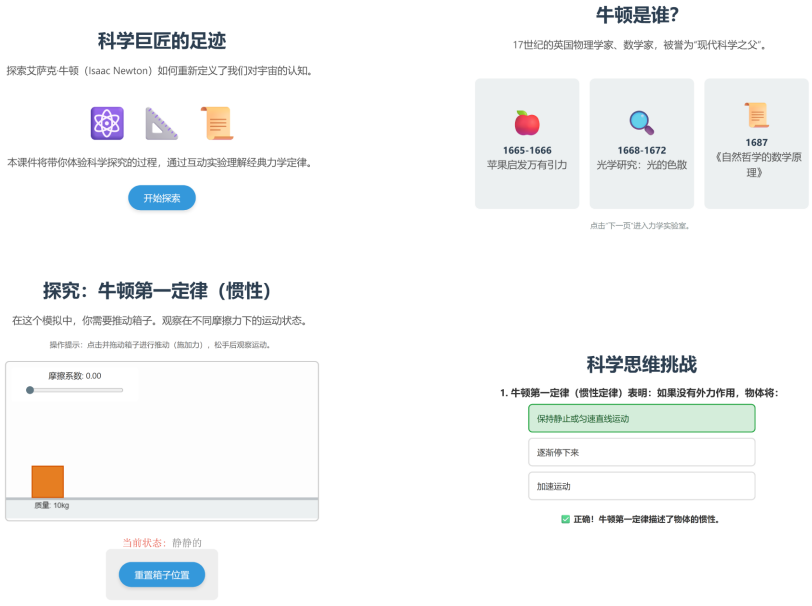
4.5 案例分析

为展示不同模型生成交互式教学内容时的行为差异, 选取“牛顿与经典力学”场景进行定性比较.

如图 9 所示, 面对“设计一个关于牛顿力学与发现的交互式探究课件”这一指令时, 两种模型的表现存在明显差异:

1) 基线模型: 生成结果以纯文本列表和静态信息框为主, 如图 9(a) 所示. 页面包含牛顿生平时间轴与三大运动定律, 但缺少动态物理仿真和交互测

(a) 基线模型生成结果
(a) Result generated by the baseline model



(b) IE-code 生成结果
(b) Result generated by IE-code

图 9 “牛顿与经典力学”交互式教学内容生成案例

Fig. 9 Case of interactive teaching content generation for Newton and classical mechanics

试, 未能提供教学场景所需的动态探究功能.

2) IE-code: 生成了包含多步引导的交互式探究应用, 如图 9(b) 所示. 该应用包含以下功能:

a) 流程控制与状态管理: 利用前端状态管理实现多个页面的逻辑串联, 通过“上一步/下一步”组件切换科学史和物理定律模块.

b) 仿真控制与评估反馈: 构建动态物理实验, 支持通过滑动条调整摩擦系数并观察运动状态变化; 同时生成基于状态判断的选择题组件, 形成“操作 $\rightarrow$ 状态更新 $\rightarrow$ 视觉验证”的反馈闭环.

该案例表明, 多模态对齐训练有助于模型由静态知识陈述转向动态交互式学习任务. 案例仅展示页面生成和运行时交互能力, 不能直接等同于真实学习效果的提升.

4.6 局限性与失败模式分析

尽管 IE-code 在总体评分、视觉呈现和交互响应方面取得较好结果, 低分样本和人工抽检结果仍反映出若干失败模式. 首先, 在复杂物理仿真或多参数耦合任务中, 模型可能生成可运行且视觉上合理的页面, 但数值更新公式、边界条件或参数单位仍可能存在偏差. 此类错误难以仅通过语法执行和首尾帧变化发现, 需要结合符号规则、领域知识库或专家审查.

其次, 在人文社会科学或语言学习任务中, 模型有时会复用理科仿真中的按钮、滑块和导航卡片等交互模板, 使交互形式与学科活动目标不匹配. 最后, 自动评估器也可能出现误判. 多模态模型可能受页面美观度或显著状态变化影响而高估教学适配性; 首尾关键帧采样可能漏检中间动画的短时错误或交互路径断裂.

4.7 实验小结

综合主实验、消融实验、小样本人类评估和案例分析结果, IE-code 的主要优势集中在视觉呈现、交互响应和教学页面结构化生成方面. 面向教育场景的数据集构建、多模态级联评估、SFT 与强化学习两阶段对齐能够提高交互式 HTML 教学材料的技术可用性. 人工偏好趋势与自动评估中 L3 和 L4 得分的提高基本一致.

IE-code 在逻辑层 L2 上仍未全面超过大规模通用模型, 复杂科学逻辑、跨学科知识迁移和隐性概念错误仍需重点改进. 视觉可用性和交互可达性的提高并不保证深层教学逻辑完全可靠, 后续需要结合符号规则校验、领域专家审查和真实课堂反馈提升模型与评估体系的可靠性.

5 结束语

针对现有教育大语言模型在动态交互式内容生成中存在的输出形态偏静态、运行时交互能力不足以及多模态质量评估困难等问题, 提出了一种面向交互式教学 HTML 生成的大语言模型构建与自动化评测方法. 所提方法将该任务建模为同时受代码可执行性、视觉可渲染性、运行时交互响应与教学目标一致性约束的闭环优化问题, 并围绕这一建模构建了 Edu-Interact 数据集、“SFT 冷启动 + GRPO 强化学习”的两阶段训练策略, 以及基于 Qwen3-VL-235B-A22B-Instruct 的多模态级联评估体系. 实验结果表明, IE-code 在交互式 HTML 教学材料生成中展现出较好的视觉呈现、交互响应和综合生成能力; 消融实验进一步验证了 SFT 冷启动、GRPO 强化学习、数据规模扩展和多维奖励设计对模型性能提升的作用.

尽管 IE-code 在 HTML 页面生成、交互代码生成和自动化评测方面取得了较明显提升, 但仍需指出, 自动评估指标主要反映生成内容的技术可用性与初步教学适配性, 包括代码能否运行、页面能否渲染、交互是否可达以及内容是否基本服务于教学目标. 这些指标是交互式教学材料进入真实教学场景前的重要前提, 但不能直接等同于真实学习成效. 更多交互元素并不必然带来更好的学习收益, 过度复杂的视觉呈现也可能增加学习者外在认知负荷. 因此, 所生成的交互式 HTML 教学资源在课堂中的实际教学效果, 仍需要在更长期、更真实的教学环境中结合学习者体验和学习结果进行验证.

未来研究可从三个方向继续推进. 第一, 结合真实教学应用开展长期验证, 包括学习者前后测、知识迁移测试、主观认知负荷量表、课堂行为数据和教师反馈等, 以系统评估生成内容对真实学习效果的影响; 第二, 引入领域知识库、符号规则校验、专家审查和学科化数据增强, 以进一步提升模型在复杂概念表达、多参数耦合任务和跨学科教学场景中的逻辑可靠性与迁移稳定性; 第三, 探索多教师模型协同合成、人工专家抽检、多风格数据增强和更细粒度的时序交互评估机制, 以降低数据合成与自动评估中的偏置风险; 同时, 后续也可拓展草图到代码、教材插图到交互课件等多模态输入形式, 并结合模型量化与轻量化部署, 推动交互式教学内容生成在平板、学习机等教育终端中的实际应用.

参考文献

1 Zhao W X, Zhou K, Li J Y, Tang T Y, Dong Z C, Hou Y P, et

al. A survey of large language models. *Frontiers of Computer Science*, 2026, **20**: Article No. 2012627

2 Google DeepMind. A new era of intelligence with Gemini 3 [Online], available: https://blog.google/products-and-platforms/products/gemini/gemini-3/, July 29, 2026

3 Yang A, Li A F, Yang B S, Zhang B C, Hui B Y, Zheng B, et al. Qwen3 technical report. arXiv preprint arXiv: 2505.09388, 2025.

4 Chen Z, Liu T Q, Tian M, Tong Q, Luo W Q, Liu Z T. Advancing math reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. arXiv preprint arXiv: 2501.14002, 2025.

5 Sweller J, van Merriënboer J J G, Paas F. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 2019, **31**(2): 261−292

6 Skulmowski A, Xu K M. Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review*, 2022, **34**(1): 171−196

7 Risko E F, Gilbert S J. Cognitive offloading. *Trends in Cognitive Sciences*, 2016, **20**(9): 676−688

8 Rutten N, van Joolingen W R, van der Veen J T. The learning effects of computer simulations in science education. *Computers & Education*, 2012, **58**(1): 136−153

9 Schnotz W, Rasch T. Enabling, facilitating, and inhibiting effects of animations in multimedia learning: Why reduction of cognitive load can have negative results on learning. *Educational Technology Research and Development*, 2005, **53**(3): 47−58

10 Wang X Y, Chen Y Y, Yuan L F, Zhang Y Z, Li Y Z, Peng H, et al. Executable code actions elicit better LLM agents. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: Proceedings of Machine Learning Research, 2024, **235**: 50208−50232

11 DeepSeek-AI, Zhu Q H, Guo D Y, Shao Z H, Yang D J, Wang P Y, et al. DeepSeek-Coder-V2: Breaking the barrier of closed-source models in code intelligence. arXiv preprint arXiv: 2406.11931, 2024.

12 Qwen Team. Qwen3-Coder: Agentic coding in the world [Online], available: https://qwenlm.github.io/blog/qwen3-coder/, July 29, 2026

13 Si C L, Zhang Y Z, Li R, Yang Z Y, Liu R B, Yang D Y. Design2Code: Benchmarking multimodal code generation for automated front-end engineering. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Albuquerque, USA: Association for Computational Linguistics, 2025. 3956−3974

14 Yang J, Prabhakar A, Narasimhan K, Yao S Y. InterCode: Standardizing and benchmarking interactive coding with execution feedback. *Advances in Neural Information Processing Systems*, 2023, **36**: 23826−23854

15 Shao Z H, Wang P Y, Zhu Q H, Xu R X, Song J X, Bi X, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv: 2402.03300, 2024.

16 Qwen Team. Qwen3-VL-235B-A22B-Instruct [Online], available: https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct, July 29, 2026

17 Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 2023, **103**: Article No. 102274

18 Kohnke L, Moorhouse B L, Zou D. ChatGPT for language teaching and learning. *RELC Journal*, 2023, **54**(2): 537−550

19 Trinh T H, Wu Y H, Le Q V, He H, Luong T. Solving olympiad geometry without human demonstrations. *Nature*, 2024, **625**(7995): 476−482

20 Hui B Y, Yang J, Cui Z Y, Yang J X, Liu D Y H, Zhang L, et al. Qwen2.5-Coder technical report. arXiv preprint arXiv: 2409.12186, 2024.

21 Zhou S Y, Xu F F, Zhu H, Zhou X H, Lo R, Sridhar A, et al. WebArena: A realistic web environment for building autonomous agents. arXiv preprint arXiv: 2307.13854, 2023.

22 Meng Y, Xia M Z, Chen D Q. SimPO: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 2024, **37**: 124198−124235

23 Zhang K C, Li G, Li J, Dong Y H, Li J, Jin Z. Focused-DPO: Enhancing code generation through focused preference optimization on error-prone points. In: Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: Association for Computational Linguistics, 2025. 9578−9591

24 Sheng G M, Zhang C, Ye Z L F, Wu X B, Zhang W, Zhang R, et al. HybridFlow: A flexible and efficient RLHF framework. In: Proceedings of the Twentieth European Conference on Computer Systems. Rotterdam, The Netherlands: Association for Computing Machinery, 2025. 1279−1297

25 Rodriguez J, Zhang H T, Puri A, Pramanik R, Feizi A, Wichmann P, et al. Rendering-aware reinforcement learning for vector graphics generation. *Advances in Neural Information Processing Systems*, 2025, **38**: 60496−60534

26 Gao L, Schulman J, Hilton J. Scaling laws for reward model overoptimization. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: Proceedings of Machine Learning Research, 2023, **202**: 10835−10866

27 Zheng L M, Chiang W L, Sheng Y, Zhuang S Y, Wu Z H, Zhuang Y H, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot arena. *Advances in Neural Information Processing Systems*, 2023, **36**: 46595−46623

28 Madaan A, Tandon N, Gupta P, Hallinan S, Gao L Y, Wiegreffe S, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 2023, **36**: 46534−46594

29 Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S Y. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 2023, **36**: 8634−8652

30 Mayer R E, Fiorella L. Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. *The Cambridge Handbook of Multimedia Learning*. New York: Cambridge University Press, 2014. 279−315

张若华 华东师范大学上海智能教育研究院、计算机科学与技术学院硕士研究生. 主要研究方向为教育大语言模型和模型高效微调.
E-mail: 51275901086@stu.ecnu.edu.cn
(**Zhang Ruo-Hua** Master student at the Shanghai Institute of AI for Education and the School of Computer Science and Technology, East China Normal University. His research interests include educational large language models and model efficient fine-tuning.)

刘 涛 华东师范大学上海智能教育研究院、计算机科学与技术学院硕士研究生. 主要研究方向为教育大语言模型和模型评测.
E-mail: 51275901030@stu.ecnu.edu.cn
(**Liu Tao** Master student at the

Shanghai Institute of AI for Education and the School of Computer Science and Technology, East China Normal University. His research interests include educational large language models and model evaluation.)

卢 烨 华东师范大学上海智能教育研究院博士研究生. 主要研究方向为教育大语言模型, 强化学习和演化计算.
E-mail: 51275901116@stu.ecnu.edu.cn
(**Lu Ye** Ph.D. candidate at the Shanghai Institute of AI for Education, East China Normal University. His research interests include educational large language models, reinforcement learning, and evolutionary computation.)

宋思宇 华东师范大学上海智能教育研究院、计算机科学与技术学院博士研究生. 主要研究方向为教育大语言模型, 强化学习和自进化.
E-mail: siyusong00@gmail.com
(**Song Si-Yu** Ph.D. candidate at the Shanghai Institute of AI for Education and the School of Computer Science and Technology, East China Normal University. His research interests include educational large language models, reinforcement learning, and self-evolution.)

刘文涛 华东师范大学上海智能教育研究院、计算机科学与技术学院与上海创智学院联合培养博士研究生. 主要研究方向为超长上下文大语言模型和多模态大模型.
E-mail: wtliu@stu.ecnu.edu.cn
(**Liu Wen-Tao** Ph.D. candidate jointly trained by the Shanghai Institute of AI for Education and the School of Computer Science and Technology at East China Normal University, and Shanghai Innovation Institute. His research interests include ultra-long-context large language models and multimodal large models.)

宋 宇 华东师范大学上海智能教育研究院研究员. 主要研究方向为智能课堂教学评价, 人机交互和多智能体协同.
E-mail: ysong@mail.ecnu.edu.cn
(**Song Yu** Researcher at the Shanghai Institute of AI for Education, East China Normal University. Her research interests include intelligent classroom teaching evaluation, human-computer interaction, and multi-agent collaboration.)

周爱民 华东师范大学上海智能教育研究院、计算机科学与技术学院教授, 上海创智学院全时导师. 主要研究方向为大语言模型, 智能体系统和演化优化与学习.
E-mail: amzhou@cs.ecnu.edu.cn
(**Zhou Ai-Min** Professor at the Shanghai Institute of AI for Education and the School of Computer Science and Technology, East China Normal University, and full-time mentor at Shanghai Innovation Institute. His research interests include large language models, agent systems, and evolutionary optimization and learning.)

郝 昊 华东师范大学上海智能教育研究院副研究员. 主要研究方向为教育大语言模型, 智能体设计和无梯度优化. 本文通信作者.
E-mail: hhao@mail.ecnu.edu.cn
(**Hao Hao** Associate researcher at the Shanghai Institute of AI for Education, East China Normal University. His research interests include educational large language models, agent design, and gradient-free optimization. Corresponding author of this paper.)