CounterTutor：基于错因诊断与反事实干预的个性化 AI 教学系统

国刚

青岛开放大学，中国·山东 青岛 266041

摘要：大语言模型已能生成即时解析和个性化反馈，但多数 AI 教学系统仍把学生错误视为待纠正的结果，较少显式建模错误背后的稳定认知规则。本文提出 CounterTutor，一个面向概念修复的错因感知反事实教学框架。系统首先根据题目、学生错误答案、解题过程与历史记录诊断潜在误概念；随后生成数值代入、规则边界、结构展开、表征转换与迁移验证等最小反事实干预，使学生发现自身错误规则的失效条件；最后通过自适应闭环更新错因状态并验证迁移效果。在错因诊断、反馈质量和学习效果三个层面评估的结果显示该框架在诊断准确性、反馈针对性、迁移正确率和误概念复发控制上优于标准解析、相似题推荐和普通 LLM 反馈。本文将 AI 教学目标从“解释正确答案”推进到“修复错误规则”，为可解释、可验证的生成式教育系统提供了一条可扩展路径。

关键词：AI 教育；智能教学系统；错因诊断；误概念建模；反事实教学；概念修复

CounterTutor: A personalized AI teaching system based on error diagnosis and counterfactual interventions

Guo Gang

Qingdao Open University, China Shandong Qingdao 266041

Abstract: Large language models can generate instant parsing and personalized feedback, but most AI teaching systems still treat student errors as results to be corrected, rarely explicitly modeling the stable cognitive rules behind the errors. This paper proposes a counterfactual teaching framework for error cause perception oriented towards concept repair. This framework first diagnoses potential misconceptions based on the question, student's incorrect answer, solution process, and historical records. Then, it generates minimal counterfactual interventions such as numerical substitution, rule boundaries, structural expansion, representation transformation, and transfer verification, enabling students to discover the failure conditions of their own error rules. Finally, it updates the error cause state through adaptive closed-loop and verifies the transfer effect. The results evaluated at three levels—error cause diagnosis, feedback quality, and learning effectiveness—show that this framework outperforms standard parsing, similar question recommendation, and ordinary LLM feedback in terms of diagnostic accuracy, feedback relevance, transfer accuracy, and misconception recurrence control. The proposed method advances the AI teaching objective from explaining correct answers to repairing error rules, providing a scalable path for interpretable and verifiable generative education systems.

Keywords: AI education; Intelligent teaching system; Error cause diagnosis; Misconception modeling; Counterfactual teaching; Concept repair

0 引言

当前 AI 教育系统已经能够自动批改、讲解答案、生成提示和推荐练习，显著降低了学生获得即时反馈的门槛。然而，学生错误通常并非随机出现，而是来自稳定的误概念。例如，将 $(a+b)^2$ 写成 a^2+b^2 表面上是漏掉 2ab，深层上可能是把“指数可分配到乘法”错误迁移到加法结构。若系统只给出正确公式，学生可能在当前题上改正，却在迁移题或延迟测试中再次应用同一错误规则。

已有智能教学系统和知识追踪方法能够预测学生对知识点的掌握程度，但通常只回答“学生会不会”，难以解释“学生为什么错”。误概念诊断把错误选项、题目语义和专家错因描述建立联系，为从知识点级反馈走向认知规则级反馈提供了基础。大语言模型虽然提升了自然语言反馈能力，但直接端到端生成容易出现反馈泛化、数学幻觉和一次性给答案等问题。

本文主张，AI 教学应从“答案解释器”转向“认知漏洞定位器”：先识别学生为什么错，再通过最小反事实样例让学生看到原规则在什么条件下失效，最后用迁移题和延迟题验证是否真正修复。基于该思想，本文提出 CounterTutor，并将论文结构整理为引言、方法、实验、讨论与结论五个核心部分。本文贡献包括：提出“错因诊断—反事实干预—修复验证”的概念修复任务；设计可验

证的 CounterTutor 框架；构建诊断、生成与学习效果三层实验方案；通过消融分析验证显式错因建模、最小反事实和质量验证器的作用。

1 方法

1.1 问题定义与整体框架

给定学生 s、题目 q、错误答案 a、可选解题过程 r 与历史记录 h，系统需要估计误概念库 M 中每个候选错因的概率 $p_\theta(m_i \mid q,a,r,h)$，并生成教学干预 I，使学生在后续迁移题中降低同类错误复发。与只追求当前题答对不同，CounterTutor 的目标是修复导致错误的潜在规则，因此最终评价应关注迁移正确率、延迟保持率和误概念复发率。

CounterTutor 包含四个核心模块：错因检索—重排诊断器、学生级错因状态、反事实干预生成器和教学质量验证器。系统流程为：学生提交答案后，诊断器从误概念库中找出最可能错因；学生状态模块融合历史错误模式；生成器围绕目标误概念构造最小反事实样例；验证器过滤数学错误、类比不当或难度失配的反馈；最后根据学生在交互和迁移题中的表现更新错因状态，见表 1。

1.2 错因诊断与学生状态建模

错因诊断采用两阶段结构。首先，将题目、学生答案、学生解释和历史摘要拼接为诊断查询，并用双塔编码器从误概念库中召回 top−K 候选；随后，交叉编码器或 LLM 评分器对候选错因进行精排，判断其是否真正解释当前错误。训练时加入随机负样本、同知识点负样本和 hard negative，以区分语义相近但教学处理不同的错因，例如“忘记二项式中间项”和“错误地将平方分配到加法上”。

学生级错因状态用于避免把单次偶然错误误判为稳定误概念。系统为每个学生保存一个动态误概念分布，并将当前诊断结果与历史错误模式融合。若某误概念在历史中反复出现，系统提高其先验；若学生在迁移题中正确解释规则边界，则降低该误概念概率。实现上，历史摘要只保存脱敏后的知识点、错误类型、时间间隔和修复状态，不保存原始隐私文本。

1.3 反事实干预生成与质量验证

反事实生成器围绕目标误概念生成五类干预：数值代入用于快速暴露矛盾，规则边界用于区分适用条件，结构展开用于显性化隐藏步骤，表征转换用于在图形、表格或程序执行中重建理解，迁移验证用于检验是否真正修复。例如对 $(x+3)^2=x^2+9$，系统先让学生比较 $(1+2)^2$ 与 1^2+2^2，再比较 $(ab)^2$ 与 $(a+b)^2$ 的差异，最后展开 $(x+3)(x+3)$ 并给出 $(x-4)^2$ 迁移题。

教学质量验证器从数学正确性、最小改动性、错因针对性和教学可用性四个维度打分。数学正确性可由计算代数或规则引擎验证，最小改动性要求反事实样例只改变暴露错因所必需的元素，针对性要求样例直接挑战目标误概念，可用性则检查表达难度、学生年级和是否保留思考空间。系统不建议完全依赖端到端大模型，而应采用“轻量检索模型 + 精排模型 + 受约束 LLM 生成 + 符号验证”的组合，见图 1。

表1 CounterTutor 方法结构

模块	输入	核心输出	方法作用
错因诊断器	题目、错误答案、过程、历史	top−K 误概念与置信度	定位“为什么错”
学生状态	历史错因分布与新诊断	动态误概念概率	区分偶然错误与稳定规则
反事实生成器	目标误概念与能力水平	最小反事实样例	暴露错误规则边界
质量验证器	候选干预	通过筛选的反馈	降低幻觉和无效解释
自适应闭环	学生反应与迁移结果	下一步教学策略	验证并巩固概念修复

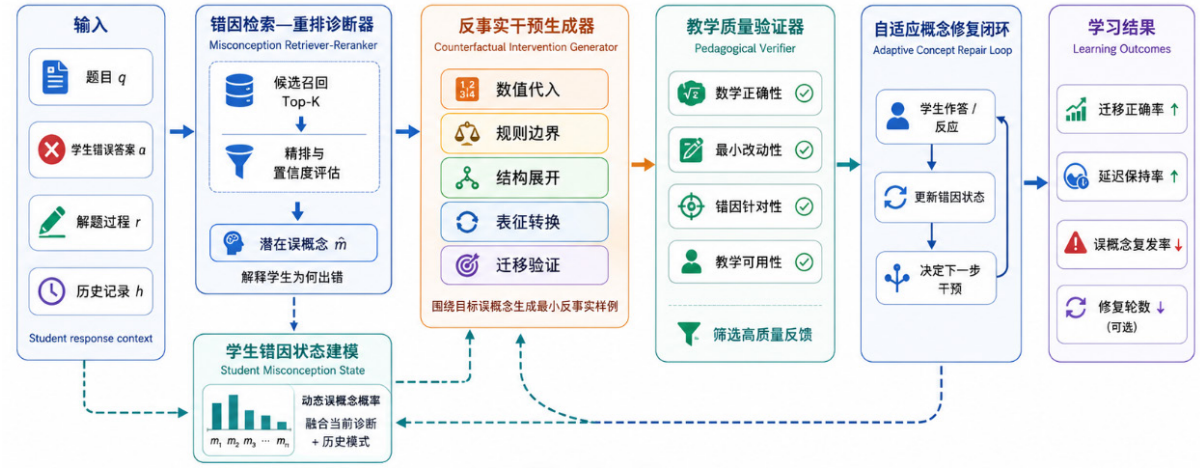

图1 方法结构

2 实验

2.1 数据集、对比方法与评价指标

实验分三层展开。第一层评估错因诊断，使用数学误概念诊断数据，样本包含 question、correct answer、wrong answer、construct 与 misconception 字段，并设置已见题、未见题和未见误概念测试。比较 BM25、Sentence-BERT、LLM zero-shot、LLM CoT、Cross-Encoder、Retriever+Reranker 与 CounterTutor Diagnosis，指标包括 MAP@25、Recall@5、Recall@25 和 MRR。

第二层评估反事实生成质量，构建小规模 Counterfactual Intervention Benchmark，覆盖代数、分式、函数、比例、几何和概率，由数学教师标注目标错因与高质量反事实样例。评价指标包括数学正确率、针对性、最小性、多样性、验证通过率和教师 1-5 分评分。第三层采用随机对照学习实验，学生先完成 pre-test，再随机接受 Answer Only、Standard Explanation、Similar Practice、Generic LLM Feedback、Socratic Hint 或 CounterTutor 干预，随后完成 post-test、3-7 天 delayed test 和迁移题。

2.2 诊断结果与 OOD 泛化

诊断结果显示，关键词 BM25 表现最弱，说明错因诊断不能只依赖表面词匹配；LLM zero-shot 与 CoT 有一定提升，但输出稳定性和细粒度区分不足；Retriever+Reranker 明显优于单阶段方法；CounterTutor Diagnosis 结合学生历史状态与 hard negative 训练后取得最高 MAP@25、Recall 和 MRR，见表 2。

2.3 反馈质量、学习效果与消融

反事实生成实验中，标准解析虽然数学正确率高，但最小性和针对性不足；相似题推荐形式接近原题，却未必暴露真实错因；普通 LLM 反馈表达自然，但存在泛化和计算错误。CounterTutor 通过错因约束与验证器筛选，在正确性、针对性、最小性和教学价值之间取得更均衡表现。学习效果实验进一步显示，普通解析和 LLM 反馈能提升即时得分，但迁移与延迟保持有限；CounterTutor 因直接挑战错误规则，在 Transfer Accuracy、Delayed Retention 和 Recurrence 上优势更明显，见表 3。

消融实验表明，去掉错因诊断下降最明显，说明反事实样例必须围绕具体误概念设计；去掉最小性约束后，反馈会引入过多新概念，学生难以聚焦；去掉验证器后，数学错误和不恰当类比增多；去掉自适应闭环后，即时表现尚可，但延迟保持与复发率变差。这些结果共同支持本文核心观点：概念修复不是更长解释，而是“定位错误规则—制造认知冲突—验证迁移”的连续过程。

3 讨论

3.1 案例分析与机制解释

以“*Expand*（$x+3$）^{2}”为例，学生回答 x^2+9，并解释“我把 x 平方，再把 3 平方”。普通反馈通常直接给出 x^2+6x+9 和公式（$a+b$）$^2=a^2+2ab+b^2$。CounterTutor 则先诊断为“指数错误分配到加法”，再构造最小数值反事实：若（$a+b$）$^2=a^2+b^2$，则（1+2）^{2} 应等于 1^2+2^2，但实际为 9 与 5。学生发现矛盾后，系统继续比较（ab）$^2=a^2b^2$ 与（$a+b$）^{2}

表2 误概念诊断与OOD结果

方法	MAP@25	Recall@5	Recall@25	MRR	Unseen Q	Unseen M
BM25	0.284	0.412	0.623	0.301	–	–
SBERT	0.371	0.536	0.741	0.398	–	–
LLM CoT	0.431	0.591	0.779	0.456	0.392	0.318
Cross-Enc.	0.468	0.633	0.812	0.491	0.421	0.276
Ret.+Rerank	0.512	0.682	0.856	0.537	0.459	0.342
CounterTutor	0.548	0.713	0.881	0.569	0.487	0.376

表3 反馈质量、学习效果与消融结果

条件/变体	针对性	最小性	Transfer	Retention	Recurrence ↓	教学价值
Std. Expl.	0.611	0.318	0.53	0.47	0.39	3.58
Sim. Practice	0.487	0.552	0.55	0.49	0.37	3.29
Generic LLM	0.642	0.463	0.59	0.52	0.34	3.62
Socratic Hint	0.684	0.588	0.62	0.56	0.31	3.89
CounterTutor	0.813	0.746	0.71	0.66	0.22	4.51
w/o Diagnosis	–	–	0.61	0.56	0.33	3.84
w/o Minimality	–	–	0.64	0.59	0.29	4.02
w/o Verifier	–	–	0.63	0.57	0.31	3.76
w/o Loop	–	–	0.65	0.60	0.28	4.11

的规则边界，并通过乘法表展开（$x+3$）（$x+3$），最后用（$x-4$）^{2} 验证迁移。该案例说明，反事实反馈的价值不在于讲得更多，而在于让学生亲自看见原规则崩溃的位置。

反事实比普通解析更适合概念修复，是因为它不是从正确公式出发要求学生记忆，而是从学生错误规则出发制造认知冲突。若学生原规则成立，某个最小变化样例也应成立；一旦样例产生矛盾，学生就必须重新审视原规则的适用边界。显式错因建模决定了“冲突”应该针对哪条规则，最小反事实约束保证学生能把新样例与原题建立联系，闭环验证则区分了短期记忆、题型熟悉和真正的概念修复。

3.2 教学价值、局限与伦理

CounterTutor 对教师的价值在于把个体错题反馈转化为可审阅的教学资源。教师端可以显示诊断置信度、候选错因排序、系统为什么选择某个反事实、学生在每轮交互中的回答以及错因概率变化；课堂层面则可统计高频误概念，自动生成“共性错因讲解卡片”，包括典型错误、最小反事实、边界对比例题和迁移练习。这样既能提升个性化反馈效率，也避免把模型输出直接等同于教学结论。

本文仍存在局限。第一，误概念库质量会限制诊断上限，描述过粗会降低针对性，描述过细会增加标注成本和分类难度。第二，不同年龄、基础能力和学习风格的学生对反事实的反应不同。第三，本文主要讨论数学场景，迁移到物理、化学、编程或医学教育时需要重新设计错因库、表征方式和验证规则。第四，复杂证明题和开放题的自动验证仍然困难。伦理上，误概念标签不应被固化为学生能力评价；学生数据必须匿名化处理，系统不确定时应提供多个可能解释，长期无法修复的问题应提示教师介入。

4 结语

本文提出 CounterTutor，一个基于错因诊断与反事实干预的个性化 AI 教学系统。与传统“错题—解析—练习”流程不同，它关注学生错误背后的潜在规则，并通过最小反事实样例暴露规则边界，再以迁移和延迟测试验证概念修复。实验结果表明，该框架有潜力提升误概念诊断、反馈针对性和长期学习效果。其核心启示是：AI 教育的关键不是让模型更会讲答案，而是让模型更会修复学生脑中的错误规则。

从系统落地角度看，CounterTutor 更适合被设计为教师可审阅的半自动反馈生成器，而不是完全替代教师的自动判分系统。学生端只呈现经过验证的引导问题和反事实样例，教师端则呈现诊断置信度、候选错因排序、干预选择理由、学生每轮回答以及错因概率变化。这样既能保证反馈效率，又能在模型不确定或学生长期无法修复时保留教师干预通道。

后续研究可以沿三条路径继续扩展：第一，跨学科迁移，将反事实干预推广到物理中的力与能量误用、编程中的循环边界与变量作用域错误，以及化学中的守恒关系误解；第二，多模态表征，将几何图形、函数图像、动画和可执行代码纳入反事实生成，使学生能够观察错误规则产生的可视化后果；第三，长期学习建模，跟踪误概念从首次出现、被干预、迁移成功到延迟复发的完整轨迹，从而更准确地区分短期记忆和真正概念修复。

参考文献：

[1] Corbett A. T., Anderson J. R. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 1994.

[2] Piech C., Bassen J., Huang J., et al. Deep Knowledge Tracing. NeurIPS, 2015.

[3] Wang Z., Lamb A., Saveliev E., et al. Diagnostic Questions: The NeurIPS 2020 Education Challenge, 2020.

[4] Eedi. Mining Misconceptions in Mathematics. Kaggle Competition, 2024.

[5] VanLehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 2011.

[6] Hattie J., Timperley H. The power of feedback. Review of Educational Research, 2007.

[7] Chi M. T. H., Bassok M., Lewis M. W., et al. Self-explanations in learning to solve problems. Cognitive Science, 1989.

[8] Pearl J. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009.

[9] Wachter S., Mittelstadt B., Russell C. Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology, 2017.

[10] Kasneci E., Sessler K., K ü chemann S., et al. ChatGPT for good? Learning and Individual Differences, 2023.

作者简介：国刚（1979.01-），男，汉族，山东省青岛市，硕士，副教授，研究方向：大数据挖掘、复杂网络，邮箱：121035458@qq.com。