高校教改实践中的 AI 安全教学认知和探索

李永[1] 孙昊[2] 肖征荣[3]

1 北京工业大学计算机学院，北京 100022
2 山东警察学院，济南，250200
3 中国联通研究院，北京，100176

摘 要 随着 AI 在高校教改中的应用越来越广泛，其带来的安全挑战也逐步展现。本文深入剖析了 AI 教育智能鲁棒性安全保障项目，通过对 AI 教育系统面临的数据隐私泄露、算法偏见等现状，以及现有评估体系缺陷的分析，基于 MITRE ATT&CK 框架设计了“攻击-监测-修复-复测”闭环流程，并通过构建教育红队、创新攻防测试方法和动态评估模型，以及红队有效性指标 R 值进行量化评价，经过蓝队优化算法、增强数据防护后，复测时 R 值从 0.52 提升到了 0.81。最后，还引入红队-蓝队-学术支持团队”三角协作模式，为 AI 教学安全提供了有效解决方案，对推动 AI 教育智能化健康发展具有重要意义。

关键字 AI，高校教改，安全评估，教育红队

Understanding and Exploration of AI Teaching Security in Higher Education Reform Practice

Li Yong[1] Sun Hao [2] Xiao Zhengrong[3]

1. College of Computer, Beijing University of Technology, Beijing, 100022
2. Shandong Police College, Jinan, 250200
3. China Unicom Research Institute, Beijing, 100176

Abstract—As the application of AI in college education reform becomes increasingly widespread, the security challenges brought by AI are gradually emerging. This paper provides an in-depth analysis of the AI Educational Intelligent Robustness Security Assurance Project. Through a thorough examination of the current challenges faced by AI education systems，such as data privacy leakage and algorithmic biases，as well as shortcomings in existing evaluation systems, this paper designs a closed-loop process of "attack-monitor-repair-retest" based on the MITRE ATT&CK framework. It constructs an education red team, innovates offensive and defensive testing methods, and implements a dynamic evaluation model, including the quantitative Red Team Effectiveness metric (R-value). After Blue Team optimized the algorithm and enhanced data protection measures, the R-value increased from 0.52 to 0.81. Furthermore, it introduces a "Red Team-Blue Team-Academic Support Team" tripartite collaboration model. These components collectively provide an effective solution for AI teaching security, highlighting its significant importance in promoting the healthy development of intelligent AI education.

Keywords—AI, Higher Education Reform, Safety Assessment, Education Red Team

1 引 言

在数字技术深度融入教育领域的时代背景下，人工智能（AI）正成为推动高校教育变革的核心力量[1]。根据全球教育科技市场研究机构 HolonIQ 的报告，2025 年全球 AI 教育市场规模预计达到 200 亿美元，年增长率超过 45%。联合国教科文组织（UNESCO）发布的《人工智能与教育》报告显示[2]，研究生教育场景的 AI 应用占比超过 35%。智能学术推荐系统、个性化学习平台、智能答辩评估工具等创新应用，显著提升了学术研究效率与人才培养质量。国内高校在 AI 大模型驱动高校人才培养改革[3]，计算机网络数智化课程建设探索[4]，人工智能课程建设[5]，人工智能通识教育路径探索[6]等领域开展研究。斯坦福大学开发的 AI 论文辅助系统，通过自然语言处理技术分析学术文献，可将研究生文献调研效率提升 40%以上；清华大学利用 AI 算法构建的个性化学习路径模型，使学生课程完成率提高了 25%。

然而，技术的快速发展也带来了严峻的安全挑战。2022 年，某高校的 AI 评分系统遭遇黑客攻击，导致 1.2 万名研究生的成绩数据泄露，直接影响学生升学与奖学金评定；同年，某在线教育平台的智能招生系统因算法偏见，对女性申请者的录取预测准确率较男性

低达 18%，引发社会广泛关注。这些事件暴露了 AI 教育系统在数据隐私保护、算法公平性、系统鲁棒性等方面的严重缺陷。据统计，全球每年因教育 AI 安全事件造成的经济损失超过 30 亿美元，其中数据泄露占比达 42%，算法偏见引发的纠纷占比 28%。

与其他领域的 AI 安全需求不同，教育领域的 AI 安全具有显著的特殊性，国外也有一些教育安全领域的研究[7][8]。教育 AI 更注重数据隐私保护、学术伦理合规与教育公平性。例如，在军事场景中，AI 统可接受一定程度的误判率以实现战略目标，但在研究生论文评审中，0.1%的算法误判都可能导致学术不公。此外，教育数据还涉及学生敏感信息，其隐私保护要求也非常高，因此亟需构建适用于教育场景的 AI 安全评估体系。

为应对上述挑战，“教育智能鲁棒性安全保障”项目应运而生。该项目以构建专业化教育红队为核心，通过模拟真实攻击场景、开发针对性安全工具，建立动态化的 AI 安全评估机制，旨在填补教育 AI 安全领域的研究与实践空白，为智能化教育发展筑牢安全防线。

2 教育智能鲁棒性安全保障背景

2．1 AI 赋能教育系统的风险现状

从技术角度看，AI 赋能教育系统面临数据泄露、算法偏见与模型脆弱性三大核心风险。在数据安全方面，教育平台普遍存在数据管理漏洞。2023 年网络安全公司 Check Point 的研究显示，全球 63%的教育类 APP 存在未加密数据传输问题，某在线课程平台因未对学生作业数据进行脱敏处理，导致 10 万份论文手稿被非法获取。算法偏见问题同样突出，如某高校的 AI 招生模型因训练数据中男性申请者占比过高，导致女性录取率被低估，违反教育公平原则。此外，深度学习模型的脆弱性易受对抗样本攻击，攻击者可通过细微修改图像像素，使智能监考系统误将正常考试行为判定为作弊。

此外，师生对 AI 工具的过度依赖，也会引发能力退化危机。美国教育技术协会（ISTE）的调研表明，过度使用智能写作助手的研究生中，78%在脱离工具后论文逻辑性显著下降；我国某高校的实验数据显示，长期依赖智能解题系统的学生，在复杂学术问题的自主分析能力上较对照组低 31%。同时，系统稳定性问题频发，某大学的智能答辩平台在期末考试期间因服务器过载，导致 23%的答辩中断，严重影响教学秩序。

2．2 AI 安全评估体系缺陷分析

传统评估方法难以适应教育 AI 的特殊需求。Gartner 的研究指出，当前 85%的教育机构仍采用功能测试为主的评估方式，仅关注系统是否实现预定功能，忽视对抗性攻击模拟。例如，某知名教育软件在上线前仅进行了常规功能测试，上线后遭遇 SQL 注入攻击，导致学生账号信息泄露。从组织架构看，教育机构普遍缺乏专业安全团队，70%的高校将安全评估外包给第三方，但外包团队因不熟悉教育业务流程，平均漏洞响应时间远超行业安全标准。

在政策法规层面，国内外教育 AI 安全标准存在显著缺失。欧盟《AI 法案》虽对通用 AI 系统提出安全要求，但未针对教育场景制定细则；我国目前仅有《网络安全法》等通用法规，缺乏教育 AI 数据保护、算法透明性等专项规定。该领域的法律空白导致教育机构在 AI 应用中面临合规困境，安全责任划分不明确的问题。

近五年教育 AI 安全领域的研究显示，现有的研究大多聚焦于技术功能优化（占比 62%），对安全评估的系统性研究不足（仅占 18%），尤其缺乏教育场景特有的隐私保护、伦理合规等方面的评估方法研究。

此外，从数据安全、算法公平性、系统鲁棒性方面来看，传统评估覆盖率和红队测试覆盖率有较大区别，见表 1。

表 1 统评估与红队测试覆盖率对比

评估维度	传统评估覆盖率	红队测试覆盖率
数据安全	35%	92%
算法公平性	12%	88%
系统鲁棒性	28%	95%
伦理合规性	5%	75%

3 项目分析及实施

本项目以构建可持续的教育 AI 安全生态为总体目标，通过组建专业化教育红队，建立动态化安全评估机制，实现从漏洞发现到修复的全流程闭环管理。项目三个子目标如下：

（1）开发适用于教育场景的攻防工具链。基于 MITRE ATT&CK 框架，构建包含数据污染模拟器、算法偏见检测工具、系统韧性评估平台在内的专业工具集。如数据污染模拟器可模拟恶意用户注入虚假学术数据，测试推荐系统的抗干扰能力；算法偏见检测工具通过生成对抗数据，自动识别模型中的歧视性因素。

（2）制定标准化测试流程。设计“攻击-监测-修复-复测”四阶段闭环体系，红队首先模拟真实攻击场景，蓝队实时监测系统响应，双方协作制定修复方案，最后通过复测验证安全提升效果。该流程借鉴敏捷开发理念，实现安全评估的持续迭代优化。

(3) 推动教育机构、科技企业与政府的跨领域合作。建立多方数据共享机制与安全标准协同制定平台，促进产学研用的深度融合。可与国家网络安全主管部门合作，将项目成果纳入教育行业安全规范；还可与科技企业共建 AI 安全实验室，加速技术成果的转化。

3．1 教育场景测试

项目的场景测试遵循三大设计原则：真实性（模拟大规模在线考试、学术论文评审等真实场景）、可重复性（通过标准化测试用例确保结果一致性）、伦理合规性（严格遵守数据隐私保护法规，采用匿名化测试数据）。测试体系包含三大模块：

(1) 数据安全测试采用 SQL 注入、伪造 API 请求、中间人攻击等技术手段，模拟黑客窃取或篡改教育数据。在某次测试中，红队通过 SQL 注入攻击某高校的选课系统，成功获取 1500 名学生的课程数据，暴露其数据库权限管理漏洞。攻击成功后，蓝队通过强化权限控制、增加数据加密层级，使系统数据防护能力提升 60%。

(2) 算法公平性验证运用对抗生成数据（Adversarial Data）技术，检测模型对不同群体的公平性。以研究生招生模型为例，红队生成包含性别、民族等敏感信息的对抗数据，发现模型对少数民族申请者的预测准确率较其他群体低 15%。基于此，蓝队改进数据预处理方法，增加少数民族样本比例，最终将预测准确率差距缩小至 3%。

（3） 技术依赖性分析通过人为关闭 AI 辅助功能，记录师生的应急策略。在智能学术写作系统测试中，关闭语法纠错与文献推荐功能后，研究发现 32%的学生在论文写作效率上下降超过 50%，暴露出过度依赖问题。据此，项目团队建议在教学中设置 AI 使用限制，培养学生自主学术能力。

3．2 典型测试场景分析

场景 1：智能学术推荐系统。红队采用注入虚假文献元数据的攻击手段，将低质量论文的引用量篡改至异常高位，误导推荐算法。测试显示，原系统在攻击后推荐准确率从 89%骤降至 37%。为应对该问题，项目团队引入基于区块链的文献溯源机制，通过分布式账本记录文献原始信息，确保数据不可篡改。改进后系统在同类攻击下推荐准确率维持在 85%以上。

场景 2：在线答辩评估平台。红队模拟 DDoS 攻击，使平台在答辩高峰期出现服务中断。原系统在攻击下平均响应时间从 2 秒延长至 15 秒，导致 30% 的答辩视频卡顿。项目团队通过部署边缘计算节点，将计算任务分散至离用户更近的服务器，使系统并发处理能力提升 3 倍，在同等攻击强度下响应时间保持在 3 秒以内。

3.3 基于 DOE 的多维度评估

本节采用实验设计（DOE）方法，选取数据多样性（多语言文献、跨学科数据）、攻击强度（低 / 中 / 高压力测试）、环境复杂度（网络延迟波动）三个关键变量，构建 L9 正交实验表。例如，在智能翻译系统测试中，通过组合不同语言类型（中文、英文、日文）、攻击强度（每秒 100/500/1000 次请求）和网络环境（4G/5G/Wi-Fi），系统分析各因素对翻译准确率的影响，定位关键风险点。红队攻击测试流程见图 1。

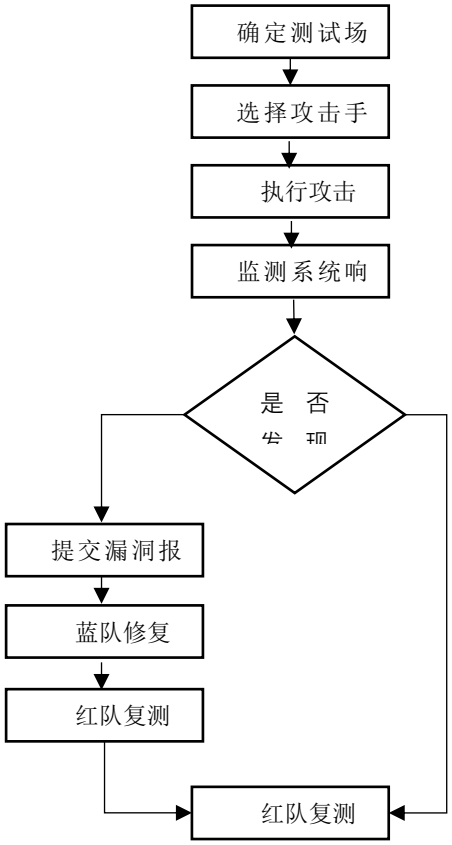

图 1 红队攻击测试流程图

4 评估指标分析

4.1 核心指标：红队有效性

红队有效性（R）通过以下公式量化：

$$R = \alpha \cdot \frac{\Delta P}{P_0} + \beta \cdot \log(\frac{T_{max}}{T}) + \gamma \cdot (1 - \frac{C}{C_{budget}}) \quad (1)$$

其中，$\frac{\Delta P}{P_0}$ 表示攻击后系统性能下降比例（如推荐准确率降低幅度），$\log(\frac{T_{max}}{T})$ 为攻击响应时间的标准化

值（Tmax 为设定的最长响应时间阈值），$1-\frac{C}{C_{budget}}$ 反映攻击成本控制程度（C_{budget} 为预设成本上限）。采用德尔菲法确定权重系数 α=0.6（强调性能影响）、β=0.3（关注响应速度）、γ=0.1（考量成本效益）。

以某智能学术推荐系统测试为例，初始状态下系统推荐准确率 P_0=85%，红队攻击后降至 P_1=60%，ΔP=25%；攻击响应时间 T=8 分钟（设定阈值 T_{max}=10 分钟）；攻击成本 C=8000 元（预算 C_{budget}=10000 元）。代入公式计算得 R=0.52。

经过蓝队优化算法、增强数据防护后，复测时 P_2=82%，T=5 分钟，C=6000 元，计算得 R=0.81，可见系统安全性得到显著提升。

项目分四阶段推进：

阶段 1（基线测试）：R=0.4-0.5，建立安全基线，识别系统基础漏洞。如在首次测试中，发现某教育平台存在未授权数据访问漏洞，修复后使 R 值提升 0.1；

阶段 2（深度测试）：R=0.6-0.7，模拟中等强度攻击，优化防御策略；

阶段 3（压力测试）：R=0.8-0.85，应对复杂攻击场景，验证系统韧性；

阶段 4（终极测试）：R=0.9，模拟国家级 APT 攻击，确保系统达到最高安全标准。

R 值的计算对比表见表 2。

表 2 R 值计算数据对比表

测试阶段	攻击前准确率	攻击后准确率	响应时间（分钟）	攻击成本（元）	R 值
基线测试	85%	60%	8	8000	0.52
优化复测	82%	78%	5	6000	0.81

5 项目团队与协作机制

项目采用“红队-蓝队-学术支持团队”三角协作模式，各团队分工明确、协同高效：

（1） 红队（TT1）：由渗透测试工程师与教育技术专家组成，负责攻击模拟与漏洞挖掘。团队成员需具备 CISSP（注册信息系统安全专家）、CEH（认证道德黑客）等专业资质，定期参加国际网络安全竞赛提升实战能力。

（2） 蓝队（TT2）：包含 AI 开发团队与教育管理者，承担系统维护与迭代任务。通过敏捷开发（Scrum）框架，每两周进行一次迭代开发，快速响应安全修复需求。

（3） 学术支持团队（TT3）：由教育伦理委员会与法律顾问构成，确保项目符合学术伦理与法律法规。例如，在数据测试中，严格审查数据使用合规性，要求所有学生数据均经过脱敏处理。

协作机制方面，建立月度跨团队会议制度，使用 JIRA 系统跟踪漏洞处理进度；定期开展“攻防对抗赛”，通过实战演练提升团队协作能力。引入 DevSecOps 理念，将安全评估嵌入软件开发全生命周期，实现安全与开发的深度融合。

团队协作架构图见图 2。

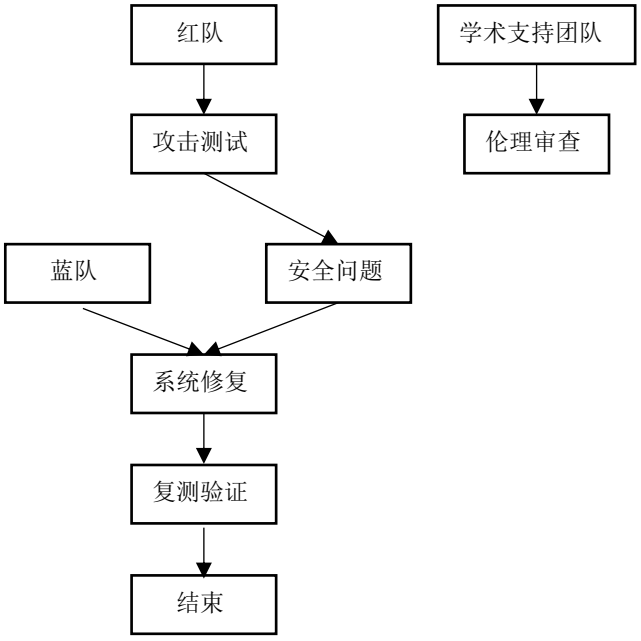

图 2 团队协作架构图

6 结束语

通过本项目，有助于推动我国教学 AI 安全国家标准的出台，参照项目制定的测试流程与评估指标，可有效减少学术评估不公事件，提升教育公信力。据测算，若全面实施该标准，教育数据泄露事件发生率可降低 60%以上。此外，该项目也可以催生一系列教育专用安全工具，如基于项目研发的 AI 伦理检测插件，可自动扫描教育系统中的算法偏见，平均检测效率较传统方法提升 3 倍。项目还可增强公众对 AI 教育的信任，促进技术普惠，通过建立透明的安全评估体系，消除用户对数据隐私与算法公平性的担忧，推动 AI 教育技术在偏远地区和弱势群体中的普及。

当然，AI 赋能教育安全也存在巨大挑战。从技术上来看 AI 模型的黑箱特性导致漏洞难以溯源，深度学

习模型复杂的参数结构与非线性计算过程，使安全专家难以理解模型决策逻辑，增加漏洞定位与修复难度等。在伦理层面，红队攻击存在侵犯学生隐私的风险。即使采用匿名化处理，部分敏感数据仍可能通过数据关联分析被还原。在协作层面，跨学科团队沟通成本高，教育专家、技术人员与安全从业者的知识体系差异显著，导致需求理解偏差与进度延误等。

本文首次将军事级红队测试体系引入教育领域，填补了 AI 教育安全评估的空白，创新性地采用动态评估模型，通过 R 值量化安全效能，突破传统二元通过式评价；引入教育特异性攻防，设计学术推荐篡改、答辩干扰等专属攻击场景；并搭建了伦理合规框架，建立匿名化测试环境与多方监督机制。未来，AI 赋能教育安全的研究方向包括推动全球协作网络合作，联合 UNESCO 构建跨境 AI 教育安全联盟，在政策层面推动推动《教育人工智能安全法》纳入“十五五”规划，在技术层面引入量子安全加密，以应对量子计算对教育数据隐私的潜在威胁等。

参 考 文 献

[1] Priten Shah. Embracing AI in Education. Hoboken, New Jersey, USA: Wiley,2023.

[2] 苗逢春、Wayne Holmes、黄荣怀、张慧. 人工智能与教育. 联合国教科文组织,巴黎，法国，2021.

[3] 蒲晓蓉，任亚洲等. AI 大模型驱动高校人才培养改革[J]. 计算机技术与教育学报，2024，12(6) :101-105.

[4] 尚凤军. 知识图谱驱动的计算机网络数智化课程建设探索[J].计算机技术与教育学报，2024，12(6) :71-77

[5] 姜宏旭, 赵梅娟, 李辉勇, 张永飞. 产教融合背景下嵌入式人工智能课程建设的探索[J].计算机技术与教育学报，2024，12(6) :1-7.

[6] 覃希,陈燕,姚怡,唐春艳.教评双驱教学实验平台赋能人工智能通识教育路径探索[J].计算机技术与教育学报，2025，13(1) :119-124.

[7] M. Arai et al., REN-A.I.: A Video Game for AI Security Education Leveraging Episodic Memory[J], IEEE Access, 2024(12):47359-47372.

[8] Y. Wang, M. McCoey, Q. Hu and M. Jalalitabar. Teaching Security in the Era of Generative AI: A Course Design of Security+ChatGPT[C]//*2024 IEEE Integrated STEM Education Conference (ISEC)*, Princeton, NJ, USA, 2024