构建未来教育安全：GAI 教育应用的多维风险与治理策略研究

Building Future Education Security: A Study on the Multidimensional Risks and Governance Strategies of GAI Educational Applications

杨博勋[1*]，张粤芳[1]，陈巧灵[1]，杨芷悦[1]，庄颖[1]
[1]华南师范大学 教育信息技术学院
[*]2292797881@qq.com

【摘要】 随着生成式人工智能在教育领域的创新性和潜力逐渐显现，同时也暴露出诸多潜在风险。如何平衡技术应用的便利性与其可能带来的安全隐患，已成为当前教育实践中的重要课题。为此，本文以 GAI 为切入点，聚焦其在教育场景中可能引发的风险，通过“风险感知-风险评估-风险应对三阶段”的逻辑路径，采用德尔菲法对风险进行了量化分析，并进一步提出规避、缓解、转移和接受四种针对性应对策略，旨在为教育者、技术开发者及政策制定者提供科学有效的指导，助力 GAI 在教育领域的安全、规范与可持续发展。
【关键词】 生成式人工智能；教育应用风险；风险评估；风险管理

Abstract: *As the innovation and potential of generative artificial intelligence in the field of education gradually emerge, many potential risks are also exposed. How to balance the convenience of technology application and the potential safety hazards it may bring has become an important issue in current educational practice. To this end, This paper uses the Delphi method to analyze these risks through three stages: risk perception, risk assessment, and risk response.This paper further proposes four targeted response strategies: avoidance, mitigation, transfer and acceptance, aiming to provide scientific and effective guidance for educators, technology developers and policymakers, and to help the safe, standardized and sustainable development of generative AI in the field of education.*
Keywords: Generative AI**,** Educational application risks**,** risk assessment**,** Risk Management

1. 引言

生成式人工智能（Generative Artificial Intelligence，GAI）在教育领域的应用已从知识生成、智能辅助向个性化学习等多元场景渗透，但其技术缺陷引发的安全性、伦理性与可靠性风险日益凸显。GAI 的“幻觉”现象可能扭曲学生认知结构，算法偏见与数据泄露威胁教育公平及隐私安全，技术过度依赖则削弱学习主体性，这些问题直接冲击教育目标的实现。当前，国内外针对人工智能在教育中的应用风险已有一些初步探索。2021 年，联合国教科文组织通过《人工智能伦理问题建议书》，明确提出在教育领域应用人工智能时需要关注透明性、隐私保护等伦理原则。我国《新一代人工智能伦理规范》也强调要加强人工智能风险的监测、预警和评估，提升风险管理能力。尽管政策和规范已开始关注 GAI 的潜在风险，但针对教育领域的具体问题尚缺乏系统性分析和针对性的解决方案。为此，本研究基于 GAI 技术特性与缺陷，构建教育风险管理框架（Education Risk Management Framework，ERMF），通过风险感知模型（Perceived Risk Model，PRM）全面识别不同类型的风险，运用风险矩阵模型（Risk Matrix Model，RMM）量化评估风险等级，最终依托风险应对策略模型（Risk Response Strategy Model，RRSM）形成规避、缓解、转移与接受的组合策略，为多利益相关者提供贯穿“识别-评估-应对”全链路的科学管理方法，促进 GAI 教育应用的可持续发展。

2. 技术基础解析：生成式人工智能的局限与潜能

2.1. “生成”与“压缩”：模型设计的内在限制

GAI 的“幻觉”现象源于深度神经网络的概率生成机制，模型通过统计模式匹配生成流畅文本，生成内容在语言上能够合理流畅，但事实上的准确性和语境一致性难以保障(Almasri, 2024)。此外，生成模型的目标函数通常侧重于流畅性和可读性，而非事实的准确性，这也进一步导致大模型会在缺乏充分的数据支撑的前提下，自主地“填补”语句中的空白，使得生成内容缺乏校准和真实性验证，进一步加剧“幻觉”现象(罗文 & 王厚峰, 2024)。在数据层面，大规模数据压缩虽解决存储失衡问题，但会丢失细节信息和上下文关联，造成生成内容溯源困难及准确性下降。

2.2. “理解”的缺席：思维逻辑空缺

GAI 本质上是一种基于语言文字的技术，通过深度学习海量文本数据模拟人类语言能力，但其本质缺乏具身认知和思想加工过程。GAI 依赖统计规律生成拟人化文本，绕过了主流语言学理论与方法(李春南 & 王山, 2024)，因缺失具身理解的“生理结构”(肖峰, 2024)，从本质上看，它并不具备人类的认知和推理能力，且缺乏对深层思想的理解和构建过程，其生成内容仅停留在语言表层匹配，无法处理对社会发展至关重要的隐性知识。

2.3. “客观表达”的逻辑悖论：语言表达的局限性

语言作为“非中立”的思想表达工具，其词汇、句式及语气选择均隐含价值观与情感倾向。其通过主动筛选实现内容建构，而这种筛选本身也意味着对未选信息的“遮蔽”。GAI 的内容生成面临双重局限，一方面，训练数据本身承载着语言文本固有的主观性与偏见，另一方面，文本生成过程必然伴随对信息的选择性呈现与省略。这种基于语言符号系统的双重筛选机制，使 GAI 的输出无法超越语言自身的选择性与主体性边界。

3. 教育领域生成式人工智能的系统化风险管理框架

GAI 在教育领域的广泛应用推动教学模式革新与教育效能提升的同时，其引发的潜在风险因具有隐蔽性、多维性等性特征而面临识别与评估困境。本文提出了教育风险管理框架，以“风险识别-风险评估-风险应对”的逻辑路径，基于教育场景特殊性的多维度风险分析机制，为规范 GAI 教育应用提供风险管理方法论支持，如图 1 所示。

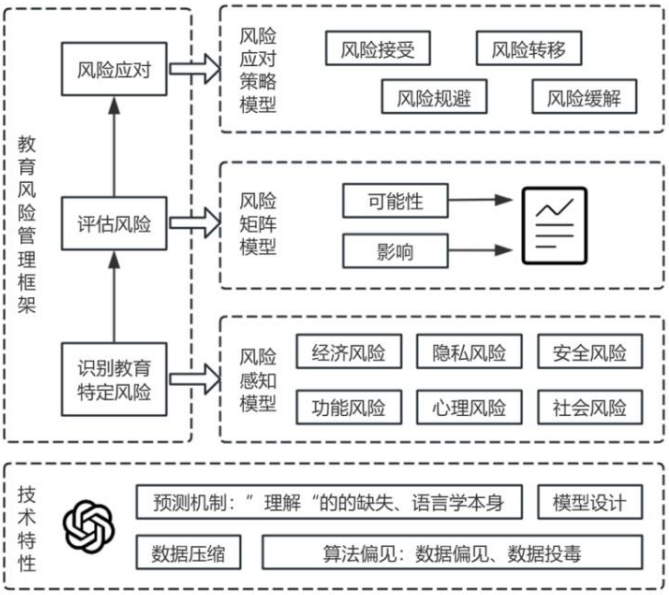

图 1 生成式人工智能教育应用的风险管理框架

GAI 在教育场景中的风险可归纳为六大核心类别，影响涵盖个体认知、教学实践与社会公平等多个层面。第一，安全风险源于 AI“幻觉”现象，可能误导学生形成错误认知，影响教学效果。第二，心理风险表现为技术依赖，削弱批判性思维，甚至引发“知识生产无需人类参与”的偏差认知。第三，功能风险体现为趋同效应，AI 输出同质化内容，限制多元视角，可能加剧社会偏见(Fan 等, 2023)。第四，社会风险涉及教师权威变化，学生直接依赖 AI 获取答案，削弱课堂管理效能，影响教育生态平衡。第五，隐私风险根植于 GAI 对数据的高需求，学生敏感信息可能因安全漏洞泄露(王佑镁等, 2023)，破坏教育技术的信任基础。第六，经济风险体现在技术部署的不均衡，发达地区可利用 GAI 提升教学，而欠发达地区因资源匮乏加剧技术鸿沟与社会分化(Dauvergne, 2021)。这些风险相互交织，唯有通过系统性框架进行动态评估与精准治理，才能在推动教育数字化的同时，实现技术创新与风险控制的平衡。

4. 风险分级分析：评估路径与结果解读

4.1. 实验设计与实施

本研究采用风险矩阵分析方法进行风险评级(朱启超,匡兴华,沈永平, 2003)，实验对象为 10 名教育技术领域的资深专家，包括高校人工智能方向的研究学者、讲师以及高级技术开发者。采取问卷调查的形式，采用 Likert 三级量表，计算每类风险在两个维度上的乘积，并计算专家意见的平均值，乘积结果与风险等级的对应关系如表 1 所示：

表 1 乘积结果与风险等级的对应关系

	低可能性（1）	中可能性（2）	高可能性（3）
低影响（1）	低风险（1）	低风险（2）	中风险（3）
中影响（2）	低风险（2）	中风险（4）	高风险（6）
高影响（3）	中风险（3）	高风险（6）	极高风险（9）

4.2. 数据分析与风险等级划分

如表 2 所示，通过风险矩阵分析，GAI 在教育应用中的风险分布呈现显著差异：隐私与安全风险被列为最高优先级，前者源于大规模数据采集与处理漏洞易致敏感信息泄露，后者则因 AI“幻觉”现象频发，二者均具有高发生概率与深远负面影响，需优先防控。心理与功能风险属中等层级，虽隐性却需持续关注——技术依赖引发的思维惰性与答案趋同化仍威胁教学创新。经济风险则因技术普及与政策支持强化，其成本负担及对教育公平的冲击呈弱化趋势，成为当前风险矩阵中相对可控的维度。

表 2 专家打分结果

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	平均分	等级
安全风险	6	9	9	6	6	6	10	8	5	7	7.2	极高
心理风险	6	2	3	4	4	3	5	1	2	3	3.3	中
功能风险	3	1	2	1	1	2	1	1	2	1	1.5	低
社会风险	4	6	6	4	4	4	5	5	3	6	4.7	高
隐私风险	6	9	9	9	6	9	8	7	7	9	7.9	极高
经济风险	3	1	2	1	1	2	3	2	1	1	1.7	低

5. 生成式人工智能风险应对

RRSM 作为一种系统化的风险管理框架，用于针对不同风险类型和等级制定科学合理的应对策略。如表 3 所示，为不同的风险等级匹配不同的应对措施，确保 GAI 教育应用的安全性

与可持续性。

表 3 *风险级别与应对措施*

风险等级	应对措施	含义
低风险	接受	风险发生可能性低，影响轻微，可接受，无需立即处理，
中风险	缓解	风险有一定可能性且有中等影响，应采取缓解措施降低风险
高风险	转移	风险可能性高或影响严重，应优先处理，采取转移措施
极高风险	规避	风险可能性极高且影响灾难性，必须立即规避。

针对不同等级的风险，应采取相应的管理策略。对于低风险，可选择接受，并通过教师复核 AI 生成内容以确保其准确性和可靠性(张惠彬 & 许蕾, 2024)。对于中高风险，应采取缓解措施，建立实时监控与反馈机制以及引入多方监督，以降低风险发生的可能性及其影响。对于高风险，可通过签订责任协议或购买保险等方式将风险部分转移至第三方，以减少教育机构的直接责任和潜在损失。对于极高风险，则需采取规避策略，严格限制 AI 在涉及隐私数据和关键性评价中的应用，以防止潜在的数据泄露和公平性问题(付睿云等, 2023)。综合而言，通过分级管理和科学应对，可有效保障 GAI 在教育领域的安全性与可持续发展。

7. 总结与展望

GAI 的应用正不断重塑教育模式，为个性化教学和智能化辅导提供新机遇，但其潜在风险可能影响教育公平性，甚至阻碍教育目标的实现。本文从风险识别、评估到应对，提出系统化的管理策略，强调技术与教育目标的协调。为实现技术与教育的深度融合，需推动多方协同治理，强化技术伦理教育，提升师生数字素养与风险意识。在技术快速演进的背景下，唯有确保风险可控，方能充分发挥 GAI 的优势，实现教育的安全、创新与可持续发展。

参考文献

付睿云, 白庆春, & 吕泰彧. (2023). 人工智能教育数据偏见成因机制及家校共治探讨. 中国信息技术教育, *24*, 75–78.

李春南 & 王山. (2024). 生成式人工智能时代的语言安全：系统性风险与治理路径. 国际安全研究, *42*(5), 135–156, 160. https://doi.org/10.14093/j.cnki.cn10-1132/d.2024.05.006

罗文 & 王厚峰. (2024). 大语言模型评测综述. 中文信息学报, *38*(1), 1–23.

王佑镁, 王旦, 梁炜怡, & 柳晨晨. (2023). ChatGPT 教育应用的伦理风险与规避进路. 开放教育研究, *29*(2), 26–35. https://doi.org/10.13966/j.cnki.kfjyyj.2023.02.004

肖峰. (2024). 大模型的理解力之争与理解观新叙事. 社会科学, *1*, 41–51. https://doi.org/10.13644/j.cnki.cn31-1112.2024.01.005

张惠彬 & 许蕾. (2024). 生成式人工智能在教育领域的伦理风险与治理路径——基于罗素大学集团的实践考察. 现代教育技术, *34*(6), 25–34.

朱启超,匡兴华,沈永平. (2003). 风险矩阵方法与应用述评. 中国工程科学, *1*, 89–94.

Almasri, F. (2024). Exploring the Impact of Artificial Intelligence in Teaching and Learning of Science: A Systematic Review of Empirical Research. *Research in Science Education*, *54*(5), 977–997. https://doi.org/10.1007/s11165-024-10176-3

Dauvergne, P. (2021). The globalization of artificial intelligence: Consequences for the politics of environmentalism. *GLOBALIZATIONS*, *18*(2), 285–299. https://doi.org/10.1080/14747731.2020.1785670

Fan, Y., Tan, Y., Raković, M., Wang, Y., Cai, Z., Shaffer, D. W., & Gašević, D. (2023). Dissecting learning tactics in MOOC using ordered network analysis. *Journal of Computer Assisted Learning*, *39*(1), 154–166. https://doi.org/10.1111/jcal.12735