Exploring Multidimensional Risk Analysis and Governance Strategies for Generative AI in Education

Nan Ke, Zhenbin Liang*, Qiao Sun

Naval University of Engineering, Wuhan, 430022, China

*1920191069@nue.edu.cn

Abstract. The rapid advancement of generative AI is profoundly transforming the modes of knowledge production, dissemination, and application. Its deep integration into the field of education brings immense pedagogical innovation to modern classrooms and offers revolutionary opportunities for educational development. However, it simultaneously introduces complex and multifaceted novel risks. This paper first reviews current research on identifying risks associated with applying generative AI in education and corresponding prevention and control strategies. Subsequently, addressing the limitations of traditional risk analysis, which often lacks focus and struggles to capture the complexity and interconnectedness of risks in generative AI educational applications, a three-dimensional analytical framework encompassing "Teaching Segments - Risk Categories - Academic Disciplines" is constructed. This framework facilitates a multi-faceted and precise identification of application risk points. Finally, considering the current implementation of higher education instruction and the characteristics of various academic disciplines, this paper explores stratified governance methods for the risks of generative AI in education and proposes discipline-specific strategies.

Keywords: Generative AI, Education and Teaching, Risk Analysis, Governance Strategies.

1 Introduction

With the rapid digitalization of education, generative AI, leveraging its powerful capabilities in knowledge comprehension and content generation, is driving education and teaching from standardized knowledge transmission towards large-scale, personalized ability cultivation. Its development is reshaping teaching relationships and classroom formats, leading the current educational model into a new stage empowered by digital intelligence. However, this empowerment also brings significant risks, including knowledge misinformation caused by AI "hallucinations," potential weakening of students' critical thinking and deep learning abilities, data privacy concerns, and exacerbated challenges related to academic integrity and the digital divide. Furthermore, influenced by technological and human factors, the application of generative AI in education can also spawn ethical risks, characterized

C. V. Garcés et al. (eds.), *Proceedings of the 6th International Conference on Internet, Education and Information Technology (IEIT 2026)*, Atlantis Highlights in Social Sciences, Education and Humanities 53,
https://doi.org/10.2991/978-94-6239-733-0_10

by the coexistence of objectivity and subjectivity, value and risk, and controllability alongside uncertainty[1]. Therefore, to foster the healthy development of generative AI in education, it is crucial to focus on analyzing application risks and researching prevention strategies. This paper will concentrate on the complexity and interconnectedness of generative AI risks, constructing a three-dimensional analytical framework of "Teaching Segments - Risk Categories - Academic Disciplines" to precisely analyze risk points and explore reasonable risk prevention strategies and methods grounded in practical contexts.

2 Literature Review

2.1 Risk Identification

Dong et al. argue that the "hallucination" problem inherent in generative AI introduces issues of knowledge authenticity and bias, potentially misleading learners, which contradicts educational principles. Simultaneously, sensitive and private information input by users during the pre-training phase faces potential leakage risks[2]. Concerning ethical issues, AI large models may provide incorrect answers regarding science, culture, history, etc., contradicting facts and leading to misleading situations, thereby causing the propagation of different ideologies[3]. Regarding intellectual property, generative AI also presents potential infringement risks, as it might use copyrighted works or trademarks without permission, leading to legal risks of intellectual property violation[4]. Epistemologically, due to insufficient understanding of the nature and characteristics of large models, existing research either conflates data security risks with personal privacy protection or focuses solely on data security risks in one specific operational link of the large model while neglecting others. Overall, a unified and systematic understanding of the data security risks associated with large models has yet to be formed[5].

2.2 Prevention and Control Measures

Chu et al. propose that to address various security risks associated with large model applications, a "human-in-the-loop" supervision strategy should be adhered to in all educational and teaching activities. Suggestions or solutions provided by AI systems require human approval before implementation. This approach allows humans to apply their expertise and judgment to AI-generated suggestions, ensuring final decisions are informed and accurate[6]. Regarding the choice of model foundation, general-purpose large models are not inherently suitable for education; the "transplantation" of technical functions may lead to a mismatch between functionality and needs. From an educational security perspective, only large models pre-trained on China's domestic educational big data can maximize the protection of data security and ideological security[7]. To enhance the data security environment, it is necessary to build highly reliable computing server clusters, improve server monitoring systems, strengthen physical and data protection, physically protect servers to prevent unauthorized access and data alteration, ensure data storage and encrypted backups, and guarantee data integrity and availability[8]. At the policy and institutional level, an access mechanism should be established, led by government and educational organizations, to review and experiment with generative AI, license large models suitable for integration into educational activities, and conduct continuous supervision of licensed models[9].

3 Construction And Analysis of the Three-Dimensional Risk Framework

Current research often focuses on single-dimensional risk analysis of generative AI in education. This paper addresses the complexity of risk scenarios and the intricate nature of disciplinary attributes in educational applications by designing a three-dimensional analytical framework of "Teaching Segments - Risk Categories - Academic Disciplines." This framework aims to precisely map abstract risks onto specific teaching segments and disciplinary contexts, reveal the interconnections and amplification effects between risk elements, and formulate targeted risk response strategies for different disciplines.

3.1 Overall Structure of the Risk Framework

The proposed three-dimensional analytical framework of "Teaching Segments - Risk Categories - Academic Disciplines" is illustrated in Fig. 1. Its core design philosophy stems from a profound understanding of the complexity of generative AI applications in education. The framework aims to construct a stereoscopic, interactive, and positionable analytical system, enabling precise "grid-based" risk assessment.

Within the context of generative AI application, the framework considers three key dimensions as an integrated whole:

1. Teaching Segment: Describes "when and where" the risk occurs, covering the entire process from pre-class preparation to post-class evaluation.

2. Risk Category: Defines "what" the risk is, encompassing four main categories: technical, ethical, academic, and psychological.

3. Academic Discipline: Determines "in which knowledge domain" the risk manifests, shaping the specific form the risk takes.

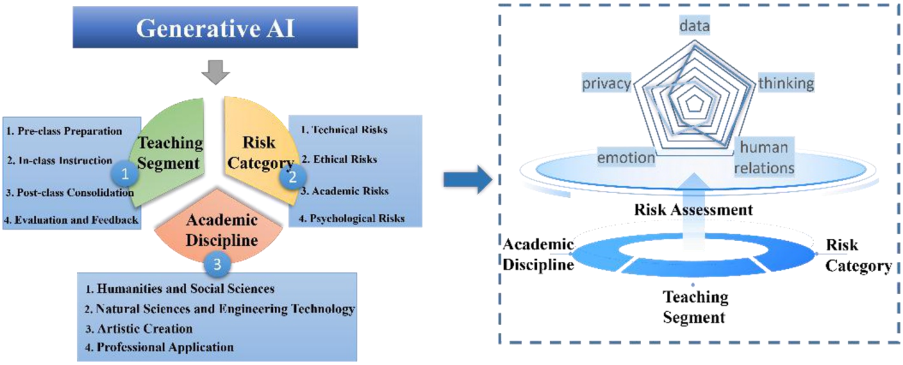

Fig. 1. Three-dimensional risk analysis framework of generative AI.

The organic integration of specific elements across these three dimensions constructs a highly contextualized and specialized mechanism for precise risk identification. Through further analysis, a grid-based diagnosis is derived, answering the question: "In which teaching segment, for which type of discipline, and facing what specific risk?" The fundamental advantage of this framework lies in its contextual precision and its ability to reveal risk heterogeneity. For instance, the same technical risk manifests, originates, and has consequences that differ significantly between an arts student writing a thesis, a science student designing an experiment, or an art student developing a creative concept. Consequently, the governance strategies must also differ.

3.2 Multidimensional Risk Analysis

The extraction and integrated analysis of risk elements are crucial prerequisites for accurately pinpointing risk points. These elements are embodied within the three dimensions of the framework, which are independent yet interwoven, collectively forming a stereoscopic coordinate system for understanding generative AI risks in education.

Teaching Segment Axis.
This dimension anchors risks to specific, operable nodes of instructional behavior based on the fundamental teaching process. The four segments are as follows:

1. Pre-class Preparation: Risks focus on “source quality” and “initial equity” (e.g., over-reliance on AI leads to homogenized lesson plans; biased or erroneous materials distort students’ cognitive structures).

2. In-class Instruction: Risks affect “process quality” and “intersubjective relationships” (e.g., AI teaching assistants weaken direct teacher-student exchange; AI hallucinations disrupt teaching rhythm and undermine teacher authority).

3. Post-class Consolidation: Risks concern “depth of learning” and “boundaries of personalization” (e.g., AI-driven personalized paths create information cocoons; reliance on AI homework tools hinders independent problem-solving).

4. Evaluation and Feedback: Risks pertain to “value orientation” and “fairness benchmarks” (e.g., AI-based evaluation fails to assess complex higher-order thinking; data biases impede accurate understanding of student engagement).

Risk Category Axis.

This axis categorizes risks by their inherent nature to support systematic attribution and governance. Four types are identified:

1. Technical Risks:Arising from AI’s inherent limitations (e.g., hallucinations, data breaches, algorithmic bias, system vulnerability, over-dependence). These are the most fundamental and often the root of other risks.

2. Ethical Risks: Involving value conflicts, rights infringements, and moral lapses (e.g., academic integrity crises, digital divide, IP disputes, blurred responsibility). These touch core educational values such as equity, integrity, and respect.

3. Academic Risks: Negatively impacting students’ cognitive development, knowledge construction, and skill cultivation (e.g., degraded critical thinking, fragmented knowledge, weakened practical skills, inhibited innovation). Addressing these safeguards holistic human development.

4. Psychological Risks: Affecting emotions, motivation, self-perception, and social development (e.g., weakened emotional bonds, learning anxiety, diminished intrinsic motivation, digital identity confusion). These remind us to maintain a human-centric stance in educational AI applications.

Academic Discipline Axis.

Academic discipline axis introduces disciplinary differences to enable precise, contextualized risk diagnosis. It answers: “In which knowledge domain does the risk manifest uniquely?” Four categories are identified:

1. Humanities and Social Sciences (e.g., literature, law, education): emphasize textual interpretation, critical thinking, and value judgment. Risk: AI’s fluency masks shallow logic and empty values, threatening academic integrity and deep thought → erosion of depth and originality.

2. Natural Sciences and Engineering Technology(e.g., math, physics, CS): emphasize logical reasoning, empirical validation, and hands-on practice. Risk: AI-generated code or formulas bypass reasoning and trial-and-error → hollowing out of practical skills and scientific inquiry.

3. Artistic Creation(e.g., fine arts, music, design): emphasize personal expression, aesthetic judgment, and innovation. Risk: AI imitation leads to stylistic copying over original voice → creative homogenization and loss of individual artistic voice (copyright issues also acute).

4. Professional Application(e.g., medicine, business, teacher education): emphasize contextual decision-making, ethical trade-offs, and interpersonal skills. Risk: AI’s

standardized solutions ignore real-world complexity → weakened judgment, adaptability, and ethical decision-making.

3.3 Three-Dimensional Integration and Localization of Risk Points

The analytical framework of "Teaching Segments - Risk Categories - Academic Disciplines" articulates which teaching segment a risk occurs in (Segment Axis), what type of issue it represents (Risk Axis), and within which disciplinary context it is highlighted and amplified (Discipline Axis). The integration of these three constructs a complete and precise risk assessment system, as illustrated in Fig. 2. Through the Cross combination of multidimensional elements, this risk analysis framework enables precise localization of application risks within specific scenarios and allows for comparative assessment of risk levels for specific risk points.

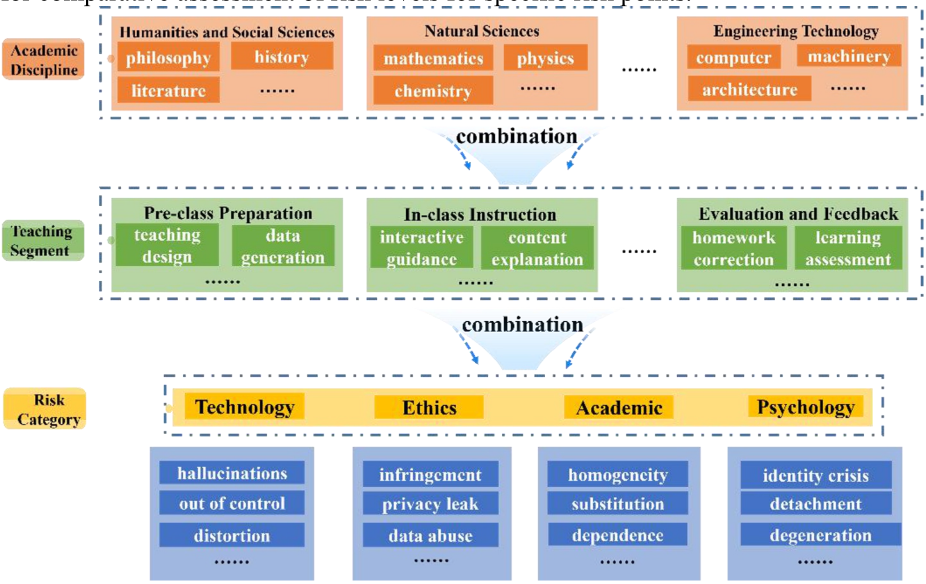

Fig. 2. Multi-dimensional risk assessment system

Case Study Risk Analysis.
Taking a science and engineering experimental teaching scenario as an example, we explore the interconnected risks involved.

1. Pre-class Preparation

Technical risk: AI-generated solutions contain hidden errors (e.g., wrong circuit parameters) that may damage equipment. Risk level: Medium (can be reduced with protective measures).

Ethical risk: Students directly copy AI-generated preview content, masking their true level of understanding. Risk level: High (widespread and hidden).

Academic risk: AI writes the experimental principles, constituting implicit plagiarism. Risk level: High (difficult to detect, prone to misuse).

Psychological risk: Over-reliance on AI weakens independent exploration motivation and resilience. Risk level: Medium (becomes obvious with long-term accumulation).

2. In-class Instruction

Technical risk: AI provides incorrect troubleshooting instructions in real time, affecting experimental safety. Risk level: High (may cause equipment or personal accidents).

Ethical risk: Some students secretly use AI, undermining team fairness. Risk level: Medium (depends on supervision intensity).

Academic risk: Using AI-generated data as fabricated experimental data, violating integrity. Risk level: High (direct falsification of the experimental process).

Psychological risk: AI's "perfect answers" trigger technical inferiority, reducing participation. Risk level: Medium (varies greatly among individuals).

3.Post-class Consolidation

Technical risk: AI-generated report analysis is logically incoherent, misleading cause-effect relationships. Risk level: Medium (can be mitigated by instructor review).

Ethical risk: Using AI to rewrite others' reports or merge multiple AI outputs to avoid deep thinking. Risk level: High (common and hard to define).

Academic risk: AI writes conclusions without acknowledgment, constituting academic misconduct. Risk level: High (similar to plagiarism, harder to detect).

Psychological risk: Facing AI's "high-scoring template" leads to self-deprecation or giving up. Risk level: Low to Medium (depends on students' psychological resilience)

4.Evaluation and Feedback

Technical risk: AI grading misjudges innovative solutions as errors. Risk level: Medium (can be reduced by human-AI collaboration).

Ethical risk: AI scoring has hidden biases (e.g., preference for a certain format), causing unfairness. Risk level: Medium (requires fairness validation).

Academic risk: Students generate "optimized answers" targeting scoring preferences, resulting in deceptive performance. Risk level: High (systemic vulnerability).

Psychological risk: High-scoring students develop "impostor syndrome"; low-scoring students experience learned helplessness. Risk level: Medium (risk increases with prolonged use).

Construction of a Three-Dimensional Risk Heatmap.
Based on the characteristics of different disciplines regarding teaching segments, ability requirements, and competency cultivation, a comparative analysis of risk levels across different teaching scenarios can be conducted. This allows for the preliminary construction of a disciplinary teaching scenario risk heatmap.

Taking the first three segments of teaching in humanities and social science and engineering technology as example disciplines, Table 1 respectively illustrate the constructed disciplinary teaching scenario risk heatmap. Symbols indicate risk levels (more ● signify higher risk). It can be observed that because teaching objectives in humanities/social sciences focus more on textual interpretation, critical thinking, and value judgment, while engineering technology emphasizes technical application and practice, the primary differences in risk between the two manifests in technical and ethical aspects. Additionally, both face significant academic risks, while psychological risks are generally smaller. Furthermore, as generative AI is often applied for knowledge and skill supplementation after class, risks across disciplines are predominantly concentrated in the post-class consolidation segment within the teaching process.

Table 1. Comparison of risk maps of teaching scenarios in different disciplines

	Pre-class Preparation		In-class Instruction		Post-class Consolidation	
	Humanities and Social Science	Engineering Technology	Humanities and Social Science	Engineering Technology	Humanities and Social Science	Engineering Technology
Technology	●●○○○	●●●●○	●○○○○	●●●●●	●●●○○	●●●○○
Ethics	●●○○○	●●●●○	●●●○○	●●○○○	●●●●●	●●●●○
Academic	●●●●○	●●●●○	●●●○○	●●●○○	●●●●●	●●●●○
Psychology	●○○○○	●○○○○	●●●○○	●●○○○	●●○○○	●●○○○

The risk heatmap reveals significant "segment-discipline" specificity in risk distribution: high-risk areas for different disciplines appear in different teaching segments. Administrators and teachers can utilize this map to quickly identify high-risk grids relevant to their discipline and teaching segment, allowing them to prioritize limited early warning and intervention resources on these most critical risk nodes, achieving precise and efficient risk management.

4 Stratified Governance System for Risks

Based on the precise assessment enabled by the three-dimensional analytical framework of "Teaching Segments - Risk Categories - Academic Disciplines," risk governance should not rely solely on single technical controls or ethical appeals. Instead, a multi-level, multi-stakeholder, and dynamically adaptive comprehensive governance system needs to be constructed. This involves exploring governance strategies from three levels—technological refinement, institutional restructuring, and cultural cultivation—while emphasizing discipline-specific implementation.

4.1 Technological Level Governance

The goal at this level is to reduce the probability of risk occurrence at its source, making generative AI a safer and more reliable educational aid. This can be achieved through domain-specific fine-tuning of models using high-quality educational materials to reduce hallucinations and enhance factual accuracy, along with plugins that automatically compare AI-generated historical content against authoritative databases and cite sources. In terms of knowledge construction, a structured core knowledge map should be established for each discipline as the logical boundary and fact anchor for AI content, while metadata tag systems (e.g., difficulty, knowledge points, ability dimensions, teaching segments) enable precise retrieval, such as finding pre-class questions on redox reaction concepts in chemistry. Regarding security and privacy, techniques like federated learning and differential privacy can utilize student data for model optimization without data leaving its domain, and explainability statements for AI decisions are crucial, so that AI-generated learning analytics reports present aggregated trend analyses and anonymized, privacy-protected individual learning path suggestions rather than raw data.

4.2 Institutional Level Governance

This level provides clear behavioral guidelines and procedural norms to integrate risk management into daily teaching administration. Universities or faculties should issue discipline-specific rules defining encouraged, restricted, or prohibited AI use and implement an “AI Use Declaration” system requiring students to state the scope of AI assistance (e.g., submitting original source code in CS courses). Teaching processes should integrate “human-AI collaboration” competency into learning objectives, design AI-mediated pedagogies, and increase the weight of process-oriented assessments to evaluate thinking and skills AI cannot replicate, such as creative sketches, inspiration logs, and oral explanations distinguishing student vision from AI output in artistic courses. Regarding academic integrity, it is essential to clearly distinguish “AI-assisted” from “AI-substituted” work, develop multimodal AI detection tools as references rather than sole determinants, and establish dispute resolution mechanisms via academic review (e.g., a review group for suspected AI-written theses using targeted defenses and original research materials).

4.3 Cultural Level Governance

This level requires cultivating a healthy and rational culture of AI application to fundamentally strengthen risk resilience. In terms of campus culture, promoting the concept of "AI as a powerful co-pilot, not an autopilot" through seminars, exemplary case studies, and teacher-student dialogues is vital. Recognizing teaching and learning cases that creatively utilize AI to solve real problems while demonstrating unique human value is important. Concerning subject formation, repeatedly emphasizing and practicing the core spirit and unique value of each discipline within teaching is crucial. Clearly establishing the teacher as the leadership of instructional design and guide of values, and the student as the active constructor of learning meaning. Regarding co-governance, establishing normalized communication and feedback mechanisms, such as a "Joint AI in Education Committee," allows frontline risk experiences of teachers and students to directly inform administrators and product managers, driving iterative improvements in technology and systems. Schools could periodically organize forums involving teachers, student representatives, and collaborating AI technology companies to discuss specific risks arising from product use in teaching segments and jointly formulate improvement plans.

5 Conclusion

Focusing on multidimensional risk analysis, this paper addressed the challenges of capturing the complexity and interdisciplinary nature of generative AI risks in education. It constructed a three-dimensional analytical framework encompassing "Teaching Segments - Risk Categories - Academic Disciplines," integrated risk analysis with teaching scenarios and disciplinary characteristics, precisely located application risk points from multiple perspectives, and built a three-dimensional risk heatmap. Through "scenario-discipline" risk identification, the heatmap facilitates rapid risk point localization, enabling the swift identification of high-risk disciplinary scenarios for subsequent warning and early intervention. Finally, considering the disciplinary and contextual nature of application risks, this paper explored stratified governance strategies for generative AI risks in education. It proposed countermeasures from the technological, institutional, and cultural levels, integrated with actual disciplinary contexts, guiding the way towards creating a safe and reliable ecosystem for the application of generative AI in education. It is worth mentioning that the case analysis results in this paper are mainly based on literature analysis and survey feedback. In the future, case studies will be conducted on more process data for generative AI teaching scenarios.

Acknowledgments

This work was supported by the Education Science Research Project of Naval Engineering University (Grant No.NUE2026ER064)

References

1. Wang Ziqi et al. Security Risks and Governance of AI Large Models [J]. Confidential Science and Technology, 2025(5), pp.22-27.
2. Dong Yun, Sun Xiaoling. Development, Application, and Challenges of "AI Large Models + Education" [J]. Information Systems Engineering, 2024(7), pp.110-113.
3. Huang Jian et al. Research on Data Security and Ethical Issues in the Era of Large Models [J]. Information Security and Communication Confidentiality, 2025(3), pp.46-53.
4. Guo Deyuan. Governance and Regulation of Generative AI in China [J]. Communications Enterprise Management, 2024(9), pp.38-41.
5. Liu Yiming et al. Data Security Risks and Legal Governance of Generative Large Models [J]. Network Security and Data Governance, 2024(12), pp.27-33.
6. Chu Leyang et al. Ethical Risks and Responses in the Application of AI Large Models in Education [J]. Journal of Suzhou University, 2024(1), pp.87-96.

7. Wu Hejiang et al. Characterization, Causes, and Governance of Ethical Risks in the Application of General Large Models in Education [J]. Tsinghua Journal of Education, 2024(2), pp.33-41.
8. Wu Yongxing et al. Security Risks and Governance Paths of Generative Artificial Intelligence Large Models [J]. China Modern Educational Equipment, 2025(9), pp.24-26.
9. Wu Di et al. Potential Risks and Avoidance of General Large Model Applications in Education: From the Perspective of Technological Ethics [J]. Journal of East China Normal University (Educational Sciences), 2024(8), pp.64-75.