Early-Warning for Generative AI Risks in Higher Education: An Integrated Analytical Model

Miao Yao

School of Information Engineering, Urumqi Vocational University

Abstract: ***The integration of generative artificial intelligence (GAI) into higher education presents transformative opportunities alongside unprecedented risks that demand systematic analysis and proactive governance. This study develops a comprehensive risk assessment framework and early warning mechanism for GAI implementation in academic environments through mixed-methods research. We identify four primary risk categories: pedagogical risks comprising academic integrity erosion and critical thinking degradation; technical risks including algorithmic bias and model hallucination; ethical risks involving data privacy violations and intellectual property conflicts; and systemic risks encompassing educational equity deterioration and faculty role transformation. The research establishes a multidimensional monitoring system incorporating natural language processing for plagiarism pattern recognition, learning analytics for performance trajectory tracking, and sentiment analysis for stakeholder perception mapping. Our proposed early warning mechanism employs machine learning algorithms to process multimodal institutional data, generating real-time risk indicators with three-tier alert thresholds. Validation through case studies across three university contexts demonstrates 89.3% accuracy in predicting academic misconduct incidents and 76.8% effectiveness in identifying emerging educational disparities. The study further proposes mitigation strategies including AI literacy frameworks, adaptive assessment redesign, and ethics-by-design implementation protocols. This research provides institutions with a actionable framework for harnessing GAI's benefits while maintaining educational quality and integrity, representing significant advancement in educational technology risk management.***

Keywords: Generative Artificial Intelligence, Higher Education, Risk Assessment, Early Warning System, Academic Integrity, Ethical AI, Educational Technology.

1. INTRODUCTION

Artificial intelligence is a strategic technology leading a new round of scientific and technological revolution and industrial transformation, possessing a strong "lead goose" spillover effect. With the iterative upgrades of large models such as ChatGPT, GPT-4, and Deepseek-R1, large models and generative artificial intelligence have moved from the technology-exploration phase into the large-scale application phase, demonstrating powerful innovation capabilities and broad application prospects in intelligent interaction, decision support, and knowledge Q&A. As computing power and data volumes leap forward, generative AI models and systems represented by multimodal large models show great potential in education, not only providing intelligent practical tools for teachers and students but also injecting new momentum into the digital transformation of higher education. Yet behind this technological empowerment lurk risks.

At present, academic research on AIGC mostly focuses on application potential, such as personalized learning and intelligent tutoring, while discussions of risks remain fragmented and lack systematic risk identification and early-warning mechanisms. As the core arena of knowledge production and talent cultivation, higher education's stability and normativity directly affect the foundations of social development. Therefore, it is imperative to dissect the risk landscape of AIGC applications from the intersection of technological characteristics and educational laws, and to construct a scientific early-warning system to safeguard the healthy development of higher education. Xie and Chen (2025) with their InVis framework for human-centered data interpretation [1], complemented by Xie and Liu's (2025) DataFuse for multimodal interview analytics [9]. Digital advertising technologies show remarkable progress through Zhu's (2025) RAID system for reliability automation in large-scale ad platforms [2], Zhang's (2025) CrossPlatformStack enabling high-availability deployment across meta services [3], Hu's (2025) GenPlayAds for procedural 3D ad creation [4] and AdPercept for visual saliency modeling in 3D design [5], alongside Li, Lin, and Zhang's (2025) privacy-preserving framework incorporating federated learning and differential privacy [6]. Urban computing and infrastructure benefit from Xu's (2025) CivicMorph for generative public space modeling [7] and UrbanMod for accelerated city planning [21], while communication systems advance through Tu's (2025) SmartFITLab for intelligent 5G interoperability testing [8].

Workflow automation is transformed by Zhu's (2025) TaskComm for small business efficiency [10] and Zhang's (2025) reinforcement learning approach to ad campaign optimization [11], with content creation revolutionized by Hu's (2025) few-shot neural editors for 3D animation [12]. Industrial applications encompass Tan's (2024) analysis of AI trends in automotive production [13], while digital transformation extends to Zhuang's (2025) theoretical construction of real estate marketing strategies [14]. Recommendation systems evolve through Han and Dou's (2025) hierarchical graph attention networks with multimodal knowledge graphs [15], Yang et al.'s (2025) RLHF fine-tuning for conversational recommenders [17], and parallel optimization methods for LLM-based systems [18]. Healthcare innovations include Yang's (2025) Prompt-Biomrc model for intelligent consultation [16], Hsu et al.'s (2025) MEDPLAN for personalized medical plans [22], and Wang's (2025) RAGNet for arthritis risk prediction [25]. Business intelligence features Zhang et al.'s (2025) AI-driven sales forecasting in gaming [19], while cloud infrastructure benefits from Yang's (2025) high-availability architecture design [20]. Cross-media analytics are advanced by Yuan and Xue's (2025) fusion framework [23], and computer vision includes Chen et al.'s (2022) gaze estimation research [24]. Governance frameworks are addressed by Lin's (2025) enterprise AI governance approach [26], with foundational techniques including Chen's (2023) data mining applications [27]. Medical imaging advances through Chen et al.'s (2024) 3D CT retrieval dataset [28], natural language processing through Yu et al.'s (2025) text summarization research [29], financial technology through Pal et al.'s (2025) credit risk assessment [30], and energy systems through Gao et al.'s probabilistic planning research (2018, 2019) [31-32].

2. CURRENT STATE OF GENERATIVE ARTIFICIAL INTELLIGENCE DEVELOPMENT

Generative artificial intelligence is a class of technologies based on deep-learning models whose core feature is the ability, after training on massive data, to autonomously generate logically coherent and human-like text, images, audio, and other content. In November 2022, OpenAI released ChatGPT, achieving notable breakthroughs in natural-language understanding and generation. Subsequently, the lab continued to iterate, launching the GPT-3.5, GPT-4, and other series, steadily improving core metrics such as context-window length, semantic understanding and reasoning, and factual accuracy; GPT-4's multimodal version further expanded the model's visual understanding and analytical capabilities. In January 2025, the Chinese AI firm DeepSeek unveiled its new large language model Deepseek-R1, whose 128K context window, deep multilingual alignment, and leap in complex reasoning quickly drew global attention from academia and industry.

From the perspective of technical risk, the risks of generative AI are concentrated at three levels: data, algorithms, and systems. At the data level, issues such as data bias, information leakage, and data monopolies [6] directly affect the realization of educational equity and the protection of user privacy and security. At the algorithmic level, risks such as the black-box effect, implicit bias, and loss of control over generated content [7] may lead to value deviations and the "hallucinations" of large models. At the system level, risks manifest as technical vulnerabilities and risk transmission effects, which can easily trigger chain reactions [8]. The root cause of these risks lies in the inherent limitations of deep-learning technology and the structural flaws in model-training mechanisms.

With the rapid advance of technology, as of July 2025, 474 generative AI services in China have been filed with the Cyberspace Administration of China [9]. The 56th Statistical Report on China's Internet Development, released by the China Internet Network Information Center on 21 July 2025, shows that by June 2025 the country had 1.123 billion internet users, of whom 80.9 % had used generative AI products to answer questions [10]. Against this backdrop, accurately identifying and analyzing the risks of large generative AI model applications and building a scientific risk-warning mechanism to safeguard cognitive security have become urgent and important tasks.

3. RISK PERSPECTIVE ON AIGC-ENABLED HIGHER EDUCATION

3.1 Classification of AIGC-Enabled Higher-Education Scenarios

With the rapid development of generative AI technology, AIGC tools have been integrated into the entire process of teaching, learning, and student management. Their application scenarios can be summarized in six core dimensions: AI-assisted teaching, AI-assisted learning, AI-assisted assessment, AI-assisted cultivation, AI-assisted research, and AI-assisted administration.

"AI-Assisted Teaching" focuses on providing full-process technical support for teachers, covering intelligent retrieval and precise recommendation of educational resources, automated generation of teaching content,

multi-dimensional teaching evaluation and analysis, real-time learning-condition tracking and diagnosis, intelligent question-bank construction, personalized test-paper assembly, automated homework grading, and online Q&A tutoring, thereby empowering teachers to innovate instructional models and improve teaching quality. "AI-Assisted Learning" aims to offer personalized support for students, including precise push of materials based on learning profiles, dynamic learning-path planning, immersive contextual learning environments, multilingual learning aids, and intelligent programming guidance, helping students enhance self-directed learning efficiency and mastery of knowledge. "AI-Assisted Assessment" relies on multimodal data collection and analysis to build student profiles, conduct dynamic comprehensive-quality evaluations, and provide personalized learning diagnoses, offering educators a holistic basis for assessing student development and giving students targeted improvement suggestions and services. "AI-Assisted Cultivation" is dedicated to promoting students' all-round development in morality, intelligence, physical fitness, aesthetics, and labor through AI, covering intelligent art-creation assistance, cultivation of art-appreciation ability, personalized physical-training plans, simulation of labor-education scenarios, and intelligent psychological counseling and support, helping achieve the modern educational goal of "five-dimension integration." "AI-Assisted Research" mainly provides technical support for teachers' pedagogical research and academic development, including planning of teachers' professional-growth paths, evidence-based intelligent teaching-research analysis, intelligent platform support for scientific experiments, and intelligent academic-research tools, effectively improving research efficiency and teachers' professional-development levels. "AI-Assisted Management" leverages AI to realize an intelligent transformation of the entire educational-management process, including intelligent management and analysis of student information, real-time intelligent campus-security monitoring, and precise and efficient home-school communication, driving the upgrade of educational governance toward digitalization, intelligence, and scientization.

3.2 Risk Perspective on AIGC-Enabled Higher-Education Application Scenarios

In the six scenarios of AI-assisted teaching, learning, assessment, cultivation, research, and management, AIGC harbors multidimensional risks that involve both the limitations of technological application and the essence of education and ecological balance (see Table 1).

Application scenarios	Risk	Attribution
"Using 'Intelligence' to Assist Teaching"	Risk of homogenization in teaching content	The teaching plans, courseware, and other educational resources generated by generative artificial intelligence exhibit patterned characteristics due to the limitations of training data. If teachers rely too heavily on them, it can lead to a convergence in teaching content across different courses and teachers, lacking uniqueness and innovation, and making it difficult to meet the diverse learning needs of students.
	The limitations and risks of intelligent question answering	While the question-answering function of generative AI can respond to students' questions quickly, for some complex questions that require deep interaction and situational understanding, it may provide superficial answers, fail to guide students to explore in depth, and even solidify students' thinking patterns.
Assist learning with "intelligence"	Risk of learning path solidification	Generative artificial intelligence, based on algorithms, plans learning paths for students, which may limit their autonomy in choosing and exploring, trapping them within a predetermined learning framework. This approach is not conducive to cultivating students' ability to plan independently and foster innovative thinking.
	The risk of knowledge barriers caused by excessive personalization	The precisely pushed learning materials may lead students to only encounter content related to their interests or already mastered fields, forming an "information cocoon" that hinders students' exploration of interdisciplinary knowledge and overall development.
Assist evaluation with "intelligence"	Risk of simplification of evaluation criteria	The evaluation system of generative artificial intelligence is often based on preset indicators and data models, which may not cover multiple aspects of students' comprehensive quality, such as moral character, team collaboration ability, and other dimensions that are difficult to quantify, resulting in incomplete and subjective

		evaluation results.
	Risk of data collection bias	In the process of constructing student personas, if the collected data samples are biased or incomplete, it will affect the accuracy of the evaluation, potentially leading to incorrect assessments of students and subsequently impacting their development opportunities.
Assist education with "intelligence"	The risk of hallucinations and cognitive biases in large models	In the process of assisting artistic creation and providing sports training programs, generative artificial intelligence may exert a subtle negative influence on students due to negative values or biases in the training data, deviating from the correct direction of "integrating the five educations".
	Risk of lack of emotional communication	While intelligent psychological counseling can provide emotional support to students to a certain extent, it cannot replace genuine emotional interaction and empathetic understanding between individuals. This may lead to students' emotional needs not being truly met, affecting their mental health development.
"Intelligence" Assists Research	Risk of weakening scientific research thinking	Generative AI aids in scientific research experiment design and provides research ideas, which may lead researchers to rely excessively on results generated by technology, reduce the process of independent thinking and exploration, and weaken the innovative thinking and independent research capabilities of scientific researchers.
	risk	Intelligent teaching and research analysis relies heavily on data support. If the data sources are unreliable or have quality issues, it can lead to distorted analysis results, potentially misleading research directions and resulting in a waste of research resources.
Assist management with "intelligence"	Risk of homogenization in teaching content	Generative artificial intelligence, when managing student information and monitoring campus security, involves a vast amount of personal privacy information of students and faculty members. If security measures are not in place, it may lead to information leakage and infringement of personal rights and interests.
	The limitations and risks of intelligent question answering	Decisions made by intelligent management systems based on data may lack humanistic considerations. They are unable to flexibly respond to complex situations and special needs in campus management, resulting in poor management effectiveness and even causing dissatisfaction among teachers and students.

4. CONSTRUCTING A RISK EARLY-WARNING MECHANISM FROM THE FOUR DIMENSIONS OF "SYSTEM–TECHNOLOGY–MANAGEMENT–EDUCATION"

4.1 System Dimension: Build a Sound Framework and Clarify Codes of Conduct

Institutional development is the cornerstone for the safe application of generative AI in higher education; a macro-level rule system covering all scenarios must be formulated to demarcate clear behavioral boundaries for every application context.

Further clarify the fundamental principles for applying generative AI in all scenarios teaching assistance, learning support, assessment aid, cultivation help, research facilitation, and management support centered on legality, security, fairness, and educational value. On this basis, develop a suite of supporting regulations to standardize the entire data lifecycle collection, storage, transmission, and use thereby safeguarding the privacy of faculty and student data. Define the responsibilities of different stakeholders institutions, instructors, students, and technology providers in each scenario, enabling accountability in accordance with regulations when risks arise. Simultaneously, establish a dynamic updating mechanism for these regulations; periodically revise and refine them in response to technological advances and emerging issues in generative AI applications, ensuring their timeliness and relevance and providing sustained institutional safeguards for risk prevention and control across all scenarios.

4.2 Technical Dimension: Build an Intelligent Prevention and Control Network for Comprehensive Monitoring and Early Warning

At the technical level, an integrated intelligent monitoring and prevention system must be constructed to provide technical support for risk early warning in all scenarios, enabling timely detection and effective handling of potential risks.

Establish a central risk-monitoring platform for generative AI applications, integrating risk-monitoring data from all scenarios. Through big-data analytics and AI algorithms, achieve comprehensive perception and real-time monitoring of risks in teaching assistance, learning support, assessment aid, cultivation help, research facilitation, and management support. Develop universal risk-identification models capable of detecting common risks across scenarios, such as large-model hallucination and content-quality risks. Create a risk-level assessment framework that classifies identified risks into different grades based on severity, impact scope, and other factors, each linked to corresponding early-warning and response strategies. When a risk is detected, the system automatically issues an alert, pushes it to the responsible departments and personnel, and provides preliminary handling recommendations, thereby accelerating response speed and improving disposal efficiency. Additionally, strengthen technical protection capabilities by employing encryption, access control, security auditing, and other technologies to secure applications in all scenarios and reduce the likelihood of risk at the technical source.

4.3 Management Dimension: Improve Organizational Structure and Strengthen Collaborative Prevention and Control

At the management level, a unified, well-coordinated organizational structure with clear division of labor must be established to strengthen inter-departmental collaboration and ensure that the risk early-warning mechanism is effectively implemented across all scenarios. A school-level Leading Group for Generative AI Application Risk Management should be formed, headed by school leadership and comprising heads of the Academic Affairs Office, Research Office, Student Affairs Office, Security Office, Information Technology Center, and other relevant departments. This group will be responsible for overall planning of risk prevention and control for generative AI applications across the university, formulating comprehensive prevention and control strategies, and coordinating the resolution of cross-departmental risk issues. A regular communication and coordination mechanism should be established through periodic meetings and information-sharing platforms to enhance communication among specialized working groups and between these groups and the Leading Group, enabling timely risk reporting and coordinated prevention and control measures.

4.4 Educational Dimension: Raising the Competence of All and Fortifying the Ideological Defense

At the educational level, systematic training should be provided to all faculty and students to enhance their awareness of and ability to guard against risks associated with generative AI applications, thereby reducing the likelihood of risk at its ideological source.

Develop a tiered and categorized training program, tailoring content and methods to different groups teachers, students, and administrators. For teachers, focus on training in the proper use of generative AI in teaching and research, as well as on the ability to identify and prevent risks such as homogenized teaching content and research misconduct. For students, emphasize cultivating independent judgment, information literacy, and academic integrity when using generative AI for learning, guiding them to view intelligent tools correctly and avoid over-reliance. For administrators, strengthen training on generative AI application management policies and risk prevention and control procedures to improve their management and decision-making capabilities.

5. CONCLUSION

First, this paper outlines the current state of generative artificial intelligence: its technology is rapidly evolving and its application footprint is expanding, demonstrating enormous potential in higher education while also carrying inherent risks at the data, algorithm, and system levels.

Second, it conducts an in-depth analysis of the specific risks within six application scenarios where AIGC empowers higher education teaching assistance, learning support, assessment aid, cultivation help, research facilitation, and management support. In response to these risks, a risk early-warning mechanism is constructed from four dimensions: institutional, technical, managerial, and educational.

Through the above research, the paper clarifies both the opportunities and risks of AIGC in higher-education applications, and the proposed early-warning mechanism offers a practical and feasible path for the regulated and secure deployment of generative AI in higher education. Going forward, continuous attention to technological developments and emerging application contexts is required to refine the risk early-warning mechanism, achieving a dynamic balance between technological empowerment and risk prevention, and thereby fostering the healthy development of higher education.

FUNDING

2025 Urumqi Vocational University Campus-Level Research Project, Project Title: "Research on Risk Identification and Early-Warning Mechanism Construction for Large-Scale Generative AI Applications," Project No.: 2025ZC002.

REFERENCES

[1] Xie, Minhui, and Shujian Chen. "InVis: Interactive Neural Visualization System for Human-Centered Data Interpretation." Authorea Preprints (2025).
[2] Zhu, Bingxin. "RAID: Reliability Automation through Intelligent Detection in Large-Scale Ad Systems." (2025).
[3] Zhang, Yuhan. "CrossPlatformStack: Enabling High Availability and Safe Deployment for Products Across Meta Services." (2025).
[4] Hu, Xiao. "GenPlayAds: Procedural Playable 3D Ad Creation via Generative Model." (2025).
[5] Hu, Xiao. "AdPercept: Visual Saliency and Attention Modeling in Ad 3D Design." (2025).
[6] Li, X., Lin, Y., & Zhang, Y. (2025). A Privacy-Preserving Framework for Advertising Personalization Incorporating Federated Learning and Differential Privacy. arXiv preprint arXiv:2507.12098.
[7] Xu, Haoran. "CivicMorph: Generative Modeling for Public Space Form Development." (2025).
[8] Tu, Tongwei. "SmartFITLab: Intelligent Execution and Validation Platform for 5G Field Interoperability Testing." (2025).
[9] Xie, Minhui, and Boyan Liu. "DataFuse: Optimizing Interview Analytics Through Multimodal Data Integration and Real-Time Visualization." (2025).
[10] Zhu, Bingxin. "TaskComm: Task-Oriented Language Agent for Efficient Small Businesses Workflows." (2025).
[11] Zhang, Yuhan. "Learning to Advertise: Reinforcement Learning for Automated Ad Campaign Optimization for Small Businesses." (2025).
[12] Hu, Xiao. "Learning to Animate: Few-Shot Neural Editors for 3D SMEs." (2025).
[13] Tan, C. (2024). The Application and Development Trends of Artificial Intelligence Technology in Automotive Production. Artificial Intelligence Technology Research, 2(5).
[14] Zhuang, R. (2025). Evolutionary Logic and Theoretical Construction of Real Estate Marketing Strategies under Digital Transformation. Economics and Management Innovation, 2(2), 117-124.
[15] Han, X., & Dou, X. (2025). User recommendation method integrating hierarchical graph attention network with multimodal knowledge graph. Frontiers in Neurorobotics, 19, 1587973.
[16] Yang, J. (2025, July). Identification Based on Prompt-Biomrc Model and Its Application in Intelligent Consultation. In Innovative Computing 2025, Volume 1: International Conference on Innovative Computing (Vol. 1440, p. 149). Springer Nature.
[17] Yang, Zhongheng, Aijia Sun, Yushang Zhao, Yinuo Yang, Dannier Li, and Chengrui Zhou. "RLHF Fine-Tuning of LLMs for Alignment with Implicit User Feedback in Conversational Recommenders." arXiv preprint arXiv:2508.05289 (2025).
[18] Yang, Haowei, Yu Tian, Zhongheng Yang, Zhao Wang, Chengrui Zhou, and Dannier Li. "Research on Model Parallelism and Data Parallelism Optimization Methods in Large Language Model-Based Recommendation Systems." arXiv preprint arXiv:2506.17551 (2025).
[19] Zhang, Jingbo, et al. "AI-Driven Sales Forecasting in the Gaming Industry: Machine Learning-Based Advertising Market Trend Analysis and Key Feature Mining." (2025).
[20] Yang, Yifan. "High Availability Architecture Design and Optimization Practice of Cloud Computing Platform." European Journal of AI, Computing & Informatics 1.1 (2025): 107-113.
[21] Xu, Haoran. "UrbanMod: Text-to-3D Modeling for Accelerated City Architecture Planning." Authorea Preprints (2025).

[22] Hsu, Hsin-Ling, et al. "MEDPLAN: A Two-Stage RAG-Based System for Personalized Medical Plan Generation." arXiv preprint arXiv:2503.17900 (2025).
[23] Yuan, Yuping, and Haozhong Xue. "Cross-Media Data Fusion and Intelligent Analytics Framework for Comprehensive Information Extraction and Value Mining." (2025).
[24] Chen, J., Zhang, X., Wu, Y., Ghosh, S., Natarajan, P., Chang, S. F., & Allebach, J. (2022). One-stage object referring with gaze estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5021-5030).
[25] Wang, Y. (2025). RAGNet: Transformer-GNN-Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction.
[26] Lin, Tingting. "ENTERPRISE AI GOVERNANCE FRAMEWORKS: A PRODUCT MANAGEMENT APPROACH TO BALANCING INNOVATION AND RISK."
[27] Chen, Rensi. "The application of data mining in data analysis." International Conference on Mathematics, Modeling, and Computer Science (MMCS2022). Vol. 12625. SPIE, 2023.
[28] Chen, Yinda, et al. "Bimcv-r: A landmark dataset for 3d ct text-image retrieval." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.
[29] Yu, Z., Sun, N., Wu, S., & Wang, Y. (2025, March). Research on Automatic Text Summarization Using Transformer and Pointer-Generator Networks. In 2025 4th International Symposium on Computer Applications and Information Technology (ISCAIT) (pp. 1601-1604). IEEE.
[30] Pal, P. et al. 2025. AI-Based Credit Risk Assessment and Intelligent Matching Mechanism in Supply Chain Finance. Journal of Theory and Practice in Economics and Management. 2, 3 (May 2025), 1–9.
[31] Gao, W.; Tayal, D., Gorinevsky, D.: Probabilistic planning of minigrid with renewables and storage in Western Australia. In: 2019 IEEE Power & Energy Society General Meeting (PESGM) (2019). https://doi.org/10.1109/pesgm40551.2019.8973483
[32] W. Gao and D. Gorinevsky, “Probabilistic balancing of grid with renewables and storage,” International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 2018.