

Research on Ethical Risks of Generative AI Applications in Higher Education

Wenbing Chang, Yunxiao Zhao, Siting Cao, Fuping Zeng, Shenghan Zhou*

School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China

zhoush@buaa.edu.cn

Abstract. This study constructs a dual-track academic risk detection system that integrates deep learning and graph theory algorithms, based on a heterogeneous dataset of documents including student assignments and academic journal papers. Driven by rapid advances in large language models, generative AI has been widely applied in higher education. While enhancing educational efficiency, it incurs severe academic ethical risks. Conventional plagiarism detection methods fail to identify AI-generated and semantically revised plagiarized content. The study establishes a multi-dimensional quantification model for AI generation probability, which integrates perplexity, burstiness, lexical features, and syntactic features to evaluate the ratio of AI-generated content in texts. Then, the paper adopts a Sentence - BERT model to map texts to high - dimensional semantic vectors, constructs a semantic similarity matrix, and builds an academic collusion network to identify covert plagiarism and improper citation behaviors. The empirical results show that about 15% of the documents in the dataset exhibit significant AI generation characteristics, and the language complexity of student assignments is significantly lower than that of academic papers, indicating a high risk of over-reliance on AI tools. Meanwhile, the semantic detection system accurately identifies multiple cases of academic misconduct, including direct plagiarism with a semantic similarity of 0.986 between documents, and reveals typical cheating patterns such as twin-star mutual plagiarism, dependent plagiarism, and free-riding group collusion. This study verifies the feasibility of localized large models for academic integrity evaluation under the premise of privacy protection, and proposes targeted governance paths for ethical risks of generative AI in higher education, providing technical and theoretical support for the construction of academic integrity systems in the era of human-machine collaboration.

Keywords: AI-Generated Content Detection; Academic Integrity; Semantic Similarity Analysis; Academic Collusion Network

1. Introduction

In recent years, generative artificial intelligence has seen disruptive breakthroughs in natural language processing, with large language models like ChatGPT and DeepSeek excelling in text generation, summarization and knowledge reasoning. In higher education, AIGC technologies have been widely applied in learning support, academic writing and data organization, advancing personalized teaching and intelligent education, lowering knowledge access barriers and improving information-processing efficiency. However, their pervasive use in coursework, reports and theses has triggered serious academic integrity and ethical concerns. High fluency and semantic consistency make AI-generated content evade conventional plagiarism detection, challenging current academic governance. Risks include impaired academic integrity, weakened critical and innovative thinking, data privacy threats and algorithmic bias. Traditional detection methods based on TF-IDF and surface similarity are ineffective, while mainstream commercial systems rely on cloud APIs, posing privacy leakage risks and lacking localized solutions for Chinese higher education.

This study explores ethical risks posed by generative AI in higher education, conducting systematic research on AI-generated content detection and academic misconduct identification. Its core contributions include: a multi-dimensional AI generation rate quantification model integrating lexical, syntactic and statistical linguistic features for accurate, interpretable academic text detection; a Sentence-BERT-based semantic framework to identify covert text spinning beyond traditional plagiarism tools; a graph theory-based academic collusion network revealing group misconduct

topological structures; and empirical validation via real academic datasets, alongside targeted governance recommendations.

2. Related Work

The ethical risks of artificial intelligence in education have become a research hotspot in the field of educational technology and AI ethics in recent years. Existing research has comprehensively explored the dual effects of AIGC on higher education, including its positive impacts and ethical risks. Hogan and Corăbian^[1] found that generative AI, although capable of producing novel content, is vulnerable to ethical problems including biased training data, and its generated text usually has extremely low perplexity for the corresponding language model, which is an important feature distinguishing it from human-written text. Zhang^[2] pointed out that generative AI is profoundly transforming the higher education ecosystem, reshaping the generation of teaching content, forms of teaching interaction, and the logic of educational evaluation, while also triggering multi-dimensional technological ethical risks including data security, algorithmic bias, and teaching responsibility. Nie et al.^[3] found that while AIGC promotes the production of interdisciplinary knowledge, it may also hinder the cultivation of students' critical thinking and independent thinking ability. Zhang and Xu^[4] elaborated on the ethical risks brought by AI in education, including data privacy and security, algorithmic discrimination, and the reduction of human interaction. Bu^[5] summarized three core ethical risks of educational AI: educational data security risks, the risk of deconstructing the teacher-student role structure and educational inequality, and so on. The past researches have proposed countermeasures for the governance of ethical risks of artificial intelligence from multiple dimensions. Li^[6] constructed an ethical framework for civil servants' application of AIGC from the aspects of criterion formulation, decision evaluation, professional ethics, and multi-subject supervision. Zhou^[7] explored the prevention countermeasures of emotional ethical risks in AI education from four levels: improving ethical risk awareness, strengthening social ethical norms, clarifying the functional positioning of education subjects, and strengthening emotional interaction practice. Liang^[8] proposed to avoid the ethical risks of AI in education by accelerating theoretical construction, promoting the deep integration of AI and education, improving ethical regulations, and enhancing the sense of subject responsibility.

These studies offer theoretical underpinnings and practical references for this research but exhibit obvious deficiencies^{[9][10]}. In ethical risk analysis, most focus on theoretical discussions, lacking empirical analysis and quantitative detection methods. In AI-generated text detection, statistical linguistic feature-based methods (with PPL as the core) and relevant indicators are widely adopted, yet single-indicator methods suffer from high false positive rates and poor adaptability to academic text scenarios. In academic plagiarism detection, research has advanced from keyword matching to deep semantic understanding (e.g., Sentence-BERT), while graph theory methods are rarely applied in Chinese higher education contexts.

To mitigate these shortcomings, this paper focuses on the following research directions: conduct empirical studies on ethical risks using real academic data; construct a multi-dimensional fusion model for AI-generated text detection; integrate Sentence-BERT and graph network analysis to identify both pairwise and group plagiarism; and propose targeted, operable governance paths for ethical risks and academic misconduct in higher education based on technical detection outcomes.

3. Methodology

3.1 Dataset and Preprocessing

This study adopts a dataset of 64 anonymized academic documents, including student course assignments and formal journal papers, with filenames numerically encoded from 0001 to 0110 for privacy protection and stored in PDF and Word formats. The dataset features notable multi-source heterogeneity: assignments show loose layout, colloquial expressions and uneven quality, while

papers have rigorous structure, standardized citations and strong professionalism, posing challenges to unified modeling and analysis. To convert unstructured raw documents into machine-readable standardized corpus, an automated ETL preprocessing pipeline is constructed. Text is extracted from different formats via pdfplumber and python-docx libraries to retain content integrity and structural information like paragraphs and sentences. Regular expression scripts are then applied for deep cleaning, removing page numbers, headers, footers, non-Chinese special symbols, redundant spaces and garbled characters, and standardizing line breaks to eliminate semantic noise.

For Chinese text processing, the high-accuracy, academically adaptable jieba tokenizer is used for word segmentation. An 86-word academic stop word list covering function words is built. Stop words, words shorter than two characters, and blank words are filtered out to retain semantically meaningful core units. TF-IDF-based keyword extraction combined with part-of-speech filtering retains nouns, verbs, adjectives, adverbs, and proper nouns, laying a solid data foundation. Additionally, a heuristic classifier based on feature word weights is developed for document classification. It scans initial and final document sections, using academic paper-related and student assignment-related feature words for weighted discrimination, with low-confidence cases manually verified via formatting. The dataset is ultimately divided into "Homework" and "Paper" classes, establishing a reliable labeling system for differentiated risk assessment.

3.2 Multi-dimensional AI Generation Rate Quantification

The past research shows that AI-generated text presents a significant 'style convergence' phenomenon in the generation process, that is, it has a relatively stable pattern in vocabulary selection, logical connection mode and abstract expression use, which is significantly different from human-written text. Based on this core feature, we constructed a multi-dimensional AI generation rate quantification model, which integrates three core modules: statistical linguistic feature detection, lexical feature detection, and syntactic feature detection, to achieve a comprehensive and accurate evaluation of the AI generation probability of the text.

The first module is the hybrid detection method based on phrase and statistical linguistic features, which includes three complementary indicators. The first indicator is perplexity (PPL) based on the language model, which is the core indicator to measure the fitting degree of the probabilistic language model to a given text sequence. For a text sequence with N tokens $X = (x_1, x_2, \dots, x_N)$, the perplexity is defined as the reciprocal of the geometric mean of the probability assigned by the model to the sequence, and the formula is:

$$PPL(X) = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(x_i | x_1, \dots, x_{i-1})\right) \quad (1)$$

where $P(x_i | x_1, \dots, x_{i-1})$ is the conditional probability of the current token x_i predicted by the language model based on the preceding context. In this study, we locally deployed the Qwen2.5-1.5B large language model for zero-shot inference to calculate the perplexity of the text. A lower perplexity value indicates that the model has a higher prediction confidence in the text, and the text sequence is more consistent with the generation pattern of the AI model.

The second indicator is the word frequency distribution entropy, which measures the concentration or dispersion of vocabulary use in the text. For a segmented text with V unique word types, let the occurrence probability of the k -th word type be p_k , then the information entropy H of the word frequency distribution is calculated as:

$$H = -\sum_{k=1}^V p_k \log_2 p_k \quad (2)$$

A higher entropy value denotes a more uniform word frequency distribution and lower word repetition, whereas a lower value signals textual confinement to cycles of specific vocabulary or phrase patterns—a typical feature of AI-generated content. The third indicator is a modified lexical diversity metric based on the Type-Token Ratio (TTR). Traditional TTR calculates the ratio of unique word types to total tokens but is highly sensitive to text length. Addressing this issue, the

log-transformed TTR (LogTTR) is employed, which mitigates the impact of increasing text length via logarithmic transformation, with the formula as follows:

$$\text{LogTTR} = \frac{\log(\text{TypeCount})}{\log(\text{TokenCount})} \quad (3)$$

Human-written text usually has a higher lexical diversity index, while AI-generated text tends to use common and safe vocabulary combinations, resulting in a relatively low diversity index. The second module focuses on lexical feature-based AI detection, which measures the match between target text and typical lexical patterns of AI-generated content. First, a cross-disciplinary AI benchmark corpus covering computer technology, biomedicine, economics and finance is built via prompt engineering to simulate AI's academic writing output and avoid subject-specific bias. Then, N-gram statistical analysis identifies high-frequency feature phrases, forming a lexicon of three categories: transitional connectives, common academic verbs and abstract nouns. Based on this lexicon, the proportion of matched phrases in the target text is calculated, with AI term density defined as the core AI term ratio score. A higher ratio indicates a greater likelihood of AI generation or deep assistance. The third module analyzes sentence-level features to assess AI generation risk, calculating a weighted comprehensive score through five metrics: the proportion of sentences starting with summarizing, causal or progressive opening words; sentence length balance measured by variation coefficient; average AI feature word density per sentence with density stability weighting; proportion of sentences containing logical connectives; and proportion of sentences using AI-favored fixed patterns. After each dimension scores are calculated, the comprehensive AI generation rate scoring model is constructed through weighted fusion, and the formula is:

$$\text{AI_Generation_Rate} = 0.5 \times \text{AI_Term_Ratio} + 0.4 \times \text{Sentence_Feature_Score} + 0.1 \times \text{Phrase_Hybrid_Test_Score} \quad (4)$$

Among them, the AI term ratio is given the highest weight of 50% as the most direct evidence, the sentence feature score is given 40% weight as the structural feature of the document, and the phrase hybrid test score is given 10% weight as an auxiliary to prevent model overfitting. We also set up a reward mechanism: when either the AI term ratio or the sentence feature score is greater than 50%, the AI generation rate increases by 5%; when both are greater than 50%, the AI generation rate increases by another 5%. Finally, we divided the AI generation risk into four levels: AI generation rate <20% is pure human writing, 20%-30% is a small amount of AI assistance, 30%-50% is mostly AI writing, and >50% is pure AI generation.

3.3 Semantic-based Improper Citation and Plagiarism Detection

For the identification of improper citation, we first constructed a multi-dimensional citation feature lexicon based on the linguistic characteristics of academic texts, covering four categories: direct citation, research basis, viewpoint reference, and literature annotation. The first three categories are citation behavior feature words used to identify citation behaviors, including 'cite', 'reference', 'according to', 'based on', 'research shows', 'point out', etc. The last category is reference annotation feature words used to verify whether there is standardized annotation, including 'references', 'bibliography' and citation symbols '[]'. We defined improper citation as the occurrence of citation feature words in the text without corresponding reference annotation. Through feature word matching and semantic verification, we accurately identified the documents with improper citation problems, counted the number of unlabeled citation feature words and the corresponding specific words, and realized the quantitative evaluation of the non-standard citation behavior.

To address the limitation that traditional methods are incompetent in identifying covert text spinning within plagiarism detection and text similarity analysis, a Sentence-BERT model rooted in the Transformer architecture is introduced, as illustrated in Fig. 1. The Sentence-BERT model is capable of mapping text of any length to a 768-dimensional high-dimensional dense semantic vector that preserves the deep semantic information of the input text. Within this semantic space,

cosine similarity is employed to quantify the logical proximity between documents, with the relevant formula presented as follows:

$$\text{CosineSimilarity}(A, B) = \frac{A \cdot B}{|A| \times |B|} \quad (5)$$

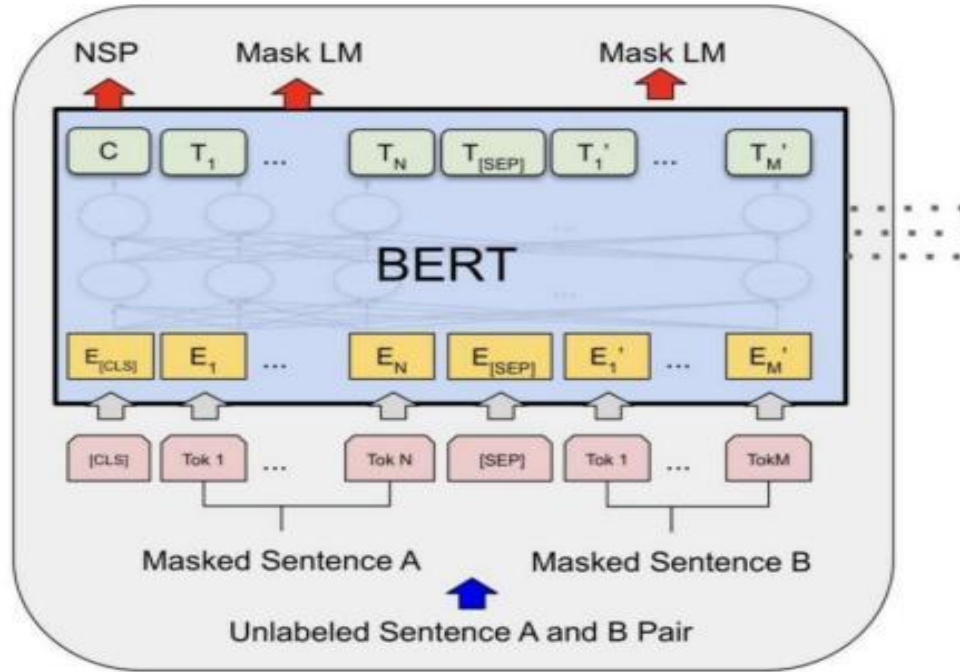


Fig. 1. Sentence-BERT model

where A and B are the semantic vectors of two documents, the numerator is the dot product of the vectors, and the denominator is the product of the modulus lengths of the vectors. The value range of cosine similarity is $[0,1]$, and the closer the value is to 1, the higher the semantic similarity between the two documents.

Based on this method, we constructed a 64×64 full semantic similarity matrix for all documents, and visualized it as a heat map. The horizontal and vertical axes of the heat map correspond to the document numbers, and the color scale transitions from light color (low similarity) to dark color (high similarity). Through the heat map, we can intuitively observe the global similarity distribution of the document set, identify the local high-brightness areas, and lock the document pairs and document clusters with high similarity. This method can effectively identify the plagiarism behavior after synonym replacement, sentence structure reorganization and other text spinning means, which is significantly better than the traditional TF-IDF based similarity calculation method.

3.4 Academic Collusion Network Analysis

To go beyond the traditional pairwise comparison perspective and reveal the group academic misconduct behavior, we used graph theory methods to transform the semantic similarity matrix into a visual network topology structure. We constructed an undirected graph $G = (V, E)$, where the node set V represents the 64 original documents, and the edge set E represents the high-confidence semantic similarity relationship between documents, with the similarity threshold set to 0.85. That is, an edge is established between two nodes only when the semantic similarity between the corresponding documents is higher than 0.85, and the weight of the edge is the similarity value.

Analysis of network topological characteristics reveals three typical academic misconduct patterns. The "twin-star" mode appears as an isolated two-node strongly connected component with a high-weight edge, representing direct mutual plagiarism between two students. The "dependent" mode features cross-domain edges between student assignment and academic paper nodes,

signifying uncredited large-scale semantic overlap via text spinning. The "free-riding" mode shows star-shaped or triangular subgraphs, reflecting group cheating through shared document templates.

4. Empirical Results and Analysis

4.1 AI Generation Risk Assessment Results

To verify the effectiveness of the AI generation rate detection model, we added 22 pure AI-generated documents as the test set. The test results show that the model has a recognition accuracy of 100% for AI-generated documents from the same source as the training corpus, and 90% of pure AI documents generated by different models are correctly identified as mostly AI writing, verifying the good detection performance of the model.

For the real dataset, the detection results show that the average AI generation rate of student course assignments is 27.58%, and the average AI generation rate of academic papers is 29.74%. About 15% of the documents have an AI generation rate of more than 35%, which are in the high AI generation risk area. Statistical analysis of perplexity shows that the median perplexity of the student assignment group is significantly lower than that of the journal paper group, with a low-value aggregation phenomenon, indicating that student assignments are the hardest hit area of AIGC abuse. We also found that some highly similar documents are also in the high AI risk area, suggesting a compound academic misconduct pattern of AI generation and direct replication.

4.2 Improper Citation and Plagiarism Detection Results

Through improper citation detection, numerous documents with citation irregularities were identified. Most such documents contain abundant citation feature words yet lack corresponding reference lists and standardized annotations, violating basic academic norms and representing a key academic integrity risk posed by generative AI. Plagiarism detection via a global similarity heat map revealed a sparse matrix pattern: outside the main diagonal, most areas have a similarity below 0.3, indicating extremely low semantic overlap between random document pairs and independent completion by most students. However, local high-brightness areas flagged high-risk pairs and clusters, including near-identical pairs showing direct plagiarism and high-similarity clusters suggesting group plagiarism or template sharing.

4.3 Academic Collusion Network Analysis

The constructed academic collusion network exhibits notable sparsity: approximately 75% of nodes are isolated with a degree of 0, indicating that most documents have no high-similarity associations with others and that the vast majority of students are not part of the plagiarism network. The identified collusion structures represent local anomalies rather than a systemic breakdown of academic integrity. Analysis of the network's topological properties validates the three typical academic misconduct patterns outlined above. The "twin-star" structure corresponds to direct peer-to-peer plagiarism between students. Cross-domain edges between assignment nodes and paper nodes reveal the "dependent" plagiarism pathway, whereby academic papers are adapted into assignments via text spinning. Star-shaped and triangular closed subgraphs reflect group cheating behaviors in which multiple students share a common document source, a phenomenon particularly prevalent in group assignments, as displayed in Fig 2.

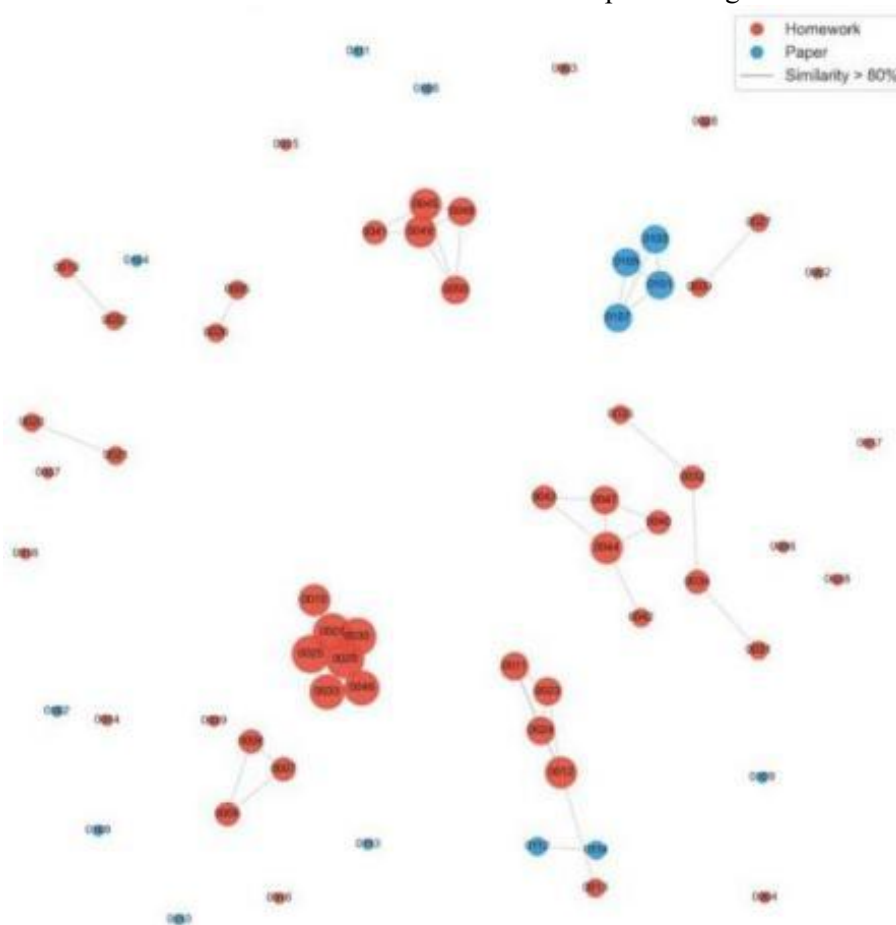


Fig. 2. Result of Academic Collusion Network Analysis.

5. Discussion and Conclusion

The empirical results confirm that generative AI (AICG) application in higher education poses significant academic integrity ethical risks, with misconduct characterized by intelligence, covertness, and grouping. Traditional word frequency-matching detection methods fail to address AI-generated text and semantic text spinning, challenging the existing academic integrity system. Root causes lie in anthropomorphic AI bypassing detection (technical), outcome-oriented assessment neglecting students' thinking processes (educational), and students' inadequate AI academic ethics awareness and over-reliance on AI tools (subjective).

This study makes theoretical and practical contributions: it constructs a multi-dimensional AI-generated text detection framework and conducts empirical analysis to fill the gap of insufficient empirical evidence in existing research; its low-cost dual-track detection system protects academic data privacy and provides feasible technical solutions, while the academic collusion network analysis offers a new perspective for group misconduct identification. Limitations include text-only detection, inadequate generalization in vertical disciplines, and a limited dataset. The key findings demonstrate that 15% of the documents within the dataset exhibit distinct AI-generated characteristics, with the majority deriving from student assignments. Associated academic misconduct encompasses pairwise plagiarism, dependent plagiarism, and group collusion. In response to these issues, targeted governance strategies should prioritize the development of multi-modal detection systems, the implementation of process-oriented assessment frameworks, and the enhancement of AI ethics education initiatives, all of which are designed to guide students in upholding the principles of academic integrity.

Acknowledgements

This research was funded by the National Natural Science Foundation of China (Grant No.72371013) and the Graduate Student Education & Development Foundation of Beihang University (Grant No. JGXY2025048C).

References

- [1] Hogan, Neil R., and Gabriela Corăbian. Generative Artificial Intelligence in Violence Risk Assessment: Emerging Technology and the Ethics of the Inevitable. *Behavioral Sciences & the Law*, 2025, 43(6): 606–615, <https://doi.org/10.1002/bsl.70014>.
- [2] Zhang Y. Research on the Construction of Digital Intelligence Competence of University Teachers and the Prevention of Technological Ethics Risks in the Era of Generative AI. *Journal of Computer Technology and Electronic Research*, 2025, 1(4), DOI:10.70767/jcter.v1i4.542
- [3] Nie, X. and Y. Tian, et al.. The impact of generative artificial intelligence on students' higher order thinking: Evidence from a three-level meta-analysis. *Education and Information Technologies*, 2025, 30 (17): 25359-25390.
- [4] Zhang, H. And Xu, L. Ethical Risks and Governance Approaches of Generative Artificial Intelligence in Education——Based on the Practical Investigation of the Russell University Group. *Modern Educational Technology*, 2024, 34 (6): 25-34.
- [5] Bu Q Y. Ethical Risks in Integrating Artificial Intelligence into Education and Potential Countermeasures. *Science Insights*, 2022, 41(2). 561–566. <https://doi.org/10.15354/si.22.re067>
- [6] Li Y. Research on Ethical Issues and Countermeasures of Civil Servants' Application of Generative Artificial Intelligence Generated Content. Beijing: Beijing University of Posts and Telecommunications, 2024.
- [7] Zhou W M. Emotional ethical risks in AI education Prevention research. Xiangtan: Xiangtan University, 2023.
- [8] Liang G Y. The Ethical Risk and Avoidance of the Application of Artificial Intelligence in Education. Taiyuan: Shanxi University, 2021.
- [9] Wang, Q. Application and Prevention of Generative AI in Graduate Research Assignments. *Software Guide*, 2025, 24(03):1-7. DOI:10.11907/rjdk.242040.
- [10] QI, J. and Y. XU, et al. The Impact of Generative Artificial Intelligence Tools on College Students' Critical Thinking and Autonomous Learning Ability. *Frontiers of Education in China* 2025, 20: 398-412.