Journal of Humanities and Social Sciences Studies
ISSN: 2663-7197
DOI: 10.32996/jhsss
Journal Homepage: www.al-kindipublisher.com/index.php/jhsss

JHSSS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Impact of AIGC Testing on Educational Evaluation and Governance Pathways

Yu Jing
Nanning vocational and technical university, Nanning, China
Corresponding Author: Yi Wang, **E-mail**: www.363489563@qq.com

| ABSTRACT

Currently, there is instability in AIGC detection, which is manifested in the field of education as significant differences in the AI generation rate of the same text at different times and platforms, which brings troubles to academic evaluation, graduation recognition, and so on. The root cause of this is the standard drift caused by technology iteration, the deviation of educational scenarios in training samples and the dependence of detection logic on superficial features. In this regard, it is recommended to build a "human-intelligence" collaborative governance framework: to reconstruct the evaluation system at the institutional level, and to establish a mechanism of "human-led + technology-assisted + process traceability"; to form a consensus on governance through collaboration between multiple parties; and to develop a proprietary model and standardize the standards at the technological level. The research aims to improve the reliability of testing, and to balance the academic and research standards. The study aims to improve the reliability of testing, balance academic integrity and innovation ability cultivation, and provide support for the stabilization of educational evaluation ecology in the digital era.

| KEYWORDS

AIGC Testing; Educational Evaluation; Technical Reliability; Generative Artificial Intelligence; Governance Framework

| ARTICLE INFORMATION

ACCEPTED: 19 July 2025 **PUBLISHED:** 14 August 2025 **DOI:** 10.32996/jhsss.2025.7.8.4

1.Introduction

With the rapid development of generative AI technology, AIGC testing is gradually becoming a tool for guarding academic integrity in education. However, its instability-the AI generation rate of the same educational text detected at different times and platforms varies significantly-is impacting on educational evaluation. From homework grading in basic education to dissertation recognition in higher education, this instability not only interferes with the normal teaching order, but also raises questions about the fairness of evaluation. In this paper, we focus on the reliability of AIGC testing in educational scenarios, analyze its impact and causes through empirical research, and finally propose a collaborative governance path of "human-intelligence" to provide reference for optimizing the educational evaluation system in the digital era.

2.Methodology

2.1 Random Sampling Method

Select five different types of academic paper texts and test them multiple times on five mainstream AIGC testing platforms, record and analyze the differences in the AI generation rate of the same text on different platforms, and quantify the volatility of the test results.

2.2 Questionnaire Survey Method

A targeted questionnaire was designed and systematically analyzed using data analysis software to understand the status quo and major problems of teachers, students and other groups in the field of education in the use of AIGC testing, and to accurately grasp the impact of AIGC testing on education evaluation.

2.3 Interview Method

Semi-interviews were conducted with some education-related personnel, such as primary and secondary school teachers, college instructors, education management, and testing technology developers, to understand their views on AIGC testing, the dilemmas they face in practice, and suggestions for improvement.

3.Definition and Characteristics

3.1 Definition of "AIGC Detection"

AIGC detection refers to the technical means of analyzing the linguistic features and structural logic of text through algorithmic models to identify whether it is created by generative artificial intelligence, such as ChatGPT, DeepSeek, and Literary Heart of the Mind, aiming at distinguishing between original human content and AI-generated content, which is mostly used for academic integrity regulation .[1]

3.2 Characteristics of the instability of the AIGC test in educational scenarios

3.2.1 Drift of detection results in the time dimension

The AI generation rate of the same educational text detected at different times showed significant fluctuations. The articles published by well-known Chinese actors in 2024 showed large differences in the detection results. in June 2024, the AI rate was 0 by Zhi.com, however, by May 23, 2025, Zhi.com retesting showed that the AI rate became 77.8%, the Wipro platform detection result was 44.5%, and the AI rate of the PaperPass platform was even as high as 91.48%.[2] Some students in the questionnaire said that when they submitted their papers for testing at the beginning of the semester, the AI rate was at a low level and met the academic requirements; however, when they were tested again at the end of the semester or retested in the next year due to graduation audits and other needs, the AI rate increased, which caught people off guard. This difference in the time dimension of the test results seriously affects the stability and authority of the academic results, making the academic evaluation as if caught in a fickle "time vortex".

3.2.2 The standardization of platform dimension

The evaluation standards of different testing platforms are significantly different, leading to the phenomenon of "platform dependence". The cross-platform testing of five educational texts shows that the highest difference in the AI generation rate of the same text reaches 76.6%. This huge difference in cross-platform testing results will cause confusion: for students and researchers, they do not know which testing platform results should prevail, so as to improve the paper; for academic journal editors, it is more difficult to make accurate judgments on the originality of the paper in the process of reviewing the manuscript, which increases the difficulty of the academic journals, especially social science journals, in the original gatekeeping.[3]

3.2.3 Selective misclassification of text types

A reporter tested for AIGC detection, uploading Zhu Ziqing's famous piece "Moonlight in a Lotus Pond" and Liu Cixin's "Wandering Earth" segments to a commonly used essay detection system. The results show that the overall degree of suspicion of the AI-generated content of these two classic works reached 62.88% and 52.88% respectively. Previously, someone else posted that the AI rate of "The Preface of the Tengwang Pavilion" was surprisingly 100%. [4]In the field of academic texts, more original and highly original papers that have been thoroughly researched and carefully written by the authors are also often misjudged as AI-generated.

4. Content of the study

4.1 Impact of the instability of AIGC testing on educational evaluation

4.1.1 Evaluation dilemma at the basic education level

The instability of AIGC testing has permeated the daily evaluation of primary and secondary schools. Some secondary school teachers have reported that the AI generation rate of the same argumentative essay varies from platform to platform, making it impossible to judge the true writing level of students and even leading to completely contradictory verdicts. Such fluctuations

make it difficult for teachers to judge the true writing level of students-if they refer to the results of a certain platform, they may misjudge students' use of AI, and if they ignore the testing data, they are worried about condoning academic misconduct. What's even more concerning is that the absurd results of the classic text test are misleading teaching evaluations, reflecting that the instability of the AIGC test has substantially interfered with primary and secondary schools' daily homework grading and classroom performance evaluation.

4.1.2 Crisis in academic evaluation at the higher education level

Graduation season, which should be the season for students to harvest the fruits of knowledge, but because of the "oolong" of the AIGC testing technology, many students are plunged into the quagmire of anxiety. Some social platforms are full of students' complaints and requests for help. Some students helplessly said that their own painstaking efforts, the originality of up to 77% of the dissertation, but was mercilessly marked as "high AI rate" by the AIGC detection system.[5]From the screenshots they shared, we can clearly see that the paragraphs that were misjudged by the system were the results of their in-depth research and repeated polishing. These students expected to demonstrate their academic ability through the dissertation, but due to the misjudgment of AI detection, they faced the dilemma of having their dissertation questioned and their graduation hindered, and were forced to spend a lot of time and energy to "lower the AI rate", and even used some absurd methods, such as replacing periods with commas, and using translation software for multiple transitions, etc., just to pass the elusive test [6]. elusive test.[6]

4.1.3 Dissolution of the credibility of educational evaluation

The instability of the AIGC testing technology may to some extent dissipate the credibility of educational assessment. In the student population, some of them attribute differences in testing results simply to the choice of platform rather than to a real gap in their own abilities, which can lead to cognitive biases in their perception of their own learning outcomes. The trust of the teacher community in the testing tools will also continue to decrease, making most of them skeptical of their reliability, and perhaps only a few of them will consider the results of the tests as the main basis for evaluating their students. This crisis of trust will have a direct impact on the core value of "student development" in educational assessment. Educational evaluation will be caught in a dilemma - over-reliance on technology may deviate from teachers' objective judgment of students' real abilities; while skepticism about technology may miss the potential of technology-enabled educational evaluation. How to balance the application of technology and the essence of education has become an urgent problem to be solved.

4.2 Analysis of the Causes of Lack of Reliability of AIGC Detection in Educational Scenarios

4.2.1 Lag between technology iteration and adaptation to educational specifications

The AIGC detection technology is iterated in weeks, while the norms of educational texts, such as the structure of academic papers and the presentation of teaching cases, need to be accumulated for a long period of time, which leads to the "standard drift" of AIGC detection. Early algorithms focused on vocabulary novelty, which is in conflict with the norms of educational texts; after the introduction of the "structural regularity" index, some texts that comply with academic norms will be given a higher AI suspicion value. For example, the AI detection rate of some standard structure writing texts is higher than that of loose structure.

4.2.2 Educational scenario bias in the training sample

Through interviews with technicians of some mainstream testing platforms, it can be seen that in the training data of the platform, the proportion of educational texts is relatively low, especially in primary and secondary education and vocational education. This imbalance in sample composition makes it difficult for the model to accurately grasp the specificity of educational texts, for example, the expression "integration of science and reality" in vocational education is often misjudged due to the inclusion of a large number of operational terminology; "contextualized cases", which combines storytelling and education, are also easily misjudged due to the lack of operational jargon in basic education. In basic education, "contextualized cases", which combine storytelling and education, are also difficult to be accurately identified because of their complex features. More critically, the language styles and expressive features of texts of different school segments differ significantly, but they are forced to be included in the same evaluation criteria, which ultimately leads to bias in the evaluation results.

4.2.3 Inadequate Adaptation of Detection Logic to Educational Thinking

The current AIGC detection technology mainly relies on superficial feature analysis, which makes it more difficult to understand the deep thinking of educational texts. The critical thinking and spiral argumentation emphasized in education is often carried out in the cyclic structure of "point of view-rebuttal-argumentation", which is in line with the laws of education, but is different from the "efficient" argumentation preferred by AI. This kind of argumentation, which is in line with the laws of education, will be recognized as AI-generated because it is different from the AI's preferred "efficient" argumentation. In addition, there are also some repeated use of professional terms, repeated sentences with emotional coloring in teaching reflections, and logical

hierarchical expressions, etc., all of which can be misjudged as traces of machine-generated, thus making the detection results insufficiently adapted to the authenticity of the creation of educational texts.

5. Results and Discussion

In view of the reliability of AIGC testing in educational scenarios, it is recommended to construct a "human-intelligence" collaborative governance framework, based on the principle of "education-oriented, technology-assisted", and to make efforts through the three dimensions of system, collaboration and technology. The three dimensions of system, synergy and technology should be utilized.

5.1 Institutional standardization: Reconstructing the educational evaluation system

Clarify the auxiliary positioning of AIGC testing, and establish the determination mechanism of "manual domination + technological assistance + process traceability": manual evaluation mainly focuses on the quality of thinking and innovation value, technological testing is used for the initial screening of obvious traces of AI, and the process traceability requires students to keep records of their creations, such as drafts of text writing, annotations, etc. as supporting evidence. The process of traceability requires students to keep records of creation, such as drafts and annotations of text writing, as supporting evidence. Basic education can develop a system of "AI instruction manuals", clearly labeling students' AIGC use scenarios, and teachers' evaluation of their reasonableness; higher education implementation of tiered standards, undergraduate AI use thresholds set a standard, postgraduate students are allowed to use moderately, but need to be labeled, supplemented by the "chain of creation traceability".

5.2 Multi-party collaboration: building an education governance community

Form a collaborative governance pattern of "colleges and universities, primary and secondary schools, enterprises and regulatory authorities". Colleges and universities can incorporate AIGC literacy into their training programs, such as ethical education on the use of AIGC in colleges and universities, AI assisted ability training and AI identification ability training, etc.; primary and secondary schools to develop the "Guidelines for the Use of AI for Homework", clarify the boundaries of the school segments, and adopt the "human-intelligence" collaborative audit to reduce misjudgment; enterprises and educational institutions to build educational text databases and develop proprietary models. Enterprises and educational institutions build educational text databases and develop proprietary models to continuously reduce the rate of misjudgment; regulators formulate AIGC testing standards for the education sector and establish a platform certification mechanism.

5.3 Technology optimization: developing an education-adapted detection system

While optimizing the AIGC detection technology, we develop an education-adapted detection system. The first is to develop a segmentation model to optimize the differentiation between basic education, which focuses on identifying children's linguistic features, and higher education, which focuses on understanding academic logic and other expressive features; the second is to unify the detection standards, specify the calculation method of AI rate for educational texts, sample distribution and thresholds for each academic section, and control the detection differences between different platforms within a reasonable range; the third is to construct a "human-machine" review mechanism, in which the detection system marks suspicious segments in the text. Thirdly, a "human-machine" review mechanism is constructed, in which the detection system marks suspicious fragments in the text, and then the teacher makes the final judgment by combining the students' historical works and the on-site defense, so as to improve the accuracy of the detection through human-machine collaboration.

6. Conclusion

The unstable situation of AIGC testing is essentially a reflection of the tension between technological rationality and the laws of education. In educational evaluation in the digital era, neither innovation should be rejected due to technical defects, nor blindly relying on tools to deviate from the essence of educating people. The "human-intelligence" collaborative governance framework clarifies humanistic attributes through institutional restructuring, gathers consensus through multi-party collaboration, and improves adaptability through technological optimization, pushing the testing from "black or white" judgment to "quality-integrated" judgment. The "combination of quality and quantity" evaluation. This can not only guard the bottom line of academic integrity, but also release the vitality of technological innovation, provide stable and scientific evaluation support for the cultivation of talents in the digital era, and adhere to the original intention of education in the midst of technological change.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Wang, Jianlei,Xie, Lingling (2025). AI traces and digital spiritual rhythms: an alternative perspective on AI-generated content. Journal of Fujian Normal University [2](Philosophy and Social Science Edition), 7(4), 98-107+171.

[2] Zhang, Minghui,Yu Lu,Yang Jun (2023). Disable or embrace: generative artificial intelligence and original gatekeeping in Chinese social science journals. Publishing and Distribution Research, 8(8), 49-55. 110.19393/j.cnki.cn11-1537/g2.2023.08.031

[3] Xue Bei. (May 10, 2025). "Defeat AI with AI? Fostering AI Literacy Needs 'Long-Termism'". Yangzi Evening News. accessed July 11, 2025 from https://epaper.yzwb.net/pc/con/202505/10/content_1446450.html.

[4] Jiang, Weichao,Wang, Zixuan,Li, Jie,et al. (May 10, 2025). "AI rate of famous articles also "exceeds the standard"? Thesis AI rate detection "mistakenly" attracts controversy". Xinhua Daily Telegraph. Accessed May 21, 2025 from https://baijiahao.baidu.com/s?id=1832722546169244197&wfr=spider&for=pc.

[5] (June 3, 2025). "AI Rate for Yang Mi Thesis Soars from 0 to 91%! Confusion over testing standards makes graduates 'break their defenses'". The River News. accessed July 11, 2025 from https://baijiahao.baidu.com/s?id=1833921881909778972&wfr=spider&for=pc.

[6] Galen Chang. (June 8, 2025). "Is it scientific to use 'AI rate' to 'veto' papers?". Science and Technology Daily. accessed July 11, 2025 from https://edu.cnr.cn/sy/sytjB/20250608/t20250608_527200557.shtml.