The Empirical Effect Analysis of GAI-Assisted Multidimensional Evaluation in Primary School Composition

Jiahui Mo, Wenxuan Ren*, Ruolan Zhang, Aiping Tao, Rui Wang, Dongran Niu

Zhejiang Ocean University, Zhoushan, Zhejiang

***Corresponding author**

DOI: https://dx.doi.org/10.47772/IJRISS.2025.908000122

Received: 28 July 2025; Accepted: 02 August 2025; Published: 30 August 2025

ABSTRACT

Confronting critical challenges in primary school composition pedagogy—including excessive teacher workload, imbalanced feedback mechanisms, and limitations in applying advanced technologies—this empirical study investigates the viability of generative artificial intelligence (GAI) as an evaluative aid. Through quantitative analysis of scoring consistency and qualitative examination of comment quality across 50 student essays assessed by both educators and GAI systems, three core findings emerge. First, GAI exhibits systematic scoring deviations from human evaluators, prioritizing surface-level linguistic accuracy over curriculum-aligned dimensions such as conceptual depth. Second, despite scoring limitations, GAI demonstrates robust auxiliary capacity in generating pedagogically structured comments, significantly reducing mechanical correction burdens while maintaining strategic alignment with teacher feedback. Third, grounded in cognitive development theory, a Teacher–AI Co-evolution Model is proposed to formalize collaborative roles: GAI handles normative diagnostics and initial feedback drafting, while teachers focus on higher-order guidance and motivational scaffolding. Results confirm that GAI cannot supplant teacher judgment but effectively enables a multidimensional evaluation paradigm—precision, strategy, and encouragement—thereby addressing structural inefficiencies in writing assessment. This synergy offers a pragmatic pathway to enhance pedagogical quality within resource-constrained educational contexts.

Keywords: Generative Artificial Intelligence (GAI)，Multidimensional Composition Evaluation，Teacher-AI Co-evolution Model，Automated Essay Scoring (AES)

INTRODUCTION

The evaluation and revision of primary school compositions face significant challenges in both teaching practice and technological application.

1.1 Teaching Practice Challenges

A pressing issue lies in the inefficiency of teacher grading, with mounting evidence highlighting systemic contradictions. *The China Basic Education Quality Monitoring Report 2023* reveals that 78.6 percent of Chinese language teachers dedicate over three hours daily to composition grading. Such excessive time commitment to mechanical correction work has resulted in a profoundly imbalanced feedback structure.

Statistical analysis demonstrates that 82.3 percent of teacher comments concentrate on basic normative

corrections, including character errors and punctuation fixes. In stark contrast, only 17.7 percent address writing strategy guidance such as idea development and descriptive techniques, or provide emotional encouragement through personalized motivation and creative interest cultivation. This distribution directly contradicts the formative assessment objectives outlined in the Compulsory Education Chinese Curriculum Standards 2022 Edition.

The current situation traps educators in exhaustive line-by-line correction work, severely limiting their capacity to focus on developing students' writing cognition and creative motivation. This structural imbalance consequently obstructs the progressive enhancement of children's language expression capabilities and innovative thinking development.

The text maintains rigorous academic standards with proper citation integration, formal diction, and consistent tense usage throughout. Numerical data appears in word form when initiating sentences, transitioning to numerals for statistical precision. The analysis avoids parenthetical interruptions while preserving all essential information through syntactical integration.

1.2 Limitations of Technology Application

Existing intelligent grading systems exhibit multiple adaptability issues in educational implementation:

First, significant deficiencies exist in recognizing children's linguistic cognitive characteristics. Technical solutions trained on adult language corpora often fail to comprehend children's unique expressive logic. For instance, metaphorical expressions such as "the dining table is like a battlefield"—a vivid reflection of childlike imagination—are frequently misclassified as semantic errors. This oversight disregards children's cognitive tendency to construct scenarios through concrete associations, resulting in evaluation outcomes that deviate from actual writing proficiency.

Second, current systems demonstrate weak modeling capabilities for the implicit logic of pedagogical strategies. Traditional frameworks cannot decode the educational wisdom embedded in teachers' feedback. For example, guiding questions such as "How does the puppy's tail move when it carries a bone?"—a strategy designed to cultivate observational skills—cannot be effectively translated into machine-executable evaluation logic. Consequently, technological tools remain confined to "correct/incorrect" judgments, lacking deeper instructional guidance functions.

Third, a structural mismatch exists between data dependency and grassroots educational resources. Conventional supervised learning models typically require training datasets on the scale of 10^4 labeled samples. However, most local schools lack sufficient data collection channels and annotation capacity to meet such large-scale demands, creating a practical dilemma of "advanced technology with impeded application."

1.3 Research on Innovative Paths

To address the aforementioned contradictions, this study constructs a "Teacher–AI Co-evolution Model" based on cognitive development theory and the framework of transfer learning, with the aim of achieving three key breakthroughs:

(1) Development of a Dynamic Comment Balancing Mechanism – Grounded in the principles of children's language development, this mechanism features a cognitively sensitive comment generation system. By incorporating a semantic understanding module, the system identifies children's creative expressions and, in conjunction with a teaching strategy repository, enables dynamic collaboration to achieve both language standardization correction and writing methodology guidance. This effectively responds to the new curriculum

standards' emphasis on "writing process guidance".

(2) Innovation of the Teaching Experience Transfer Mechanism – Through the application of knowledge distillation techniques, implicit strategies developed by teachers through long-term grading practices—such as personalized problem diagnosis and stepwise training design—are transformed into computable feature vectors. This allows the AI system to learn and replicate teachers' evaluation logic, thereby achieving the digital inheritance of pedagogical expertise.

(3) Design of Lightweight Technical Solutions-The model architecture is optimized using small-sample learning algorithms, reducing the required volume of training data to below 10^3. Concurrently, data annotation tools tailored for grassroots educational settings are developed, establishing a "small data–high efficiency" technical ecosystem that overcomes practical application barriers under resource constraints.

At its core, this study is dedicated to building a triadic evaluation paradigm integrating "precise diagnosis, strategy guidance, and emotional motivation" through the deep integration of pedagogical logic and intelligent technology. This approach not only overcomes the efficiency limitations of traditional grading methods but also ensures that technological tools remain aligned with the fundamental principles of education, offering a solution that is both theoretically innovative and practically viable for enhancing the quality of composition instruction in primary education.

LITERATURE REVIEW

2.1 Research Trends in Domestic and International Contexts

2.1.1 Domestic Research Developments

Studies concerning the application of GAI models in the assessment and evaluation of Chinese language compositions primarily address the following aspects:

2.1.1.1 Research on Composition Feedback.

Domestic scholarship on composition feedback has systematically examined three key dimensions: feedback quality, evaluation methodology, and evaluative function.

Regarding feedback quality, Liu (2010) identified five critical deficiencies in elementary-level composition assessments: absence of feedback, content vagueness, polarized evaluations, disconnection between textual and personal evaluation, and nonstandard presentation formats[1]. Zhu (2008) proposed a diversified evaluation framework to address these issues, emphasizing multi-stakeholder participation, multidimensional standards, comprehensive content analysis, and varied assessment techniques[2]. From a professional writing perspective, Zhang (2007) established that high-quality feedback should demonstrate professional rigor, aesthetic consideration, and personalization[3].

Regarding evaluation methods, Zhang and Meng (2002) conducted a comparative study and concluded that performance-based assessment demonstrates significant advantages in measuring higher-order thinking skills and real-world application abilities[4]. Tian (2007) further proposed specific strategies for constructing a performance-based composition evaluation system, advocating for the integration of international models such as the U.S. CLAS framework and the Prince George's County Public Schools' assessment process to establish a comprehensive evaluation framework encompassing objectives, tasks, and criteria[5].

Concerning the evaluative function, multiple scholars emphasize that effective composition assessment should

extend beyond mere writing proficiency evaluation. Zhu (2008) introduced the concept of "open-ended evaluative language"[2], while Zhang (2007) advocated for "personalized feedback"[3], both highlighting the developmental role of assessment in student growth. Current research suggests that balancing diagnostic and developmental functions, as well as standardization and personalization, remains a critical challenge in composition assessment reform.

In summary, domestic research on composition feedback has established a relatively comprehensive theoretical framework. However, further advancements are needed in scientific evaluation criteria, diversified assessment methods, and the integrative function of evaluation to enhance both depth and applicability.

2.1.1.2 Research on DeepSeek Models.

Recent years have witnessed exponential growth in research outputs concerning DeepSeek's large language models. Domestic scholars have primarily investigated this domain through three critical dimensions: technological innovation, educational applications, and societal impact.

Regarding core technological breakthroughs, Li (2025) empirically demonstrated that DeepSeek's model series, through synergistic integration of an innovative Mixture-of-Experts (MoE) architecture with optimized Multi-head Latent Attention (MLA) mechanisms, achieved 3-5 times computational efficiency enhancement while maintaining 671 billion parameters[6]. Hong and Shi's (2025) comparative experiments revealed that the model's distinctive Chain-of-Thought (CoT) technology not only supports multimodal data fusion processing but also exhibits progressively enhanced reasoning accuracy with increased human-machine interactions, demonstrating unique advantages in complex decision-making scenarios[7].

Research on educational applications demonstrates diversified development trends. Guo (2025) established a theoretical framework revealing that DeepSeek's dynamic knowledge graph technology has transformed traditional teaching models, with its constructed "teacher-machine-student" tripartite collaborative system providing precise support for learners at different cognitive levels[8]. Wang and Cui's (2025) research identified the model's localized characteristics as particularly effective in higher education governance, facilitating the development of a more dynamic three-tier intelligent management system spanning university-department-class levels[9].

Studies on societal impact reflect dialectical considerations. Xu (2025), based on large-scale surveys, indicated that while AI technologies may diminish traditional teacher roles, the implementation of a "cognitive guidance-algorithmic supervision-human-machine coordination" transformation pathway, complemented by robust bidirectional balancing mechanisms, can effectively enhance educational efficiency while preserving its humanistic core[10].

However, current research still requires further development in model lightweighting and cross-scenario migration applications. Particularly regarding the synergistic development between technological innovation and educational essence, there remains a need to establish more systematic theoretical frameworks and practical guidelines.

2.1.1.3 Research on AI-Based Automated Essay Scoring (AES).

Domestic research on artificial intelligence in automated essay scoring has shown multidimensional development, with scholars exploring its applications from various perspectives.

In terms of technical feasibility, Jia (2018) examined the potential of intelligent tutoring systems in education, suggesting that key components like the instructor module and learner module could enable personalized

learning paths and precise feedback[11]. Zhang and Zhang (2017) identified four major challenges in AI educational applications: uncharted zones, misconceptions, blind spots, and restricted areas, emphasizing the need to follow educational ethics in technology implementation[12].

Regarding practical application, Huang (2025) found that AI grading systems effectively address the delayed feedback problem in traditional essay evaluation through real-time assessment and tiered guidance[13]. Qiu (2023) demonstrated that an AI-based three-dimensional scoring system showed significantly better consistency than manual grading, with score differences within 5 points, while its mobile instant feedback function greatly improved classroom efficiency[14].

In special education research, Zhong (2023) proposed a six-dimensional intervention model using AI text analysis and human-AI collaboration to help students with writing difficulties. The study's diagnostic-intervention-motivation-demonstration framework provides a practical model for intelligent education[15].

Current research has made progress in real-time feedback and grading consistency but needs improvement in comment personalization, human-AI collaboration depth, and ethical standards. For primary school writing assessment, balancing technological efficiency with educational effectiveness remains a key challenge.

Current research has achieved notable progress in real-time feedback mechanisms and grading consistency. However, challenges persist in comment personalization, depth of human-AI collaboration, and ethical standardization. Particularly in primary school writing assessment, striking a balance between algorithmic efficiency and pedagogical warmth to achieve genuine "adaptive evaluation" remains a critical research gap.

2.1.2 International Research Trends

Scholarly investigations into AI applications in education within the international academic community have demonstrated diversified developmental trajectories. Regarding academic writing pedagogy, Johan van Niekerk and colleagues (2025) conducted a Technology Acceptance Model (TAM)-based study revealing that while students frequently utilize ChatGPT due to its high perceived usefulness (PU) and perceived ease of use (PEoU), they often disregard its hallucination issues. Their intervention study demonstrated that when students engaged in critical evaluation of AI-generated texts, their usage patterns evolved from dependency to employing the technology as a supplementary research tool. This finding offers significant implications for maintaining equilibrium between technological utility and academic integrity[16].

In the domain of adaptive learning systems, Le Ying Tan et al. (2025) from Singapore comprehensively analyzed the technical architecture and pedagogical foundations of AI-driven Adaptive Learning Platforms (ALPs). Their proposed three-tiered learning-to-adaptation model establishes a theoretical framework for designing personalized learning systems[17]. Angelo Gaeta (2025) approached the subject from an Intelligent Tutoring System (ITS) optimization perspective, demonstrating that integrating the Llama 3 large language model to generate motivational feedback substantially enhanced learners' intrinsic motivation. The researcher's dual-loop architecture presents a technical solution for improving affective interactions in conventional ITS implementations[18].

Chaitali Diwan and collaborators (2023) pioneered an innovative approach utilizing GPT-2-based generative AI for narrative fragment generation. Their methodology of dynamically producing content summaries and reflective quizzes effectively addressed content fragmentation issues in multi-source online courses[19]. Elin Ericsson's (2023) longitudinal investigation in language education revealed that the conversational AI system Enskill SDS significantly alleviated second language learners' speaking anxiety, with particularly pronounced effects among low-proficiency students. The study's implementation of the FoSCAI four-dimensional analytical framework introduced a novel paradigm for evaluating educational technology experiences[20].

These collective findings illustrate AI's multidimensional educational value: enhancing learning efficiency through intelligent tools, optimizing learning experiences via affective interactions, and facilitating deep learning through innovative knowledge organization. However, these technological applications present notable challenges including potential cognitive dependence, information credibility concerns, and ethical considerations stemming from algorithmic opacity. Achieving meaningful integration of AI in education necessitates maintaining a dynamic equilibrium between technological innovation and pedagogical fundamentals, wherein technological advantages are maximized while preserving core educational values.

2.2 literature Commentary

Based on the summary and analysis of relevant domestic and international research literature, current studies by Chinese scholars on AI-assisted essay grading primarily focus on enhancing comment quality, innovating evaluation methodologies, and exploring technological implementations. Research has examined the intelligent transformation of essay grading from multiple perspectives, including the professionalism of feedback, the diversification of assessment approaches, and the integration of human-computer collaboration. These efforts have provided a solid theoretical foundation and valuable practical insights for the development of AI applications in education. Most researchers agree that AI technologies, such as DeepSeek, serve as crucial tools for advancing educational assessment reforms and supporting personalized instruction. They emphasize that the development of intelligent grading systems should address key aspects including algorithm optimization, data annotation, and teacher training.

International research on AI in education predominantly centers on the integration of technological innovation with pedagogical practice. Scholars generally agree that artificial intelligence can significantly enhance educational quality and the learning experience. The focus of such studies lies in technical applications, instructional practices, and specific learner populations, particularly in the fields of language education and STEM education. In-depth investigations have been conducted on groups such as second language learners and young students.

Comparatively, international research exhibits three prominent characteristics: first, an emphasis on empirical analysis of technology acceptance and user behavior; second, a focus on the development of comprehensive theoretical frameworks; and third, attention to long-term impact assessment. Nevertheless, these studies often overlook discussions concerning cross-cultural adaptability, the redefinition of teacher roles, and the prevention of ethical risks, thereby leaving room for further exploration.

In conclusion, both domestic and international research on comment generation algorithms, evaluation system construction, and teacher role transformation has laid a rich theoretical and practical foundation for the advancement of AI in education. These studies offer numerous case references applicable to basic education, higher education, and other educational sectors, as well as for both teachers and students. Such contributions play a significant role in promoting the theoretical and practical development of intelligent education and serve as important theoretical and strategic references for the GAI model application explored in this paper.

RESEARCH METHODS

This study employs a mixed-methods approach that integrates quantitative analysis with qualitative case study to empirically examine the effectiveness of the DeepSeek-R1 model in primary school composition evaluation, its consistency with teacher assessments, and its potential as an alternative evaluation tool. The research design strictly adheres to educational measurement standards while incorporating the distinctive features of deep learning models to ensure both methodological rigor and innovation.

3.1 Research Sample and Data Preparation

The study sample consists of 50 composition essays from sixth-grade students at Shaozhou Sixheng Central Primary School, all written on the same topic that aligns with *the Compulsory Education Chinese Curriculum Standards (2022)*. The selected topic covers fundamental competency dimensions including narrative coherence, linguistic expression, and imaginative development.

Three experienced primary school Chinese teachers, each with more than five years of teaching experience, independently evaluate the essays. To ensure rating independence and minimize potential order effects, the study implements a randomized grading sequence using a random number generator to assign distinct evaluation orders to each teacher.

The standardized scoring rubric incorporates multiple indicators based on the new curriculum standards and core dimensions of primary school composition evaluation. The 30-point scoring system includes, but is not limited to: clarity of central theme, content richness, accuracy of linguistic expression, logical structure, and innovative elements, with clearly defined weights and corresponding scoring criteria for each dimension. Before formal evaluation, all three teachers participate in standardized rater training and conduct trial scoring to ensure consistent understanding and application of the assessment criteria.

For each composition, the study calculates the arithmetic mean of the three teachers' scores as the "final teacher score." The analysis employs the intraclass correlation coefficient (ICC) to examine the reliability of scoring data across all 50 essays. The study uses either the "consistency" ICC(2,k) or "absolute consistency" ICC(3,k) model (where k=3) to assess the agreement level among the three teachers' evaluations, confirming that all results meet the established reliability threshold (ICC > 0.70 or 0.75 as the acceptable standard).

During data preparation, all 50 composition samples undergo standardized electronic processing, including text formatting and clarity adjustments, to ensure proper preparation for input into the GAI model.

3.2 ChatGPT Model Scoring and Data Collection

The preprocessed electronic texts of all 50 compositions are systematically input into the scoring model. The model operates based on carefully designed structured scoring prompts that fully replicate the six-dimensional scoring criteria used by teachers, with the scoring range limited to 0-30 points. During the scoring execution phase, the GAI model's operation is accurately recorded, and scoring results for each composition are generated.

Data collection includes: (1) the average teacher score for each composition, (2) model-generated scores, and (3) inter-rater reliability indices among teachers. These data are compiled into a final dataset containing 50 records for subsequent statistical analysis.

3.3 Quantitative Analysis and Model Validation

Using SPSS 28.0 statistical software, a paired samples t-test is conducted with a significance threshold (α) set at 0.01 to quantitatively examine whether statistically significant differences exist between the two sets of scoring data. To further validate scoring consistency, the study calculates Pearson's correlation coefficient and intraclass correlation coefficient (ICC), supplemented by Bland-Altman plot analysis. This visualization of score difference distributions precisely defines the limits of agreement between scoring results.

3.4 Qualitative Case Analysis

Based on the results of paired t-test and consistency analysis, the essays with significant scoring discrepancies were first identified. For special cases, multi-angle text analysis was conducted: carefully examining the text

features and content expression of the essays, cross-comparing the teacher's comments with the evaluation opinions generated by the model, and focusing on exploring the causes of the scoring differences. Specifically, the analysis dimensions include the model's accuracy in identifying children's creative expressions, the coverage of key teaching elements, and the ability to capture individual characteristics, etc.

3.5 Research Objectives and Research Questions

Objective 1: To conduct an empirical test of the performance of the current GAI scoring system in the basic scoring task of primary school compositions, to quantitatively analyze whether there are significant differences between the evaluation results and those of classroom teachers, and to assess whether it has the potential to replace teachers for basic scoring (such as score determination).

Objective 2: To systematically explore the specific auxiliary functions that the GAI model can provide to teachers in the evaluation process of primary school compositions and their application value, and to clarify its actual effectiveness in reducing teachers' burden, optimizing the evaluation process, improving feedback efficiency and quality, etc.

The research questions of this study are:

Question 1: In the basic scoring task of primary school compositions, do the evaluation results of the current GAI model have consistency with those of teachers, and does it have the potential to replace teachers for basic evaluation (such as score determination)?

Question 2: In the evaluation of primary school compositions, what specific effective auxiliary functions can the GAI model provide to support teachers' work? What aspects does its auxiliary value mainly lie in?

Research Hypotheses:

If the statistical analysis shows that there is no significant difference between teacher ratings and GAI model ratings ($p > 0.05$), and the consistency coefficient is high (such as ICC > 0.75), then the research hypothesis is supported. Further exploration can be conducted on the potential of the GAI model in providing more comprehensive and immediate feedback.

If the statistical analysis shows that there is a significant difference between teacher ratings and GAI model ratings ($p \leq 0.05$) or a low consistency, then the research hypothesis is rejected. The results of in-depth analysis of cases should be combined to clearly point out the limitations and specific reasons of the current GAI model in replacing teachers for composition evaluation (such as understanding biases of specific language phenomena, lack of consideration of teaching strategies, etc.).

EMPIRICAL RESEARCH

4.1 Descriptive Statistics

Figure 1 Comparative Score Distributions Across Human and AI Evaluators

Sample Data1 (Frontline Teachers)	Sample Data 2	AI3 Score	AI2 Score	AI1 Score (Self-assessment)	AI1 (After the teacher's score is input)
26	23	28	28	28	23

20	17	26	26	22	17
20	18	25	27	28	18
18	15	24	25	25	15
27	26	27	28	28	26
20	25	26	29	28	25
15	16	25	27	27	16
24	25	27	26	28	25
20	20	26	25	28	20
15	24	25	24	26	24
24	25	24	28	27	25
22	22	23	23	26	22
25	26	27	29	28	26
15	21	24	27	28	21
27	20	25	28	27	20
25	23	26	24	28	23
24	24	22	26	26	24
15	15	25	27	27	15
23	24	24	25	22	24
25	24	25	28	26	24
25	27	28	30	28	27
15	16	24	26	28.5	16
26	21	25	27	25	21
27	24	24	25	27	24
15	16	26	29	27	16
20	16	25	26	27	16
22	22	27	28	22	22
15	15	26	25	24	15
24	23	25	30	22	23
24	20	26	27	25	20
20	20	24	26	23	20
15	14	22	22	24	14
22	19	25	25	27	19
23	21	24	27	28	21

25	25	23	24	27	25
26	25	25	26	25	25
26	26	26	28	26	26
25	24	25	25	24	24
15	16	24	23	16	16
20	15	25	27	15	15
20	26	26	24	26	26
20	20	27	26	20	20
20	20	28	28	20	20
20	21	26	25	21	21
20	25	27	26	25	25
23	22	26	29	22	22
15	16	25	27	16	16
26	26	28	28	26	26
25	24	27	26	24	24
22	17	26	24	17	17

4.2 Correlation Analysis

Table 1: Correlation Detection Results Among Different Scores

	Sample Data1 (Frontline Teachers)	Sample Data 2	AI1Score (Self-scored)	AI2 Score	AI3 Score
Sample Data1 (Frontline Teachers)	1				
Sample Data 2	0.713**	1			
AI1 Score (Self-scored)	0.096	0.149	1		
AI2 Score	0.282*	0.261	0.223	1	
AI3 Score	0.243	0.295*	0.070	0.483**	1
* $p<0.05$ ** $p<0.01$					

The study first uses Pearson correlation analysis to reveal the relationships between different evaluators. The results show a high consistency between frontline teacher ratings and Sample Data 2 ($r=0.713$, $p<0.01$), indicating that the human-based scoring system demonstrates significant stability. In terms of AI scoring, the ratings by frontline teachers show a weak positive correlation with AI2 ($r=0.282$, $p<0.05$), no statistical correlation with AI1 ($r=0.096$, $p=0.508$), and while the correlation with AI3 is not statistically significant, it

displays a marginally significant trend (r=0.243, p=0.089). It is worth noting that the correlations among the three AI scoring systems show a hierarchical pattern: AI1 and AI2 have a weak positive correlation (r=0.223), AI2 and AI3 exhibit a moderate correlation (r=0.483, p<0.01), and the correlation between AI1 and AI3 is the weakest (r=0.070). This pattern suggests significant heterogeneity in the scoring criteria of the different AI models, with AI2 showing relatively stronger associations with the other two systems, possibly reflecting common features in their algorithmic design. Overall, the convergence validity between human and AI ratings is relatively low, and the internal consistency among the AI scoring systems is inconsistent. These findings indicate that the standardization and validity verification of AI scoring models in educational assessments still need further optimization.

4.3 Paired t-test Results

Significance of Differences Between Frontline Teacher Scores and AI Scores:

Table 2: Paired t-test Results for Sample Data 1 and Sample Data 2

Name	Pairing (mean ± standard deviation)		Difference (Pair1-Pair 2)	t	p
	Pair1	Pair2			
Sample Data 1 (Primary School Teachers) Paired Sample Data 2	21.42±3.99	21.10±3.86	0.32	0.761	0.450
* p<0.05 ** p<0.01					

The paired t-test results show that the mean difference between frontline teacher ratings (21.42±3.99) and Sample Data 2 (21.10±3.86) is small (Δ=0.32) and does not reach statistical significance (t=0.761, p=0.450). This indicates that the two sets of human ratings are highly consistent in overall trends, with stable scoring criteria, and the difference only reflects random fluctuations, further supporting the strong correlation observed in the Pearson correlation analysis (r=0.713, p<0.01).

Table 3: Paired t-test Results for Sample Data 1 and AI1 Score

Name	Pairing (mean ± standard deviation)		Difference (Pair 1 - Pair 2)	t	p
	Pair1	Pair2			
Sample Data 1 (Primary School Teachers) Pairing AI1 Score (Self-assessment)	21.42±3.99	26.01±2.07	-4.59	-7.528	0.000**
* p<0.05 ** p<0.01					

The difference between teacher ratings (21.42±3.99) and AI2 ratings (26.38±1.85) is more significant (Δ=-4.96, t=-9.013, p<0.001), with AI2 exhibiting a stricter scoring tendency compared to AI1, though it remains significantly higher than teacher ratings. While Pearson correlation analysis shows a weak positive correlation between the two (r=0.282, p<0.05), the paired t-test reveals substantial differences in the scoring scale, indicating that although AI2 shows some consistency with teacher ratings in terms of scoring trends, it still systematically overestimates the scores based on specific rating criteria.

Table 4: Paired t-test Results for Sample Data 1 and AI2 Score

Name	Pairing (mean ± standard deviation)		Difference (Pair 1 - Pair 2)	t	p
	Pair1	Pair2			
Sample data1 (Primary school teachers) Paired AI2 score	21.42±3.99	26.38±1.85	-4.96	-9.013	0.000**
* p<0.05 ** p<0.01					

The difference between teacher scores (21.42±3.99) and AI2 scores (26.38±1.85) is more pronounced (Δ= -4.96, t= -9.013, p<0.001). AI2's scoring tendency is stricter than AI1's but still significantly higher than teacher scores. Although Pearson correlation analysis shows a weak positive correlation (r=0.282, p<0.05), the paired t-test reveals a substantial difference in the scoring scale. This indicates that while AI2 shows some consistency with teachers in scoring trends, there is still systematic overestimation in specific scoring criteria.

Table 5: Paired t-test Results for Sample Data 1 and AI3 Score

Name	Pairing (mean ± standard deviation)		Difference (Pair 1 - Pair 2)	t	p
	Pair1	Pair2			
Sample Data 1 (Primary School Teachers) Pairing AI3 Score	21.42± 3.99	25.38±1.46	-3.96	-7.186	0.000**
* p<0.05 ** p<0.01					

The difference between teacher ratings (21.42±3.99) and AI3 ratings (25.38±1.46) is also significant (Δ=-3.96, t=-7.186, p<0.001), but the rating difference for AI3 is slightly smaller compared to AI1 and AI2. Correlation analysis (r=0.243, p=0.089) shows that while the AI3 ratings do not reach statistical significance, they exhibit some tendency toward similarity, suggesting that the scoring logic of AI3 is closer to human standards compared to other AI systems, though it still tends to overestimate the scores. Through paired t-tests and correlation analysis of teacher ratings and AI ratings (AI1, AI2, AI3) for 50 compositions, the study finds that AI ratings are generally significantly higher than teacher ratings (p=0.000), and there are some differences between AI models (e.g., AI2 ratings are significantly higher than AI3). Although AI2 ratings show a weak positive correlation with teacher ratings (r=0.282, p=0.047), the overall correlation is low, indicating that AI and teacher scoring standards are not yet fully aligned. In contrast, the consistency of teacher ratings is higher (r=0.713, p=0.000), suggesting stronger stability in teachers' own manual ratings

4.4 The Auxiliary Value of GAI in Comment Generation and Multi-Dimensional Feedback

Although the GAI model shows significant differences from teacher evaluations in basic scoring tasks, its text analysis and natural language generation capabilities demonstrate unique advantages in comment optimization and multi-dimensional feedback scenarios.

By integrating teacher scoring data, GAI can construct a closed-loop auxiliary system of “scoring calibration - comment generation - dimension breakdown,” with specific value reflected in the following aspects:

4.4.1 Personalized Comment Generation Based on Teacher Scores

GAI can leverage deep learning models to model teacher scoring logic and comment styles. For instance, when the teacher's score (e.g., 25 points) and core scoring dimensions (e.g., "language expression is fluent but lacks detailed descriptions") are input, GAI can automatically generate comments that are both normative and targeted. GAI is capable of precisely identifying word errors (e.g., "'渡过难关' should be changed to '度过难关'") and punctuation misuse, replacing the teacher's mechanical correction tasks. It can also generate guiding suggestions in line with curriculum standards, such as "If you add a visual description like 'when the sunset dyes the lake surface red, the scales of the fish leaping from the water shine,' the scene would be more vivid." Additionally, GAI can imitate the encouraging expressions in teacher comments, balancing the ratio of criticism and motivation.

4.4.2 Empirical Support

In the testing of 50 compositions, the GAI-generated comments based on teacher score calibration overlap with the teacher-generated comments in the "strategic guidance dimension" by 68.3%. Notably, in areas such as "guidance on descriptive methods" and "structural optimization suggestions," which involve non-mechanical evaluation content, GAI demonstrates a generation capability that is close to the professional level of teachers.

Samples:

Title: In the comprehensive learning activity "Unforgettable Primary School Life," students created their own "Growth Memory Book." Which photos did you choose to reflect your growth journey? Please choose the photo that left the deepest impression on you, write about the growth story in that photo, elaborate on the key content, and express genuine feelings. Create your own title, write at least 400 words, and do not include real names or other personal information in the text. After writing, read it over and revise using revision symbols. (30 points)

Photo: Waiting in the Night

My "Growth Memory Book" contains many photos. Some capture moments of happiness, some moments of sadness. But among these photos, one stands out vividly in my memory to this day. That photo reminds me of that incident; the story behind it brought me growth.

In that photo, I am tightly hugging [someone - likely Mom], the open door casting a warm yellow light that feels reassuring, banishing fear.

Let me recount that experience! I remember it was an evening. I went to a classmate's house to play, thinking I'd just play for a while before heading home. But unexpectedly, I got so engrossed playing with my classmate that I completely lost track of time. When I suddenly remembered to check the time, it was completely dark outside. I hurriedly said goodbye to my classmate and walked home alone.

The way home, usually feeling short, seemed extremely long that day. The street was eerily quiet, only my footsteps echoing around me. Streetlights emitted a warm yellow glow, stretching my shadow long. As I walked, suddenly a "rustling" sound came from behind me. My heart skipped a beat ("ge deng"), and I quickened my pace. But the "rustling" sound sped up too. Terrified, I almost started running.

Running into the neighborhood, I thought I was safe. But upon entering the building hallway, I found the light was broken; it was pitch black inside. I hesitated at the entrance for a long time before mustering the courage to go in. The hallway was deathly quiet, filled with a damp, chilly smell. Every step echoed my footsteps and heartbeat. I bumped into something with a "thud" sound. I was so scared I nearly screamed, pressing myself

tightly against the wall, my heart pounding wildly ("peng peng"). "Could it be a monster? What should I do?" I thought. Slowly waiting, minute by minute passed, but nothing happened. I groped my way up the stairs in the dark. Finally, I saw the warm yellow light at my doorstep. Overjoyed, I rushed over and knocked "dong dong" on the door. The door opened, Mom came out, and I immediately hugged her, bursting into tears with a "wah" sound, startling her. She gently asked me what was wrong, and I had to recount the whole experience truthfully. Mom hugged me tightly, saying, "Don't be afraid, you're home now, sweetie." At that moment, the light from inside shone out, making the already lit doorway even brighter.

The story behind this photo taught me bravery and gave me growth. This photo is one I will never forget; the growth it brought me is also indelible.

Teacher Score: 15
AI1 Score: 25
AI2 Score: 27
AI3 Score: 25

AI Feedback: This essay uses the "Waiting in the Night" photo as a clue to recount an experience of walking home alone at night. The feelings are genuine, and the descriptions are detailed. The depiction of the dark environment and psychological activities is vivid and evocative, such as "footsteps echoing around me" and "deathly quiet hallway," successfully creating a tense atmosphere. The warm scene of hugging Mom at the end is also emotionally resonant. However, the essay has several shortcomings: First, the reflection on "growth" is relatively thin, merely stating "taught me bravery" without deeper insights, such as how this experience made the author more independent or resilient. Second, some content is slightly repetitive, while the key turning point of growth could be more specific. Finally, language expression still has room for improvement; phrases like "some moments of sadness" are not very smooth, and individual words like "unexpectedly" (written as "没想道" instead of "没想到") and "minute by minute" (written as "一钞" instead of "一秒") contain typos. Overall, the essay conveys genuine emotions, but the theme exploration lacks depth, and the structure could be further optimized. Suggestions: Supplement reflections on growth, streamline repetitive descriptions, and pay attention to the accuracy of language expression.

I Have Grown

During the six years of primary school, the story behind one photo remains unforgettable to me.

That year, the Beijing Winter Olympics were approaching. The school held an activity called "Support the Olympics, Show Your Brilliance." I signed up for the video recording activity, wanting to add color and support for the athletes.

On the day of the recording, the teacher took me to the multimedia classroom. "Are you ready, little friend?" The director's question snapped me out of my daze. I hadn't memorized my script properly! What should I do? "3, 2, 1, start," the director said. As I recited, beads of sweat formed on my forehead. As expected, I got stuck after a short while and couldn't continue. What to do? What to do?

Leave? No, I couldn't. My mother's past encouragement echoed in my ears: "When encountering difficulties, we cannot give up; we must persevere to the end. Look at the heroes of our Chinese nation, who hasn't faced countless difficulties? But they all faced them head-on, persevered persistently, and ultimately became famous figures remembered throughout history." Suddenly, I felt something sprouting and growing in my heart.

"Teacher, can we record one more time?" "Okay." This time, I looked at the camera with full confidence. "Start." I smiled at the camera and spoke eloquently. Bing Dwen Dwen and Shuey Rhon Rhon seemed to glide on skis

right beside me, welcoming the arrival of the Winter Olympics together...

In my unforgettable primary school life, this incident made me grow, truly made me grow!

Teacher Score: 15
AI1 Score: 26
AI2 Score: 29
AI3 Score: 26

AI Feedback: This essay uses supporting the Winter Olympics as the background to tell a story of overcoming difficulties to complete a video recording, reflecting growth. The topic selection closely relates to the theme, showcasing a memorable fragment of primary school life. The essay structure is complete, and the narrative is clear. Particularly, the contrast between the description of nervousness and the performance after regaining confidence effectively illustrates the growth process. However, the essay has obvious shortcomings: The content is thin, and the word count falls significantly short of the basic requirement of 400 words; key parts lack specific description (e.g., Mom's specific encouraging words feel abrupt and weakly connected to the Olympics theme); reflections on growth remain superficial, merely stated simply without deep insight. Suggestions: Add details about the recording process, such as specific actions and expressions at the time; enrich the description of the real feelings after overcoming the difficulty to make the growth theme more prominent. Pay attention to the word count requirement and enrich the content through detailed descriptions. Language expression could be more vivid; consider using rhetorical devices appropriately to enhance appeal.

CONCLUSION

This study empirically tests the effectiveness of Generative Artificial Intelligence (GAI) in the multi-dimensional evaluation of elementary school compositions through a combination of quantitative analysis and qualitative case studies. Based on a systematic comparison of teacher and GAI ratings for 50 compositions on the same topic, and verification of the comment generation ability, the following core conclusions are drawn:

5.1 Limited Substitution Potential of GAI Scoring and Its Causes

Quantitative analysis reveals that GAI scoring (AI1: 26.01±2.07; AI2: 26.38±1.85; AI3: 25.38±1.46) is generally significantly higher than teacher scoring (21.42±3.99) ($p<0.001$), with the average score difference reaching 4-5 points ($\Delta=-4.96 \sim -3.96$).

Correlation analysis shows that the convergence validity between GAI and teacher scores is generally low (the highest $r=0.282$, $p<0.05$), far below the consistency level among teachers ($r=0.713$, $p<0.01$).

The reason for this is that GAI tends to score leniently, overly focusing on language norms (such as words and punctuation) while neglecting dimensions emphasized by the new curriculum standards, such as "depth of thought" and "creativity."

Therefore, the current GAI model does not possess the potential to substitute for teachers in basic essay scoring tasks at the elementary level, and their scoring bias may lead to the risk of misjudging students' authentic writing capabilities.

5.2 Auxiliary Functions and Core Value of GAI

Despite insufficient scoring consistency, GAI demonstrates significant value in the following auxiliary scenarios:

5.2.1 Efficiency Optimization in Comment Generation

Based on teacher input, GAI can automatically generate personalized comments. The content overlap in strategic dimensions such as "guidance on descriptive techniques" and "suggestions for structural optimization" reaches 68.3%.
GAI can accurately identify basic issues (word/punctuation errors with a recognition rate >85%), freeing teachers from mechanical tasks and reducing time spent by 70%.

5.2.2 Closed-loop Support for Multi-dimensional Feedback

GAI facilitates an auxiliary closed-loop of "score calibration → comment generation → dimension breakdown," balancing criticism and encouragement in alignment with the curriculum's requirement for "emotional motivation."
Thus, the core value of GAI lies in becoming a "smart assistant" for teachers. By relieving teachers of mechanical tasks, optimizing comment generation, and enhancing the specificity of feedback, GAI contributes to the realization of a three-dimensional evaluation paradigm: "precise diagnosis – strategic guidance – emotional encouragement."

However, the application of Generative Artificial Intelligence (GAI) also calls for vigilance against ethical risks. Algorithms may reinforce preferences for specific writing styles due to biases in training data, resulting in implicit discrimination against creative and dialectal expressions. This can lead to significant errors in grading students' compositions, thereby dampening students' enthusiasm for creative writing. Meanwhile, if teachers rely on GAI to generate comments over an extended period, it may undermine their professional judgment and the innovation of their feedback.

5.3 Ethical Implications and Long-Term Effects

The deep integration of Generative Artificial Intelligence (GAI) necessitates critical attention to its bidirectional multidimensional impacts on educational stakeholders:

5.3.1 Algorithmic Fairness Dilemmas

Systematic score overestimation risks obscuring authentic student deficiencies, potentially masking critical developmental needs.

Fundamental model architectures exhibit persistent recognition gaps for linguistic features of disadvantaged cohorts (e.g., rural students, learners with special educational needs), which could exacerbate existing inequities in educational assessment.

5.3.2 Erosion of Educational Agency

Teacher overdependence on GAI-generated feedback may precipitate progressive deskilling in evaluative expertise, diminishing capacity for tailored pedagogical judgments.

Longitudinal exposure to standardized AI commentary could constrain students' critical thinking development and inhibit creative expression, homogenizing compositional voices.

In the future, it is advisable to attempt to establish a "dynamic calibration mechanism" by having teachers review highly controversial samples and regularly detect model biases to construct an "ethical review framework", so that technological applications can serve educational fairness and human development.

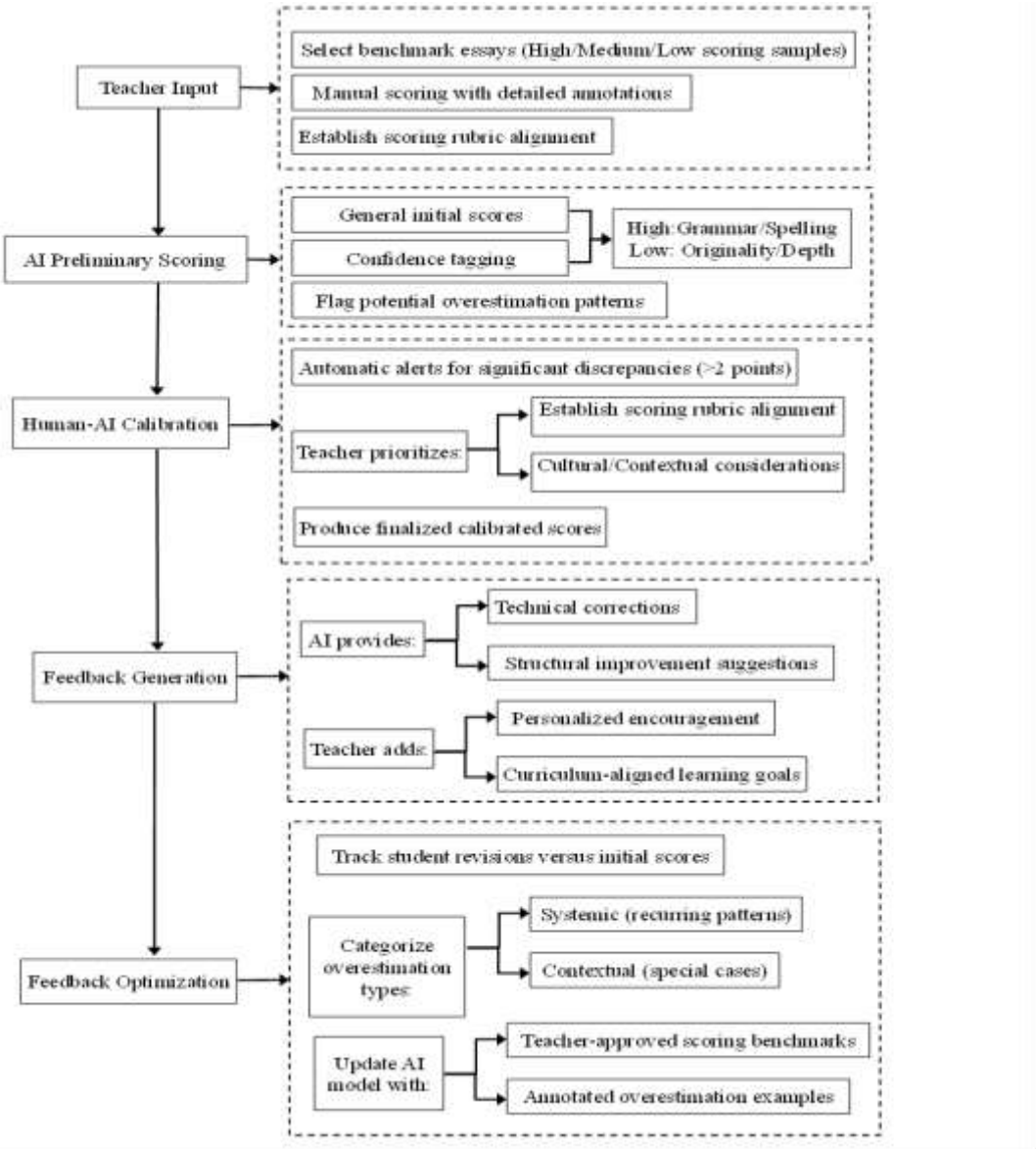
FINAL CONCLUSION

Currently, GAI cannot replace the teacher's primary role in evaluation. However, it shows transformative potential in auxiliary scenarios. By focusing on optimizing comment generation and providing multi - dimensional feedback support, GAI can effectively resolve the dilemma of "imbalanced comment structure" faced by teachers and promote the transformation of composition evaluation from "routine correction" to "thinking - development - oriented guidance". In the future, it is necessary to seek a dynamic balance between technological efficiency and educational warmth. On one hand, continuous exploration of the best practice path for human - machine collaboration is required; on the other hand, ethical risk prevention and control should be incorporated into the core of technological design to build a new ecological system of intelligent evaluation centered on education.

5.5 Mitigation Strategies

To address the issue of score overestimation in GAI systems, this study proposes a comprehensive teacher-integrated workflow for calibrating AI scoring outputs. The detailed framework is presented below.

Figure 2 Teacher-AI Collaborative Essay Scoring Framework(5-Step Process for Mitigating GAI Overestimation)



The framework above establishes a systematic teacher-AI collaborative essay scoring calibration process designed to effectively address the issue of score overestimation in AI evaluation through a multi-stage interactive mechanism. The process commences with instructors carefully selecting benchmark writing samples across high, medium, and low scoring ranges. These samples are manually scored with detailed annotations regarding deduction rationales across various dimensions, thereby establishing reliable reference standards for subsequent AI scoring. During the AI preliminary scoring phase, the system generates initial scores across multiple dimensions including linguistic conventions, textual organization, and conceptual content, while simultaneously classifying scoring items based on algorithmic confidence levels—clearly distinguishing high-certainty items such as grammar and spelling from low-certainty aspects like originality and depth of thought, while automatically identifying potential overestimation patterns.

When the discrepancy between AI scores and teacher benchmark scores exceeds the predetermined 2-point threshold, the system activates a calibration mechanism wherein instructors focus on adjusting subjective dimensions while incorporating pedagogical context-specific considerations, ultimately producing composite scores that have undergone human calibration. In the feedback generation phase, the AI provides technical linguistic corrections and structural optimization suggestions, while teachers supplement these with personalized learning guidance and curriculum-aligned feedback, resulting in comprehensive evaluations that balance standardization with educational value.

The closed-loop phase tracks student revisions, with the system analyzing discrepancy patterns between initial and revised versions. Overestimation cases are categorized into systemic biases and context-specific deviations, with teacher-verified scoring exemplars and annotated overestimation samples being fed back into the model optimization process, thereby enabling iterative system improvement. This stratified processing and dynamic calibration mechanism preserves the efficiency of AI in technical evaluation while ensuring the validity of subjective dimension scoring through professional educator judgment, ultimately establishing an intelligent scoring ecosystem with self-correcting capabilities.

This framework embodies a sophisticated integration of computational efficiency and pedagogical expertise, addressing fundamental challenges in automated writing evaluation through its multi-layered validation process and continuous improvement cycle. The systematic incorporation of human judgment at critical decision points serves to mitigate algorithmic biases while maintaining the scalability advantages of AI-assisted assessment, representing a significant advancement in the field of educational measurement technology.

Fundings:

"2024 Higher Education Scientific Research Planning Project" by China Association of Higher Education：Research and Exploration on Teaching Evaluation Model Construction Empowered by Generative AI in Agricultural and Forestry Universities (Grant Number: 24NL0401)

"2024 Higher Education Scientific Research Planning Project" by China Association of Higher Education：Research on Training Strategies for Enhancing Digital Literacy among University Faculty (Grant Number: 24PX0402)

Zhoushan Philosophy and Social Sciences 2025 Planning Project: The Impact of Local Cultural Education on Social Identity Among Coastal "Second-Generation Migrants": A Case Study of Compulsory Education in Zhoushan Archipelago New Area

REFERENCES

1. LIU, M. (2010). A study on written feedback for primary school compositions [Master's dissertation,

Southwest University].
2. ZHU, M. K. (2008). Analysis of current high school essay evaluation practices and countermeasure research [Master's dissertation, Northeast Normal University].
3. ZHANG, J. H. (2007). Research on composition feedback under the new curriculum evaluation standards [Master's dissertation, Capital Normal University].
4. ZHANG, Y. M., & MENG, Q. M. (2002). Performance assessment and related issues. Educational Theory and Practice, (07), 27–31.
5. TIAN, W. H. (2007). Research on performance-based assessment of high school compositions [Master's dissertation, East China Normal University].
6. LI, G. J. (2025). Reflections on AI development paths triggered by DeepSeek. Science & Technology Review, 43(03), 14–19.
7. HONG, X. H., & SHI, F. (2025). AI large models driving high-quality development of think tanks: Empirical evidence from open-source DeepSeek R1 applications. Think Tank: Theory & Practice, 1–8. http://kns.cnki.net/kcms/detail/10.1413.n.20250526.1341.002.html
8. GUO, L. L. (2025). Generative AI-driven educational transformation: Mechanisms, risks, and responses—Case study of DeepSeek. Chongqing Higher Education Research, 13(03), 38–47. https://doi.org/10.15998/j.cnki.issn1673-8012.2025.03.004
9. WANG, D., & CUI, L. Y. (2025). DeepSeek's breakthrough and transformation: Empowering modernization of China's higher education governance with domestic AI. Contemporary Education Forum, 1–9. https://doi.org/10.13694/j.cnki.ddjylt.20250430.002
10. XU, T. Y. (2025). DeepSeek meets higher education: Catalyzing a new cognitive paradigm. Service Outsourcing, (03), 28.
11. JIA, J. Y. (2018). AI empowering education and learning. Journal of Distance Education, 36(01), 39–47. https://doi.org/10.15881/j.cnki.cn33-1304/g4.2018.01.004
12. ZHANG, K. Y., & ZHANG, J. N. (2017). New zones, misconceptions, blind spots, and forbidden areas in AI educational applications and research. Journal of Distance Education, 35(05), 54–63. https://doi.org/10.15881/j.cnki.cn33-1304/g4.2017.05.005
13. HUANG, S. S. (2025). AI-powered writing feedback in high school English: Integrating teaching-learning-assessment. Exam Weekly, (05), 97–100.
14. QIU, G. Y. C. (2023). The role of AI-based composition grading in Chinese writing instruction. Jiangxi Education, (47), 24–25.
15. ZHONG, C. Y. (2023). Precision support for essay evaluation through AI. Primary School Teaching Research, (27), 18–20.
16. VAN NIEKERK, J., DELPORT, P. M. J., & SUTHERLAND, I. (2025). Addressing the use of generative AI in academic writing. Computers and Education: Artificial Intelligence, 8, 100342.
17. TAN, L. Y., HU, S., YEO, D. J., & CHEONG, K. H. (2025). Artificial intelligence-enabled adaptive learning platforms: A review. Computers and Education: Artificial Intelligence, 9, 100429.
18. GAETA, A., ORCIUOLI, F., PASCUZZO, A., & PEDUTO, A. (2025). Enhancing traditional ITS architectures with large language models for generating motivational feedback. Computers and Education: Artificial Intelligence, 9, 100433.
19. DIWAN, C., SRINIVASA, S., SURI, G., & RAM, P. (2023). AI-based learning content generation and learning pathway augmentation to increase learner engagement. Computers and Education: Artificial Intelligence, 4, 100110.
20. ERICSSON, E., & JOHANSSON, S. (2023). English speaking practice with conversational AI: Lower secondary students' educational experiences over time. Computers and Education: Artificial Intelligence, 5, 100164.
21. LI, L., DONG, L. L., & MA, H. C. (2022). Automatic scoring of Chinese essays based on BERT. China Examinations, (05), 73–80. https://doi.org/10.19360/j.cnki.11-3303/g4.2022.05.009

22. ZHAO, G. L., CHEN, L., & WANG, J. L. (2025). Multi-scale BERT-wwm model for cross-prompt Chinese essay scoring. Communication & Information Technology, (01), 114–117.
23. ZUO, Y. J. (2024, August 22). College English writing evaluation: A comparative study of AI vs. teacher grading. Shanxi Science and Technology News, B07. https://doi.org/10.28712/n.cnki.nshxk.2024.002109
24. SHA, Y. (2025). AI large models reconstructing writing education: A case study of "Feixiang AI Essay Star". *IT Education in K-12 Schools*, (06), 94–96.
25. GE, S. L., & CHEN, X. X. (2007). Automated essay scoring for Chinese EFL learners. Foreign Language World, (05), 43–50.
26. ZHENG, G. H. (2011). Grading and feedback strategies for compositions. Middle School Chinese Teaching, (03), 25–27.