International Education Forum, 2026, Volume 4, Issue 10
https://bbwpublisher.com/journal/IEF
Online ISSN: 2981-8605
Print ISSN: 3083-4902

Exploring Human-AI Collaborative Process Assessment in English Writing in College Education

Jiehui Li*

School of General Education, Guangzhou Civil Aviation College, Guangzhou 510403, Guangdong, China

Author to whom correspondence should be addressed.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: To address the declining validity of English writing assessment in higher vocational education caused by generative AI and the difficulty of reflecting students' true language and critical thinking abilities through final drafts alone, this study localizes and adapts the AISAR process evidence collection framework to construct a process evidence collection mechanism comprising AI prompt chains, content selection and rejection rationales, revision marks, and reflection reports. It establishes a human-AI collaborative assessment workflow of "AI standardized preliminary evaluation + teacher higher-order judgment." Through classroom action research, it is found that human-AI collaborative assessment based on process evidence can guide students' writing revisions from surface-level language correction toward deeper thinking optimization, help teachers capture the cognitive trajectories of student-AI interaction, and achieve an expansion from single final-draft evaluation to multidimensional assessment of "textual output + cognitive process." The study also identifies practical dilemmas encountered in implementation and proposes targeted optimization pathways, with the aim of providing practical references for intelligent assessment reform in higher vocational English education.

Keywords: College English; Human-AI collaboration; Process assessment; Process evidence; Writing evaluation

Online publication: November 11, 2026

1. Introduction and research background

The rapid iteration of large language models has driven the deep integration of artificial intelligence into assessment in higher vocational foreign language education, with intelligent assessment platforms such as iTEST being implemented in multiple higher vocational institutions, promoting the normalization of computerized English testing. However, most current intelligent assessments only address surface-level problems such as slow feedback and low grading efficiency in traditional assessment, without breaking through the inherent limitations of summative evaluation; the paradigm transformation of foreign

language assessment thus remains at a superficial stage. The widespread adoption of generative AI has brought a significant validity crisis to English writing assessment. Students can use AI to rapidly produce grammatically accurate and structurally complete English texts, resulting in a severe misalignment between final drafts and students' actual language proficiency and critical thinking abilities. The traditional assessment model centered on final drafts can no longer adapt to the realities of teaching. Artificial intelligence is not only a challenge to assessment fairness but also an opportunity for innovation in evaluation models; its value should not be limited to automated scoring but should lie in reshaping evaluation concepts and promoting a shift from "result-oriented" to "process-empowered" assessment.

In this study, human-AI collaborative process assessment refers to a formative assessment model in which, in AI-assisted writing scenarios, artificial intelligence undertakes standardized objective evaluation tasks such as grammar, format, and discourse structure; teachers focus on qualitative judgment of students' AI interaction behaviors, textual decision-making logic, and levels of thinking; and complete process evidence is simultaneously collected to integrate both textual output and cognitive development—distinct from summative writing assessment that relies solely on final drafts.

Different researchers have gained substantial theoretical and practical experience in evaluating foreign languages with the help of AI. Liu Jianda developed a three-level analysis system for language ability in the age of AI, stressing that the evaluation should include cognitive functions such as discrimination, judgment, and selection used by learners in AI applications [1]. Liang Jie established the AISAR process evidence collection framework, suggesting the collection of instruction prompts, content selection records, revision trails, and self-reflections to incorporate the cognitive process of AI interactions into the evaluation, which provides a theoretical basis for evidence collection in human-AI cooperative writing [2]. Zhuang Meifang indicated that the joint evaluation of humans and machines is more efficient in determining students' thinking levels [3]. Zhu Hong carried out formative assessment based on iTEST in college English, confirming that intelligent systems can enhance assessment efficiency [4]. Li et al. found that large models can provide certain score reliability as classroom assessment tools, but higher-order critical thinking feedback shows template restrictions, indicating the necessity for teachers' subjective judgment [5]. Moreover, studies such as He Xiaoyang focused on the combination of artificial intelligence in foreign language teaching and assessment [6]. Bao Xiaomei and Deng Hui investigated the improvement of college English teaching effectiveness through the use of artificial intelligence [7–8]. Sun Litao et al. and Huang Lan separately examined the application of artificial intelligence technology in English listening and speaking tests and in essay scoring [9–10]. Ye Chunju proposed an English evaluation model for vocational education based on the integration of industrial education and AI strengths along with Problem-Oriented Assessment (POA) [11]. Yang Wei and Yu Peng analyzed the current situation and corresponding measures for the evaluation of AI-supported college English classroom teaching [12]. Zhang Lina explored a three-dimensional evaluation mechanism for AI-aided high school English writing assessment [13]. Liao Xing and Hou Jiangting put forward practical strategies and optimization methods for the reform of AI-aided assessment in secondary vocational English [14]. Kaddouri et al. pointed out that General Artificial Intelligence (GAI) offers opportunities to improve educational assessment, but due consideration should be given to the ethical, pedagogical, and technical aspects during its implementation [15]. Leng Jing et al. demonstrated that ChatGPT has high efficiency in capturing and evaluating features such as directness and explicitness in critical thinking [16].

In conclusion, previous studies have pointed out the importance of integrating AI interaction cognition

in the evaluation process, but there are still some research deficiencies: firstly, most empirical investigations do not provide specific practices for the learning environment of higher vocational education; secondly, most pilot projects are too short, mainly focusing on single writing tasks, and lack continuous assessment experiences over a long period of time; thirdly, the issues related to the control of process evidence quality, authenticity verification, and teacher's workload have not been discussed sufficiently. Therefore, this research carries out a 10-week long-term teaching pilot based on the iTEST platform, exploring: How can a long-term human-AI cooperative evaluation procedure for English writing be established according to the AISAR framework and adapted to the higher vocational educational conditions? What are the results and practical difficulties of this assessment model? How can the mechanism be improved to solve the problems of process evidence quality, authenticity, and teachers' assessment workload?

2. Research design

2.1. Research subjects

Two vocational English classes in the institute, including 86 non-English major students, were chosen as the experimental group without a control group. The participants showed significant differences in their English proficiency, similar to the usual learning situations of non-English majors in higher vocational education.

2.2. Research instruments

An intelligent assessment cloud system is able to provide several writing assignments, gather the related documents, and evaluate the essays automatically using a quantitative method.

Collect evidence requirements for the writing tasks: Four kinds of evidence were needed in each round of work, which were complete AI prompt sequences, reasons for selecting and rejecting AI content, text revisions, and approximately 200-word reflection reports. With the advancement of teaching, the completeness of evidence and the depth of reflection were gradually increased.

Teacher qualitative judgment dimensions: Based on the core of AISAR, three evaluation aspects were separately established: the first is the awareness of judgment, that is, to check if the students can distinguish AI content; the second is the basis of judgment, to evaluate the consistency between the selection logic and the writing context; the third is the depth of judgment, to assess the cognitive evaluation level of the reflection reports.

A self-designed short questionnaire containing 5 multiple-choice questions for learning conditions.

2.3. Writing task design

An eleven-week plan includes a series of argumentative writing assignments concerning campus life, current social issues, and professional competence. The present research aims to study a selected mid-term assignment, that is, whether college students should attend physical education classes, about 180 words.

2.4. Implementation procedure

In the modification stage, the teacher introduced the idea of cooperative writing between humans and AI and the rules for submitting four kinds of evidence, gave some simple composition assignments, and showed examples of qualified and unqualified evidence. In the improvement stage, several levels of successive composition tasks were carried out; the students had to submit their final drafts and the evidence of the

process, and an evaluation method consisting of "AI quantitative preliminary assessment plus teacher's superior judgment" was applied. After each round, class discussions were held to solve the common problems in the process of producing evidence. The data for the research came from the composition task in Week 8. In the conclusion and investigation stage, the results from various tasks were summarized, questionnaires were distributed to obtain the students' subjective opinions, and the practical results and difficulties were examined.

3. Research results

3.1. Observation of student writing behavior

Under the constraints of process evidence, some students no longer limited themselves to vocabulary and grammar correction but began to focus on argument reasonableness, logical coherence, and contextual appropriateness; they were able to discriminate AI output and make selections, and arbitrary subjective revisions decreased. However, some students with weak foundations still directly relied on AI-generated paragraphs.

3.2. Observation of teacher assessment work

With complete process evidence, teachers could reconstruct the full cognitive chain from conception and AI interaction to revision and finalization, distinguishing students' true abilities masked by final drafts. For the 86 samples in Week 8 alone, teacher qualitative judgment took a cumulative 8 hours; after multiple rounds of tasks over 10 weeks, teachers' overall workload was significantly higher than in traditional writing instruction. When faced with cases such as "good process but mediocre final draft" or "excellent final draft but perfunctory evidence," teachers found it difficult to calibrate evaluation standards.

3.3. Student questionnaire feedback

Out of the 78 valid questionnaires, 70% (n=54) of the students considered that maintaining records helped them to examine the AI output more carefully and reduced their dependence on technology; 60% (n=47) approved of the advantage of reflective review in improving their composition; 50% (n=39) regarded the educational value of the tasks; 30% (n=23) thought the procedure was boring. The majority of the students reported that their refusal to the tasks became less as they got used to them step by step.

3.4. Practical dilemmas

Several practical problems found in the teaching experience are as follows. Firstly, although some students have received much training, they still show incomplete data, lack of logical reasoning, and superficial reflection; without standardized scoring criteria, it is hard to grade the process data fairly, and if it is not considered in the grading, it will not attract sufficient attention from the students. Secondly, the time for teachers to judge is quite long, and some teachers still follow the idea of evaluating results only, which is not professional enough to interpret the evidence of the interaction process with intelligent aid. Lastly, the iTEST system is good at evaluating the grammatical structure and format, but the feedback on the depth of argumentation and logical innovation is based on templates and does not provide individualized inspirational guidance.

4. Discussion

This action research shows that the implementation of human-AI collaborative process assessment can be done via existing teaching platforms. The practical outcomes agree with the theoretical opinions of Liu Jianda and Liang Jie, indicating that integrating AI interaction cognition into assessment can broaden the evaluation range from a single final draft to "textual output plus cognitive process" and is applicable in the classroom. From the students' viewpoint, after continuous training, the required preservation of process evidence makes the students handle the AI-generated content more carefully, decreasing the behavior of direct copying; the revision of writing moves from superficial corrections to optimizing thoughts, and the selection of writing shifts from subjective impression to rational judgment, which is consistent with the findings of Zhuang Meifang et al. In addition, the research discovers that there is a great difference among the learning conditions of higher vocational students, and it is hard to eliminate the quality polarization of the evidence even after prolonged training. From the teachers' perspective, the process evidence can help overcome the mistaken evaluations due to "evaluation based on final draft only" and provide a basis for individualized instruction. Nevertheless, the new model brings new competence requirements for the teachers who should not only evaluate the textual language quality but also interpret the logical thinking behind the student-AI interaction — a skill which cannot be covered by the traditional essay grading experience. The study also verifies the conclusion of Li et al.: AI is appropriate for the standardized evaluation of surface-level, but higher-order evaluation such as critical thinking cannot be assigned to machines, and the teacher's qualitative judgment is irreplaceable.

In order to solve these problems, the research proposes the following optimization techniques: First, the study should arrange the tasks gradually and progressively. Initially, the study should decrease the difficulty of the evidence items to lower the threshold and then improve the completeness and depth of the evidence step by step during the teaching process to reach an appropriate balance between the teaching results and the academic burden. Secondly, the study should establish an evaluation standard and use it reasonably for formative assessment. The evidence should take up 10%–15% of the total score of the course formative assessment; a grading scale should be set according to the level of judgment, the basis of decision, and the depth of thinking so that the qualitative evidence can be quantified. Thirdly, special teaching research should be carried out in the teaching and research departments. Discussions should concentrate on the interpretation of the process evidence, the discrimination of AI feedback and the students' cognitive judgments in order to update the teachers' evaluation concepts and enhance the ability of collaborative assessment between humans and machines. Fourthly, a classroom oral spot-check verification system should be set up. Students should be randomly chosen to explain their understanding of the writing and the reasons for selecting AI on the spot, which can supplement the lack of written materials and reduce fabrication afterward. Finally, the study should promote the continuous improvement of the iTEST platform functions. Structured sections for collecting process evidence should be developed to automatically store the materials; the AI higher-level feedback module should be optimized to offer personalized guidance for logical reasoning and viewpoint innovation.

5. Conclusion and outlook

Generative AI brings both difficulties and advantages to the traditional final-draft evaluation method, and it also provides an opportunity for assessment reform. Based on the AISAR framework, a 10-week teaching

experiment was carried out, and several progressive writing tasks were designed. The process evidence, including prompt chains, content selection and rejection reasons, revision marks, and reflection reports, was collected, and a human-aided assessment system named "AI preliminary evaluation plus teacher's higher judgment" was established. From the practice, it is found that the continuous provision of process evidence can improve students' ability to use AI critically, which helps to change the writing from correction of language to optimization of thinking; assists the teachers in reconstructing the entire cognitive process of the interaction between students and AI; and achieves multi-dimensional evaluation by combining the process and the results. This model can be implemented using existing systems, but some practical issues exist, such as the polarization of evidence quality, heavy burden on teachers, and the lack of sufficient higher-level feedback from AI.

Future research may conduct large-sample, long-duration controlled experiments, improve higher-vocational-specific human-AI collaborative assessment rubrics, distinguish performance differences among students of different English proficiency levels, and explore technical pathways for platforms to automatically collect AI interaction traces, thereby continuously refining the assessment system. In the intelligent era, higher vocational English assessment reform should be grounded in classroom practice with gradual innovation, reasonably utilize technological tools, return to the essence of foreign language education, and contribute to improving the quality of talent cultivation in higher vocational foreign language education.

Funding

Foreign Language Teaching Advisory Board for Vocational Colleges, Ministry of Education of China: "Digital-Intelligent Empowerment and Industry-Education Collaboration: Building Foreign Language Education Pathways in Civil Aviation Vocational Colleges" (WYJZW2026-55); Special Innovation Project, Department of Education of Guangdong Province: "Empowerment Mechanism and Pathways of Workplace English for Civil Aviation Maintenance Personnel in Guangdong under Industrial Upgrading and Competency Dual Drive" (26Z5451)

Disclosure statement

The author declares no conflict of interest.

References

[1] Liu JD, 2026, "Changes" and "Unchanges" in Language Assessment in the Era of Artificial Intelligence. Foreign Language Testing and Teaching, 2026(2): 2–10.

[2] Liang J, 2026, Generative Artificial Intelligence Empowering Formative Assessment of College English Writing: Construction of the AISAR Assessment Framework. English Square, 2026(8): 66–68.

[3] Zhuang MF, 2025, A Study on Human-Machine Collaborative Evaluation of High School Students' English Essays Based on SOLO Theory, thesis, Nanchang University.

[4] Zhu H, 2026, An Exploration of Formative Assessment in College English Based on the iTEST Intelligent Platform. Modern English, 2026(1): 67–69.

[5] Li R, Liu Y, Gao N, et al., 2025, On the Effectiveness of Formative Assessment Method Assisted by Artificial Intelligence in College Education. In 2025 5th International Conference on Artificial Intelligence and Education (ICAIE), 670–674.

[6] He XY, Qu LY, 2025, Summary of the 8th Symposium on English Teaching and Assessment. China Examinations, 2025(12): 106–110.

[7] Bao XM, 2026, Pathways for Enhancing the Effectiveness of College English Teaching Empowered by Artificial Intelligence. English Square, 2026(21): 111–114.

[8] Deng H, 2025, Research on the Reform and Development Path of English-Speaking Assessment Based on Artificial Intelligence. Economist, 2025(9): 231–232.

[9] Sun LT, Li Q, Lu HR, 2025, Application of Artificial Intelligence Technology in English Listening and Speaking Tests. Shanghai Municipal Educational Examinations Authority. Proceedings of the 9th Educational Examination and Evaluation Symposium (SEEE2025). Shanghai Translation Publishing House Co., Ltd., 359–368.

[10] Huang L, 2024, Research on the Application of Artificial Intelligence Assessment Technology in College English Essay Grading. Hubei Admissions and Examinations, 2024(5): 61–63.

[11] Ye CJ, Chen R, 2025, Research on Higher Vocational English Assessment Model under the Guidance of Industry-Education Integration, AI Empowerment, and POA. Industry and Technology Forum, 24(17): 252–255.

[12] Yang W, Yu P, 2024, Current Situation and Countermeasures of AI-Empowered College English Classroom Teaching Evaluation. Journal of Hubei Open Vocational College, 37(8): 139–141.

[13] Zhang LN, 2026, AI-Empowered High School English Writing Assessment: Practical Exploration of a Three-Dimensional Evaluation Mechanism. Henan Education (Teacher Education), 2026(10): 74–75.

[14] Liao X, Hou JT, 2026, Practical Pathways and Optimization Strategies for AI-Empowered Assessment System Reform in Secondary Vocational English. Overseas English, 2026(11): 153–155.

[15] Kaddouri M, Mhamdi K, Jabri A, et al., 2025, Generative Artificial Intelligence and Assessment of Learning: Challenges and Perspectives, Effective Instructional Design Informed by AI. IGI Global, Pennsylvania, 277–325.

[16] Leng J, Lu HH, Dai L, 2024, Generative Artificial Intelligence Empowering Critical Thinking Assessment: An Experimental Application Based on ChatGPT. Modern Distance Education Research, 36(6): 102–111.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.