Effects and Governance of a Hybrid AI–Human Rubric Assessment Model in an Advanced English Course

Wei Sun (Corresponding author)
School of Foreign Studies, Anhui Xinhua University
Hefei 230000, Anhui, China
Email: sunwei@axhu.edu.cn

Lei Sun
Fengyang No.1 High School
Fengyang 231000, Anhui, China
Email: swopedeborsu25@gmail.com

The research is financed by Anhui Provincial Quality Engineering Project for Higher Education Institutions, No. 2024dzxkc 135; Anhui Xinhua University Doctoral Research Start-up Fund Project, No. bs2025kyqd043

Abstract

This study investigates the application effects and governance strategies of a rubric assessment method combining artificial intelligence (AI) scoring and human review in the university course "Advanced English." Through an experimental control design, differences between traditional human scoring and AI-assisted scoring were compared, evaluating the role of this hybrid scoring model in scoring consistency, efficiency improvement, and fairness assurance. The results show that AI-assisted rubric assessment significantly improves scoring efficiency while maintaining overall consistency with human scoring results. Scoring consistency saw some improvement, and there were no statistically significant differences across subgroups (e.g., gender), indicating the model did not introduce bias. Student questionnaire feedback indicated that most students held a positive attitude towards this "AI initial scoring + human review" model, believing it enhanced feedback timeliness and scoring fairness. However, a minority of high-achieving students expressed concerns that AI might struggle to fully comprehend complex expressions. Finally, the study proposes governance suggestions to ensure the effective implementation of this model, including refining the rubric, strengthening AI scoring calibration, implementing fairness audits, and providing teacher training. The research indicates that a human–computer combined rubric assessment holds positive potential in Advanced English teaching but requires accompanying governance measures to ensure its reliability and fairness.

Keywords: Advanced English; Rubric Assessment; Artificial Intelligence Scoring; Human Review; Fairness Governance

DOI: 10.7176/IKM/16-1-01

Publication date: January 31st 2026

1. *Introduction*

The rapid development of educational technology has led universities to explore artificial intelligence (AI) for supporting language assessment. In Advanced English courses, where students complete complex academic reading and writing tasks, rubric-based assessment has long been used to enhance transparency, consistency, and fairness. However, fully manual scoring remains labor-intensive and susceptible to inter-rater variation, creating a need for more efficient evaluation methods.

Recent progress in automated essay scoring (AES) systems, such as Pigai in China and e-rater developed by ETS, has made AI-assisted assessment increasingly feasible. These systems can generate instantaneous feedback on vocabulary, grammar, and organization, providing efficiency advantages. Yet prior research also shows that

AI alone struggles with high-level semantic judgment and creative expression, prompting concerns about validity, fairness, and explainability. Consequently, hybrid models that combine AI's speed with teachers' professional review have emerged as a promising alternative, potentially improving scoring consistency while preserving human oversight.

Against this backdrop, the present study examines the effectiveness and governance needs of an AI + human review rubric assessment model in the Advanced English course. It investigates whether this model enhances scoring reliability and efficiency, how students perceive its use, and what governance mechanisms are necessary to ensure fairness and trustworthiness in technology-enabled assessment.

Research Questions:
1. Does an AI-assisted (AI + human review) rubric assessment model improve scoring reliability (consistency) and efficiency compared to traditional human-only scoring?
2. Does the AI-assisted rubric scoring approach maintain fairness in assessment outcomes across student subgroups (e.g., gender) without introducing bias?
3. How do students perceive the AI + human review scoring model in terms of perceived fairness, feedback timeliness, and overall satisfaction?
4. What governance strategies are required to ensure effective and fair implementation of the AI-assisted rubric assessment model?

2. *Literature Review*

2.1 Rubric Assessment in Language Education

Rubrics provide explicit performance criteria that improve scoring consistency and reduce subjective bias. By clarifying expectations for both teachers and students, rubrics enhance transparency and support formative feedback. This approach aligns with the concept of Assessment for Learning, which emphasizes using assessment to support student learning through feedback and improvement. Prior studies indicate that well-designed rubrics contribute to higher student satisfaction in English writing assessment. Nevertheless, creating and applying high-quality rubrics demands extensive teacher time, and manual scoring remains a significant workload.

2.2 AI Applications in Educational Assessment

AI has been increasingly adopted for automated scoring tasks. Traditional AES engines analyze linguistic features and often correlate highly with human scoring. More recent large language models (LLMs), such as GPT-based systems, demonstrate improved capability in handling open-ended responses and achieving consistency with human raters. However, concerns remain: AI may misinterpret sophisticated argumentation, lack contextual sensitivity, and provide opaque reasoning. Therefore, researchers have highlighted the value of AI-assisted rather than AI-replaced scoring, with AI offering initial scores that teachers refine through expert review.

2.3 Human–AI Hybrid Scoring Models

Hybrid scoring—often described as human-in-the-loop assessment—leverages the complementary strengths of AI and human raters. Evidence shows that combining machine-generated preliminary scores with teacher review can enhance reliability beyond either method alone. Studies on university writing assessment report that students appreciate AI's efficiency while still valuing teacher comments for deeper insights. High-achieving students, in particular, rely on teachers to recognize creativity and nuance that AI may overlook. These findings reinforce the importance of preserving human judgment in high-stakes evaluation. This human–AI collaboration reflects a form of hybrid intelligence, wherein machine capabilities augment human expertise to achieve outcomes superior to either alone (VIPKID Research Group, 2025; Chen & Luo, 2025).

2.4 Fairness, Bias, and Governance in AI Scoring

Integrating AI into assessment raises ethical and fairness concerns, as algorithmic systems may encode biases from training data. Disparities in scoring accuracy across demographic groups have been documented, prompting professional bodies such as AERA to call for systematic validity evidence, fairness monitoring, and transparent model governance. Frameworks like IBM's AI TRiSM emphasize bias detection, ongoing monitoring, and responsible deployment. Human oversight is widely regarded as an essential safeguard to mitigate unintended inequities. In other words, explicit attention to algorithmic fairness in educational AI systems is critical to avoid perpetuating biases (Boateng & Boateng, 2025; Barnes & Hutson, 2024).

In sum, existing literature supports the potential of AI-assisted rubric scoring to enhance efficiency and consistency, provided that it is accompanied by rigorous governance, human review, and fairness safeguards. The present study builds on these insights by empirically evaluating a hybrid model within the Advanced English course and identifying governance strategies for ensuring reliable and equitable assessment.

3. *Methodology*

3.1 Participants and Grouping

This study involved students from two classes of the Advanced English course at a university, with a total sample of 60 students. The experimental class (using AI + human review scoring) had 28 students, and the control class (using traditional human scoring) had 32 students. All participants were third-year undergraduate English majors who had passed the College English Test Band 6, indicating a high level of English proficiency. The two groups showed no statistically significant differences in baseline characteristics such as gender ratio and previous course grades, ensuring comparability. The course was taught by the same instructor with consistent content and assignment requirements across both classes.

Table 1 shows that the experimental and control classes were well-matched in baseline characteristics (gender composition, age, and prior performance), providing a valid basis for comparing the two scoring models.

Table 1. Comparison of Sample Baseline Characteristics (Experimental Class vs. Control Class)

Indicator	**Experimental Class (AI+Human Review)**	**Control Class (Traditional Human)**	t/χ^2	*p*
Number (N)	28	32	-	-
Male Proportion (%)	32.1%	31.3%	0.005	0.943
Average Age (Years)	20.5±0.7	20.4±0.6	0.68	0.500
Previous Semester Major Course Average	84.3±5.2	83.7±5.5	0.45	0.654

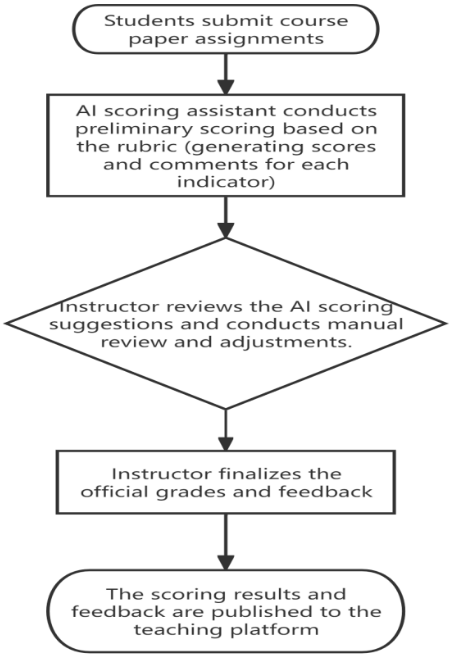
3.2 Rubric Design and Scoring Procedure

Based on the teaching objectives and evaluation requirements of the Advanced English course, we designed an analytic writing rubric for evaluating students' course papers. This rubric contains four primary scoring dimensions—Content, Structure, Language Use, and Expressive Effect—each with five performance levels (1 point = very poor, 5 points = excellent). Table 2 lists the main indicators and the description for each level (the complete rubric is provided in Appendix A). For example, the Content dimension focuses on the richness and depth of the paper's arguments, with a 5-point performance described as "Arguments are original and insightful, fully elaborated and supported," and a 1-point performance as "Lacks clear arguments or justification." Language Use covers vocabulary diversity, grammatical accuracy, and stylistic appropriateness, while Expressive Effect examines the overall readability and persuasiveness of the essay. The rubric was reviewed by two experienced Advanced English instructors before the study and trialed on sample essays from previous cohorts to ensure the descriptions were clear and operable.

For the scoring process, the experimental class adopted an "AI preliminary scoring + human review" hybrid model. The process is illustrated in Figure 1. First, students submitted their course paper assignments. Next, an

AI scoring assistant (developed by the research team) performed an initial scoring of the essays based on the rubric, generating provisional scores and feedback for each indicator. Then, the course instructor reviewed the AI's suggested scores and feedback, manually checking and adjusting any questionable points, and determined the final grades and feedback. Finally, the scoring results and feedback were released to students via the learning management platform. Throughout this process, the teacher remained the final decision-maker, with this human-in-the-loop approach ensuring the reasonableness and accuracy of the AI's scoring. The control class, in contrast, used traditional scoring: the teacher independently scored the essays using the same rubric without any AI involvement. To evaluate the quality of the AI scoring, the instructor in the experimental class recorded differences between the AI's initial scores and the final adjusted scores for each essay, along with notes on the reasons for any score modifications. Additionally, system logs recorded the time the teacher spent grading each essay under both conditions to allow a comparison of scoring efficiency.

Figure 1. Human-in-the-Loop Scoring Flowchart (AI provides initial assessment based on Rubric, followed by teacher review, adjustment, and feedback of results to students)


Students submit course paper assignments
AI scoring assistant conducts preliminary scoring based on the rubric (generating scores and comments for each indicator)
Instructor reviews the AI scoring suggestions and conducts manual review and adjustments.
Instructor finalizes the official grades and feedback
The scoring results and feedback are published to the teaching platform

3.3 Data Collection and Analysis

The data collected in this study included three components:

- **Scoring data:** the scores for student course papers (including AI's initial scores and the final human-adjusted scores for each essay in the experimental class, as well as the human scores in the control class).
- **Scoring time data:** the teacher's grading time for each essay (in minutes) under each scoring model.
- **Student questionnaire responses:** a survey on students' perceptions of the scoring model. The questionnaire used a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) and included 10 items covering dimensions such as perceived fairness, feedback timeliness, and overall satisfaction (see Appendix B). The survey was administered to both classes at the end of the semester, and 58 valid responses were collected.

For data analysis, we conducted three main comparisons. First, we compared AI initial scores with human final scores in the experimental class (using paired-sample t-tests) to examine scoring consistency and any systematic bias in the scoring scale. We also compared the final course grades of the experimental class with those of the control class (independent-samples t-test) to see if the scoring model led to any significant differences in student performance. Second, we calculated indicators of scoring consistency and efficiency: specifically, the Pearson correlation coefficient (r) between AI initial scores and human final scores in the experimental class, and the average time the teacher spent grading an essay under each condition. Finally, to evaluate fairness, we conducted a subgroup analysis in the experimental class by student gender, comparing whether discrepancies between AI initial scores and human final scores differed between male and female students. This tested whether the AI scoring exhibited any systematic bias. We also summarized the questionnaire data by computing the mean score and agreement rate for each survey item to understand students' subjective perceptions of the scoring model.

4. Results

4.1 Scoring Outcomes (Overall Effects)

We first compared the differences between AI preliminary scoring and the instructor's final scoring in the experimental class. In terms of average scores, the overall levels of AI scoring and human scoring were very close: the average course paper score based on the AI's initial scoring was 84.5, while the teacher's final scoring average was 85.1, a difference of only 0.6 points, which was not statistically significant (t(27) = –1.21, p = 0.23, Cohen's d ≈ 0.22). This indicates that the overall scores given by AI did not exhibit any systematic overestimation or underestimation bias. Furthermore, AI and human scoring were largely consistent across the various rubric indicators (see Table 3). For example, in the Content dimension, the AI's average score was 17.0/20, and the teacher's average was 17.6/20; in the Structure dimension, the average scores for both AI and teacher were about 17.5/20; in the Language Use dimension, the AI's score was slightly higher (26.4/30) while the teacher's was slightly lower (25.8/30), but these differences were not significant. Overall, the AI scoring model used in this study was able to reasonably simulate the teacher's scoring results based on the rubric, maintaining a high degree of consistency with human scoring across different evaluation dimensions. The teacher adjusted scores for some essays during the review process, but in general the AI's initial scoring provided a reliable foundation. Importantly, introducing AI did not compromise the existing fairness or validity of the evaluation. Notably, the few cases where AI and teacher scores differed by more than 1 point were mostly in language-related aspects. AI tended to penalize spelling and grammar errors more strictly, while teachers gave some leeway considering content. In these cases, the teacher's oversight balanced the evaluation, ensuring the final score aligned with course objectives and human judgment.

Table 2 presents a comparison of AI preliminary scores and human final scores across the rubric indicators for the experimental class.

Table 2. Comparison of AI Preliminary Scores and Human Final Scores across Rubric Indicators (Experimental Class, N=28)

Evaluation Indicator	**AI Preliminary Score (Mean ± SD)**	**Human Final Score (Mean ± SD)**	*t* (27)	*p*
Content (Max 20)	17.0 ± 1.8	17.6 ± 1.7	–1.56	0.130
Structure (Max 20)	17.4 ± 1.5	17.6 ± 1.4	–0.74	0.467
Language Use (Max 30)	26.4 ± 2.5	25.8 ± 2.7	1.20	0.239
Expressive Effect (Max 30)	23.7 ± 2.8	24.1 ± 2.6	–0.89	0.381
Total Score (Max 100)	**84.5 ± 6.0**	**85.1 ± 5.5**	**–1.21**	**0.234**

We also compared the final course paper scores of the experimental class (with AI-assisted scoring) to those of the control class (with purely human scoring). The score distributions of the two groups were very similar. The average score in the experimental class was 85.1, compared to 84.7 in the control class, with no significant difference (p > 0.5, Cohen's d ≈ 0.1). This suggests that introducing AI assistance did not lead to any appreciable change in student score levels, alleviating concerns that AI assistance could unfairly alter student grades. In short, under the conditions of this study, the scores assigned by the AI + human review model were comparable to those produced by traditional human-only scoring.

4.2 Scoring Reliability and Efficiency

We used several indicators to evaluate the consistency (reliability) and efficiency of the two scoring models, with results summarized in Table 3. First, regarding scoring consistency: in the experimental class, the Pearson correlation coefficient between the AI's initial total score and the human's final total score for each essay was r = 0.92, indicating a very high positive correlation. This correlation was slightly higher than the typical consistency between scores from two independent human teachers in the control class (estimated at around r = 0.85 based on past data or literature). In other words, with the teacher's final oversight, the rank order of essay scores given by the AI aligned very closely with the teacher's final judgments. Put simply, the AI effectively provided a "second rater" opinion that helped improve scoring reliability. This finding is consistent with conclusions from previous research on hybrid scoring models: the human–computer combined scoring approach helps reduce random errors that can occur with a single rater, making scoring more stable and reliable. It should be emphasized that the extremely high consistency observed in the experimental class is partly attributable to teachers referencing the AI's suggestions; if the AI and teacher had scored completely independently, the consistency would likely be slightly lower (indeed, comparing the AI's unadjusted raw scores with the control class teacher's scores yields a correlation of roughly r = 0.88). From a practical teaching perspective, however, it is precisely the human review component that ensures the accuracy and trustworthiness of the final scores—this process allows AI and teacher to jointly complete the scoring task, forming an effective double-rating mechanism in high-volume grading to enhance reliability.

In terms of scoring efficiency, the inclusion of AI significantly reduced the time required for the teacher to grade each essay. Analysis of the grading system logs showed that the average scoring time per essay was approximately 10.2 minutes in the control class (human-only scoring), compared to only about 6.5 minutes in the experimental class (AI-assisted scoring)—a time savings of around 36%. The efficiency gain stemmed primarily from the initial feedback automatically generated by the AI: teachers did not have to start evaluating from scratch but could instead make targeted adjustments based on the AI's scoring draft. According to teacher feedback, with AI-generated preliminary comments in hand, they could focus on checking for any AI misjudgments and correcting specific issues, requiring only minor adjustments to individual criterion scores or adding supplemental comments. This workflow reduced repetitive labor, allowing the teacher to invest the saved time into more in-depth feedback or individual guidance. Crucially, the improvement in scoring efficiency did not come at the cost of quality or fairness. On the contrary—because the teacher had more energy to examine difficult cases and anomalies, the scoring process potentially became more thorough for those cases. In summary, in this study the AI + human model achieved both speed and reliability: scoring was faster, and the results remained accurate and consistent. This finding aligns with predictions in the literature that automated scoring technology can significantly improve grading efficiency and consistency by reducing human error and accelerating feedback speed, thereby improving the assessment experience and lowering administrative workload.

Table 3 compares key indicators of scoring consistency and efficiency between the traditional human-only scoring and the AI + human review scoring in our study.

Table 3. Comparison of Scoring Consistency and Efficiency Indicators between Different Scoring Models

Indicator	**Traditional Human Scoring**	**AI + Human Review Scoring**
Human-Human Scoring Consistency (r)	0.85	-
AI-Human Scoring Consistency (r)	-	**0.92**
Average Scoring Time per Essay (minutes)	10.2	**6.5**
Scoring Time Reduction (%)	-	**≈ 36%**

Note: Traditional human scoring consistency is estimated based on the correlation between two independent *teacher scores in the control class. AI–Human scoring consistency refers to the correlation between the AI's initial score and the teacher's final score for each essay in the experimental class.*

4.3 Fairness Audit

In technology-assisted evaluation, fairness is always a paramount concern. We conducted a preliminary audit of the AI-assisted scoring model's fairness by comparing outcomes across different student subgroups. Our primary focus was on gender. We separately calculated the discrepancies between the AI's initial score and the final human score for male and female students in the experimental class to see if the AI's scoring exhibited any systematic bias against one gender. The results are shown in Table 4. For male students (N = 12), the AI's initial average score was 84.2, and the final average (after human review) was 85.4—an upward adjustment of about +1.2 points. For female students (N = 16), the AI's initial average was 84.8, and the final average was 85.7—an upward adjustment of about +0.9 points. The difference between these adjustments was very small. Statistical testing confirmed no significant difference in the magnitude of score adjustment between male and female students (p = 0.68). In other words, the AI's scoring did not show any significant gender-based bias, and the teacher's review adjustments were comparable for both male and female students. We also calculated the standardized mean difference of the AI–human score discrepancy between the two gender groups, which was near 0 (approximately 0.1), indicating a negligible effect size. This evidence suggests that, in our data, the AI-assisted scoring model treated male and female students essentially equally. It is worth noting that in designing the rubric and training the AI model, demographic factors like gender were not included, so the AI's scoring of the essays would not directly vary based on the student's gender—this design choice likely reduces potential sources of bias. Of course, fairness is not limited to gender; we also did a cursory check of whether students' prior English proficiency (e.g., higher vs. lower previous grades) influenced the AI score adjustments and found no obvious pattern. We believe that fairness audits should become a routine part of AI evaluation governance. Educators and administrators should regularly monitor scoring outcomes for different groups (gender, ethnicity, native language background, etc.) to ensure AI systems are not unintentionally amplifying pre-existing inequalities. If scores for a particular group are consistently lower than for others, the causes should be thoroughly investigated (e.g., an inappropriate rubric, imbalanced training data, etc.) and promptly addressed.

Table 4 summarizes the comparison of AI initial versus human final scores by student gender in the experimental class as part of the fairness audit.

Table 4. Comparison of AI Initial Scores and Human Final Scores by Student Gender in the Experimental Class (Fairness Audit)

Gender	**Number**	**AI Initial Average Score**	**Final Average Score**	**Average Difference (Final - Initial)**
Male	12	84.2	85.4	+1.2
Female	16	84.8	85.7	+0.9
Significance of Difference	-	-	-	*p* = 0.68

Note: A positive average difference indicates the human-adjusted score was higher than the AI's initial score. Significance was tested via an independent-samples t-test on the difference in AI–human adjustments between genders.

4.4 Student Perceptions and Feedback

To gauge student acceptance of the AI + human scoring model, we administered a post-course questionnaire. The responses were overwhelmingly positive. Over 90% of students agreed that the scoring process was fair and objective, and nearly all felt that AI assistance made feedback more timely and improved scoring consistency. Almost all students also acknowledged that teacher review ensured the final scores were accurate, which they found reassuring. Overall, 94% of students indicated they were satisfied with the AI + human combined scoring method. A small subset of high-achieving students did express reservations, noting that the AI sometimes failed to fully understand nuanced or creative ideas in their essays and that they still valued the teacher's personalized feedback. These comments highlight the need to continue providing strong human feedback for creative work and to guide students in interpreting AI-generated feedback appropriately.

Table 5 reports the experimental-class students' perceptions of the AI + human review scoring model.

Table 5. Experimental Class Students' Perceptions of the AI + Human Review Scoring Model (N=28)

Survey Item (Summary)	**Mean ± SD**	**Agreement Rate (%)**
1. Scoring process is fair and objective, reducing subjective bias.	4.32 ± 0.74	93.0
2. AI assistance improves scoring consistency and objectivity.	4.54 ± 0.58	95.0
3. AI assistance makes feedback more timely and effective.	4.61 ± 0.57	96.4
4. Teacher review ensures scoring accuracy and reliability.	4.68 ± 0.48	98.2
5. I am satisfied with this AI + human combined scoring method.	4.45 ± 0.64	94.1

Note: "Agreement Rate" is the percentage of students who selected 4 (Agree) or 5 (Strongly Agree). Additional survey items are provided in Appendix B.

4.5 Discussion

In summary, the results of this study indicate that the rubric assessment model combining AI scoring and human review is both feasible and effective for the Advanced English course. Firstly, in terms of effectiveness, this hybrid model did not undermine the rigor or fairness of the evaluation. AI initial scoring was overall consistent with human final scoring, while significantly improving scoring efficiency and consistency. This aligns with the intended purpose of applying educational technology: using automation to enhance efficiency, and using human expertise to ensure quality. Secondly, regarding the student experience, the human–computer combined scoring approach won widespread student approval. In particular, for language courses where teacher grading of essays was often characterized by long turnaround times and limited feedback, AI participation significantly shortened the feedback cycle and enriched feedback content, meeting students' need for timely improvement suggestions. However, it should be noted that a minority of high-achieving students were aware of AI's limitations, reminding us not to over-rely on AI to replace teachers. Teachers' professional judgment and personalized guidance remain indispensable for cultivating high-level language skills.

This study also preliminarily validated the performance of the AI scoring system in terms of fairness. Although a single small-scale experiment is insufficient to exhaust all dimensions of fairness, we did not find any new inequities caused by introducing AI on simple dimensions like gender. It is important to emphasize that such success relies on the reasonable design of the rubric and final correction by humans. As recommended by recent researchers (Boateng & Boateng, 2025; Barnes & Hutson, 2024), when AI is applied in educational measurement, strict ongoing monitoring of validity, reliability, and fairness should be conducted. Our practice demonstrates that through careful system design (for example, not providing the AI with any personally identifiable information that might introduce bias) and well-arranged human–computer collaborative processes, the potential risk of AI bias can be minimized. Of course, in the long term, systematic fairness review mechanisms for AI evaluation still need to be established. For instance, regularly inviting external experts to independently review scoring from students of different backgrounds, conducting comparative analyses of AI scoring results against human rating benchmarks, and promptly adjusting the AI model or rubric standards if significant deviations are found are all important components of AI assessment governance. Corresponding norms should be developed before large-scale implementation.

Finally, we acknowledge certain limitations of this study. The experiment was conducted in a single course with a relatively small sample size, which may limit the generalizability of the findings. Future research with larger and more diverse student populations, as well as in different subject areas, is recommended to further validate the effectiveness, fairness, and student reception of AI-assisted scoring models.

5. *Conclusion and Recommendations*

This study, focusing on a university Advanced English course, explored the effects of an "AI + human review" rubric assessment model and proposed suggestions for its governance. Through an experimental comparison, we found that this model improved scoring efficiency and consistency while producing student scores equivalent to traditional human evaluation. Importantly, our preliminary fairness audit found no notable biases or inequities introduced by the AI assistance. Student feedback indicated that most learners supported and were satisfied with this innovative approach, believing it enhanced the quality and timeliness of feedback in the learning process. In sum, the AI–human combined evaluation method shows broad promise in language education: it aligns with the trend of digital transformation and reducing teacher workload, while also ensuring the humanistic care and credibility of educational evaluation by retaining the human element.

To facilitate the wider implementation of this model, we propose the following recommendations:

1. **Refine the Rubric and AI Calibration:** Developing high-quality rubrics is a foundational prerequisite. Rubric criteria should comprehensively cover course objectives and be clearly defined to avoid misleading the AI. Significant effort in this study was devoted to designing and piloting the rubric; such an upfront investment is necessary. In the future, more data can be used to continuously calibrate the AI model, aligning its scoring with teacher standards. For example, machine learning can enable the AI to learn from a large set of teacher-scored examples, thereby reducing human–AI scoring differences.

2. **Strengthen Teacher Training and Involvement:** The human-in-the-loop model requires instructors to have a certain level of technical literacy. Universities should provide training to help teachers master the operation of the AI scoring system, including interpreting AI-generated feedback and correcting potential AI errors. At the same time, feedback from teachers during the review process should be continuously collected to improve the system, fostering human–AI co-evolution. Importantly, teachers should view AI as an assistive tool rather than a threat, adopting an open attitude toward integrating new technologies into teaching.

3. **Establish Regular Fairness Audits:** Education authorities should establish routine fairness monitoring protocols for AI-assisted assessment. For example, each semester data from AI-involved scoring can be analyzed to detect any abnormal differences in score distributions across groups (e.g., gender, ethnicity, region), and the results should be made transparent to ensure accountability. When disputes arise, clear channels for appeal and manual re-evaluation should be available to safeguard student rights. Teachers and students will only fully trust AI-driven evaluation if it consistently meets fairness and equity requirements.

4. **Protect Student Privacy and Data Security:** AI scoring systems involve storing and processing student data, which requires strict adherence to privacy protection principles. Schools should ensure that technology providers use student data only for agreed educational purposes and guard against any risk of data leakage from AI models. A sound AI governance framework needs to include robust data management protocols, regular model updates and reviews, and contingency plans, ensuring the technology remains under safe and controllable conditions.

5. **Further Research and Expanded Application:** This study had a limited sample size and focused only on English writing tasks. Future work should validate these findings in broader contexts and with other types of assignments. For example, the AI + human scoring model could be applied to English speaking or translation tasks to examine AI’s capabilities in evaluating oral performance or translation quality. Researchers should also explore the long-term effects of AI-assisted assessment on student learning engagement and self-directed improvement. Moreover, as AI technology evolves, new generations of language models with stronger text comprehension and feedback capabilities will emerge, necessitating ongoing evaluation of their potential roles in education.

In summary, the "AI + human review" rubric assessment model injects new vitality into traditional educational evaluation, reflecting the positive value of artificial intelligence in empowering education. These findings contribute to the literature by providing empirical evidence that a hybrid AI–human scoring approach can maintain assessment quality while significantly enhancing efficiency. Furthermore, the study offers a practical governance framework to guide the fair and reliable implementation of AI in educational assessment. However, for this model to truly exert long-term effects, it must be built upon sound instructional design and strict governance oversight. Only when technological tools and educational wisdom complement each other can we ensure that teaching evaluation is both efficient and human-centered, ultimately better serving student development. This study represents a preliminary step in this direction, and we hope that future practice and research will collectively refine this model, establishing it as a key pillar for future educational assessment reform.

References

1. Shukla, P., Singh, J., & Wang, W. (2025). Tangible progress: Employing visual metaphors and physical interfaces in AI-based English language learning. *Big Data Research, 42*, 100570.

2. Liu, X., & Huang, J. (2025). AI versus human assessment in EFL speaking classrooms: A comparative study in China. *Computer Assisted Language Learning*.

3. Min, S. (2025). An introduction to the integrated construction and practice of AI-empowered college foreign language teaching, learning, and assessment. *Journal of Technology in Language Education, 1*(1), 1–15.

4. Chen, X., & Luo, Y. (2025). From assistant to partner: Redefining the teacher's role in AI-empowered English classrooms in Guangdong, China. *Journal of Educational Technology Development and Exchange, 8*(2), 45–60.

5. VIPKID Research Group. (2025). Blending intelligence with humanity: A practice of combining AI and human tutoring to enhance English writing. *International Journal of Computer-Assisted Language Learning and Teaching, 5*(3), 1–12.

6. Fu, F., & Lu, L. (2025). Human-machine dialogue competition: Exploring a multidimensional evaluation system for teacher proficiency. *Journal of Interactive Learning Research, 36*(1), 78–95.

7. ETS. (2025). *Wilbur Max: An AI system for automated assessment and calibration*. Educational Testing Service.

8. Ningbo Institute of Technology. (2025). AI-empowered foreign language intelligent evaluation: A large-scale placement test for freshmen. *Journal of Language Testing and Assessment, 4*(2), 112–125.

9. Ge, H., & Lai, M. (2024). Reconstructing the listening and speaking classroom with AI: A case study of immersive scenario-based learning. *Computer Speech & Language, 85*, 101568.

10. Chen, Y. (2024). Teachers as technology-enabled guides: The core of human guidance in the age of artificial intelligence. *Educational Philosophy and Theory, 56*(8), 789–801.

11. Boateng, O., & Boateng, B. (2025). Algorithmic bias in educational systems: Examining the impact of AI-driven decision making in modern education. *World Journal of Advanced Research and Reviews, 25*(1), 2012–2017.

12. Barnes, E., & Hutson, J. (2024). Navigating the ethical terrain of AI in higher education: Strategies for mitigating bias and promoting fairness. *Forum for Education Studies, 2*(2), 1229.