Comparing AI and Human Assessment of Academic Writing Skills: A Kappa Analysis

Yeni Anistyasari[1*], *Shintami C* Hidayati[2], *Suparji* Suparji[1], *Ekohariadi* Ekohariadi[1], and *Dyah Ayu* Kusumaningtyas[3]

[1]Faculty of Engineering, Universitas Negeri Surabaya, 60231 Surabaya, Indonesia
[2]Department of Informatics, Institut Teknologi Sepuluh Nopember, 60111 Surabaya, Indonesia
[3]Department of Business and Management, Southern Taiwan University of Science and Technology, 71005 Tainan City, Taiwan

Abstract. This study aimed to examine the reliability and accuracy of ChatGPT in scoring reflective essays. The study took place in a course on Teaching and Learning in English for Academic Writing. Students completed a task to write a systematic literature review on computing education. The main goal was to compare ChatGPT scores with scores from human tutors. ChatGPT received assessment rubrics and prompt instructions. The research team verified ChatGPT's understanding of the rubric. ChatGPT then scored the essays independently. The research used 72 student reflections as the dataset. Expert tutors also scored the same essays. The comparison used statistical tests. Cohen's Kappa and Intraclass Correlation Coefficient (ICC) showed low agreement. The results indicated poor consistency between ChatGPT and human scores. Error analysis showed a consistent pattern. ChatGPT gave higher scores than human raters. The difference was statistically significant. The study found that ChatGPT worked efficiently and gave fast results. However, ChatGPT lacked the ability to interpret subtle meaning. ChatGPT also did not always follow the rubric as expected. The study concludes that ChatGPT can support formative feedback. However, ChatGPT should not replace human judgment in academic writing assessment.

1 Introduction

Global academic institutions often discuss the role of AI in assessment. Many institutions focus more on academic integrity than on AI's potential to improve assessment reliability. They rarely compare AI performance directly with human grading in terms of consistency. In Indonesia, several institutions are exploring the use of AI to support alternative assessment methods. These institutions aim to improve consistency in grading and provide personalized feedback. They are also examining the ethical, intellectual, and legal aspects of AI in education. Some educators remain hesitant to use AI in assessments. They prefer traditional assessment methods because they are concerned about academic integrity [1]–[3].

* Corresponding author: yenian@unesa.ac.id

In recent decades, educational scholars have recorded the difficulties encountered in the development of automated essay grading (AES) systems. Notwithstanding extensive efforts, these systems faced several constraints. A significant challenge, such as the conversion of phrases into vector representation has been identified, an essential phase in feature extraction and the training of machine learning models. Moreover, existing AES systems do not evaluate the overall completeness of an essay and lack the capability to provide personalized feedback on student answers. They also had difficulties in recognizing consistency among writings [4]. Moreover, these systems faced challenges from irrelevant or antagonistic student answers, prompting inquiries about their efficacy. Despite advancements, substantial obstacles persisted in the development of completely effective and dependable automated essay grading systems.

Nonetheless, because to the rapid advancement of technology and the complexity of machine learning, generative AI can identify language and text. Consequently, it is essential to investigate methods through which generative AI can enhance assessment quality, uphold academic integrity, expedite assessment turnaround, or reduce costs for educational institutions and faculty, especially concerning marking, grading, and delivering formative feedback on students' written evaluations. Presently, little information exists about the use of AI to enhance the dependability of the assessment process, including the grading of students' written evaluations. Moreover, the influence on the grading of written assessments, beyond conventional human evaluation, often results in subjective and inconsistent grading outcomes. The intrinsic subjectivity of the human marking process presents a considerable issue, since discrepancies in marks awarded to students may arise in big classrooms, especially when numerous markers are engaged [5]–[8].

Evaluating written evaluations may be costly and labour intensive, particularly when examiners are responsible for many courses. The absence of standardized grading standards and the difficulties in delivering comments exacerbate the issues. Therefore, the inquiry is, “Can AI tools such as ChatGPT effectively evaluate scientific writing as well as humans?” The research aimed to investigate the efficacy of generative AI tools in evaluating reflective essays. The study focused on ChatGPT. The study examined whether ChatGPT could provide quicker, more consistent, and more objective judgments than human grading. The research also aimed to identify the obstacles and constraints in using AI for assessment. The goal was to contribute to the ongoing discussion about the role of AI in education. The study explored how AI could improve the consistency of student evaluations.

2 Methods

This inquiry is part of a course called Teaching and Learning in English for Academic Writing. The rubric assessment is adopting from Scientific Writing Assessment [9]. The course required students to write a comprehensive literature review on computer education. The researchers used algorithms trained with student reflections and rubrics. The trained algorithms evaluated reflective essays independently. The main goal of the study was to compare tutor ratings with AI-generated scores. The study aimed to assess the effectiveness of the AI program. The study also aimed to identify limitations in human grading. The researchers used a dataset collected from a completed project.

The research focused on measuring the accuracy and effectiveness of ChatGPT in scoring reflective essays. The process involved giving ChatGPT assessment instructions. The researchers provided detailed rubrics to guide the scoring. They verified ChatGPT’s understanding of the rubrics. They tested its scoring ability using essays graded by expert tutors. They extracted key features from each essay. They used consistent prompt formats. They refined ChatGPT’s prompts to help it distinguish different performance levels. ChatGPT and expert tutors separately rated the same set of randomized essay scripts,

reducing bias and permitting direct comparison of human and AI performance. The final assessment used untrained student feedback to assess ChatGPT's generalization skills and provide a fair comparison. The study used a sample of 72 written reflection evaluations, each with thorough and well-documented rubrics. All student entries went through independent moderation and extensive marking. ChatGPT's ability to score independently and deliver individualized feedback was pre-tested on a sample of moderated scripts and rubrics rather than a broad dataset for learning evaluation and feedback giving.

3 Results and Discussion

3.1 Descriptive Statistics

Descriptive statistics show the characteristic differences between ChatGPT and human rater scores. The mean score of ChatGPT is 4.42. The mean score of human raters is 3.93. The ChatGPT score has a standard deviation of 0.38. The human rater score has a standard deviation of 0.74. The range of ChatGPT scores is 1.5. The range of human rater scores is 2.6. ChatGPT gives scores that tend to be higher and more consistent. Human raters give more varied scores. The median score of ChatGPT is higher than the human rater score. These results indicate that ChatGPT tends to give more positive ratings with a narrow spread. Human raters show greater accuracy in distinguishing writing quality.

Table 1. Descriptive Stataistics Results.

Statistic	ChatGPT Score	Human Rater Score
N (Count)	72	72
Mean	4.42	3.93
Standard Deviation	0.38	0.74
Minimum	3.50	2.10
25th Percentile (Q1)	4.26	3.50
Median (Q2)	4.48	4.30
75th Percentile (Q3)	4.70	4.40
Maximum	5.00	4.70

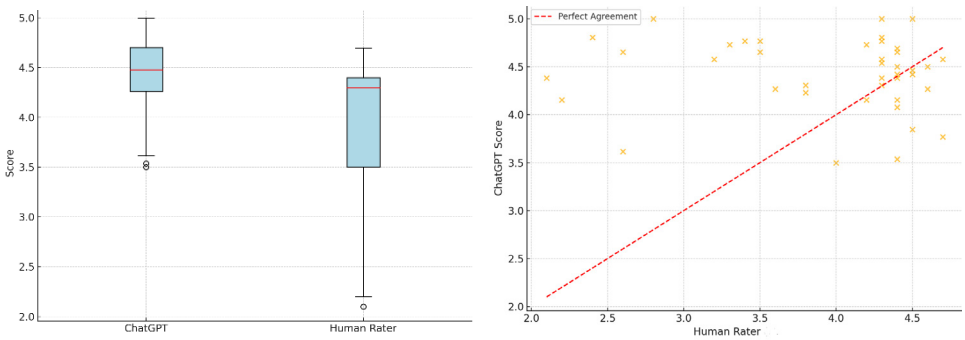

Fig. 1. Boxplot and Scatter plot of ChatGPT vs Human Rater.

3.2 Consistency and Reliability Results

In an effort to understand the extent of the alignment between artificial intelligence-based assessment (ChatGPT) and human raters on students' academic writing, this study uses two analytical approaches: Cohen's Kappa and Intraclass Correlation Coefficient (ICC). Cohen's Kappa is calculated after the numerical scores of both raters are categorized into five performance levels (A to E). The results show a kappa value of -0.048, which reflects a very low level of agreement, even worse than the agreement that can occur by chance. This negative value indicates a fairly striking difference in perspective between the automated system and human assessment in categorizing the quality of writing.

Furthermore, the consistency of the assessment was also tested using a continuous score-based ICC with a two-way random effects model. The ICC(2,1) value was obtained at -0.054 and ICC(2,k) at -0.115, both of which are below the minimum threshold of reliability. This shows that the difference between the scores given by ChatGPT and human raters is even greater than the variation between the students' writings themselves. In other words, there is not enough consistency between the two assessment sources.

The Pearson correlation test results produced a value of -0.089. This value indicates a very weak linear relationship between ChatGPT scores and human raters. This test is not statistically significant with a p-value of 0.587. The Spearman correlation test results showed a value of -0.176. This value indicates a weak ranking relationship between the two raters. This test was also not statistically significant with a p-value of 0.278. Both of these results indicate that there is no significant correlation between ChatGPT assessments and human assessments. Inter-rater reliability is stated to be low based on both correlation approaches.

These findings are not to discredit the role of AI, but rather serve as a reminder that while artificial intelligence has great potential to support educational processes, human intervention remains crucial in nuanced and complex assessment contexts such as academic writing. A more integrative approach is needed going forward—for example, training standardized rubric-based models, or combining human and machine assessments in a complementary way—for AI to become a more reliable partner in supporting equitable learning and evaluation processes.

Table 2. Consistency and Reliability Results.

Test Method	Value	Interpretation
Cohen’s Kappa	-0.048	No agreement; worse than chance
ICC (2,1) – Single Rater	-0.054	Not reliable if using only one rater
ICC (2,k) – Average Raters	-0.115	Not reliable even when averaging scores from both raters
Pearson Correlation	-0.089	Very weak linear correlation; not significant (p = 0.587)
Spearman Correlation	-0.176	Weak rank-order correlation; not significant (p = 0.278)

3.3 Error Analysis

Analysis of variance was used to measure the difference in scores between ChatGPT and human raters. The difference was calculated by subtracting the human rater's score from the ChatGPT score. The average difference was 0.497. This value indicates that ChatGPT gave a higher score than the human rater. The standard deviation of the difference was 0.864. This value indicates that the difference in scores between individuals was quite varied. A paired t-test was conducted to determine whether this difference in scores was statistically significant. The test results showed a t-value of 3.63 and a p-value of 0.0008. This result is significant at the 99 percent confidence level. This means that the difference between

ChatGPT and human rater scores did not occur by chance. Visualization using the Bland-Altman plot shows that most of the points are within the 95 percent confidence limit. However, the wide spread of points indicates that there is a large variation in the difference in scores. This result indicates that ChatGPT has a systematic tendency to give higher and less consistent scores when compared to human raters.

Table 3. Error Analysis.

Metric	Value	Interpretation
Mean Difference (Bias)	+0.497	ChatGPT tends to assign higher scores
Standard Deviation of Error	0.864	Score differences are moderately variable
Paired t-test (t-value)	3.63	Significant difference between ChatGPT and human scores
p-value	0.0008	Statistically significant at the 0.01 level
Direction of Bias	Positive	Systematic overestimation by ChatGPT

Bland-Altman plot is used to evaluate the agreement between ChatGPT and human rater scores. The horizontal axis shows the average of both scores. The vertical axis shows the difference between ChatGPT and human rater scores. The middle line shows the average difference of 0.497. The upper and lower lines represent the 95 percent confidence limits. Most of the points fall within the confidence limits. The points are quite widely spread around the middle line. This indicates that the difference in scores varies considerably between papers. The points that are not symmetrically distributed indicate a systematic trend. ChatGPT consistently gives higher scores than human raters. This plot confirms the statistical test results that there is bias in ChatGPT scoring. ChatGPT tends to give consistently higher scores. Although most of the scores are within acceptable limits, the discrepancies are still large enough to be noticed.

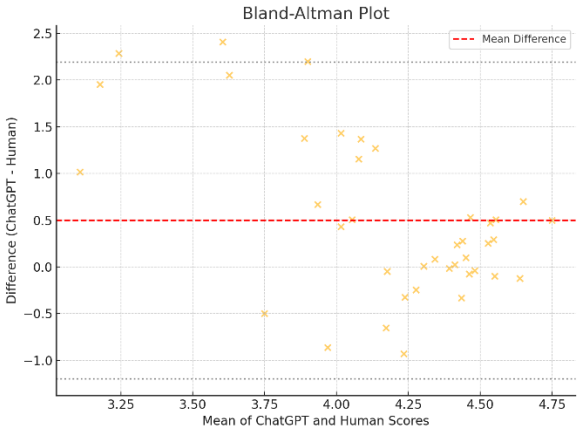

Fig. 2. Bland-Altman Plot Comparing ChatGPT and Human Rater Scores

The finding is in line with several previous studies that show that AI-based scoring systems still have limitations in imitating complex and contextual human judgment. ChatGPT performed well in generating texts, but was less accurate in evaluating logical coherence and argument authenticity. AI tends to give higher scores to long and technical texts, without considering the overall strength of the content. On the other hand, previous researches revealed the opposite outcomes. ChatGPT could deliver essay evaluations with

the same consistency as human raters in the setting of formative assessment in science and technology. The research indicated that ChatGPT has the potential to be an effective assessment tool, particularly when combined with precise rubric rules. ChatGPT in the reflective assignment evaluation process improves fairness and speeds up feedback since AI is unaffected by personal biases that may occur in human raters [9]–[14].

4 Conclusion

This study demonstrates that ChatGPT has advantages and disadvantages in grading academic writing. ChatGPT provides speed, technological consistency, and efficiency. The system generates assessments swiftly. ChatGPT does not suffer weariness or personal prejudice. These properties make it ideal for automated evaluation. However, the findings indicate that ChatGPT is less trustworthy than human assessors. ChatGPT lacks the capacity to evaluate writing at the same level. It does badly in terms of context, arguments, and logic. In several areas, ChatGPT falls short of human judgment. Low to negative Cohen's Kappa and ICC scores suggest a very low degree of agreement. Furthermore, ChatGPT has a systematic propensity to give higher ratings than human assessors, which may result in positive bias in the assessment. Given these findings, potential next steps include developing an AI-based assessment system trained on a structured rubric, displaying instances of expert-assessed student replies, and using hybrid assessments that combine AI evaluation with human verification. More research is also needed to evaluate the consistency of ChatGPT in different sorts of writing assignments and disciplines of study, as well as to investigate the degree to which ChatGPT can give formative feedback that is beneficial to students' learning processes. With a progressive and data-driven approach, AI, such as ChatGPT, may be guided to become a partner in the learning environment, supporting rather than replacing lecturers' roles.

References

[1] L. Wang, X. Chen, C. Wang, L. Xu, R. Shadiev, dan Y. Li, "ChatGPT's capabilities in providing feedback on undergraduate students' argumentation: A case study," *Think. Ski. Creat.*, vol. 51, no. August 2023, hal. 101440, 2024, doi: 10.1016/j.tsc.2023.101440.

[2] A. S. Espartinez, "Exploring student and teacher perceptions of ChatGPT use in higher education: A Q-Methodology study," *Comput. Educ. Artif. Intell.*, vol. 7, no. June, hal. 100264, 2024, doi: 10.1016/j.caeai.2024.100264.

[3] S. Groothuijsen, A. van den Beemt, J. C. Remmers, dan L. W. van Meeuwen, "AI chatbots in programming education: Students' use in a scientific computing course and consequences for learning," *Comput. Educ. Artif. Intell.*, vol. 7, no. September 2023, hal. 100290, 2024, doi: 10.1016/j.caeai.2024.100290.

[4] V. GECKİN, E. KIZILTAŞ, dan Ç. ÇINAR, "Assessing second-language academic writing: AI vs. Human raters," *J. Educ. Technol. Online Learn.*, vol. 6, no. 4, hal. 1096–1108, 2023, doi: 10.31681/jetol.1336599.

[5] M. F. Teng, "'ChatGPT is the companion, not enemies': EFL learners' perceptions and experiences in using ChatGPT for feedback in writing," *Comput. Educ. Artif. Intell.*, vol. 7, no. July, hal. 100270, 2024, doi: 10.1016/j.caeai.2024.100270.

[6] J. Steiss *dkk.*, "Comparing the quality of human and ChatGPT feedback of students' writing," *Learn. Instr.*, vol. 91, no. March, hal. 101894, 2024, doi: 10.1016/j.learninstruc.2024.101894.

[7] Y. Jiang, J. Hao, M. Fauss, dan C. Li, "Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers?," *Comput. Educ.*, vol. 217, no. November 2023, hal. 105070, 2024, doi: 10.1016/j.compedu.2024.105070.

[8] M. Asadi, S. Ebadi, dan L. Mohammadi, "The Impact of Integrating ChatGPT with Teachers' Feedback on EFL Writing Skills," *Think. Ski. Creat.*, hal. 101766, 2025, doi: 10.1016/j.tsc.2025.101766.

[9] M. D. C. Hampton, R. Rosenblum, C. D. Hill-Williams, L. Creighton-Wong, dan W. A. Randall, "Scientific writing development: Improve DNP student skill and writing efficiency," *Nurse Educ. Today*, vol. 112, 2022, doi: 10.1016/j.nedt.2022.105334.

[10] S. Toscu, "ChatGPT : Is It Reliable as an Automated Writing Evaluation Tool ?," *Anadolu J. Educ. Sci. Int.*, vol. 15, no. 1, hal. 329–349, 2024, doi: 10.18039/ajesi.1463503.

[11] T. Yamasaki dan A. Hiramatsu, "Experiments of Automatic Scoring Using Generative AI in a Summary Essay Learning System," *tters Institutional Res.*, vol. 004, no. LIR294, 2024, doi: https://doi.org/10.52731/lir.v004.294 Experiments.

[12] M. Liu dan H. Reinders, "Do AI chatbots impact motivation? Insights from a preliminary longitudinal study," *System*, vol. 128, no. March 2024, hal. 103544, 2025, doi: 10.1016/j.system.2024.103544.

[13] S. Tobler, "Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments," *MethodsX*, vol. 12, no. November 2023, hal. 102531, 2024, doi: 10.1016/j.mex.2023.102531.

[14] A. Pack, A. Barrett, dan J. Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," *Comput. Educ. Artif. Intell.*, vol. 6, no. April, hal. 100234, 2024, doi: 10.1016/j.caeai.2024.100234.