Evaluating the Reliability of AI-Assisted Writing Assessment: A Comparative Analysis of *WritingGPT*, *ChatGPT*, and Human Raters in EFL Contexts

Emre ARTUT[1] **Kübra ŞIK**[2] **Yasin TURAN**[3]

The use of artificial intelligence (AI) in language education has substantially changed how writing is taught, learned, and assessed. The most recent advancement in large language models (LLMs) has led to a progression from automated grammar checkers and superficial error checkers to dynamically operating, context-aware writing evaluation systems (Rost, 2024). These advances hold great promise for writing instruction in the form of immediate feedback, scalable assessment, and personalized learning experiences (Li et al., 2024). AI-based writing assessment tools, in particular, are increasingly being advocated for their efficiency, cost-effectiveness, and potential to eliminate human bias. However, it remains an open question whether these tools yield pedagogically reliable, valid, and interpretable evaluations that correspond to human assessment (Pilicita & Barra, 2025; Tajik, 2025).

[1]Lect., Abdullah Gül University, School of Foreign Languages, Department of Foreign Languages, E-mail: emre.artut@agu.edu.tr, ORCID: 0000-0003-0642-2784.

[2]Dr., Abdullah Gül University, School of Foreign Languages, Department of Foreign Languages, E-mail: kubra.keser@agu.edu.tr, ORCID: 0000-0003-1492-5883.

[3]Dr., Abdullah Gül University, School of Foreign Languages, Department of Foreign Languages, E-mail: yasin.turan@agu.edu.tr, ORCID: 0000-0002-8252-569X.

Furthermore, the role of AI in writing assessments extends beyond efficiency, as it challenges assumptions about what writing competence entails and how it should be evaluated. In English as a Foreign Language (EFL) contexts, writing is not simply a product to assess; it is also a process that supports linguistic and cognitive development (Toscu, 2023). Thus, replacing or even supplementing human raters with AI will directly affect both the pedagogical and ethical aspects of educational instruction. As AI-generated feedback takes center stage in the development of writing among learners, teachers, and researchers have to consider if that feedback can encourage learning, rather than simply optimizing scores.

While initial studies have shown that AI tools can provide consistent and criterion-based feedback, fostering learner engagement (Chen & He, 2023), questions regarding validity in comparison to human raters with training remain unsettled. Likewise, AI's ability to interpret the organization of discourse, the effectiveness of rhetorical moves, and fulfillment of tasks remains severely lacking (Geckin et al., 2023). Notably, too much reliance on AI feedback contributes to a learner's lack of engagement with metacognitive/reflective aspects, which undermines the qualities of reflexivity and criticality in writing development (Stevenston & Phakiti, 2023). In the case of EFL preparatory programs, which place a significant emphasis on formative writing assessment in the development of academic literacy, these conflicting premises are particularly problematic.

The present study will examine the correlation between human raters and AI-based scoring systems in an authentic academic writing context. More specifically, it will examine the reliability of *WritingGPT*—a model that has been fine-tuned on the rubric and corpus of a university's undergraduate writing course—in comparison to standard *ChatGPT* and expert human raters. *WritingGPT* was trained by researchers to more closely reflect the academic writing expectations on campus and the institution's assessment rubrics. The study will next outline the comparative examination through the context of two research questions:

1. To what extent do AI raters (*WritingGPT* and *ChatGPT*) align with human raters in evaluating EFL writing performance?
2. Can fine-tuned AI tools be considered pedagogically valid and ethically responsible alternatives or supplements to human scoring in higher education?

This study contributes to the increased dialogue surrounding responsible AI adoption in education by considering these questions. It aims to serve as a guide for institutions engaging with policy development, to foster evidence-based policies that allow for a balance between efficiency and pedagogical fidelity. The study's recommendations are expected to inform the design of evaluation instruments that hybridize human judgment and AI efficiency while offering considerations for fairness, transparency, and validity. Ultimately, this work reiterates that the study does not aim to replace teachers, but rather to expand human evaluative capacity through the ethical and contextual deployment of AI.

Literature Review

AI-Assisted Writing Assessment

Automated writing evaluation (AWE) has evolved from early systems that relied on rules to more sophisticated, neural-network-driven platforms. Early systems, such as *e-rater* and *IntelliMetric*, relied on linguistic and statistical features to train human scores (Attali & Burstein, 2006). Early AWE models were primarily concerned with surface-level features (grammar, mechanics, and vocabulary range), but did not account for aspects such as discourse coherence, argument quality, and language creativity, which are essential in a university writing environment (Ding & Zou, 2024).

With the technology known as LLM, this began to shift. Distinct from traditional AWE systems, LLMs are capable of representing meaning, context, and rhetorical intent through the examination of text across large corpora. LLMs not only score but also generate and adjust text for writers in context, resulting in an interactive, process-oriented feedback loop. This

has changed how AI systems evaluate a wider variety of characteristics of writing quality typically used in educational settings—cohesion, lexical richness, task achievement, and stylistic appropriateness (Crossley, 2020). For example, *ChatGPT* provides narrative, argumentative, or expository responses as a reaction to the prompt and context in which the user presented the information, while also being able to offer examples and revisions in real-time.

Still, researchers stress the importance of domain-specific fine-tuning to lessen bias and promote fairness. Generic models that have been trained on English-dominant, large-scale corpora may misinterpret some of the unique features of non-native writing, or may inadvertently penalize valid and appropriate English as a foreign language (EFL) variation. Domain-specific AWE systems (eg, *WritingGPT*) utilized in the present study are trained through locally relevant rubrics and learner corpora to provide more contextually sensitive scoring. Despite these significant advancements, there is a lack of empirical literature examining how AI scoring aligns with human judgment across a diverse range of EFL contexts. Most validation studies have been conducted in English-dominant contexts, where learners' linguistic practices often differ from those of EFL learners, leaving the extent to which findings can be transferred unknown (Li et al., 2024). Lastly, while AI scoring has been shown to have high correlations with human raters in standardized test settings (eg, *TOEFL iBT*), the findings may not generalize to the classroom-based assessment context, which emphasizes formative feedback and learner autonomy.

Human–AI Alignment and Reliability

Reliability and validity are fundamental aspects of any assessment mode. Inter-rater reliability between human raters has long been a concern in writing assessment, with differences often attributed to variations in interpretation, experience, and even fatigue (Toscu, 2023). The involvement of AI raters complicates the issue further, as machine consistency has to be balanced against the interpretive flexibility available

to a human rater. Recent work has explored this intersection from psychometric and pedagogical interests.

Chen and He (2023) found that after controlling for contextual variables, AI scores correlated strongly with human scores, particularly in terms of surface-level features (grammar, lexical accuracy). However, Geckin et al. (2023) found that human raters weigh rhetorical structure and argument coherence more heavily than AI systems, resulting in differences in scores. Working with a large-scale meta-analysis, Pilicita and Barra (2025) additionally concluded that AI scoring tends to be more lenient than human scoring, as AI issued higher average scores and often assessed student writing based on linguistic fluency rather than conceptual depth. These studies indicate that AI is evaluating writing more for its form than its content.

Another aspect of alignment is interpretability, which refers to the extent to which human raters can understand and trust the decisions made by AI. Rost (2024) notes that unless teachers understand the process by which AI generates scores, the use of this technology has the potential to decrease pedagogical transparency. For example, teachers may struggle to explain the AI feedback to their learners, which raises questions about the formative function of assessment. Another consideration is bias: AI models trained on unbalanced datasets can trigger linguistic or cultural bias, which may impact non-native speakers.

Reliability measures, such as Pearson correlations or Cohen's kappa, provide some statistical evidence that scores are aligned; however, they do not account for the qualitative dimension of trust between raters and the AI. In practice, teachers' trust in AI scoring is important, even as the metrics drive some reporting of agreement. Pilicita and Barra (2025) explain that teachers are more likely to consider an AI evaluation favorable when they feel it augments their judgment, rather than replaces it. This relational trust forms the basis of hybrid assessments, which combine human intuition and AI scoring in reference to alignment.

Pedagogical and Ethical Considerations

Apart from its practical effectiveness, introducing AI writing assessment raises considerable pedagogical and ethical issues for educators and researchers. From a pedagogical perspective, AI feedback appears to foster learner autonomy more effectively than traditional means, as it provides nearly immediate, transparent, and criterion-referenced feedback on writing drafts that emphasizes the iterative process of learning (Lui, 2024). The nature of AI feedback enables learners to revise their drafts independently, allowing them to develop a more self-regulated writing style. When employed in a formative assessment context, this type of feedback may even further empower learners by connecting what they learn in the classroom and their own writing (Lui, 2024; p. 239).

However, there are dangers in an overreliance on AI writing assessments that could dilute critical thinking and creativity. In a case that learners become accustomed to "help" through automated feedback, they may focus on checking for mechanical correctness rather than developing the rhetorical aspects of their writing. They could respond to AI feedback without carefully formulating their responses. In contexts such as EFL, where metacognitive awareness can facilitate not only language but also learning, there may be negative consequences to providing prompts instead of responses that encourage reflection.

Considering AI as an ethical issue raises the problem of data privacy, algorithmic transparency, and fairness (Zawacki-Richter et al., 2023). While many AI tooling, particularly LLMs, function in the black box, or opaque, architecture of Machine Learning, determining how they process learner data or evaluate learning is far more difficult. It is thus critical to discuss the implications of both explainability and accountability in teachers and institutions in determining the responsible and ethical use of AI and AI feedback with students for educational purposes. Further, considerations of fairness in scoring for second language or learner populations will require monitoring for potential systematic disadvantage. Many researchers on AI assessment tools endorse a hybrid model of AI–

human assessments. In hybrid models, AI would provide scoring or feedback in a consistent manner based on data, addressing surface-level accuracy, while teachers and/or other human raters would focus on conceptual depth, discourse-focused assessment, and the contextualized appropriateness of the assessment. This method could serve efficiency and educational validity in the process.

Nevertheless, while the teacher transitions from an evaluator to a mediator (Pangrazio et al., 2020), interpreters' feedback for students prompts students to engage with AI's feedback critically. This mediating role as a pedagogy keeps institutions in control of the use of educational technology: students are supposed to learn with educational technology, rather than the educational technology governing the learning. AI-supported writing evaluation has the potential to be transformative for EFL. However, the integration of AI must be adequately contextual and undertaken with caution. The literature highlights three recurrent challenges: (1) the need for psychometric reliability, comparability to human grading, (2) pedagogical integrity, in terms of developing metacognitive and critical engagement with the AI, and (3) ethics of transparency and fairness. Developing these issues requires empirically valid evidence, in authentic, classroom-based instruction versus lab-based tasks.

The present study compares *WritingGPT*, *ChatGPT*, and human raters in a writing program at a university in this effort. The study extends our empirical and theoretical understanding of how AI can ethically enhance writing assessment in EFL higher education—connecting the innovation of technology with human-centered pedagogy.

Methodology

Research Design

A within-subjects quantitative comparison design was employed to evaluate the performance of two AI models, *WritingGPT* (fine-tuned and domain-specific) and *ChatGPT* (general-purpose), alongside the ratings of

human raters in the EFL academic writing environment. Using the within-subject quantitative comparison design, every paper is rated by each type of rater, ensuring that the ratings of the raters are directly comparable for the same written work. This particular research design not only eliminates the variance in ratings caused by the subjects' proficiency levels but also provides maximum power with a small sample size.

Data Collection Instruments and Procedure

The dataset consisted of 29 sample essays written for a single task in a course on Academic English Writing at a state university in Türkiye. The course is nested within the English as a Foreign Language (EFL) preparation program and focuses on enhancing learners' knowledge of academic English writing skills, enabling them to write well-structured and well-supported English essays. The essays used in the study were typical coursework and not specially designed studies for experimentation, thus lending a high degree of construct validity to the findings, as the work represents the performance of learners within the course environment. The essays were written by different learners, and the authors were not paired with any individual learner's information, as the objective of the study was to compare scoring performance across the three raters.

The raters consisted of three instructors with experience and knowledge of Academic English and Writing at the same institution. The three raters also possessed a significant level of familiarity and understanding of the course outcomes and the rubric used within the program to carry out the summative and formative evaluations. Prior to the rating task, the raters participated in a brief and formal orientation procedure designed to establish a common understanding of the study's procedures and the importance of various aspects, including confidentiality and the objective and unbiased nature of the ratings.

Essays and Rubric: The essays were graded using the same rubric (0-100 points) that is used in the writing course. The rubric captures the essential aspects of the writing performance criteria, including task satisfaction and

the relevance of the written work, organization and cohesion, vocabulary and accuracy, and mechanics (such as spelling, punctuation, and layout). Although the final score of the essays provides a holistic judgment of the work, the scorers are required to take into account each of the aspects of the work being considered for the final judgment. The rubric was collectively reviewed with the three instructors for the course to ensure a common understanding of the rubric's performance criteria and the characteristics that define low-, mid-, and high-performing essays. This enabled the scorers to align local standards with the institution's shared criteria.

Human Scoring: The essays were anonymized by removing all personally identifiable information and then further randomized so that the scorers could not determine the original order of submission or the identities of the authors. This served the dual purpose of reducing the risk of any bias and allowing the essays to be scored purely within the merits of the written product itself. Each of the three scorers individually scored the entire set of 29 essays, not referring to the work of the other scorers and not having access to any AI-computed information. Each of the essays received a single score within the range of 0 to 100, based on the rubric criteria. Once the scoring concluded, the three scores were then combined into a Human Mean, which would serve as the standard for measuring AI performance. Before applying the Human Mean to further analysis, inter-rater agreement among the three scorers of the essays and the derived Human Mean standard itself would be assessed to ensure that the standard itself possessed adequate inter-rater reliability to be used for the purpose of comparison.

The interaction between the candidate's writing performance and the fine-tuned, domain-specific model of the writing generator (*WritingGPT*) only occurred after the finalization of the grading scores, ensuring that the AI model scores could not influence the final scoring procedure. The fine-tuned model, specifically the writing generator (*WritingGPT*), focuses on a locally representative dataset of student writing and related scores from

the same institution, aiming to replicate the standards of a similar writing program in a comparable educational organization. The procedure of the submission of each candidate's work (29 essays) to the writing generator (*WritingGPT*) followed a specific pattern of common questions that included (a) the description of the scoring task (EFL academic writing performance), (b) a summary of the scoring criteria of the specific rubric (0-100), and (c) a scoring task of the model to derive a single score within the (0-100) range. *ChatGPT* served for comparison. The model responded to the same essays and the same rubric-aligned prompt, but, unlike the *WritingGPT* model, the training of the model was not adjusted to the local context and rubric. This difference enabled a comparison of a domain-specialized AI model and a general-purpose model operating within the same scoring task instructions. Both AI models were accessed through their respective interfaces, and each essay was scored only once using each model. Additional information generated during scoring and text produced during the scoring task are considered irrelevant to the study and therefore not utilized; only the score itself is retained and used for analytical comparison. The scoring of both AI models using rubric-aligned prompts enabled a comparison of model performance that could be attributed to model training and specialization, rather than to scoring task instructions.

Data Analysis

The assumption of normality for score distributions was assessed using the Shapiro–Wilk test, which is recommended for small- to moderate-sized samples ($n < 50$) (Ghasemi & Zahediasl, 2012). As the distributions for all three rater groups did not deviate significantly from normality ($p > .05$), the data satisfied the assumption required for parametric analyses.

Descriptive statistical tests (mean, standard deviation, minimum, maximum) were carried out for each rater group (Human Mean, *WritingGPT*, *ChatGPT*) to obtain a summary of each of the raters' distribution of scores and a first impression of the raters' level of leniency vs. severity. The calculations of the descriptive statistics further enabled the inspection of the data and its distribution, which has implications for

whether the rater tends to assign scores that are bunched close together or whether the rater has the ability to discriminate between the essays (see Table 1 below and further discussion below).

Given the within-subject design, where each essay received ratings from both human and AI raters, paired-samples t-tests were employed to determine whether AI-generated ratings systematically differed from the human benchmark. The dependent nature of the observations makes this test suitable for comparing the ratings of the same essays across raters (Field, 2018).

To examine the degree of linear association between AI-generated scores and human ratings, Pearson correlation coefficients (r) were computed. Pearson's r is appropriate for interval-level continuous data under parametric conditions, enabling the assessment of how closely AI and human raters align in the rank ordering of essays (Dancey & Reidy, 2017).

Because correlation reflects association but not agreement or bias, error metrics were also computed:

- **Mean Difference (Bias)** — to assess systematic over- or under-scoring by AI relative to human ratings.
- **Mean Absolute Error (MAE)** — to quantify the average magnitude of score deviations irrespective of direction.
- **Root Mean Square Error (RMSE)** — to weight larger deviations more heavily, highlighting the impact of outlier discrepancies.

To further evaluate agreement — especially the consistency across different raters (human and AI) — the Intraclass Correlation Coefficient (ICC) was calculated using a two-way random-effects, absolute-agreement model. Both ICC(2,1) (single-rater reliability) and ICC(2,k) (average-rating reliability) were reported following established guidelines (Koo & Li, 2016).

Additionally, the agreement between AI and human ratings was visualized using Bland–Altman plots, which depict the mean difference and limits of

agreement to examine bias and variability across the scoring range—a recommended method for comparing measurement methods (Giavarina, 2015).

Finally, to determine whether AI-generated grades could be considered practically equivalent to human grades for educational purposes, a Two One-Sided Tests (TOST) procedure was conducted with a predefined equivalence margin of ±5 points. In line with recommendations for equivalence testing, the equivalence bounds were set a priori to reflect the smallest difference deemed educationally meaningful in this grading context, that is, a range within which score differences would not be expected to alter substantive grading decisions (Lakens, 2017; Meyners, 2012; Wiens, 2002).

The predictive validity of the AI-generated ratings was examined using linear regression analyses, with the Human Mean serving as the dependent variable and each AI's scores as predictors. This approach allowed the assessment of how much variance in human ratings can be explained by the AI scores. Regression provides evidence of predictive alignment—that is, the extent to which AI scores can estimate human scores linearly, rather than merely reflecting association or mean-level equivalence. Similar regression-based approaches are widely employed in automated essay scoring research, where models are explicitly trained to predict human essay scores from linguistic features or embeddings (Attali & Burstein, 2006; Attali, 2008; Crossley, 2020; Li & Ng, 2024).

Results

The Results section is organised into four parts: (a) descriptive statistics for the three rater types, (b) associations and agreement between AI and human scores, (c) predictive validity analyses, and (d) equivalence testing. As a starting point, Table 1 presents the descriptive statistics for the Human Mean, *WritingGPT*, and *ChatGPT* scores, all on a 0–100 scale. Overall, *WritingGPT* tended to assign slightly higher scores than human

raters, whereas *ChatGPT*'s mean scores were closer to the human benchmark but showed different patterns of variability.

Table 1

Descriptive Statistics for Rater Scores (0–100)

Rater	***Mean***	***SD***	***Min***	***Max***
Human	64.09	7.79	50	82.5
WritingGPT	67.69	8.53	54	82
ChatGPT	63.86	6.92	46	72

Human = mean of three calibrated human raters.

Table 1 summarises the distribution of scores assigned by the three rater types. Human raters awarded on average 64.09 points (SD = 7.79), whereas *WritingGPT* produced higher scores (M = 67.69, SD = 8.53). *ChatGPT*'s mean score (M = 63.86, SD = 6.92) is almost identical to the Human Mean. These values indicate a leniency bias for *WritingGPT* at the level of central tendency, while *ChatGPT* aligns with humans in terms of mean severity. The larger standard deviation for *WritingGPT* indicates that it utilizes a wider range of the scale and discriminates more strongly between essays than either the human raters or *ChatGPT*.

Table 2

Mean Bias vs. Human (Paired Comparisons)

Rater	***Mean Difference (AI − Human)***	***t (28)***	***p***	***Max***
WritingGPT − Human	+3.60	4.43	0.00013	82.5
ChatGPT − Human	−0.22	−0.16	0.874	82

Table 2 presents the mean differences between AI scores and the Human Mean, together with paired-samples *t*-tests. *WritingGPT* graded essays

significantly higher than human raters, with an average positive bias of 3.60 points, $t(28) = 4.43$, $p = .00013$. This result shows a systematic leniency effect rather than random fluctuation. By contrast, *ChatGPT*'s scores do not differ statistically from human grades ($MD = -0.22$), $t(28) = -0.16$, $p = .874$, indicating the absence of a consistent positive or negative bias. In terms of mean severity, *WritingGPT* functions as a lenient rater, whereas *ChatGPT* is equivalent to the Human Mean.

Table 3

Correlations with Human Scores

	Pearson r
WritingGPT ↔ Human	0.860
ChatGPT ↔ Human	0.480

Table 3 reports Pearson correlations between AI and human ratings. *WritingGPT* exhibits a very strong positive association with human scores ($r = 0.860$), demonstrating that it closely reproduces the human rank ordering of essays: essays that receive higher scores from human raters also receive higher scores from *WritingGPT*. The correlation between *ChatGPT* and human ratings is moderate ($r = .480$), indicating a weaker alignment in ranking and more frequent deviations from human judgment. Overall, *WritingGPT* approximates the scoring pattern of human raters to a substantially greater extent than the general-purpose *ChatGPT* model.

Table 4

Error Metrics Relative to Human Grading

Model	***Mean Difference***	***MAE***	***RMSE***
WritingGPT	+3.60	4.76	5.61
ChatGPT	−0.22	6.07	7.41

Table 4 presents absolute and squared error indices relative to the Human Mean. *WritingGPT* shows a positive mean difference (+3.60), but its MAE (4.76) and RMSE (5.61) are smaller than those of *ChatGPT* (MAE = 6.07, RMSE = 7.41). These values indicate that individual scores produced by *WritingGPT* remain closer to human ratings, with fewer large discrepancies, even though *WritingGPT* is consistently higher on average. ChatGPT produces larger and more variable errors around the Human Mean. Consequently, in terms of point-by-point accuracy, WritingGPT provides a closer numerical approximation to human grading than ChatGPT.

Table 5

Regression Models Predicting Human Scores from AI Scores

Predictor (AI)	***Intercept***	***Slope***	***R²***	***RMSE***
WritingGPT	10.98	0.785	.739	4.05
ChatGPT	—	—	.230	6.96

Table 5 summarises the regression models predicting human scores from AI scores. The model with *WritingGPT* as a predictor explains 73.9% of the variance in human ratings (R^2 = .739) with an RMSE of 4.05, indicating high predictive validity. The slope of 0.785, which is below 1, shows that WritingGPT inflates scores in the lower and middle range and compresses scores at the top of the scale, consistent with the observed leniency. The regression model based on ChatGPT accounts for only 23.0% of the variance in human scores (R^2 = .230) and yields a considerably larger RMSE (6.96). These results demonstrate that WritingGPT can be used to estimate human ratings with substantially greater precision than *ChatGPT*.

Table 6

Inter-Rater Reliability

ICC Type	*Value*
ICC (2,1)	0.563
ICC (2, k)	0.794

Table 6 reports the intraclass correlation coefficients for overall agreement across the three rater types. The single-rater coefficient ICC(2,1) = 0.563 reflects moderate reliability when any one rater is considered in isolation. The average-measure coefficient ICC(2,k) = 0.794 indicates good reliability when ratings are averaged across raters. From an assessment perspective, these coefficients show that combining scores from multiple raters—such as human raters with *WritingGPT*—produces more stable and dependable grading than relying on a single source of scores.

Table 7

Bland–Altman Agreement Summary

	Mean Difference	*Spread (approx.)*	*Pattern*
WritingGPT − Human	+3 to +4	Narrower	Predictable leniency
ChatGPT − Human	~0	±15 (wider)	Less precise agreement

Table 7 summarises the Bland–Altman analyses. For *WritingGPT*, the mean difference relative to the Human Mean is consistently between +3 and +4 points, with a relatively narrow dispersion. This pattern indicates a stable agreement structure characterised by a constant leniency offset: WritingGPT grades higher than humans by a predictable amount across the scale. For *ChatGPT*, the mean difference is close to zero, but the spread of differences is approximately ±15 points, revealing wide limits of agreement. This wide dispersion indicates that *ChatGPT* occasionally

assigns scores that are substantially higher or lower than human ratings, resulting in imprecise agreement despite the absence of systematic bias.

Table 8

Equivalence Testing vs. Human (TOST, ±5 Points, 90% CI)

Comparison	***Mean Difference***	***90% CI***
WritingGPT − Human	+3.60	[2.22, 4.99]
ChatGPT − Human	−0.22	[−2.61, 2.16]

Table 8 presents the results of the TOST procedure with an equivalence margin of ±5 points. For WritingGPT, the 90% confidence interval of the mean difference [2.22, 4.99] lies entirely within the equivalence bounds, indicating that its average scores are statistically equivalent to human scores within this tolerance. The same holds for *ChatGPT*, whose confidence interval [−2.61, 2.16] also falls inside the ±5-point region. Thus, both AI models can be regarded as producing scores that are, on average, practically equivalent to human grades. However, when the equivalence results are interpreted in conjunction with the error indices and correlations, *WritingGPT* exhibits a more coherent pattern of equivalence and reliability, whereas *ChatGPT* achieves equivalence only at the mean level and displays considerable variability at the level of individual essays.

Discussion & Conclusion

This research investigates whether or not AI-supported writing assessment produce reliable and valid results by using three types of evaluative approaches: 1) A Fine-tuned AI Model called *WritingGPT* (WWN); 2) a General-purpose AI Model like *ChatGPT* (API); and 3) the Evaluator; Analyzing the three approaches produces quantitative analysis and comparisons of scores distribution and correlations and Ross et. al (2021)

provide detailed information regarding error metrics. From our analyses, the research offers insights into evaluating academic writing (English as a Second Language [ESL]) using Artificial Intelligence (AI); however, our primary goal here is to shed light on the potential and/or limitations of AI for academic writing assessment.

The results of this research suggest that *WritingGPT* had a higher level of correlation with human judgments than did either the API or evaluator approach; the Pearson Correlation Coefficient of *WritingGPT* was 0.86. The mean absolute and maximum errors indicate that the score produced by *WritingGPT* approximates the evaluation produced by an evaluator. This supports the conclusion that *WritingGPT* assessment scores closely resemble those of a human evaluator. As another form of support for *WritingGPT*'s high level of reliability, the statistical analysis performed in this research (Statistical Equivalent Test) showed that *WritingGPT*'s assessment scores were statistically equivalent to those produced by human evaluators within an acceptable range of ±5 points of the maximum score. While *WritingGPT*'s tendency to score more leniently than a human evaluator was consistent and systematic in its scoring, it was not random; therefore, this tendency could be easily adjusted and/or interpreted.

ChatGPT's performance was the least consistent and most variable across samples compared to the other two tools. The correlation coefficient between *ChatGPT*'s ratings and those of human raters was lower ($r = .48$), indicating that *ChatGPT*'s ratings were less stable and less reflective of its ability to evaluate writing quality in relation to the assignment context. While *ChatGPT* tended to rate writing more highly based on its surface-level linguistic properties (i.e., grammar, spelling, complexity, etc.) and less on the overall quality of the writing, including how well the argument was presented, how coherent and structured the text was, how effectively the author used rhetorical elements, and, ultimately, how well the reader understood the text (Crossley & McNamara, 2020). Thus, *ChatGPT* faces a significantly greater challenge than general-purpose AI models when it comes to providing feedback on a writing assignment that is highly

specific and based on a detailed set of criteria, such as standardized writing assessments and the conventions of writing in a particular discipline (Biber & Gray, 2016). The results of this study indicate a substantial difference in the abilities of AI and generalized writing tools to rate writing accurately, and that *WritingGPT*'s successful ratings can be explained by the fact that it was specifically trained to support the evaluation of academic writing in the English as a Foreign Language (EFL) context and was trained based on the criteria used by human raters to evaluate EFL writing (Li et al., 2023). *WritingGPT* is thus able to simulate the way a human would reason through the quality of a text much better than a generic AI was able to eliminate generative programming.

One major discovery related to predictive validity. *WritingGPT* accounted for nearly 74% of the variance seen in human scores. As a result, *WritingGPT*'s evaluations correspond well to the features that human raters consider important, and therefore, it was able to predict writing proficiency by integrating not only measurable elements but also overall and interstate writing proficiency as interpreted through human judgment (Weigle 2002; Knoch 2009). Conversely, *ChatGPT* only explained about 23% of the variance found in human scores, indicating that a significant portion of its scoring methodology did not correspond to the constructs that are valued in the evaluations made by human raters. Therefore, this supports that the reliability and validity of AI tools do not occur naturally, but rather must be established by creating reliable AI tools through careful planning, training, and contextualization (Chapelle & Voss, 2021).

The results of this research also provide insight into the potential applications of AI in the field of education. Given that *WritingGPT* has a high degree of reliability, AI tools can be used effectively to augment human raters in the context of large-scale assessments, when resources are constrained. Teachers may use AI score evaluations to identify outliers more effectively and review certain papers more closely. AI-based systems could also provide formative feedback to students, highlighting

strengths and weaknesses more rapidly than traditional grading methods, thereby supporting iterative learning and timely revision cycles.

Nevertheless, the findings also reveal significant warnings. First, if AI were to achieve results similar to those of a human rater, one should not view it as an alternative to human judgment. The slight difference in leniency between *WritingGPT*'s scores and those of a human is indicative of a systematic bias that could favour or disfavour certain groups of learners, particularly within the context of EFL, where the linguistic features of both groups are often significantly different (Ortega, 2020). If these types of biases are not monitored by a human being, they can build upon themselves, and over time, they can result in a substantial difference between the two ratings. Second, AI systems are unable to provide the same level of contextual empathy and interpretive flexibility that human raters naturally offer in their responses. Writing is not only an assembly of vocabulary and grammar; writing is an act of communication that carries intent, nuance, and awareness of the audience. A model like WritingGPT, although sophisticated, is unable to account for the rhetorical or emotional aspects of a text that would be valued by a human reader (Hyland & Jiang, 2021). Thus, it would likely misinterpret the creative or non-traditional choices made in a piece of writing that a human evaluator may be able to recognise as valid.

Thirdly, AI will offer fast feedback; however, both teachers and learners must have access to and be able to interpret AI's comments as well as understand how AI arrives at evaluation scores. Suppose a learner does not understand why they received a specific evaluation score or is unsure how to proceed based on feedback provided by AI. In that case, the value of AI as an educational assessment tool has been diminished. For this reason, creating clarity and user understanding, as well as developing AI evaluation systems that meet the needs of teaching and learning, will be the highest priority, alongside creating accurate systems.

The findings of this study also impact teacher practice and school or institutional policy regarding assessment. Teachers can utilize AI as a co-

evaluator to enable them to focus on higher-order skills in their feedback, such as argumentation, organization, or creativity. Additionally, policymakers at the institutional level should establish transparent, accountable, and fair AI-supported assessments (Reich & Ito, 2023). Regular alignment between the human scoring and the AI scoring will help support fairness and consistency in scored assessments.

In addition to establishing reliability, the current research provides guidance for future development. This future development could be supported by additional research in the area of how the continual fine-tuning of AI using locally generated students' writing samples impacts its performance over time. A longitudinal study on how students react to the Artificial Intelligence (AI) feedback, and what impact this has on students' levels of autonomy, improvement in revisions, and perspectives of fairness for graded materials. Furthermore, the addition of multimodal features, such as audio prompts or visual cues, could increase the use of automated assessment across all communication modalities.

Ultimately, the study underscores the importance of a balanced integration of AI in educational evaluation. While many of the advantages of integrating AI into educational evaluation (e.g., increased efficiency, more consistent grading, and enhanced scalability) are apparent, the human element of teacher evaluation cannot be fully substituted by AI alone. In addition to grading students' writing, teachers interpret what their students' writing means, identify the uniqueness of each student, and provide emotional and motivational support; these areas cannot be addressed by AI (Cummins, 2021). Thus, the collaboration of both the teacher and AI—the model of the future—would likely be the most effective. Specifically, the AI would handle the routine and more technical aspects of the grading process, while the educator would ensure fairness, understandability, and pedagogical value.

In conclusion, this research demonstrates that when AI models are customised and applied in the correct context, these models can replicate the level of accuracy found in human assessment of academic writing. In

the case of *WritingGPT*, reliability, validity, and predictability were extremely high, while the results produced by *ChatGPT* indicated the inherent limitations of general-use AI for educational applications. In the future, we will see a cooperative relationship between human experts and Artificial Intelligence that will produce writing assessments that are more efficient, timely, and fair. However, we must approach this cooperative relationship with extreme caution, ensuring that automation continues to enhance the human connection to education. The goal of using AI should be to empower educators rather than replace them. To accomplish this, AI can provide greater transparency, immediate feedback, and equitable assessment processes while maintaining the inherent subjective and ethical factors that can only be derived from human judgment.

REFERENCES

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*(3), 1–30.

Attali, Y. (2008). A developmental writing scale. *ETS Research Report Series*, *2008*(1), i–56. https://doi.org/10.1002/j.2333-8504.2008.tb02105.x

Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.

Chapelle, C. A., & Voss, E. (2021). Introduction to validity argument in language testing and assessment. In *Validity argument in language testing: Case studies of validation research* (pp. 1–16). https://doi.org/10.1177/026553222092794

Chen, Y., & He, L. (2023). Comparing AI and human essay evaluation: Reliability and construct validity in second language writing assessment. *Assessing Writing*, *58*, 100767.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, *11*(3), 415–443.

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment*, *7*(1), 1–19.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2025). The evolution of automated writing evaluation: From surface features to deep learning. *Language Testing*, *39*(1), 3–25.

Cummins, J. (2021). *Rethinking the education of multilingual learners: A critical analysis of theoretical concepts*. Multilingual Matters.

Dancey, C., & Reidy, J. (2017). *Statistics without maths for psychology* (7th ed.). Pearson.

Ding, L., & Zou, D. (2024). Automated writing evaluation systems: A systematic review of Grammarly, Pigai, and Criterion with a perspective on future directions in the age of generative artificial intelligence. *Education and Information Technologies*, *29*(11), 14151–14203. https://doi.org/10.1007/s10639-023-12402-3

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage.

Geckin, V., Kızıltaş, E., & Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. *Journal of Educational Technology and Online Learning*, *6*(4), 1096–1108. https://doi.org/10.31681/jetol.1336599

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Giavarina, D. (2015). Understanding Bland–Altman analysis. *Biochemia Medica*, *25*(2), 141–151. https://doi.org/10.11613/BM.2015.015

Hyland, K. (2019). *Second language writing*. Cambridge University Press.

Hyland, K., & Jiang, F. K. (2021). *Academic discourse and global publishing: Disciplinary persuasion in changing times*. Routledge.

Hyland, K., & Zou, M. (2024). AI-supported feedback and teacher mediation in EFL writing classrooms. *System*, *122*, 103692.

Kao, C. W., & Reynolds, B. L. (2024). Timed second language writing performance: Effects of perceived teacher vs. perceived automated feedback. *Humanities and Social Sciences Communications*, *11*(1), Article 14. https://doi.org/10.1057/s41599-024-03522-3

Knoch, U. (2009). *Diagnostic assessment of writing: A comparison of two rating scales* (Vol. 28). Peter Lang.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Li, S., & Ng, V. (2024). Automated essay scoring: Recent successes and future directions. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)* (pp. 8114–8122). International Joint Conferences on Artificial Intelligence Organization.

Li, J., & Ranalli, J. (2024). Meta-analysis of AI-assisted writing assessment: Validity, bias, and pedagogical implications. *Computers & Education*, *208*, 105983.

Li, Y., Liu, Q., & Tang, C. (2024). Assessing English writing through large language models: Potential and pitfalls. *Language Assessment Quarterly*, *21*(2), 113–131.

Liu, W. (2024). A systematic review of automated writing evaluation feedback: Validity, effects, and students' engagement. *Language Teaching Research Quarterly*, *45*, 86–105. https://eric.ed.gov/?id=EJ1456678

Link, S., Ryu, M., & Hegelheimer, V. (2023). AI-mediated feedback and learner autonomy in academic writing. *ReCALL*, *35*(4), 410–429.

Lu, X. (2023). Natural language processing for automated writing assessment: Advances and challenges. *Language Learning & Technology*, *27*(2), 1–20.

Meyners, M. (2012). Equivalence tests: A review. *Food Quality and Preference*, *26*(2), 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Ortega, L. (2020). The study of heritage language development from a bilingualism and social justice perspective. *Language Learning*, *70*(Suppl. 1), 15–53.

Pilicita, A., & Barra, E. (2025). LLMs in education: Evaluating GPT and BERT models in student comment classification. *Multimodal Technologies and Interaction*, *9*(5), Article 44. https://doi.org/10.3390/mti9050044

Ranalli, J., & Link, S. (2023). Re-thinking automated feedback in academic writing: Pedagogical affordances and risks. *Journal of Second Language Writing*, *61*, 101024.

Reich, J., & Ito, M. (2023). *From good intentions to real outcomes: Equity by design in learning technologies*. MIT Press.

Rost, M. (2024). *Teaching and researching listening*. Routledge.

Sun, Z. (2024). Human–AI differences in assessing EFL writing: A dimension-based analysis. *Language Testing in Asia*, *14*(2), 45–61.

Tajik, A. (2025). Exploring the role of AI-driven dynamic writing platforms in improving EFL learners' writing skills and fostering their motivation. *Research Square*. https://doi.org/10.21203/rs.3.rs-5788599/v1

Toscu, S. (2023). Assessing writing in EFL context. *ELT Research Journal*, *12*(2), 174–192.

Weigle, S. C. (2020). *Assessing writing* (2nd ed.). Routledge.

Wiens, B. L. (2002). Choosing an equivalence limit for noninferiority or equivalence studies. *Controlled Clinical Trials*, *23*(1), 2–14. https://doi.org/10.1016/S0197-2456(01)00196-9

Yoon, H., & Lee, H. (2024). AI and fairness in second language writing assessment: A critical review. *Language Testing*, *41*(1), 65–84.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2023). Systematic review of research on artificial intelligence applications in higher education. *International Journal of Educational Technology in Higher Education*, *20*, Article 11. https://doi.org/10.1186/s41239-019-0171-0

Zhang, Y., & Hyland, F. (2024). ChatGPT as an assessor: Investigating AI evaluation of EFL student writing. *Assessing Writing*, *59*, 100781.

Ethical Declaration and Committee Approval

This study was conducted in accordance with the principles of scientific research and publication ethics. Consent forms were collected from students.

Proportion of Authors' Contribution

Author 1: Conceptualization, Methodology, Results, Data Collection, Formal Analysis.

Author 2: Conceptualization, Review & Editing, Visualization, Data Collection, Writing – Original Draft.

Author 3: Conceptualization, Review & Editing, Writing - Last Draft