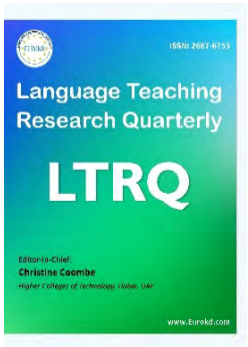


Language Teaching Research Quarterly

2024, Vol. 43, 22–42

Thematic Collection[1]

The Intersection of AI and Language Assessment: A Study on the Reliability of ChatGPT in Grading IELTS Writing Task 2

Osama Koraishi

Bursa Uludağ University, Türkiye

Received *14 February 2024* ***Accepted*** *07 August 2024*

Abstract

This study conducts a comprehensive quantitative evaluation of OpenAI's language model, ChatGPT 4, for grading Task 2 writing of the IELTS exam. The objective is to assess the alignment between ChatGPT's grading and that of official human raters. The analysis encompassed a multifaceted approach, including a comparison of means and reliability measures such as Cohen's weighted kappa and intraclass correlation. The results revealed a high agreement in means and substantial reliability between the two grading methods on the level of the majority of texts. However, individual discrepancies and outliers were also identified, underscoring the nuanced nature of grading. While ChatGPT demonstrated efficiency and general alignment with human grading, the study concludes that it should not replace human judgment, particularly due to these observed inconsistencies. The findings contribute valuable insights into the potential and limitations of AI in educational grading and emphasize the importance of a comprehensive quantitative evaluation.

Keywords: *Artificial Intelligence (AI), Artificial Intelligence in Education (AIEd), CALL, ChatGPT, IELTS, Natural Language Processing (NLP)*

How to cite this article (APA 7th Edition):

Koraishi, O. (2024). The intersection of AI and language assessment: A study on the reliability of ChatGPT in grading IELTS writing task 2. *Language Teaching Research Quarterly*, 43, 22-42. https://doi.org/10.32038/ltrq.2024.43.02

[1] This paper is part of a thematic collection (2024, 43) entitled: Artificial Intelligence and ChatGPT for Language Education: A Research Agenda.

* Corresponding author.
E-mail address: osama.koraishi@gmail.com https://doi.org/10.32038/ltrq.2024.43.02

Introduction

Artificial Intelligence (AI) has significantly impacted various sectors, including education, revolutionizing learning, curriculum design, autonomy and even assessment methods across disciplines (Talan & Kalinkara, 2023; Wang et al., 2023). AI's potential has opened new avenues, particularly in language learning and teaching, which is being increasingly explored across the globe (Fitria, 2021; Luo, 2022). The advent of automated essay scoring (AES) systems, which harness machine learning and natural language processing techniques, offers the potential for immediate feedback, reduces manual scoring burdens, and maintains scoring consistency which does away with the subjectivity and biases that human raters might have (Ludwig et al., 2021). This paper focuses on the International English Language Testing System (IELTS), a globally recognized English proficiency test, and explores the potential role of a specific AI model or LLM (Large Language Model), OpenAI's ChatGPT 4.0, in assessing IELTS Writing Task 2.

Holding a significant position in the global context, IELTS serves as a benchmark for English proficiency among non-native English speakers. It is widely accepted by universities, employers, and immigration authorities worldwide (Read, 2022). The exam is structured into four sections: Listening, Reading, Writing, and Speaking, with the Writing section, specifically Task 2 (essay writing), being considered among the most challenging parts of the test because of its academic nature.

Traditionally, IELTS assessments are conducted by human examiners from Cambridge Assessment English, i.e. the official body that grades all the IELTS exams worldwide, following a detailed rubric *(Writing Band Descriptors, 2023).* Despite successfully deploying a range of assessment techniques, teachers, including examiners, often express a lack of confidence in their knowledge of assessment (Berry et al. 2019). Moreover, this process is subject to human bias and error and can be time-consuming, raising questions about its efficiency and consistency (Shirazi, 2019).

In contrast, AI technologies, such as ChatGPT 4.0, a state-of-the-art language model developed by OpenAI, have shown potential in various educational applications, including AES (Woo et al. 2024). ChatGPT 4.0 has demonstrated remarkable capabilities in understanding and generating human-like text, raising the intriguing possibility of employing AI models for language assessment tasks, and potentially offering a more efficient and consistent grading system. It is able to generate high-quality responses to human input and can self-correct based on subsequent conversations (Qin et al. 2023). As highlighted by Kumar and Boulanger (2020), deep learning has shown promise in automated essay scoring, providing personalized, formative, and fine-grained feedback to students. In fact, the TOEFL test, provided by ETS, is also being graded by using their own AI system when it comes to the speaking section (Xi et al. 2008) and the writing section (Chen et al. 2017). However, the reliability and accuracy of AI Large Language Models in assessing language proficiency tasks, specifically the IELTS essay, remains an open question that creates a gap in the literature. This research aims to fill that gap by addressing the following question:

RQ: How reliable is ChatGPT 4.0 in assessing IELTS essays in comparison to real official raters?

Automated Grading Systems in EFL

The evolution of automated grading systems has been marked by significant advancements in technology. An early example includes the automated language proficiency test assembly system introduced at the Lackland Air Force Base Defense Language Institute English Language Center, which laid down the foundation of modern AES systems (Henning et al., 1994). These systems continued developing to reach adaptive English reading programs, indicating a place for technology in EFL education, especially among young students (Shamir & Johnson, 2012). Current applications in EFL include various methodologies and algorithms, such as corpus tools that contribute to EFL writing (Lai, 2015).

Automated grading systems offer several benefits in EFL, including improved consistency, efficiency, and the ability to handle large volumes of assessments (Palmer et al., 2002). they have evolved to become an essential part of modern language assessment. However, challenges and drawbacks such as biases, limitations and the need for alignment with human grading standards (Celik et al., 2022; Zhao & Huang, 2020) do exist.

Recent studies have further advanced the field of automated grading systems. For instance, studies by Liu (2012) and Smith et al. (2020) highlighted the significant improvements in the accuracy and reliability of automated grading systems due to advancements in natural language processing (NLP) and machine learning. Their findings indicate that modern automated grading systems can handle nuanced linguistic features more effectively introducing a new era of NLP-based systems, which is crucial for assessing EFL learners. Moreover, these systems are even more beneficial in settings where large volumes of student work need to be assessed consistently and efficiently, leaving educators with a reduced workload allowing them to focus more on more human-oriented tasks such as personalized instruction and scaffolding for their students since the reality of large classes make manual grading an impractical and time-consuming endeavor. Using NLP/Machine learning-based systems was the start of using AI in this capacity.

Despite these advancements, bias and fairness remain significant concerns in automated grading systems. A study by Devi et al. (2023) explored the ethical implications of using AI for grading, learning and teaching, emphasizing that biases in training data can lead to unfair assessments. Leaving the answer to this dilemma open for the coming years of research. It is, then, recommended to continuously monitor and update AI models to mitigate these biases and ensure fair treatment of all students.

Artificial Intelligence in Language Proficiency Education and Testing

The integration of AI models in English as a Foreign Language (EFL) education has become increasingly prominent (Jiang, 2022; Marzuki et al., 2023). AI-driven tools have been used to automatically assess the quality of learners' compositions, providing feedback on their performance and assisting in grading (Wu, 2023). Furthermore, Automated Writing Evaluation programs, which employ AI technologies such as machine learning and natural language processing, have been widely implemented in EFL writing instruction (Gayed et al., 2022). Perhaps more importantly, AI Automated Essay Scoring systems for high-stakes language proficiency tests such the TOEFL have emerged as an aid in verifying the reliability of manual scoring and reassessing essay quality when needed (Chodorow & Burstein, 2004). In fact, the

use of AI in language proficiency testing has also been linked to the need for more precise estimates of English language proficiency across various domains (Powers & Powers, 2014).

Thus, AI's role in language proficiency testing, particularly in the context of EFL, represents a dynamic and evolving field. While offering innovative solutions and enhancing the accuracy and efficiency of assessments, AI also presents challenges that must be carefully navigated. The Duolingo English Test, for instance, has revolutionized the testing landscape by integrating AI and machine learning at every step with a huge amount of research backing up its accuracy and validity (Ye, 2014; Maris, 2020). This approach allows for adaptive testing, which adjusts the difficulty of questions based on the test-taker's performance, providing a more personalized and accurate assessment of language skills (Settles & LaFlair, 2022).

On the other hand, AI-driven tools such as ChatGPT and other LLMs offer personalized learning experiences by adapting content to individual proficiency levels and learning styles. These tools generate feedback on various aspects of language proficiency, including pronunciation, grammar, and vocabulary, which helps learners improve more effectively Moreover, AI-powered systems minimize human subjectivity and provide more objective assessments. AI can analyze large datasets and identify patterns, ensuring consistent and reliable evaluations. This reduces the potential for bias and inconsistency that can occur with human graders (Austin et al., 2023).

AI's role extends to spoken language assessment through advanced speech recognition technologies that evaluate pronunciation, fluency, and intonation. Additionally, AI systems enhance test security by detecting and preventing cheating, ensuring a fair testing environment (Isbell et al., 2023).

These systems integrate various AI technologies such as natural language processing (NLP), machine learning, and pattern recognition to evaluate student responses consistently and impartially. The conceptual framework for these systems is built on several key components: data collection, analysis, feedback generation, and continuous improvement (Cuayáhuitl et al., 2019).

IELTS as a Tool for EFL Assessment

The International English Language Testing System (IELTS) is a globally recognized tool for assessing English proficiency. It serves various purposes, including assessing English language skills for academic and professional purposes and controlling immigration, though these purposes may sometimes be incompatible (Chaloff & Lemaître, 2009). It is so influential that it has been suggested as a necessary addition to English language curriculum and modern teaching methods (Al-Mously et al., 2013).

The IELTS Writing section, specifically Task 2 (Essay Writing), is a critical component of the assessment. It evaluates test-takers' ability to present ideas, support arguments, and use language effectively. This is quite clear in the rubric of Writing Task 2 provided by Cambridge Assessment (*Writing Band Descriptors*, 2023). Moreover, the grading criteria and rubrics used in IELTS not only serve as tools for evaluation but also guide curriculum design and teaching methodologies.

The role of IELTS in EFL education extends beyond mere assessment. Higher IELTS marks at entry have been found to translate into higher grade point averages (GPAs), indicating its influence on academic performance (Thorpe et al., 2017). Even with its shortcomings in

relation to ensuring the reliability, validity, fairness, grading and the broader social, economic, and political dimensions of international high-stakes English language testing (Hall, 2010), it is still one of the most popular exams in use today with a recognition of 11000 organizations (Read, 2022). Moreover, research on the washback effect of IELTS—how the test influences teaching and learning—reveals both positive and negative impacts. A study conducted in Japan found that while IELTS preparation can motivate students and provide clear language learning goals, it can also lead to anxiety and a narrow focus on test-taking strategies rather than broader language competence (Allen, 2016). This underscores the need for balanced preparation strategies that incorporate both test-specific skills and general language development. Due to the worldwide recognition of this examination and its importance in global education, there is a demand for dependable and effective automated essay-evaluation systems for assessing the writing section. These systems have the potential to replace or at least aid human scoring efforts, thereby reducing manual labor involved in the process.

Methods

Research Design

This research employs a quantitative approach for its analysis. The correlation and alignment between the grades generated by ChatGPT 4 and those given by official human raters. This is done in order to evaluate whether ChatGPT's scores are adequately reliable.

Instruments

There was no specific instrument that was used in this study, the official grades are based on the rubric provided by IELTS Assessment (*Writing Band Descriptors, 2023*). Furthermore, the grades generated by ChatGPT are based on its internal rubric, shown in Appendix A.

Participants and Data Collection

The data corpus comprises 55 writing samples taken from real examinations, thus ensuring their authenticity. These samples were sourced from three publicly accessible sources: the IELTS IDP website (*IELTS Test Preparation & Practice Materials*, n.d.), the official IELTS website (*Sample Test Questions*, n.d.), and the Cambridge IELTS Exam Practice series (Cambridge Press & Assessment, n.d.). The samples are completely anonymized and no personal or demographic information of the participants is revealed. The choice for the convenience sampling of 55 samples was dictated by the availability of public data i.e. there are no other available samples accessible to the general public. There are, however, many websites that have "sample answers" that were generated and graded by those who run the websites (Braveman, n.d.); these will not be suitable in this case. In other words, these are unofficial sources and might prove detrimental to the reliability of the study since the official grades are considered the gold standard for the investigation. Thus, only official samples were used.

To facilitate data analysis, Optical Character Recognition (OCR) technology was employed to scan the documents and extract raw texts from the aforementioned sources (Hamad & Kaya, 2016). This process enabled the extraction of the examination questions, candidates' responses, their scores, and the human raters' comments on each text, all of which were subsequently

compiled into an Excel database. The researcher revised all of the extracted text and ascertained that they are identical to the original.

Procedures

Each text was individually inputted into ChatGPT 4, preceded by the prompt: "*Act as an official IELTS examiner. I will send you a text written by a student in relation to writing Task 2. You have to give me an accurate grade according to the IELTS rubric for Task 2. The question is: [question]. The text is: [the candidate's essay]*". The bracketed words were replaced with the actual data from the database i.e. the essays written by the candidates and the questions in the test sample. ChatGPT's output, which included an analysis and an overall score, was then added to the Excel database. It is worth noting that the assessment took place in a new "chat" that did not have previous conversations to minimize hallucinations (Alkaissi et al. 2023) which are inaccuracies that ChatGPT start fabricating regardless of the data it was provided to it. In ChatGPT, opening a new "chat" is akin to wiping out the history of the conversation and starting from scratch. Also, the same prompt was always used to ensure consistency. Also, as seen in the prompt, the researcher has mentioned the rubric because ChatGPT has a version of the IELTS rubric, or at least a very similar one, in its database. When asked "*In a table, provide the rubric of the writing Task 2 of the IELTS exam*", it provided a very accurate rubric albeit less detailed than the official one. The rubric can be seen in Appendix A.

Lastly, at first glance, the corpus itself might seem to be of a limited proficiency range i.e. the vast majority of the scores are between 4 and 8. In this context of the IELTS exam, this phenomenon is not uncommon and is, in fact, representative of the broader population of test-takers since according to data provided by IELTS.org (Demographic Data, 2022), an exceedingly small proportion of individuals who participate in the IELTS exam achieve scores above 8 or below 4. This distribution has consequently shaped the composition of the sample utilized in this study, reflecting a realistic and, hopefully, representative spectrum of proficiency levels.

Data Analysis

Upon the completion of data collection, JASP was used to analyze the grades generated by human raters and those produced by ChatGPT. Three different tests were run by JASP in order to compare the two sets of scores: a Wilcoxon test which is a non-parametric equivalent to the t-test to compare the means of each data set (Rosner et al., 2005), an intra-class correlation (ICC) test to assess the agreement between human raters and ChatGPT scores as it was similarly used in other studies (Khademi, 2023), and a Rater agreement test to examine the consistency of identical scores among the two sets of scores (Mancar & Gülleroğlu, 2022), specifically Cohen's Kappa which is used to measure rater reliability that takes chance into consideration. (McHugh, 2012).

Finally, The Noteable platform was employed in this study to enhance the analytical process, serving as a tool for data visualization and representation. It uses Python in order to calculate and map out charts accurately (Embarak, 2018). It was used to generate charts of comparative data. Moreover, it is crucial to acknowledge that, according to the standard practice adopted by Cambridge Assessment Services, student scores are conventionally rounded up to the nearest half (*Understanding and Explaining IELTS Scores*, n.d.) i.e. if a

student's score is (7.25), it is rounded to (7.5). However, in the context of this study, ChatGPT recommended to round up at times and round down at others. If the scores were always rounded up, as the standard practice suggests, it could potentially skew the results and misrepresent the accuracy of ChatGPT's scoring. Thus, the researcher adhered to the rounding recommendations proposed by ChatGPT, whether upward or downward, to ensure a more precise evaluation and representation of its performance.

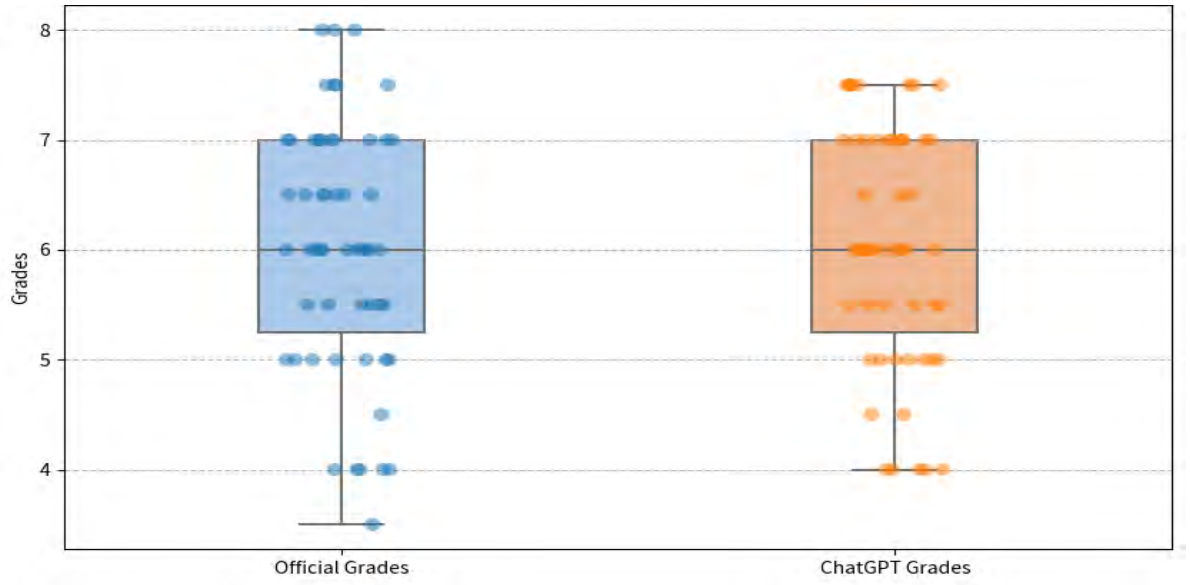
Results

The grades at hand are not normally distributed as the P-value was (0.016) on the Shapiro-Wilk test of normality (Ghasemi & Zahediasl, 2012). Thus, all the performed tests are non-parametric tests. The comparison of the grades assigned by ChatGPT and the official IELTS assessors revealed an, unusually, perfect alignment in the mean, with both ChatGPT and the official assessors assigning a mean grade of (6.027) and a standard deviation of (1.136) for the Official grades compared to (1.087) for ChatGPT grades. This also resulted in an incredibly high P-value of (0.91) in the Wilcoxon test comparing the two set of grades. The second test, the Intraclass Correlation Coefficient (ICC) test, offers a more nuanced assessment by quantifying the consistency or agreement between two sets of grades (Shrout & Fleiss, 1979). Specifically, it measures how strongly the grades assigned by ChatGPT resemble the grades assigned by official human assessors across the same subjects/texts. A high ICC value would indicate strong agreement (Koo & Li, 2016).

The Intraclass Correlation Coefficient (ICC) test conducted on the grades assigned by ChatGPT and the official human assessors yielded a point estimate of (0.814), with a 95% confidence interval ranging from (0.702 to 0.887). This ICC value suggests a high degree of agreement between the two sets of grades, suggesting that ChatGPT's grading aligns closely with that of human raters at the individual level. The confidence interval further reinforces this finding, as the entire range falls within the bounds that signify strong agreement. This can be seen clearer in Figure 1 which complements this statistical analysis by providing a visual representation of both the distribution and individual grades. The individual scatter points reveal the alignment between each pair of grades, while the boxplot captures the overall distribution, including the mean, quartiles, and potential outliers. Together, the ICC value and the scatter boxplot paint a comprehensive picture of the general consistency between ChatGPT's grading and human grading.

Figure 1

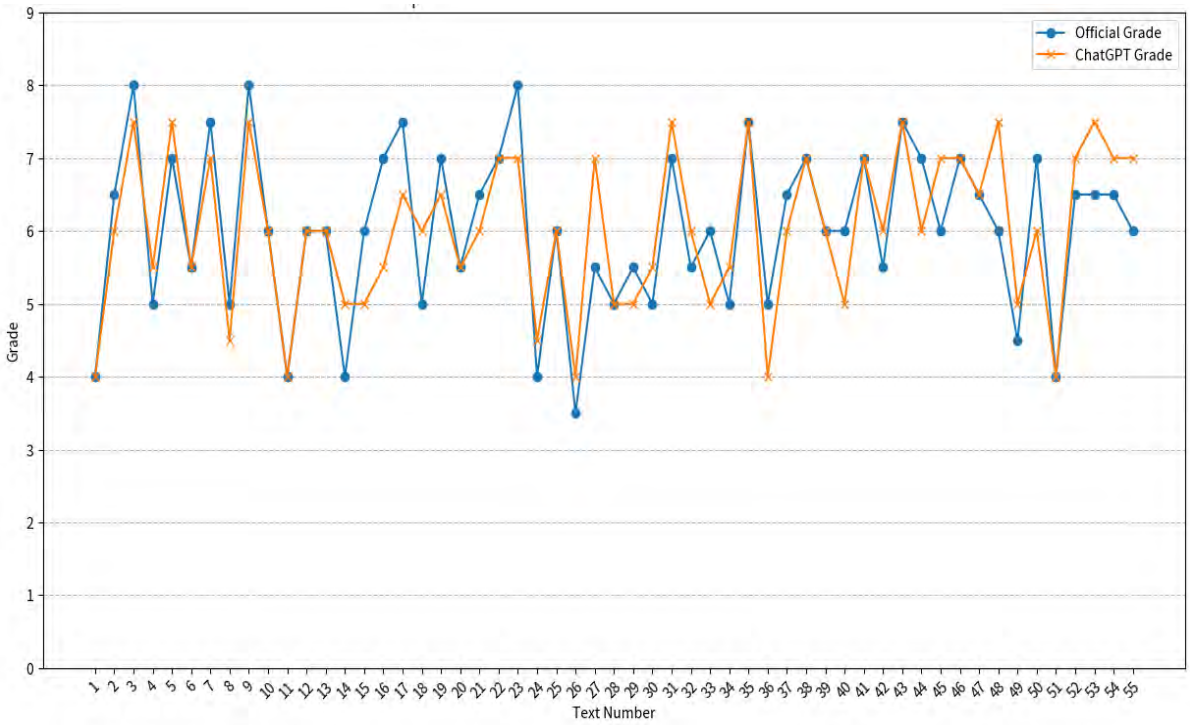
Scatterplot and Boxplot Comparison of Official Grades and ChatGPT Grades



However, upon further examination, some discrepancies can be seen. If we further inspect them in line graph that compares the two sets of scores, it becomes evident that although overall the scoring is quite similar as seen in previous tests, certain individual assessments exhibit substantial differences. Some scores such as scores of texts:14, 16, 23, 27... manifest marked deviation, as illustrated in Figure 2. A list of all the scores of the corpus can be found in Appendix B for further inspection.

Figure 2

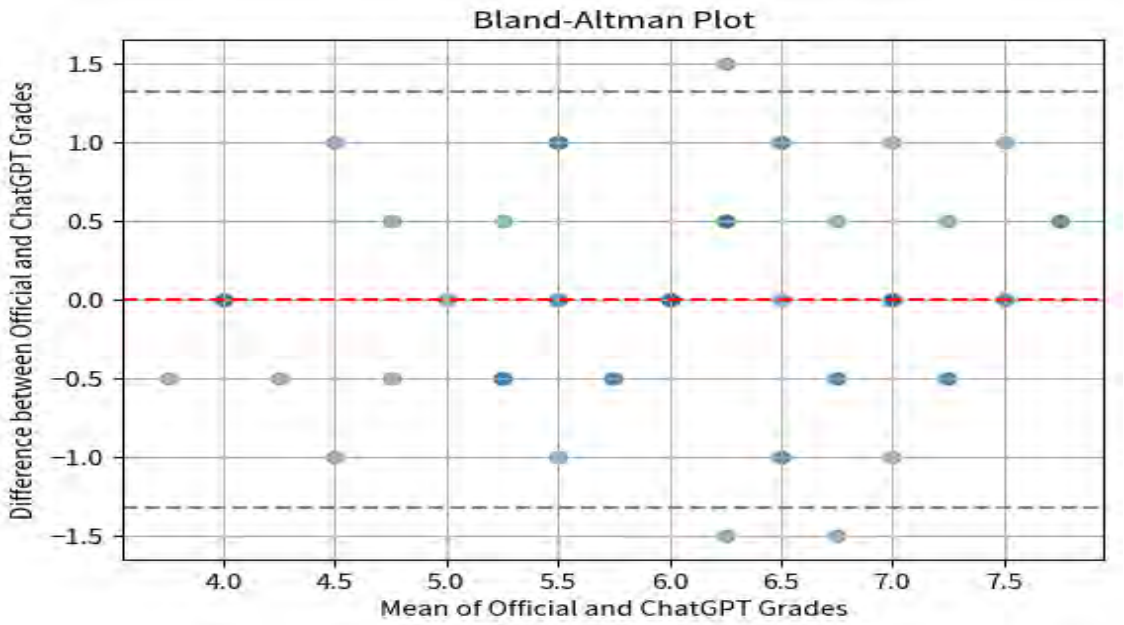
Comparison of Official Grades and ChatGPT Grades for each Text



Therefore, it is important to proceed with a Rater Agreement test (Mancar & Gülleroğlu, 2022). This statistical approach allows us to quantify the degree of concordance between the individual grades assigned by ChatGPT and the official human assessors. By examining both the overall alignment and the specific agreements and disagreements at the individual level, the Rater Agreement test provides a comprehensive assessment of the reliability of ChatGPT's grading in comparison to human judgment. The resulting weighted kappa value of (0.811) indicates a strong agreement between the two sets of grades, which is further supported by the (95%) confidence interval ranging from (0.726) to (0.896). This interval falls within the range typically considered substantial agreement according to McHugh (2012). The Bland-Altman plot presented in Figure 3 offers a graphical analysis of the agreement between the "Official Grades" and "ChatGPT Grades" (Giavarina, 2015). On the x-axis, the mean of the two grades for each subject is plotted, while the y-axis represents the corresponding difference between the two grades. The red line indicates the mean difference, serving as a measure of systematic bias between the two grading methods. The grey dashed lines clarify the limits of agreement, calculated as the mean difference (± 1.96) times the standard deviation of the differences, encompassing the range within which (95%) of the differences are expected to fall.

Figure 3

Bland-Altman Plot Assessing Agreement between Official Grades and ChatGPT Grades



These outliers represent instances where the difference between the "Official Grades" and "ChatGPT Grades" is significantly larger than the average difference observed across the dataset.

Discussion

The aim of this study was evaluating the efficacy of OpenAI's language model, ChatGPT 4, in grading Task 2 writing of the IELTS exam, and compare its performance to the official human raters. The results of the analysis provide compelling evidence of the potential of AI Large Language Models in writing assessment. The findings of this study suggest that OpenAI's

ChatGPT 4 exhibits a significant degree of alignment with human raters when grading Task 2 of the IELTS exam. Notably, the identical mean grades assigned by both ChatGPT and human assessors indicate a potential for AI models in educational assessment. However, while the statistical similarity is promising, it is crucial to delve beyond aggregate measures to understand the efficacy of AI grading fully and examine it under further scrutiny on the individual grade level.

While this identical mean suggests that ChatGPT's grading aligns perfectly with that of human raters, it is essential to recognize that this analysis is likely to be misleading. This could be a mere coincidence. Focusing solely on the means does not account for potential differences between individual scores, and further analysis would be required to fully understand the consistency and alignment between ChatGPT's grading and that of human raters. To that end, an Intraclass Correlation test was introduced to the analysis. This specific test examined not just when raters agree, but how close their ratings are within the context of the variability present in the scores. It revealed that when individual scores are considered, significant discrepancies emerge. This highlights the importance of taking different approaches to assessing AI grading systems and not be content with only one promising evaluation. Furthermore, the high value of the Intraclass Correlation Coefficient (ICC) indeed indicates high agreement between ChatGPT and human ratings which suggests that, at a macro level, ChatGPT could reliably approximate human grading practices. This analysis is vital in establishing the reliability and potential applicability of ChatGPT in educational assessment.

However, the presence of outliers in the data- as shown in the line graph and the Bland-Altman plot- raises questions about the consistency of ChatGPT's assessment. They show that though ChatGPT may perform well on average, its application might not yet be suitable for precise and individualized grading where it is critical, such as in the case of the IELTS exam. Moreover, the selection of Cohen's weighted kappa as an additional measure of agreement was guided by the specific characteristics of the grading data. Unlike simple measures of agreement that consider only exact matches, the weighted kappa accounts for the ordinal nature of the grades and allows for partial agreement when the ratings are close but not exactly the same. This is particularly pertinent in the context of grading, where differences between adjacent grades may not be equally significant across the scale. The application of Cohen's weighted Kappa indicated that partial agreements between the independent grades of the same text are accounted for, and similar to the ICC test value, the Weighted Cohen Kappa still does not show perfect alignment with the human raters because of these discrepancies.

However, it is important to note that even the official grades of the IELTS writing assessment do not have a perfect alignment, and indeed has been criticized before about the inter-rater reliability of the official assessment (Veerappan & Sulaiman, 2012). That said, it is fair to say that although ChatGPT's reliability coefficient of (0.811) that this study shows is not perfect, it is competing with the official reliability coefficient of (0.92) according to their *Test Statistics* page (2022), and not a perfect alignment.

ChatGPT has been examined in general writing assessment and was beneficial as a tool to help leaners and provide instant feedback (Parker et al., 2023), but as can see in this study, when it comes to IELTS, a high-stakes exam, there were outliers in the numerical value of ChatGPT's assessments, so when it comes to the outliers, they are not mere statistical anomalies. They underscore the importance of a multifaceted evaluation approach, recognizing

that overall measures of agreement may mask individual discrepancies. The combination of high reliability but variable agreement suggests that ChatGPT's grading aligns closely with human grading on average but may differ drastically in specific cases. This has implications for how ChatGPT might be used in educational assessment, indicating that it may be suitable for general grading tasks but might require human oversight or additional validation for critical assessments.

The application of AI in educational contexts presents significant implications. For example, the ability of AI systems to provide consistent and immediate feedback can reduce the workload on teachers and potentially enhance learning outcomes by offering timely support (Ming, 2005). This also provides an equitable access to such high quality "tutor" since numerous students might not have the financial means of hiring an IELTS tutor, somewhat alleviating the gap in access to information and education (Blanden et al., 2023). Of course, as aforementioned, these scores might not be completely reliable, but they can serve as preliminary scores that students can receive to approximate the scores in the real test. Furthermore, though recently some other services such as Cambridge's Write and Improve (Karpova, 2020) which added an IELTS section recently and Upscore AI (Yurik BV, n.d.) do use AI for writing assessment, ChatGPT is iterative and interactive which gives it an advantage over them. In other words, instead of just providing the score, it can in fact provide qualitative feedback pointing out mistakes and how to improve on them especially that, as we have already established, it is well-aware of the IELTS rubric within its dataset. This evident is other studies as well, for example González et al (2021) state that AI works best with formative assessment i.e. iterative assessment to improve the students' output throughout the learning process. This also aligns with what Messer et al. (2024) found sating that automated grading tools enhance learning by offering instant feedback and supporting multiple resubmissions, increasing student satisfaction.

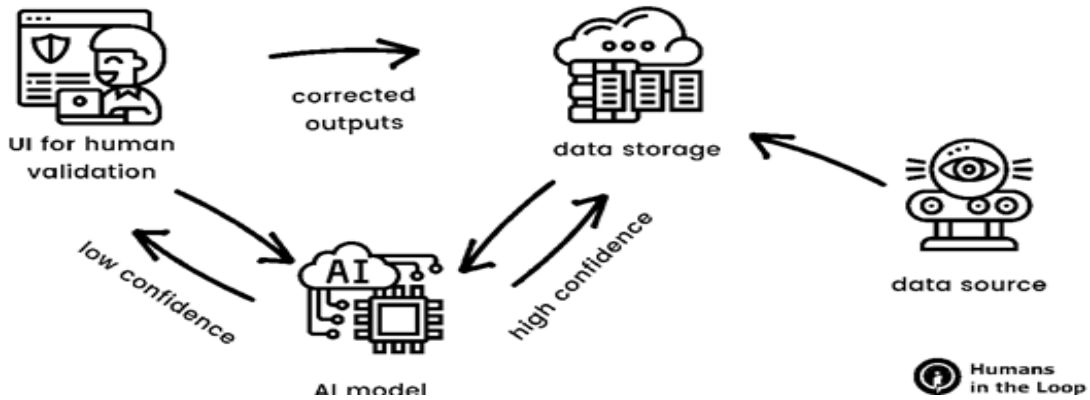
What is more is that the learner can debate his/her mistakes, ask for clarification, explanations, examples and even a new re-written essay based on their essay that is predicted to score a better band in the IELTS, so ChatGPT's interactive nature (Lo, 2023) and its ability for text manipulations (Koraishi, 2023) is crucial in this regard. This can be further exploited after the introduction ChatGPT4 Omni (OpenAI, 2024) which can analyze visual and audio inputs making it possible for students to show their work to ChatGPT and have a conversation about improving it without resorting to even type their essay making it a more natural way of communication.

Moreover, concerns about bias and fairness in AI grading must be addressed. Studies have shown that AI models can inadvertently perpetuate biases present in their training data that were primarily consisted of "native English", which can be problematic in a diverse educational environment. For example, AI-generated detectors are infamous to be biased against non-native users of English (De Zwart, 2024) who are, of course, the main target of proficiency exams such as the IELTS. Ensuring that AI grading systems are fair and unbiased requires continuous monitoring and improvement of the models which further highlights the need of a human-in-the-loop system (Chen et al., 2023) which is a system that always have a human component(s) to ensure oversight over the automated system were these automations to become more mainstream. In a human-in-the-loop system, the AI model will provide a confidence rate of its output. If the confidence rate is high enough, then it is deemed credible, but if it is low,

then the output [grade] would be reviewed by a human expert. The process is illustrated by the diagram in Figure 4 from (HITL Team, 2021).

Figure 4

How HITL Works



Although this framework was originally designed for general purposes, the author sees that it can apply in educational assessment since it has already been somewhat implemented in various areas in education (Memarian & Doleck, 2024). In our case, the outliers would probably be among the low confidence output that the AI model would give, and the human review would mitigate this discrepancy. From another point of view, demanding more explanation of the process that the AI had to extrapolate its evaluation is also a valid way of maintaining accuracy. For example, explainable AI is an evolving field that makes it possible for humans to evaluate machine learning models for their correctness, fairness, and reliability (Zahoor et al., 2024).

The reception of AI grading by students and educators is another important factor. Research indicates that while students might appreciate the immediacy of AI feedback, there are concerns about the accuracy and fairness of such systems (Zakaria & Ningrum, 2023). Also, some studies warned about the potential over-reliance on these AI tools which can lead to less social learning and the loss of the social skills that students gain through interactions with the rest of the class as well as their teachers (Hurst et al., 2013). Teachers, on the other hand, may view AI as a tool that can assist in grading but are anxious about the potential of being fully replaced by machine assessment as what recently happened in Texas (Weatherbed, 2024; Marion & Cisneros, 2023) where many teachers were relieved of their jobs because an AI assessment system replaced them in assessing the STAAR tests. The reliability of such systems is also source of worry for educators since it is still difficult to validate whether the system is working or not once implemented. In fact, the very same case regarding Texas State faced a severe backlash because of the accuracy of their implemented system where only 25% of the tests will be graded by human teachers. This, of course, seems to be a continuing trend as many other educational institutions and governmental bodies are leaning towards AI for assessment (Klein, 2024; Glanville, 2023) for budgetary concerns among others though they might not be completely ready for implementation (Merod, 2023).

Data privacy presents another significant point of contention in the use of AI-based grading systems especially if they were based on LLMs. Currently, IELTS candidates share their

information and written outputs exclusively with Cambridge Assessment and/or whoever is training them to do the exam. However, when student data is entered into ChatGPT, there is uncertainty regarding how OpenAI which is a for-profit company might utilize such data. Although OpenAI has previously claimed that ChatGPT is not trained on user inputs, there have been instances where this claim has been questioned, particularly given the proprietary nature of ChatGPT's codebase and it being a close-source technology. Consequently, if such systems were to be implemented in schools, institutions, and even teachers might need to exercise caution and obtain explicit consent from students before their essays are shared with OpenAI. Even if educational institutions use their own systems but still utilizing OpenAI's API, which is a sort of key that enables a different system to utilize the capabilities of a service [ChatGPT in this case] in the background (MuleSoft Videos, 2015), the same problem still persists. Another option could be the use of "Silo" AIs, such as GPT-4ALL, which is similar to ChatGPT but trained on specific, localized data and operates entirely on local machines. While this approach may entail higher initial costs, it would significantly mitigate privacy concerns regarding student data.

Future research should focus on longitudinal studies to assess the long-term impact of AI grading on student performance and explore the integration of AI with other educational technologies (Shrungare, 2022), perhaps leading to a fully automated IELTS that mirrors the TOEFL IBT in that regard. Additionally, investigating the use of AI in different cultural and educational contexts can provide a more comprehensive understanding of its applicability and effectiveness (Ndukwe et al, 2019). Finally, is important to note that ChatGPT is a general-purpose Large Language Model i.e. it was trained on various kinds of datasets. Other LLMs such as Khamingo (Khan, 2023) which was trained for educational purposes would fare much better in their context. Thus, GPT-based LLMs might be even more accurate and reliable for IELTS assessment if they were trained on specific datasets in relation to IELTS. Recently, there has been some advances in this respect since GPTs were released (OpenAI, 2022) which are mini ChatGPT models that are linked to specific services or trained on specific data that makes them efficient in a narrow spectrum of application instead of being general-purpose such as ConsencusGPT (Consensus, 2024) which tabs into many scholarly databases to provide sources, references, summaries and answers for research purposes and World of Words GPT (King, 2023) which helps the user explore the vocabulary of the English language. The narrowness of such GPTs might make it much more accurate and reliable as aforementioned. Thus, future teachers/researchers might consider, at the very least, a specialized GPT for this purpose.

Conclusion

This study represents an attempt to add robust empirical evidence supporting the efficacy of OpenAI's language model, ChatGPT, in assessing IELTS writing tasks. In the evaluation of OpenAI's language model, ChatGPT 4, for grading Task 2 writing of the IELTS exam, a multifaceted analysis was conducted to assess its alignment with official human raters. ChatGPT demonstrated a significant statistical agreement with human raters. While the central tendencies closely align, suggesting similar overall grading patterns, observed outliers and some inconsistencies indicate the need for careful consideration in individual assessments which is why the study still advocates for human oversight in high-stakes evaluations. These

findings lead to important implications for the application of ChatGPT in educational assessment. While not perfect, ChatGPT emerges as a relatively good, fast, and efficient way to grade the IELTS writing task 2. Its alignment, though not perfect, with human grading, supports its potential utility. However, an IELTS expert's oversight of what ChatGPT produces remains essential to ensure the nuanced evaluation that human judgment provides. Furthermore, with such degree of reliability, ChatGPT and similar AI models could potentially serve as publicly available automated assessment tools, providing relatively consistent grading aligned with official rubrics. Also, the AI's continuous availability promises the advantage of providing immediate and detailed feedback to students, enhancing their learning experience by allowing instant review and improvement.

In conclusion, the current analysis suggests that ChatGPT should not be implemented as an official rater, at least not yet. The complexities of human grading and the individual variations observed in the analysis warrant a more cautious approach. Nevertheless, ChatGPT can be very useful in streamlining the grading process reducing the workload on human raters and increasing grading efficiency, thus exemplifying the potential for resource saving in the educational assessment industry. By combining the efficiency of AI with the nuanced understanding of human experts in a HIT framework, ChatGPT offers a promising avenue for enhancing educational assessment practices, balancing speed and accuracy with the critical oversight that ensures quality and fairness. In light of these findings and implications, it is suggested that future research could explore the application of AI models in other domains of educational assessment, investigating the limits of their capabilities. This could also include investigating and crafting strategies for their effective integration into the assessment process. Better yet, experts, in collaboration with Cambridge Assessment, might train specific new LLMs that have the sole purpose of assessment which, in turn, would increase its accuracy many folds. Moreover, the scope of this paper investigated OpenAI’s ChatGPT which leaves other modern Large Language Models relatively unexplored in the area of assessment. Future researchers could consider replicating the study with Google’s Bard or any other LLMs that have become more commonplace in recently.

While this study provides valuable insights into the potential of AI, specifically ChatGPT, in grading IELTS Writing Task 2, it is not without limitations. Firstly, the sample size of 55 essays, although dictated by the availability of public data, is relatively small. A larger sample size could provide a more comprehensive understanding of ChatGPT's grading capabilities and potentially reveal patterns or inconsistencies not apparent in a smaller dataset. Secondly, the study only focuses on IELTS Writing Task 2, limiting the scope of the research. The performance of ChatGPT in grading other tasks, such as Writing Task 1 or the Speaking section, remain unexplored. Finally, the AI-generated data were generated by the version of ChatGPT 4 available in November, 2023. Because of the rapid and iterative development of ChatGPT, there is no guarantee that future models would result in the exact same numbers. Future research could address these limitations by expanding the sample size, exploring ChatGPT's performance in other IELTS tasks and examining the ethical considerations of using AI in educational assessment.

ORCID

https://orcid.org/0009-0008-1670-3436

Acknowledgements
Not applicable.
Funding
Not applicable.
Ethics Declarations
Competing Interests
No, there are no conflicting interests.
Rights and Permissions
Open Access

References

Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus 15*(2), Article e35179. https://doi.org/10.7759/cureus.35179

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia*, *6*(1), Article 7. https://doi.org/10.1186/s40468-016-0030-z

Al-Mously, N., Salem, R. O., & Al-Hamdan, N. (2013). The impact of gender and English language on the academic performance of students: An experience from new Saudi medical school. *Journal of Contemporary Medical Education*, *1*(3), 170-176. https://doi.org/10.5455/jcme.20130226121358

Austin, T., Rawal, B., Diehl, A., & Cosme, J. (2023). AI for equity: Unpacking potential human bias in decision making in higher education. *AI, Computer Science and Robotics Technology, 2023*(2), 1-17. https://doi.org/10.5772/acrt.20

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal*, *73*(2), 113–123. https://doi.org/10.1093/elt/ccy055

Blanden, J., Doepke, M., & Stuhler, J. (2023). Educational inequality. In *Handbook of the Economics of Education* (Vol. 6, pp. 405-497). Elsevier. https://doi.org/10.1016/bs.hesedu.2022.11.003

Braveman, S. (n.d.) IELTS essay, topic: Computers instead of teachers. *IELTS-Blog*. https://www.ielts-blog.com/ielts-writing-samples/ielts-essays-band-8/ielts-essay-topic-computers-instead-of-teachers/

Cambridge Press & Assessment. (n.d.). *IELTS 1-17*. Cambridge University Press. https://www.cambridge.org/gb/cambridgeenglish/catalog/cambridge-english-exams-ielts/ielts

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends*, *66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chaloff, J., & Lemaître, G. (2009). *Managing highly-skilled labour migration: A comparative analysis of migration policies and challenges in OECD countries*. OECD Social, Employment and Migration Working Papers (79). OECD Publishing. https://doi.org/10.1787/225505346577

Chelaine, M & Cisneros, J. (2023, December 1). *STAAR RLA update, assessment development division* [Conference presentation]. Texas Assessment Conference, Texas, United States of America. https://tea.texas.gov/student-assessment/testing/student-assessment-overview/2023-tac-staar-rla-update.pdf

Chen, J., Zhang, M., & Bejar, I. I. (2017). An investigation of the *e-rater* automated scoring engine's grammar, usage, mechanics, and style microfeatures and their aggregation model. *ETS Research Report Series*, *2017*(1), 1–14. https://doi.org/10.1002/ets2.12131

Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "Human-in-the-Loop" system. *Systems*, *11*(11), Article 548. https://doi.org/10.3390/systems11110548

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays. *ETS Research Report Series. 2004*(1), i-38. https://doi.org/10.1002/j.2333-8504.2004.tb01931.x

Consensus. (2024, January 10). Introducing: Consensus GPT, your AI research assistant. *The Consensus blog*. https://consensus.app/home/blog/introducing-researchgpt-by-consensus/

Cuayáhuitl, H., Lee, D., Ryu, S., Cho, Y., Choi, S., Indurthi, S. R., Yu, S., Choi, H., Hwang, I., & Kim, J. (2019). Ensemble-based deep reinforcement learning for chatbots. *Neurocomputing*, *366*, 118–130. https://doi.org/10.1016/j.neucom.2019.08.007

De Zwart, H. (2024, March 4). *Racist technology in action: ChatGPT detectors are biased against non-native English writers.* Racism and Technology Center. https://racismandtechnology.center/2024/03/04/racist-technology-in-action-chatgpt-detectors-are-biased-against-non-native-english-writers/

Demographic data. (2022). IELTS.org. Retrieved August 19, 2024, from https://www.ielts.org/for-researchers/test-statistics/demographic-data

Devi, S., Boruah, A. S., Nirban, S., Nimavat, D., & Bajaj, K. K. (2023). Ethical considerations in using artificial intelligence to improve teaching and learning. *Tuijin Jishu/Journal of Propulsion Technology*, *44*(4), 1031–1038. https://doi.org/10.52783/tjjpt.v44.i4.966

Embarak, O. (2018). Data gathering and cleaning. *Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems (pp. 205-241)*. Apress. https://doi.org/10.1007/978-1-4842-4109-7

Fitria, T. N. (2021). The use of technology based on artificial intelligence in English teaching and learning. *ELT Echo : The Journal of English Language Teaching in Foreign Language Context*, *6*(2), 213-223. https://doi.org/10.24235/eltecho.v6i2.9299

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers & Education: Artificial Intelligence*, *3*, Article 100055. https://doi.org/10.1016/j.caeai.2022.100055

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

Glanville, M. (2023, February). Artificial intelligence in IB assessment and education: a crisis or an opportunity? *The IB Community Blog*. https://blogs.ibo.org/2023/02/27/artificial-intelligence-ai-in-ib-assessment-and-education-a-crisis-or-an-opportunity/

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), Article 5467. https://doi.org/10.3390/app11125467

Hall, G. (2010). International English language testing: A critical response. *ELT Journal*, *64*(3), 321-328. https://doi.org/10.1093/elt/ccp054

Hamad, K. A., & Kaya, M. (2016). A detailed analysis of optical character recognition technology. *International Journal of Applied Mathematics, Electronics and Computers*, *4*(Special Issue-1), 244-249. https://doi.org/10.18100/ijamec.270374

Henning, G., Johnson, P. J., Boutin, A. J., & Rice, H. R. (1994). Automated assembly of pre-equated language proficiency tests. *Language Testing*, *11*(1), 15–28. https://doi.org/10.1177/026553229401100103

HITL Team. (2021, March 24). What is a human in the loop? *Humans in the Loop Blog*. https://humansintheloop.org/what-is-a-human-in-the-loop/

Hurst, B., Wallace, R. R., & Nixon, S. B. (2013). The impact of social interaction on student learning. *Reading Horizons*, *52*(4), 375–398. https://bearworks.missouristate.edu/articles-coe/23/

IELTS Band 8 essay Samples. (n.d.). IELTS Buddy. https://www.ieltsbuddy.com/ielts-band-8-essay-samples.html

IELTS scoring in detail. (n.d.). IELTS.org. https://www.ielts.org/for-organisations/ielts-scoring-in-detail

IELTS Test Preparation & Practice Materials (n.d.). IDP IELTS Australia. https://ielts.com.au/australia/prepare/ielts-test-preparation-material/writing-and-ielts-practice-tests

IELTS Writing band descriptors. (n.d.). IELTS.org. https://www.ielts.org/-/media/pdfs/ielts-writing-band-descriptors.ashx

Isbell, D. R., Crowther, D., & Nishizawa, H. (2023). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English test to the university. *Language Testing*, *41*(2), 233–262. https://doi.org/10.1177/02655322231165984

Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Frontiers in Psychology*, *13,* Article 1049401. https://doi.org/10.3389/fpsyg.2022.1049401

Karpova, K. (2020). Integration of "Write and Improve" AWE tool into EFL at higher educational establishment: Case study. *Celtic*, *7*(2), 137–150. https://doi.org/10.22219/celtic.v7i2.14036

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning & Teaching,* *6*(1), 75-80. https://doi.org/10.37074/jalt.2023.6.1.28

Khan, S. [Khan Academy]. (2023, March 14). Khan Academy announces GPT-4 powered learning guide [Video]. YouTube. https://www.youtube.com/watch?v=yEgHrxvLsz0

King, B. (2023, November 10). *My first GPT: The World of Words.* LinkedIn. https://www.linkedin.com/posts/byronkingjhb_chatgpt-a-world-of-words-activity-7128659394078089216-EKxU/

Klein, A. (2024, March 27). AI may be coming for standardized testing. *Education Week*. https://www.edweek.org/teaching-learning/ai-may-be-coming-for-standardized-testing/2024/03

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

Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education & Technology (LET Journal)*, *3*(1), 55–72. https://langedutech.com/letjournal/index.php/let/article/view/48

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, *5,* Article 572367. https://doi.org/10.3389/feduc.2020.572367

Lai, S. (2015). EFL students' perceptions of corpus-tools as writing references. *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 356-361). Research-publishing.net. https://doi.org/10.14705/rpnet.2015.000355

Liu, H. H. (2012). The application of natural language processing and automated scoring in second language assessment. *Studies in Applied Linguistics and TESOL*, *12*(2), 38–40. https://doi.org/10.7916/d8hx1r9h

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Ludwig, S., Mayer, C., Hansen, C. L., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, *3*(4), 897–915. https://doi.org/10.3390/psych3040056

Luo, X. (2022). Practice of artificial intelligence and virtual reality technology in college English dialogue scene simulation. *Wireless Communications and Mobile Computing*, *2022*, Article 4922675. https://doi.org/10.1155/2022/4922675

Mancar, S. A., & Gülleroğlu, H. D. (2022). Comparison of inter-rater reliability techniques in performance-based assessment. *International Journal of Assessment Tools in Education*, *9*(2), 515–533. https://doi.org/10.21449/ijate.993805

Maris, G. (2020). *The Duolingo English test: Psychometric considerations*. Technical Report DRR-20-02, Duolingo. https://doi.org/10.46999/mfkw9830

Marzuki, N., Widiati, U., Rusdin, D., Darwin, D., & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, *10*(2), Article 2236469. https://doi.org/10.1080/2331186x.2023.2236469

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Memarian, B., & Doleck, T. (2024). Human-in-the-loop in artificial intelligence in education: A review and entity-relationship (ER) analysis. *Computers in Human Behavior. Artificial Humans*, *2*(1), Article 100053. https://doi.org/10.1016/j.chbah.2024.100053

Merod, A. (2023, November 1). Just 2 states have released guidance on artificial intelligence for schools. *K-12 Dive*. https://www.k12dive.com/news/state-ai-guidance-schools-crpe/698402/

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM transactions on Computing Education*, *24*(1), 1–43. https://doi.org/10.1145/3636515

Ming, N. L. S. (2005). Reduction of teacher workload in a formative assessment environment through use of online technology. In *2005 6th International Conference on Information Technology Based Higher Education and Training* (pp. F4A-18). IEEE. https://doi.org/10.1109/ithet.2005.1560302

MuleSoft Videos. (2015, June 19). *What is an API?* [Video]. YouTube. https://www.youtube.com/watch?v=s7wmiS2mSXY

Ndukwe, I. G., Daniel, B. K., & Amadi, C. E. (2019). A machine learning grading system using chatbots In *Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part II 20* (pp. 365-368). Springer International Publishing. https://doi.org/10.1007/978-3-030-23207-8

OpenAI. (2022, November). *Introducing GPTs*. Openai.com. https://openai.com/index/introducing-gpts/

OpenAI. (2024, May). *Hello GPT-4o*. Openai.com. https://openai.com/index/hello-gpt-4o/

Palmer, J., Williams, R. E., & Dreher, H. (2002). Automated essay grading system applied to a first year university subject - How can we do it better? In *proceedings of IS2002 Informing Science and IT Education Conference* (pp. 1221-1229). Informing Science Institute. https://doi.org/10.28945/2553

Parker, J., Becker, K. D., & Carroca, C. (2023). ChatGPT for automated writing evaluation in scholarly writing instruction. *Journal of Nursing Education*, *62*(12), 721–727. https://doi.org/10.3928/01484834-20231006-02

Powers, D. E., & Powers, A. F. (2014). The incremental contribution of TOEIC® listening, reading, speaking, and writing tests to predicting performance on real-life English language tasks. *Language Testing*, *32*(2), 151–167. https://doi.org/10.1177/0265532214551855

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 1339–1384). Association for Computational Linguistics. https://doi.org/ 10.18653/v1/2023.emnlp-main.85

Read, J. (2022). Test Review: The international English language testing system (IELTS). *Language Testing*, *39*(4), 679–694. https://doi.org/10.1177/02655322221086211

Rosner, B., Glynn, R. J., & Lee, M. T. (2005). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, *62*(1), 185–192. https://doi.org/10.1111/j.1541-0420.2005.00389.x

Sample test questions. (n.d.). IELTS.org. https://ielts.org/take-a-test/preparation-resources/sample-test-questions

Settles, B., & LaFlair, G. T. (2020, April 30). The Duolingo English test: AI-driven language assessment. *Duolingo Blog*. https://blog.duolingo.com/the-duolingo-english-test-ai-driven-language-assessment/

Shamir, H., & Johnson, E. (2012). The effectiveness of computer-based EFL instruction among primary school students in Israel. *Educational Media International*, *49*(1), 49–61. https://doi.org/10.1080/09523987.2012.662624

Shirazi, M. A. (2019). For a greater good: bias analysis in writing assessment. *SAGE Open*, *9*(1), Article 215824401882237. https://doi.org/10.1177/2158244018822377

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Shrungare, J. (2023). AI in education. *Crossroads*, *29*(3), 63–65. https://doi.org/10.1145/3589657

Smith, G. G., Haworth, R., & Žitnik, S. (2020). Computer science meets education: Natural language processing for automatic grading of open-ended questions in ebooks. *Journal of Educational Computing Research*, *58*(7), 1227–1255. https://doi.org/10.1177/0735633120927486

Talan, T., & Kalinkara, Y. (2023). The role of artificial intelligence in higher education: ChatGPT assessment for anatomy course. *International Journal of Management Information Systems and Computer Science*, *7*(1), 33–40. https://doi.org/10.33461/uybisbbd.1244777

Test statistics. (2022). IELTS.org. https://ielts.org/researchers/our-research/test-statistics

Thorpe, A., Snell, M., Davey-Evans, S., & Talman, R. (2016). Improving the academic performance of non-native English-speaking students: The contribution of pre-sessional English language programmes. *Higher Education Quarterly*, *71*(1), 5–32. https://doi.org/10.1111/hequ.12109

Understanding and explaining IELTS scores (n.d.). British Council. https://takeielts.britishcouncil.org/teach-ielts/test-information/ielts-scores-explained

Veerappan, V., & Sulaiman, T. (2012). A review on IELTS writing test, its test results and inter-rater reliability. *Theory and Practice in Language Studies*, *2*(1), 138-143. https://doi.org/10.4304/tpls.2.1.138-143

Wang, X., He, X., Wei, J., Liu, J., Li, Y., & Liu, X. (2023). Application of artificial intelligence to the public health education. *Frontiers in Public Health*, *10*, Article 1087174. https://doi.org/10.3389/fpubh.2022.1087174

Weatherbed, J. (2024, April 10). Texas is replacing thousands of human exam graders with AI. *The Verge*. https://www.theverge.com/2024/4/10/24126206/texas-staar-exam-graders-ai-automated-scoring-engine

Woo, D. J., Susanto, H., Yeung, C. H., Guo, K., & Fung, A. K. Y. (2024). Exploring AI-generated text in student writing: How does AI help? *Language Learning & Technology*, 28(2), 183–209. https://hdl.handle.net/10125/73577

Writing Band Descriptors. (2023, May). British Council. https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). Automated scoring of spontaneous speech using Speechrater V1.0. *ETS Research Report Series*, *2008*(2), i–102. https://doi.org/10.1002/j.2333-8504.2008.tb02148.x

Ye, F. (2014). *Validity, reliability, and concordance of the Duolingo English Test*. Technical report, University of Pittsburgh, May 2014. https://doi.org/10.46999/eixn6416

Yurik BV (n.d.) *Upscore* [online software]. https://upscore.ai/trainer

Zahoor, K., Bawany, N. Z., & Qamar, T. (2024). Evaluating text classification with explainable artificial intelligence. *IAES International Journal of Artificial Intelligence*, *13*(1), 278-286. https://doi.org/10.11591/ijai.v13.i1.pp278-286

Zakaria, & Ningrum, S. (2023). ChatGPT's impact: The AI revolution in EFL writing. *Borneo Engineering & Advanced Multidisciplinary International Journal*, *2*(Special Issue (TECHON 2023)), 32-37. https://beam.pmu.edu.my/index.php/beam/article/view/109

Wu, L. (2023, February). Automatic English essay scoring algorithm based on machine learning. In *2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1-4). IEEE. https://doi.org/10.1109/ICICACS57338.2023.10099945

Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation*, *67*, Article 100911. https://doi.org/10.1016/j.stueduc.2020.100911

Appendix A

ChatGPT-generated Rubric

Task Response

Band Score	Description
9	Fully addresses all parts of the task with a fully developed position, well-supported ideas, and examples.
8	Addresses all parts of the task with a clear position, supporting ideas, and relevant examples.
7	Addresses all parts of the task, presents a clear position, and provides relevant examples. Some ideas may lack full development.
6	Addresses the task with a relevant position but may lack supporting details or examples.
5	Addresses the task but may have irrelevant details or lack clear focus.
4	Responds to the task only partially; the format may be inappropriate.
3	Provides an unclear or unrelated response to the task.
2	Contains little or no response to the task.
1	Contains no response to the task.
0	Did not attempt the task.

Coherence and Cohesion

Band Score	Description
9	Uses cohesion effectively and presents a fully coherent, well-structured response.
8	Sequences information and ideas logically, with a clear overall progression.
7	Presents information coherently with clear paragraphing and some effective cohesive devices.
6	Presents information with some organization but may lack logical sequencing or clear paragraphing.
5	Presents information with some lack of clarity, and ideas may be disconnected.
4	Presents information in a disorganized manner, lacking clear connections between ideas.
3	Presents information incoherently with little or no clear structure.
2	Presents information with serious incoherence and lack of structure.
1	Contains no clear organization or structure.
0	Did not attempt the task.

Lexical Resource

Band Score	Description
9	Uses a wide range of vocabulary fluently and flexibly, with full awareness of nuances and collocations.
8	Uses a wide range of vocabulary readily and flexibly, with some awareness of style and collocation.
7	Uses a sufficient range of vocabulary to express ideas clearly, with some awareness of style and collocation.
6	Uses an adequate range of vocabulary but may have some inaccuracies and inappropriate word choices.
5	Uses a limited range of vocabulary and may have frequent inappropriate word choices.
4	Uses a very limited range of vocabulary with frequent repetition and errors.
3	Uses extremely limited vocabulary with little understanding of word choice.
2	Uses isolated words or short phrases with little relevance to the task.
1	Uses a few isolated words.
0	Did not attempt the task.

Grammatical Range and Accuracy

Band Score	Description
9	Uses a wide range of structures with full flexibility and accuracy.
8	Uses a wide range of structures with full control; may have occasional minor errors.
7	Uses a variety of complex structures with some errors.
6	Uses a mix of simple and complex structures but may have frequent errors.
5	Uses a limited range of structures with frequent errors.
4	Uses only basic sentence structures with limited control of grammar.
3	Uses very limited sentence structures with little control of grammar.
2	Uses isolated phrases or sentences with serious grammatical errors.
1	Uses isolated words or phrases with no control of grammar.
0	Did not attempt the task.

Appendix B

ChatGPT-generated Grades and Official Grades

Text Number:	*Official Grade:*	*ChatGPT Grade:*
Text 1	4	4
Text 2	6.5	6
Text 3	8	7.5
Text 4	5	5.5
Text 5	7	7.5
Text 6	5.5	5.5
Text 7	7.5	7
Text 8	5	4.5
Text 9	8	7.5
Text 10	6	6
Text 11	4	4
Text 12	6	6
Text 13	6	6
Text 14	4	5
Text 15	6	5
Text 16	7	5.5
Text 17	7.5	6.5
Text 18	5	6
Text 19	7	6.5
Text 20	5.5	5.5
Text 21	6.5	6
Text 22	7	7
Text 23	8	7
Text 24	4	4.5
Text 25	6	6
Text 26	3.5	4
Text 27	5.5	7
Text 28	5	5
Text 29	5.5	5
Text 30	5	5.5

Text Number:	*Official Grade:*	*ChatGPT Grade:*
Text 31	7	7.5
Text 32	5.5	6
Text 33	6	5
Text 34	5	5.5
Text 35	7.5	7.5
Text 36	5	4
Text 37	6.5	6
Text 38	7	7
Text 39	6	6
Text 40	6	5
Text 41	7	7
Text 42	5.5	6
Text 43	7.5	7.5
Text 44	7	6
Text 45	6	7
Text 46	7	7
Text 47	6.5	6.5
Text 48	6	7.5
Text 49	4.5	5
Text 50	7	6
Text 51	4	4
Text 52	6.5	7
Text 53	6.5	7.5
Text 54	6.5	7
Text 55	6	7