International Journal of English Language and Pedagogy (IJELP)
Vol. 3, No. 2, pp. 44 - 59
ISSN: 2964-6545, DOI: 10.33830/ijelp.v3i2.13551

Reliability of ChatGPT-5.0 as an Automated Essay Scoring Tool: What Matters?

Thanh T.G. Trinh

Faculty of Foreign Languages; Ho Chi Minh City University of Technology and Education

Article Info

Article history:

Received: August 16th, 2025
Revised: Sept 20th, 2025
Accepted: October 15th, 2025

Keywords:

Essay scoring
Rubric
Reliability measurements
Prompting engineering
Data feeding

ABSTRACT

The aims of the present study were twofold: Exploring the variabilities of ChatGPT-5.0's capabilities of rubric-based essay scoring across three prompting designs and two essay feeding methods; and testing the reliability of ChatGPT-generated scores against human ratings. Drawing upon three reliability measurements, including: Spearman's correlations, Intraclass Correlations and quadratic weighted kappa (QWK), the findings revealed that although the reliability coefficients ranged from moderate to substantial, the essay scoring abilities of ChatGPT-5.0 depended greatly upon users' expertise to engineer prompts and their choices of essay feeding. This study highlights the importance of continued effort in the validation of this technology as an automated essay scoring tool and emphasizes the irreplaceability of human professional judgment in this field.

Corresponding Author:

Thanh T.G. Trinh
Email: thanhttg@hcmute.edu.vn

1. INTRODUCTION

In Vietnam, high-stakes English language proficiency tests, such as the International English Language Testing System (IELTS), function as a crucial gate-keeping mechanism in academic and professional contexts worldwide. Specifically, the IELTS has been employed for a wide range of purposes, such as academic admissions, graduation requirements as well as job opportunities, and has become an influential factor in the academic and professional success of Vietnamese undergraduate students. Writing is an important skill that contributes to the successful performance of learners in this crucial test. With the advancement of AI-driven technologies like ChatGPT, the learning and teaching of this skill has undergone substantial transformation, from the initial drafting to final revision (Barrot, 2018). Particularly, the practices of providing feedback and assessment of L2 writing in the era of ChatGPT are subject to debate among practitioners and researchers due to the potential and pitfalls of this technology.

Recently, conceptual reviews together with empirical studies have attempted to draw a conclusion as to ChatGPT's assessment capabilities in terms of accuracy, qualitative feedback and rubric-based scorings. There is still disagreement on its consistency, accuracy, and reliability. A number of studies which focused on error detection ability of ChatGPT as compared to human raters revealed

its potential and a high degree of accuracy in identifying errors made by human students (Pfau et al., 2023; Coyne et al., 2023; Algaraady & Mahyoob, 2023). Quantitative and qualitative research also investigated the qualitive feedback that ChatGPT provided on students' essays, concluding that its feedback outperformed the human raters in breadth, relevance, specificity, and accuracy of the feedback. In rubric-based scoring, studies found strong correlations between ChatGPT's scores and those assigned by human raters, such as Mizumoto and Eguchi (2023), Yancey et al. (2023), Koraishi (2024), to name just a few, whereas others produced a reverse result such as Bui and Barrot (2025). The common concern widely discussed in these studies is whether ChatGPT can replace human feedback, given ethical implications, risks of hallucinations, contextual sensitivity, unpredictable data training effect, insufficient understanding of nuances in student writings, teacher's loss of evaluative role leading to misrepresenting learner performance. A closer examination reveals that methodological variations in these empirical studies left a big question regarding the reliability and accuracy of ChatGPT. These differences include data input methods, prompting strategies, the presence or absence of calibration examples, etc. To this end, few investigations have examined these factors from the perspective of classroom teachers, many of whom might possess limited technical expertise in working with ChatGPT.

2. LITERATURE REVIEW

2.1. ChatGPT: Nature and affordances

ChatGPT (Chat Generative Pre-trained Transformers) is an AI-driven conversational bot developed by OpenAI. ChatGPT is powered by a Large Language Model (LLM) trained on massive datasets. ChatGPT is constantly updated through pre-training and fine-tuning. One feature of ChatGPT is its randomness, resulting in divergent outcomes for the same prompt since the system uses probabilistic prediction of words. Therefore, ChatGPT possesses the ability of linking words and concepts cohesively, leading to contextual sensitivity. Although it is AI-driven, there is not completely devoid of biases possibly caused by training data. The underlying algorithm is subject to change, altering the system's patterns and interpretations.

Generally, ChatGPT could understand prompts in context, enabling it to generate human-like responses across a wide range of topics and maintaining meaningful interactions. The integration of AI-driven ChatGPT can transform the practices of language education in many aspects. ChatGPT enables personalized and differentiated language instruction (Sim, 2025), assists in curriculum development and text adaptation (Nguyen, 2024). In L2 writing pedagogy, ChatGPT could be a valuable tool throughout the writing process (Barrot, 2018). Especially, when used as an L2 writing assessment tool, ChatGPT provides instant and timely feedback in terms of error detection, qualitative feedback on different dimensions of L2 writing and generating specific scores based on writing rubrics or criteria based on which it is trained (Sim, 2025).

2.2. Empirical research on ChatGPT as an L2 writing assessment tool

2.2.1. Error detection

Back in 2023, a couple of studies aimed at exploring the error detection capabilities of ChatGPT across models (Pfau et al., 2023; Coyne et al., 2023). In the first research, ChatGPT 3.5 Turbo's ability to identify errors in a corpus of essays of varying proficiency levels produced by Greek learners of English was compared with that of human raters (Pfau et al., 2023). Despite some errors being ignored, there was a strong correlation between ChatGPT's performance and that of human raters ($r=0.97$). Coyne et al. (2023)'s finding revealed the satisfactory ability of GPT-4 to detect errors and lower temperatures enhanced the system's performance. A common position among researchers is that ChatGPT can be an effective tool for error identification but not a sole benchmark for error analysis since it fails to perform effectively with nuances and complexities of human writing. Human instructors' feedback is still more accurate (Algaraady & Mahyoob, 2023).

2.2.2. Qualitative feedback

There is no doubt that ChatGPT provides more extensive and instant feedback on students' essays than human teachers. Studies aimed at evaluating ChatGPT's capacity in terms of its qualitative feedback indicated that the system produced more relevant, accurate or specific feedback on the EFL

essays than its human counterpart with great consistency (Li et al., 2024; Saricaoglu & Bilki, 2025). However, it occasionally misclassified errors in Grammar and Lexical Resources (Saricaoglu & Bilki, 2025).

2.2.3. Automated essay scoring

Another big consideration of ChatGPT as an L2 writing assessment tool is its essay scoring reliability. Extensive research has inconclusive findings as to the trustworthiness of ChatGPT's automated essay scoring. Bui and Barrot (2025) found a weak correlation between ChatGPT3.5-assigned scores and human ratings (Pearson's R = 0.1-0.3) and the lack of consistencies in the scores generated by ChatGPT3.5 across multiple time- of scorings (ICC=0.3-0.5 vs. 0.8-0.9). Their corpus included argumentative essays of varying proficiency levels (A2-B2) written by students from 10 Asian countries and randomly chosen from the publicly accessible ICNALE-Written Corpus. They also argued for their choice of only one experienced human rater as their benchmark against ChatGPT3.5. They developed their own writing scoring rubric (Claim, Development, Audience, Cohesion, Style and Conventions) and trained ChatGPT3.5 to apply their newly-developed rubric via one single prompting strategy. In terms of inputting methods, a conversation on ChatGPT was opened for one single essay.

On the other hand, other studies found a strong correlation between ChatGPT-assigned scores and human ratings. Mizumoto and Eguchi (2023) examined the potentials of GPT-3 text-davinci-003 model as an automated essay scoring tool for more than 1,000 TOEFL essays. Their findings suggested that ChatGPT achieved acceptable levels of reliability (quadratic weighted kappa~=0.38); however, the incorporation of a number of other computational metrics (e.g. Lexical measures, Syntactic complexity measures, etc.) improved the scores to a QWK of 0.6. Technical considerations of this study included the application of IELTS writing task 2 band descriptors (public version) with 0-9 bandscores to train ChatGPT to assess TOEFL essays, the prompt design used Python 3.8.5. No calibration examples were fed. They used the levels of TOEFL for benchmarking. Yancey et al. (2023) compared the performance of ChatGPT3.5 and ChatGPT4.0 under three experiments. Their data were 1,000 responses for a Duolingo Writing Test which required students to write a short essay within 5 minutes and human raters scored these essays using CEFR scoring rubric. Calibration examples were taken from the data and developed by two researchers. ChatGPT scores were compared with both human baseline and AWE baseline. Another study employed the corpus of 56 IELTS Task 2 Writing sample essays officially published by Cambridge University Press and compared ChatGPT's scores with the official scores assigned by certified IELTS Writing Task 2 examiners, revealing a strong agreement between the two sets of grades (QWK = 0.811) (Koraishi, 2024).

The inconsistent findings about the accuracy and reliability of ChatGPT's scoring performance can be attributed to a wide range of factors, i.e. methods of data inputting, prompting strategies, presence of calibration examples, temperature control or fine-tuning, etc. For example, studies involved a modest number of essays, with researchers manually opening a new chat for each essay entry (Bui & Barrot, 2025; Koraishi, 2024) whereas others employed API-based or other computational approaches to process a large corpus of essays (Mizumoto & Eguchi, 2023; Yancey et al., 2023). A number of studies investigated the effects of prompting strategies (Xu et al., 2024; Poole & Coss, 2023) or tested whether fine-tuning could enhance the reliability of ChatGPT-assigned scores (Wang & Gayed, 2024) or emphasized the importance of calibration examples (Yancey et al., 2023).

The review of previous studies reveals that ChatGPT is highly context-sensitive and that few studies have examined it from the perspective of classroom teachers. This gap is particularly salient, since many teachers often have limited technical knowledge of ChatGPT and are constrained by the financial costs of API- or token-based research. Therefore, the aim of the current study is to evaluate the extent to which different prompting engineering strategies and data inputting methods affect the reliability of ChatGPT-generated scores compared to human ratings. Importantly, the prompting engineering strategies and data inputting methods adopted in this study required no advanced technical knowledge and ChatGPT underwent a data training process similar to that of training human raters. Furthermore, no literature of this line of inquiry has been found for ChatGPT-5.0, the most updated version of this GenAI technology.

Research questions:

1. Does the reliability of ChatGPT-generated scores vary depending on prompting design?
2. Does the reliability of ChatGPT writing scores vary depending on methods of data input?
3. To what extent are ChatGPT-generated scores compatible with human ratings?

3. METHOD

3.1. Participants and data collection

The data of the current study was collected from 05 Vietnamese students (4 female, 1 male) enrolled in an IELTS Writing preparation course. They did not have any prior experience with IELTS. The entry test for the course revealed no disparity in their overall English skills. At the beginning of the course, they were assessed to have upper-intermediate English proficiency level and also expressed their ambition to achieve an IELTS band score of 7.0 or above. The course lasted about 4 months, with two 2-hour class meetings per week.

The learner corpus included 56 IELTS Task 2 essays covering common writing topics of this international test, i.e., traditions and cultures, crime, tourism, studying abroad, technology, etc. The tutor of the enrolled course gave instructions on how to outline and deal with each task type: discursive essays, opinion essays, argumentative essays, etc. The students wrote the essays as their homework and no corrective feedback was made. The total number of words was 18,741.

Table 1. Descriptions of the Writing Corpus

Student	Gender	Number of essays written	Total number of words	Min	Max
Student 1	Female	11	3,026	239	319
Student 2	Female	09	3,072	247	521
Student 3	Female	12	4,341	279	463
Student 4	Female	13	4,455	309	373
Student 5	Male	11	3,847	293	456

3.2. Procedures

The procedures were conducted from the simplest prompting strategy (1) to the most complex prompting strategy (3), as suggested by Poole & Coss (2024):

(S1) Prompting strategy 1: Prompt with ChatGPT internal rubric for IELTS Writing Task 2 as in Koraishi (2024). See Appendix A for the internal rubric of ChatGPT.

(S2) Prompting strategy 2: Prompt with IELTS official rubric. See Appendix B for the official rubric.

(S3) Prompting strategy 3: Prompt with IELTS official rubric and one calibration example for each bandscore (Yancey et al., 2023). The calibration examples were selected from the series of IELTS Practice Tests 9-20. The essay samples published in this series are followed by Examiner's evaluations with a designated score. After a collection of essays were randomly chosen for each bandscore from 4.0 to 8.5, they were fed into ChatGPT after the prompt. (See Appendix C for the sample essays)

The essays were fed into ChatGPT via two different methods:

(M1) Method 1: the whole Word files containing 9-11 essays were uploaded onto ChatGPT5.0 for essay scoring. The result was produced in a summary table with scores for each criterion and overall scores. For each file, a new chat was opened, the prompt was copied and a new file was attached.

(M2) Method 2: The prompt was copied into the text box of ChatGPT5.0, followed by an essay. For each essay, a new was opened and the procedure was repeated until 56 essays were inputted for assessment.

3.3. Data analysis

All of the essay scores under different prompting strategies and two essay inputting methods were fed into SPSS26 for analysis. Test of normality revealed that none of the measures were normally distributed with every significance value of 0 (for both Kolmogorov–Smirnov and Shapiro–Wilk). Therefore, in this study, the non-parametric tests were chosen.

To test the reliability of ChatGPT-generated scores compared to those of human raters across prompting strategies and input methods, three reliability measurements were employed, including: Spearman's correlations, Intraclass Correlations (Koraishi, 2024; Bui & Barrot, 2025) and quadratic weighted kappa (QWK) (Mizumoto & Eguchi, 2023; Poole & Coss, 2024). The first two measurements were conducted on SPSS26 whereas Python was used for QWK.

4. FINDINGS

The current study investigated the consistency and reliability of ChatGPT5.0 as an assessment tool depending on prompt design and essay feeding methods. Specifically, I aimed to explore whether different prompting strategies and data inputting methods significantly impacted ChatGPT-generated scores as well as to find out under which condition ChatGPT produced better scores in relation to human ratings. To achieve these objectives, I employed a combination of descriptive and reliability statistics.

Table 2. Scores Assigned by ChatGPT5.0 and Human Raters

Strategy	Method	Mean	Std	min	max
S1	M1	6.125	0.218	6.0	6.5
	M2	6.491	0.481	5.5	7.0
S2	M1	6.643	0.401	6.0	7.5
	M2	6.892	0.483	5.5	8.0
S3	M1	6.009	0.442	5.0	6.5
	M2	6.741	0.513	5.5	8.0
HUM1		6.964	0.485	6.0	8.0
HUM2		6.625	0.702	5.5	8.0

Table 2 summaries the scores assigned by ChatGPT (under three prompting strategies and two data inputting methods) and a human rater. The data indicated that the mean scores by the human raters were higher than those assigned by ChatGPT. Upon closer examination of the mean scores of each criterion, the human raters tended to assign higher scores for grammatical range and lexical resources.

Table 3 presents the correlations between ChatGPT's scores and human ratings, indicating that essay scores generated by ChatGPT5.0 under three prompting strategies and two inputting methods were moderately correlated, with the exception to that of Prompt 1 and Method 1. Method 2 outperformed Method 1 in the values of Spearman's Rhos.

Table 3. Correlations between ChatGPT's scores and human ratings

	Strategy 1				Strategy 2				Strategy 3			
	Method 1		Method 2		Method 1		Method 2		Method 1		Method 2	
	Spearman's rho	p	Spearman's rho	p	Spearman's rho	p	Spearman's rho	p	Spearman's rho	p	Spearman's rho	p
AV_HUM1	0,184	0,175	,544**	0,000	,579**	0,000	,627**	0,000	,503**	0,000	,608**	0,000
AV_HUM2	-0,034	0,802	,687**	0,000	,630**	0,000	,600**	0,000	,286*	0,033	,715**	0,000

Table 4 illustrates high levels of agreement and consistency between ChatGPT and human raters in assigning scores for Lexical Resources as well as Coherence and Cohesion across the prompting

strategies and data feeding methods, except for Prompt 1x Method 1. For the former category, intraclass correlations reached 0.8 depending on the prompt design whereas the latter achieved the value of 0.7. Lower levels of reliability fell under the dimensions of Grammatical range and Accuracy and Task response, but the values still indicated moderate to substantial levels of agreement.

Table 4. Intraclass correlations between the scores assigned by ChatGPT5.0 and the two human raters.

Strategy	Method	Average score	Task response	Coherence & Cohesion	Lexical Resources	Grammatical range & Accuracy
S1	Method 1	0.483	0.484	0.420	0.521	0.309
S1	Method 2	0.783	0.671	0.730	0.834	0.709
S2	Method 1	0.759	0.641	0.727	0.789	0.662
S2	Method 2	0.778	0.705	0.763	0.818	0.673
S3	Method 1	0.678	0.578	0.752	0.705	0.588
S3	Method 2	0.786	0.703	0.744	0.799	0.704

All correlations are significant at the 0.05 level (2-tailed)

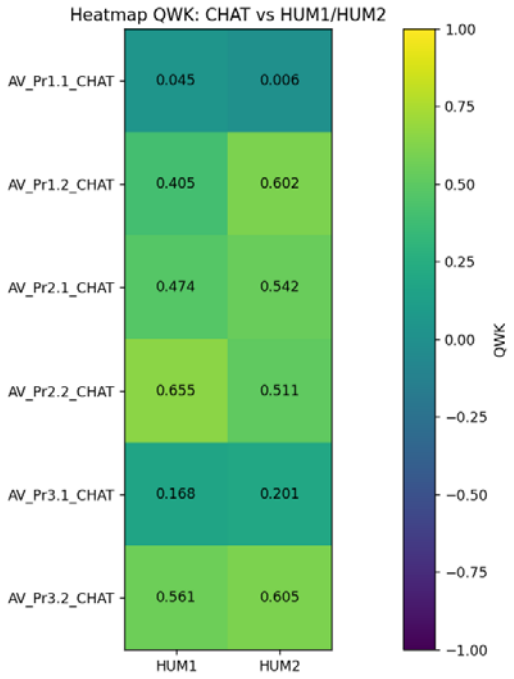

Figure 1. QWK Heatmap

As illustrated in the heatmap, there were different levels of agreement between the CHAT-automated scores and the two human raters (HUM1 and HUM2). While S1 x M1 (QWK = 0.045 with HUM1, 0.006 with HUM2) and S3xM1 (QWK = 0.168 with HUM1, 0.201 with HUM2) demonstrated negligible agreement, S2xM2 (QWK = 0.655 with HUM1, 0.511 with HUM2) and S3xM2 (QWK = 0.561 with HUM1, 0.605 with HUM2) achieved the highest levels of agreement, from moderate to substantial. It should also be noted that S1xM2 achieved a moderate and substantial agreement with HUM1 and HUM2, respectively.

Three conclusions were drawn: (1) Input method 2 brought about more reliable scores than Input Method 1 as compared to human raters; (2) Prompting strategies with the involvement of official rubric tended to produce higher reliability than the use of ChatGPT5.0's internal IELTS Writing Task 2 rubric; and (3) The involvement of calibration examples failed to increase the agreement level between ChatGPT scores and human ratings.

4. DISCUSSIONS

Firstly, the findings suggest that prompting strategies that incorporated the official IELTS rubric (S2, S3) achieved higher levels of agreement (i.e. higher ICC and QWK) than the use of ChatGPT's internal rubric. A closer examination of ChatGPT's internal rubric revealed that it omitted some construct-defining elements within each criterion across bandscores. This limitation explains why the official IELTS rubric better provides construct-relevant guidance, thereby enhancing construct validity. If ChatGPT is trained with a construct-aligned rubric (official rubric), the risk of construct-irrelevant variance may be reduced.

Secondly, the findings reveal that Input Method 2 consistently produced more reliable results than Method 1 when compared with human raters. In M1, a cluster of essays were fed into the system simultaneously and ChatGPT produced a summary scoring table for all of the essays in one single conversation. This may cause cross-essay bias, reducing reliability. In M2, each essay was processed in a new conversation with the same prompt. This approach helped reduce order/comparison effects between essays, prompting ChatGPT to focus on the features of each individual essay. Spearman correlations, ICC and QWK exhibited higher values in Method 2 than in Method 1, entailing that how essays are fed into ChatGPT should be considered as a decisive factor affecting reliability.

In this study, calibration examples were included in the prompts to guide ChatGPT in interpreting the official IELTS rubric in a more accurate way, similar to the norming process used to train human raters in assessing writing performance. The inclusion of calibration examples is considered as "few-shot" prompting technique. However, the findings suggest that calibration examples did not help ChatGPT improve the reliability in essay scoring. One possible explanation can be that ChatGPT does not have a cognitive calibration mechanism just like a human rater. When human raters are provided with calibration examples, their interpretations of the scoring rubric or construct understandings are clarified, substantiated and reinforced. In other words, they internalize their scoring framework via calibration examples, and learning does take place. It seems that ChatGPT does not actually learn from calibration examples. In their system, few-shot prompting operates like a pattern matching in a conversation, without leading to better construct understanding. However, several studies did not share similar results. For example, in Poole & Coss (2024)'s findings, ChatGPT3.5 performed best when multiple examples together with detailed band descriptors were fed into the system. Or Yancey et al. (2023) concluded that "GPT-4 only required one calibration example per rating category to achieve near optimal performance" (p. 580) but the inclusion of detailed rubric "contributed negligible effect" (p. 579). These conflicting results reveal that the prompts do not behave in a similar way across models (GPT-4.0 vs. GPT-5.0).

Taken together, a long and complex prompt increases contextual load for ChatGPT. The incorporation of rubric, calibration examples, multiple essays at once (S3-M1) forced the system to juggle between multiple elements, reducing the attention to construct-relevant features of each essay and leading to measurement noise. The results showed that reducing prompt load by employing S2 and M2 improved reliability. This is also exemplified by Wang & Gayed (2024)'s study confirming that the best-performing model was the one finetuned by only one prompt. When it comes to prompting engineering, "less is more" seems workable.

The correlations between ChatGPT and human raters were found to be varied according to on input and prompt design. Specifically, the approach of feeding one single essay combined with the incorporation of the official IELTS rubric produced better results. The reliability coefficients ranging from moderate to substantial suggested that ChatGPT can used as an assessment tool with some degree of accuracy. These findings are, to some extent, similar to previous studies (Li et al., 2024; Koraishi, 2024; Mizumoto & Eguchi, 2023). It should be noted that these previous studies employed IELTS rubrics in assessing the reliability of ChatGPT-assigning scores. However, the findings of this study contradict with those of Bui & Barrot (2025) who prompted ChatGPT3.5 with a newly built writing scoring rubric for 200 argumentative essays across levels.

Looking closer at Table 4, we can draw conclusions for the reliability in scoring specific dimensions of essays. There were high levels of agreement and consistency between ChatGPT and human raters in assigning scores for Lexical Resources and Coherence and Cohesion across the prompting strategies and data feeding methods. This may be attributed to the clarity and observability of these two criteria. As for the other two criteria, ChatGPT5.0 achieved lower reliability indices. Several factors may have contributed to this low correlation.

ChatGPT5.0 could identify grammar errors better than human raters, possibly causing biases in its assessments. On the other hand, human raters tend to detect subtle errors requiring nuanced understanding of complex aspects of writing (Algaraady & Mahyoob, 2023). In addition, human raters with expertise and experience seem to be more lenient with errors typically made by non-native speakers of English (Bui & Barrot, 2025), providing that these errors are compensated with other well-performed aspects of writing such as intelligibility or writer's intents. Essay quality is measured multidimensionally. Some subtle dimensions such as Task Response are involved with the message, content, creativity and criticality expressed in students' essays. However, ChatGPT5.0's scoring algorithms are not advanced and capable enough to evaluate such subtle aspects of argument quality or idea development of human-produced essays (Yancey et al., 2023; Bui & Barrot, 2025).

5. CONCLUSION

This study aimed at exploring the impact of different prompting strategies and inputting methods on the reliability of ChatGPT's essay scoring abilities with human ratings as a benchmark. The findings revealed that the reliability of ChatGPT5.0's scores was significantly affected by data feeding methods and prompt design, with the levels of agreement ranging from moderate to strong across the four criteria. Specifically, reliability measurements between ChatGPT and human raters were higher for Lexical Resources and Coherence and Cohesion than for Task Response and Grammatical Range and Accuracy. Moreover, the impacts of data feeding and prompt design were substantial. The results showed that feeding a single entry for each produced higher consistency in essay scoring than inputting a cluster of essays at one time, underscoring the importance of data feeding methods affecting reliability; incorporating official rubric in the engineering of prompts can reduce the risk of construct-irrelevant variance; the presence of calibration examples in the prompts proved that ChatGPT does not have similar learning mechanism to human raters and prompts in different models of ChatGPT fail to produce consistent results.

The conclusion is that although ChatGPT can serve as an assessment tool to reduce teachers' workload, its lack of consistency and dependence on know-how factors may distort the accuracy and reliability. Therefore, GenAI-based language assessment practices should be limited to classroom use as a supporting tool for teachers rather than as a mainstream assessment instrument. For high-stakes exams which require transparency and fairness, the era of GenAI is not yet to come. This study is a gentle reminder that human professional judgment is irreplaceable.

This study contributes to the growing body of literature using AI as an official assessment tool. Continuous validation on the effectiveness of GenAI technologies would be inevitable, given the robust development of this field. Furthermore, it is essential to extend the strand of research exploring calibration examples, prompting engineering, etc.

REFERENCES

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's Capabilities in Spotting and Analyzing Writing Errors Experienced by EFL Learners. *Arab World English Journal*, *9*, 3–17. https://doi.org/10.24093/awej/call9.1

Barrot, J. S. (2018). Using the Sociocognitive-Transformative Approach in Writing Classrooms: Effects on L2 Learners' Writing Performance. *Reading & Writing Quarterly, 34*(2), 187–201. https://doi.org/10.1080/10573569.2017.1387631

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies*, *30*(2), 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M., & Inui, K. (2023). *Analyzing the Performance of GPT-3.5 and GPT-4 in Grammatical Error Correction* (No. arXiv:2303.14342). arXiv. https://doi.org/10.48550/arXiv.2303.14342

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

Li, J., Huang, J., Wu, W., & Whipple, P. B. (2024). Evaluating the role of ChatGPT in enhancing EFL writing assessments in classroom settings: A preliminary investigation. *Humanities and Social Sciences Communications*, *11*(1), 1268. https://doi.org/10.1057/s41599-024-03755-2

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nguyen, T. Q. Y. (2024). Unraveling the Potential of ChatGPT: Investigating the Efficacy of Reading Text Adaptation. *Proceedings of the AsiaCALL International Conference*, *4*, 159–169. https://doi.org/10.54855/paic.23412

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal*, *73*(2), 197–206. https://doi.org/10.1093/elt/ccz006

Pfau, A., Polio, C., & Xu, Y. (2023). Exploring the potential of ChatGPT in assessing L2 writing accuracy for research purposes. *Research Methods in Applied Linguistics*, *2*(3), 100083. https://doi.org/10.1016/j.rmal.2023.100083

Poole, F.J. & Coss, M.D. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt (s). *Journal of Technology & Chinese Language Teaching*, 1-24.

S. Barrot, J. (2024). Leveraging ChatGPT in the Writing Classrooms: Theoretical and Practical insights. *Language Teaching Research Quarterly*, *43*, 43. https://doi.org/10.32038/ltrq.2024.43.03

Saricaoglu, A. & Bilki, Z. (2025). The capacity of ChatGPT-4 for L2 writing assessment: A closer look at accuracy, specificity, and relevance. *Annual Review of Applied Linguistics*, 1-21.

Sim, J. J. (2025). The GALL of it all: Grading and teaching in the age of GenAI-assisted language learning. *Language Teaching*, 1–8. https://doi.org/10.1017/S0261444825100785

Wang, Q., & Gayed, J. M. (2024). Effectiveness of large language models in automated evaluation of argumentative essays: Finetuning vs. zero-shot prompting. *Computer Assisted Language Learning*, 1–29. https://doi.org/10.1080/09588221.2024.2371395

Xu, Y., Polio, C., & Pfau, A. (2024). Optimizing AI for Assessing L2 Writing Accuracy: An Exploration of Temperatures and Prompts. In C. Chapelle, G. Beckett, & J. Ranalli (Eds.), *Exploring AI in Applied Linguistics* (pp. 151–174). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.10

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating Short L2 Essays on the CEFR Scale with GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 576–584. https://doi.org/10.18653/v1/2023.bea-1.49

APPENDIX A: INTERNAL RUBRIC BY CHATGPT-5.0

Prompt: In a table, provide the rubric of the writing Task 2 of the IELTS exam.

IELTS Writing Task 2 Assessment Rubric

Band Descriptor	Task Response	Coherence and Cohesion	Lexical Resource	Grammatical Range and Accuracy
9 (Expert)	Fully addresses all parts of the task; presents a fully developed position with relevant, fully extended, and well-supported ideas	Uses cohesion effortlessly and logically; paragraphs are skillfully managed	Uses a wide range of vocabulary naturally and precisely; rare minor errors	Uses a wide range of structures with full flexibility and accuracy; rare slips
8 (Very Good)	Covers all requirements; presents a well-developed response with relevant, extended, and supported ideas	Sequences information and ideas logically; cohesion well-managed; paragraphing is clear and appropriate	Wide range of vocabulary used flexibly and precisely; occasional inaccuracies	Wide range of structures with majority error-free sentences; very occasional mistakes
7 (Good)	Addresses all parts of the task; presents a clear position with relevant main ideas, though some may lack full development	Logically organizes information; clear progression; uses cohesive devices effectively though may be mechanical at times; adequate paragraphing	Sufficient range of vocabulary; some less common vocabulary used appropriately; occasional errors in word choice/spelling	Variety of complex structures used; frequent error-free sentences; some grammatical errors persist
6 (Competent)	Addresses task only partly or unevenly; presents a relevant position but may be unclear or repetitive; some ideas not developed	Information arranged coherently but progression may be faulty; cohesion may be inadequate, over/underused; paragraphing may be faulty	Adequate range of vocabulary but with noticeable repetition; some errors in word formation and collocation	Mix of simple and complex structures; errors are frequent but rarely cause misunderstanding
5 (Modest)	Addresses task only partially; may present a position but unclear, irrelevant, or repetitive; limited ideas and support	Some organization but lack of overall progression; cohesion may be mechanical or faulty; inadequate paragraphing	Limited range of vocabulary; frequent errors in word choice, spelling, word formation	Limited range of structures; frequent grammatical errors that may cause misunderstanding
4 (Limited)	Responds minimally; position unclear; ideas very limited, irrelevant, or undeveloped	Very limited organization; cohesion ineffective; no clear paragraphing	Very limited vocabulary; errors predominate	Very limited structures; frequent errors cause strain for the reader

Band Descriptor	Task Response	Coherence and Cohesion	Lexical Resource	Grammatical Range and Accuracy
3 (Extremely Limited)	Does not adequately address the task; no clear position; ideas largely irrelevant	No clear organization; very limited cohesion	Very limited vocabulary with frequent errors; communication often unclear	Very limited control of grammar; errors predominate, communication often breaks down
2 (Intermittent)	Barely responds to the task; no position; no relevant ideas	No logical organization; no cohesion	Essentially no vocabulary control	Essentially no grammatical control
1 (Non-User)	Completely fails to address the task	No organization	No vocabulary except a few isolated words	No control of grammar
0	Did not attempt the task or answer not related to task	–	–	–

APPENDIX B: PROMPTING STRATEGIES

S1 x M1: Act as an official IELTS examiner. I will send you a file including texts written by students in relation to writing Task 2. You have to give me an accurate grade according to the IELTS rubric for Task 2. Give a score per criterion. You should provide the result without any justifications.

S1 x M2: 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. Give a score per criterion. Provide the result in a horizontal table format without any justifications. The text is:

[Insert an essay]

You should provide the result without any justifications.

S2 x M1: Act as an official IELTS examiner. I will send you a file including texts written by students in relation to writing Task 2. You have to give me an accurate grade according to the IELTS rubric for Task 2. Give a score per criterion.

##IELTS official rubric is as follows:

[Insert the official rubric]

You should provide the result without any justifications.

S2 x M2: 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. Give a score per criterion.

##IELTS official rubric is as follows:

[Insert the official rubric]

##You should provide the result without any justifications.

##The text is: *<Insert the text>*

S3xM1: Act as an official IELTS examiner. I will send you a file including texts written by students in relation to writing Task 2. You have to give me an accurate grade according to the IELTS rubric for IELTS Writing Task 2. Give a score per criterion.

##IELTS Writing Task 2 official rubric is as follows:

[Insert the official rubric]

Below are the calibration examples for each overall band score. These essays were taken from IELTS Practice Tests published by Cambridge ESOL with the official and certified IELTS examiners' scores given.

The essay below is scored at 4.0 <*Insert the text*>

The essay below is scored at 4.5 <*Insert the text*>

The essay below is scored at 5.0 <*Insert the text*>

The essay below is scored at 5.5 <*Insert the text*>

The essay below is scored at 6.0 <*Insert the text*>

The essay below is scored at 6.5 <*Insert the text*>

The essay below is scored at 7.0 <*Insert the text*>

The essay below is scored at 7.5 <*Insert the text*>

The essay below is scored at 8.0 <*Insert the text*>

The essay below is scored at 8.5 <*Insert the text*>

##You should provide the result without any justifications.

S3xM2: Act as an official IELTS examiner. I will send you a text including texts written by students in relation to writing Task 2. You have to give me an accurate grade according to the IELTS rubric for IELTS Writing Task 2. Give a score per criterion.

##IELTS Writing Task 2 official rubric is as follows:

[Insert the official rubric]

Below are the calibration examples for each overall band score. These essays were taken from IELTS Practice Tests published by Cambridge ESOL with the official and certified IELTS examiners' scores given.

The essay below is scored at 4.0 <*Insert the text*>

The essay below is scored at 4.5 <*Insert the text*>

The essay below is scored at 5.0 <*Insert the text*>

The essay below is scored at 5.5 <*Insert the text*>

The essay below is scored at 6.0 <*Insert the text*>

The essay below is scored at 6.5 <*Insert the text*>

The essay below is scored at 7.0 <*Insert the text*>

The essay below is scored at 7.5 <*Insert the text*>

The essay below is scored at 8.0 <*Insert the text*>

The essay below is scored at 8.5 <*Insert the text*>

##You should provide the result without any justifications.

##The text is: <*Insert the text*>

APPENDIX C: CALIBRATION EXAMPLES

 # The essay below is scored at 4.0

The Advantage of Driverless Vehicles

First of all number of vehicles is increase day after day which means every day the world gets more drivers than before. If we admit that a lots of people prefer to use public transport we do not have any doubts that many people use the vehicles because of advantage of driving.

The history shows us that the human like to move from place to another for many reasons and the always felt pleased when the rid. This days people have all kind of vehicles bikes, cars, motor… etc because they all have a different advantage.

People needs also can not meet at be found in one place. for that reason people need to move from a place to another place to meet their needs which means the advantage of moving from point to another point will be exist for ever.

World has bee changed a lot and many people have got great jobs with big salaries. The can easly fund their vehicl and because people get feeling boring if the used to some thing they always prefere to chang their vehicle from time to time.

Finally I think it is very hard to believe that the driverless vehicles will outweigh the disadvantages because people always find drive more and more give their life meaning and add more advantage to it all kind of vehicles give happiness to a lot of people that they can not think about lossing it.

 # The essay below is scored at 4.5

In their advertising, businesses nowadays sometimes stress that their products are new in some way. From my point of view, some businesses want to have good products to give to the people, but usually they worry about their products are newer than some other's businesses products.

I think it is a negative development, because when businesses stress about the quality of their products, sometimes they do something wrong while they are producing them. It is good when the businesses take care of and look after their products but with a limit. According to some experts, when you take a lot of care of something, you will probably do some things, about it, wrong.

From my own experience, I was trying to make three school projects, which my teachers asked me to do, and despite my hard work and because I was stressed about the projects I had to do, I finally failed because I had made a lot of mistakes.

To sum up, businesses nowadays should not stress about their products being new in some way. Besides that they should calm down and be careful on what they are producing.

 # The essay below is scored at 5.0

Nowadays, the people of some countries that have the young people more than the old people. Some people thinks when their contries have the yonge people more than the old people will be good because, that could increases the population in the future. Another people thinks it not good due to some countries limit the population, if that have more young children, it will over limit. This essay will discuss the advantages and disadvantages about in some countries have the young people more than the old people.

One of advantages is increasing the population. In some countries support the family to have more children because that can increases the population in the future. For example, in Singapore, Philippine and so on. What is more chancing to improve the educations as when they have a lot of young generation, the government could improve a good education. Also, they can develop the systems include the quilified teachers, the good atmosphere.

One of disadvantages is the place for study. If the young generation still a lot, the school will not enough for the study, the government should construct more school. Also, when they have the new schools, the teacher will not enough to teach them. The university should get more student to study about teaching education. Another disadvantages is the quality of education. If the many students learn in the classroom, the teachers can not take care all. For instance, when they have a problem they will need some help from the teachers. Furthermore, when they grow up, the unemploye problem will happen because the company can not receive everybody to get a job.

In conclusion, in some countries that have the young population more than the old population, the government should manage the education system. Moreover, they should prepare the plans for solving unemploye problems which can happen in the future.

 # The essay below is scored at 5.5

I completly disagree with the written statment. I believe that most of the people in the world have more information about their health and also about how they can improve their healthy conditions.

Nowadays, information about how harmful is to smoke for our bodies can be seen in many packets of cigars. This is a clear example how things can change from our recent past. There is a clear trend in the diminishing of smokers and if this continues it will have a positive impact in our health.

On the other hand, the alimentation habits are changing all over the world and this can affect people's health. However every one can choose what to eat every day. Mostly everybody, from developed societies, know the importance of having a healthy diet. Advances such as the information showed in the menus of fast food restaurants will help people to have a clever choice before they choose what to eat.
Another important issue that I would like to mention is how medicine is changing. There are new discovers and treatments almost every week and that is an inequivoc sintom of how things are changing in order to improve the world's health.

 # The essay below is scored at 6.0
Sharing information is actual issue in our world where it has strong influence on people. There are various spheres of our life where information is more or less important for people working there on out of this. For this reason some people consider that it is good to share information while others think in opposite way.
For example, practically all scientists are glad to share information with ordinary people or other scientist. There is no competition in this sphere. Sometimes it is bad for government which scientist share the secret information with international spy but it will not hurt information.
There are some simple rules in academic world which limite informational sources between people. If person is interested in theme discussing with you and you are ready to keep talking then the person gives you all information what he knows for free. On the other hand, if the person knows much and he knows that you can not give him actual or new information then he will share information with you just for money. For example, student pay for his learning while two students can cooperate and share information with each other. It is obviously that sharing information in business world can followes by releases. There is large competition and it may takes much costs for companies. Companies loose their profit every day because some one can not keep silence especially IT companies.
To sum up all above it is neccessery to say that there are some spheres in which sharing information is a crime. In my opinion, in many cases information can be too important or sharing at all.

 # The essay below is scored at 6.5
It is said that taking risks brings a lot of benefits. However, it also gives us some drawbacks.
First of all, it is obvious that taking risks will cause a great loss if people do it and fail. In personal life, this loss might not be so harmful. However, it will be really harmful in professional life, because people take a responsibility not only for themselves but also others such as colleagues, customers, and their families. It will even damage the society from the economic point.
On the other hand, we can receive huge benefits by taking risks. Firstly, we can learn how to prepare for one goal through this process. In order to achieve the aim, people will make all the efforts to think about it and try to find more efficient way. If they do this in the professional circumstances, they will recognise the responsibility and importance of cooperation.
Also, it will be completely meaningful even though people can't achieve the goal after taking risks. They will learn the reason why they have failed and how to change it. The failure will enable them to improve their skills and to achieve their object next time.
As I mentioned, it is true that taking risks give us both advantages and disadvantages. However, it can be argued that the benefits outweighed the drawbacks in that we can obtain advantages not only from the result but also from the process of taking risks.

 # The essay below is scored at 7.0
Saving money for the future is always a very good idea. First of all money is something that is needed in almost all areas in life. Whether you are young or old you need money to buy everyday things like food, clothing, etc. living etc.
Young people are often full of enthusiasm about their future. They are looking forward to their first job, to meeting new people or to getting to know as much of the world as they can. Many tend to live in the present rather than in the future so that they don't always plan ahead.
When people get older and settle down they realize that buying a house, starting a family or caring for your health takes up a considerable amount of money and everybody who began saving money in younger years is glad to have done so. However, saving money is not always possible. Sometimes unforeseen expenses cannot be avoided, life situations are suddenly changing or there is never even enough money available even for the most necessary things. So how could you save money for the future in this case? In general, you have to ask yourself what your priorities in life are. What are the things you cherish most? Is it more important for you to plan ahead or do you prefer to just enjoy the moment that you live in?
Everybody has to make own choices and to consider what is really essential for him or herself.
In any age taking a moment to reflect on your life and looking back at the things you have already done is always a good thing to do.

If you know yourself well and all about all the things that really make you happy you will be able to make the right decisions in financial issues as well as other areas in life.
In what way money plays an important role will be easy to be found out then. Perhaps you need less than you first thought years ago.

 # The essay below is scored at 7.5

For many people around the world, the preferred method of transportation is high-speed rail. Commuters travelling to and from work rely on the safety and efficiency, whilst tourists appreciate the convenience and novelty that trains provide. Others believe that highways, busses and regular trains should be improved before new, high-speed lines are added.
Safety is chief among concerns for those who travel to work or school on a regular basis. If one drives a car, they have to concentrate on the road not only to avoid accidents but also to prevent other drivers from causing a problem on the road. High-speed rail allows the commuter to leave the driving to the professional controlling the train, allowing them to get some work done while getting to work safely.
In addition, people tend to move further and further away from city centres, where land and houses are more affordable. High-speed rail allows these commuters to travel greater distances in a shorter ammount of time. There is a flow-on effect here, because if we can reduce the number of cars on the road, we can also cut down on traffic jams and road delays.
On the other hand, high-speed trains are expensive, and some believe this money could be spend on repairing motorways which are used by cars, busses and motorcycles. Another possibility would be to use this money to build more regular commuter trains and busses to service the ever-expanding urban populations. Moreover, boats and ferries could benefit from a budget which focuses more on existing forms of transport.
In the end, public transport is an issue which affects us all. The taxes which we pay should be spent on the type of transport which will have the most benefit to all citizens. In addition, we need to take into account how much the environment is damaged by fossil fuels and pollution. Therefore, I believe in order to move forward, we need to embrace the idea of high-speed rail so that future generations can continue to live safely and efficiently.

 # The essay below is scored at 8.0

It has been suggested that high school students should be involved in unpaid community services as a compulsory part of high school programmes. Most of the colleges are already providing opportunities to gain work experience, however these are not compulsory. In my opinion, sending students to work in community services is a good idea as it can provide them with many lots of valuable skills.
Life skills are very important and by doing voluntary work, students can learn how to communicate with others and work in a team but also how to manage their time and improve their organisational skills.
Nowadays, unfortunately, teenagers do not have many after-school activities. After-school clubs are no longer that popular and students mostly go home and sit in front of the TV, browse internet or play video games.
By giving them Compulsory work activities with charitable or community organisations, they will be encouraged to do something more creative. Skills gained through compulsory work will not only be an asset on their CV but also increase their employability. Students will also gain more respect towards work and money as they will realise that it is not that easy to earn them and hopefully will learn to spend them in a more practical way.
Healthy life balance and exercise are strongly promoted by the NHS, and therefore any kind of spare time charity work will prevent from sitting and doing nothing. It could also possibly reduce the crime level in the high school age group. If students have activities to do, they will not be bored and come up with silly ideas which can be dangerous for them or their surroundings.
In conclusion, I think this is a very good idea, and I hope this programme will be put into action for high schools/colleges shortly.

 # The essay below is scored at 8.5

I do agree to the statement that children brought up in poor families are better prepared to deal with the problems of adult life than children brought up in wealthy parents.
Children of poor parents are prematurely exposed to the problems of adult life e.g. earning a living and learning to survive on a low family income, sacrificing luxuries for essential items. These children begin to see the ‘realities’ of life in their home or social environment. Their parents’ own struggles serve as an example to them.
These children are taught necessary skills for survival as an adult from a very early age. Many children e.g. work in the weekends or holidays to either collect some pocket money or even contribute to their families’ income. A good example is the many children who accompany their parents to sell produce at the market. They are making a direct contribution to their families in terms of labor or income.
Children of poor families also are highly motivated. They tend to set high goals to improve their economic & social situation. A relevant example would be Mr Bill Gates (Founder of Microsoft Corporation). He had an

impoverished background but he used his talent and motivation to set up the world's largest computer organisation.
However, there are some problems that children from poor backgrounds do encounter. Many of these children, who are 'robbed' of their childhood while working, may feel cheated. They often turn to crime. This however, is a small group.
In summing up, children with impoverished backgrounds are able to deal with problems of adult life because of early exposure, family role models and sheer motivation.