

Evaluating Generative AI as an Automated Essay Scorer: A Comparative Study of GPT-5 and Human Examiner Ratings in MUET Writing

Noor Aishah Idrishah¹, Nor Izzati Md Jaafar², Nurzatul Shazreena Zainol Abidin³

¹ English Department, Centre for Languages and Foundation Studies, Universiti Sultan Azlan Shah, Kuala Kangsar, Perak, Malaysia

² English Department, Centre for Languages and Foundation Studies, Universiti Sultan Azlan Shah, Kuala Kangsar, Perak, Malaysia

³ English Language and Literature Department, Faculty of Languages and Communication, Universiti Pendidikan Sultan Idris, Jalan Utama UPSI, Tanjung Malim, Perak, Malaysia

¹ Email: nooraishah@usas.edu.my

² Email: izzati.jaafar@usas.edu.my

³ Email: nurzatul@fbk.upsi.edu.my

Proceedings of the Kalijaga International Conference on Education 2026

Abstract. This exploratory study investigates the predictive validity and inter-rater reliability of automated essay scoring using GPT-5 against certified human examiner ratings in the high-stakes Malaysian University English Test (MUET). Employing a purposive sample of 23 student essays, the writing scripts were evaluated by GPT-5 utilizing a zero-shot prompting pipeline embedded with the official 90-mark MUET analytic rubric (Task Fulfilment, Language, and Organisation). The resulting automated scores were benchmarked against the candidates' official examiner marks. Statistical analyses revealed a severe upward scoring bias, with GPT-5 awarding significantly higher marks than trained human examiners (mean difference = 23.52, $t(22) = 6.55$, $p < .001$, $d = 1.37$), overpredicting outcomes in the vast majority of cases. Furthermore, absolute agreement between the two rating sets was negligible ($ICC = -.07$), and linear regression confirmed that raw GPT-5 scores failed to reliably predict official MUET outcomes ($F(1, 21) = 1.45$, $p = .243$, $R^2 = .06$). These findings demonstrate that, in its default and uncalibrated configuration, GPT-5 applies the shared rubric with extreme leniency and lacks the centralized calibration of certified human raters. The study cautions against deploying general-purpose Large Language Models as predictive or summative tools in high-stakes writing assessments, though it leaves open their potential utility for moderated, formative pedagogical feedback.

Keywords: Artificial Intelligence (AI); Automated essay scoring; GPT-5; MUET; Predictive validity; Writing

Abbreviations: Malaysian University English Test (MUET), Common European Framework of Reference for Languages (CEFR), Large Language Models (LLM)

INTRODUCTION

The Malaysian University English Test (MUET), administered by the Malaysian Examinations Council (MEC), is a pivotal high-stakes assessment of English proficiency for pre-university and undergraduate students [1]. This comprehensive examination evaluates a candidate's proficiency across four language domains: listening, speaking, reading, and writing [2]. Among these, the Writing component holds particular significance, contributing substantially to the overall band score. Specifically, the MUET Writing test assesses the candidate's capacity to present, organise, and articulate ideas, construct coherent arguments, and apply appropriate linguistic resources [3]. These skills are evaluated through a structured rubric calibrated in alignment with the Common European Framework of Reference for Languages (CEFR), ensuring a consistent and internationally recognised standard of assessment. Because MUET functions as a compulsory gatekeeping instrument for admission and progression in Malaysian public universities, the outcomes on this component carry high-stakes consequences for candidates [4], [5].

Given the mandatory nature of the examination, the immense volume of writing scripts generated each year makes manual hand-scoring highly resource-intensive and time-consuming [6], [7]. To address these workload bottlenecks, there has been a rapid emergence of automated essay scoring (AES) systems powered by Large Language Models (LLMs) [8]. Recent studies have demonstrated the potential of these AI language models to evaluate English language learner writing across CEFR proficiency groups with varying degrees of

accuracy and reliability [9], [10]. Consequently, large language models are already being utilised informally by lecturers and candidates to gauge likely MUET Writing performance.

However, the accuracy and validity of using general-purpose LLMs as evaluative tools remains a critical technical issue, particularly concerning rater bias and score variance [10]. Current literature indicates that where AI prompting is zero-shot and uncalibrated, model scores diverge systematically from those of trained human raters in both variance and severity [11], [10]. Uncalibrated zero-shot prompting frequently produces a severity or leniency mismatch, exhibiting inconsistent performance that renders it currently unsuitable for high-stakes assessment replacement [11]. Research has highlighted that while zero-shot models may achieve moderate adjacent agreement, they struggle with exact human-AI score matching; this exact-versus-adjacent gap can only be substantially narrowed through few-shot anchoring with exemplar scripts [12], [13].

Despite the growing integration of AI in educational settings, the empirical correspondence between zero-shot LLM evaluation and official human examiner outcomes has not been established in the Malaysian context. Addressing this gap, this study investigates the predictive validity, inter-rater reliability, and systematic scoring bias of GPT-5 against certified human MUET examiners. By doing so, it seeks to determine whether uncalibrated AI scoring can serve as a reliable proxy for official band outcomes in high-stakes, CEFR-aligned writing assessments.

literature review

While human raters provide valuable contextual judgment, the reliance on manual grading presents numerous challenges [14]. These challenges include the resource intensive nature of the process, its time-consuming demands, and inherent inter-rater variability [15]. In response to these limitations, the implementation of Automated Essay Scoring systems offers promising advantages in terms of efficiency, consistency, and cost effectiveness [16]. Furthermore, these systems enable educators to deliver targeted feedback, particularly in cases where student performance is marginal or ambiguous [17].

Previous research on Automated Essay Scoring in Malaysia has primarily concentrated on adapting machine learning based scoring systems to the specific requirements of the Malaysian University English Test. For instance, the Automated Essay Scoring Feedback prototype was developed to align directly with the rubrics of the examination. This prototype provided immediate scoring and formative feedback, utilizing the open source LightSIDE platform to create predictive models for the writing component [18], [17]. However, these systems focus primarily on concurrent validity by comparing automated scores to human generated scores for the same essays. Furthermore, they were developed prior to the advent of modern large language models and have not tested whether automated scores can predict actual performance in the official examination.

Recent advancements in large language models such as GPT-4 and GPT-5 have sparked growing interest in their potential. These models are now explored not only to perform human-like writing tasks but also to evaluate written text with a level of sophistication comparable to that of human raters [19], [9]. As these models continue to evolve, their capacity to provide reliable, consistent, and scalable evaluations presents a new direction in the assessment of academic writing. The integration of large language models into language testing could change how educational assessments are conducted and offer scalable responses to the challenges faced by traditional human grading systems [20].

Recent studies on GPT-3.5 and GPT-4 in Automated Essay Scoring show promising results, with substantial agreement between AI and human raters [21], [22]. For instance, GPT-4 has achieved Quadratic Weighted Kappa scores above 0.75, indicating high reliability [23]. Past research also highlights the importance of prompt design and rubric specificity for consistency in scoring [24]. In English as a Foreign Language context, GPT based scoring has shown comparable results to traditional Automated Writing Evaluation systems, such as those used for IELTS and TOEFL tasks. However, most studies are cross sectional, comparing AI scores with human scores concurrently. There is limited evidence on whether GPT can predict future performance in high stakes assessments.

To date, few published studies have investigated whether scores generated by large language models can reliably predict candidate performance in official examination bands, particularly when applied to pre-exam essays. Existing studies on Automated Essay Scoring for the Malaysian University English Test are limited to pre LLM technologies and focus on concurrent validity, which assesses agreement between human and machine generated scores for the same set of essays. Furthermore, international studies exploring the use of GPT for automated scoring have not examined the predictive validity of these models in high stakes contexts such as similar Malaysian assessments.

This study addresses a critical gap in the literature by evaluating the predictive validity of GPT-5 in scoring practice essays administered prior to the actual writing test. Holding task format and scoring rubric constant across scripts, the study aims to determine whether GPT-5 can serve as a reliable forecasting tool for

official exam performance. By investigating this predictive capacity, this research contributes to the growing body of knowledge on the application of AI in high stakes language assessments and offers practical insights into its potential as a tool for pre-exam performance prediction. Specifically, this research is guided by the following questions:

Research Question

What is the inter-rater reliability between GPT-5 and human examiners in scoring MUET writing?

Does GPT-5 exhibit significant scoring leniency compared to human benchmark scores?

How effectively can GPT-5 raw scores model and predict actual MUET candidate marks?

The corresponding null hypotheses are:

H01: There is no statistically significant inter-rater reliability between the writing marks generated by GPT-5 and the official marks awarded to the same candidates.

H02: GPT-5 does not exhibit a statistically significant scoring leniency or severity bias when compared to human benchmark scores.

H03: GPT-5 generated marks do not significantly predict official writing marks for the candidates.

METHODS

Research Design and Corpus

This study employed a quantitative research approach to analyse and predict selected participants MUET Writing band scores using artificial intelligence. A total of 23 undergraduate students were selected through purposive sampling. No a priori power analysis was conducted; the sample size was bounded by the number of eligible candidates and the selection was guided by specific inclusion criteria whereby participants were required to be active Universiti Sultan Azlan Shah students officially registered for the MUET 2025 examination. The resulting corpus consisted of 23 student essays encompassing both Task 1 Guided Writing and Task 2 Extended Writing genres, covering a diverse thematic distribution representative of the actual examination. An overview of the research was disclosed to the attending participants, and their informed consent was obtained prior to participation. Ethical approval for the study was granted by the relevant institutional review board.

Scoring Rubric and Human Rating Protocol

The student answer scripts were evaluated using the official 90-mark MUET analytic rubric. The assessment criteria covered key domains of the MUET Writing component, specifically Task Fulfilment, Structure, Lexis, and Organisation. The scoring framework allocated a maximum of 30 marks for Task 1 and 60 marks for Task 2, combining for the total 90 marks. Content validity of the applied criteria was established through expert review by experienced MUET examiners from the Malaysian Examinations Council. For the human rating protocol, the baseline benchmark scores utilized in this study were the official MUET Writing results obtained directly from the examination portal. These official scores were awarded by certified and well trained MUET examiners operating under standard benchmarking procedures, ensuring high reliability of the criterion measure.

AI Prompting Pipeline and Procedures

The students handwritten answer scripts were collected, scanned, and converted into digital formats. To realise the first research objective, these files were uploaded to GPT-5. The AI evaluation utilized a strict zero shot prompting pipeline. The exact system prompt instructed the model to act as a certified MUET examiner and grade the scripts based on the provided criteria. The prompt embedded the specific rubric input structure detailing the bands for Task Fulfilment, Language, and Organisation. Temperature settings were held constant at a default analytical level to ensure deterministic and reproducible scoring behavior across all scripts without the use of few shot exemplars. The generated data, consisting of the predicted MUET scores and detailed analytical feedback, were recorded in a spreadsheet to assist with the second research objective. These AI generated scores were then compared with the official human marked results to evaluate scoring alignment.

Data Analysis

The collected quantitative data were processed and analysed using statistical software to evaluate the predictive validity and reliability of the GPT-5 scoring pipeline. Descriptive statistics including means, standard deviations, minimum, and maximum values were calculated to summarise both the AI predicted scores and the actual human assigned scores. Normality of the score distributions and paired differences was examined using the Shapiro Wilk test. To determine whether a significant severity bias existed between the

two evaluation methods, a paired samples t test was conducted, supported by a Wilcoxon signed rank test for robust confirmation. Furthermore, inter-rater reliability between GPT-5 and human examiners was assessed using Pearson and Spearman correlations alongside a two-way mixed effect, absolute agreement, single measures intraclass correlation coefficient. Finally, a simple linear regression analysis was performed to test the predictive power of the AI scores in forecasting official MUET examination results. Effect sizes were reported alongside significance tests, utilizing Cohen d for the mean difference and R squared for the regression model.

RESULTS AND DISCUSSION

Descriptive Statistics and Normality Testing

The descriptive statistics for the scores awarded by certified human examiners and the zero-shot GPT-5 evaluation across the N = 23 student essays on the 90-mark Malaysian University English Test (MUET) writing rubric are presented in Table 1.

Table 1. Descriptive statistics for MUET Writing scores

Evaluator	N	Mean (M)	Std. Deviation (SD)	Minimum	Maximum
Human Examiner	23	36.96	7.41	22.00	56.00
GPT-5	23	60.48	13.77	33.00	80.00

Certified human raters demonstrated a constrained score distribution spanning 22.00 to 56.00 (M = 36.96, SD = 7.41). In contrast, GPT-5 produced a broader score distribution spanning 33.00 to 80.00 (M = 60.48, SD = 13.77).

Prior to inferential testing, normality was assessed using the Shapiro-Wilk test, as summarized in Table 2.

Table 2. Shapiro-Wilk test of normality

Variable	Statistic (W)	df	Significance (p)
Human Score	.956	23	.381
GPT-5 Score	.955	23	.368
Paired Difference Score	.950	23	.299

Both human scores (W = .956, p = .381) and GPT-5 scores (W = .955, p = .368) met the assumption of normality (p > .05). Furthermore, the paired difference scores met the normality assumption required for paired parametric analysis (W = .950, p = .299).

RQ1: Inter-Rater Reliability between GPT-5 and Human Examiners

To address Research Question 1, Bivariate Correlation Coefficients and an Intraclass Correlation Coefficient (ICC) were computed to evaluate both rank consistency and absolute agreement across raters. The statistical output is displayed in Table 3.

Table 3. Inter-rater reliability and agreement analysis

Metric	Value	95% Confidence Interval	Significance (p)
Pearson Correlation (r)	-.254	—	.243
Spearman Rank Correlation (p)	-.225	—	.303
Intraclass Correlation (ICC 3,1, Single Measures)	-.066	[-.190, .166]	.840
Intraclass Correlation (ICC 3, k, Average Measures)	-.141	[-.470, .284]	.840

Bivariate Pearson correlation analysis indicated a weak, negative, and non-significant linear relationship between human examiner scores and GPT-5 ratings, $r(21) = -.254$, $p = .243$. The non-parametric Spearman rank correlation confirmed this lack of monotonic association, $p(21) = -.225$, $p = .303$.

Furthermore, the single-measures absolute agreement Intraclass Correlation Coefficient was negative and non-significant, ICC (3,1) = -.066, 95% CI [-.190, .166], $F(22, 22) = 0.650$, $p = .840$. These results demonstrate an absence of inter-rater reliability, indicating that GPT-5 ratings did not align with or maintain the proficiency ranking established by certified examiners. Consequently, the null hypothesis H01 is retained.

RQ2: Scoring Leniency and Severity Bias

To address Research Question 2, a paired-samples t-test was conducted alongside an error-direction analysis to determine whether GPT-5 exhibited systematic severity or leniency bias. The comparative results are presented in Table 4 and Table 5.

Table 4. Paired-Samples t-Test comparing GPT-5 and human benchmark scores

Evaluator Pair	Mean Diff.	SD	SE Mean	95% CI Lower	95% CI Upper	t	df	p (two-tailed)	Cohen's d
GPT-5 – Human	23.52	17.22	3.59	16.08	30.97	6.552	22	< .001	1.37

Table 5. Directional breakdown of AI scoring discrepancies

Error Classification	Operational Definition	Number of Cases (n)	Percentage (%)
Overprediction	GPT-5 Score > Official Human Score	19	82.61%
Underprediction	GPT-5 Score < Official Human Score	2	8.70%
Exact Agreement	GPT-5 Score = Official Human Score	2	8.70%

The paired-samples t-test revealed a statistically significant scoring difference between the evaluators, $t(22) = 6.552$, $p < .001$. GPT-5 awarded marks that were on average 23.52 points higher than those assigned by certified examiners (95% CI [16.08, 30.97]). This difference represents a large effect size (Cohen's $d = 1.37$).

As detailed in Table 5, GPT-5 overpredicted candidate marks in 82.61% of all cases ($n = 19$), while producing exact score matches in only two instances (8.70%). Because a statistically significant upward leniency bias was observed, the null hypothesis $H02$ is rejected.

RQ3: Simple Linear Regression Model

To address Research Question 3, a simple linear regression analysis was performed to evaluate whether raw zero-shot GPT-5 scores could predict official human examiner marks. The regression coefficients and model summary are detailed in Table 6.

Table 6. Simple linear regression predicting human MUET scores from GPT-5 ratings

Predictor	Unstandardize d B	Std. Error (SE)	t	Significance (p)	95% CI Lower	95% CI Upper
Constant	45.22	7.04	6.424	< .001	30.58	59.86
GPT-5 Score	-0.14	0.11	-1.202	.243	-0.37	0.10

The simple linear regression model was not statistically significant, $F(1, 21) = 1.446$, $p = .243$. GPT-5 scores accounted for 6.4% of the variance in official human scores ($R^2 = .064$, Adjusted $R^2 = .020$), leaving 93.6% of the variance unexplained. The slope coefficient was non-significant ($B = -0.14$, $SE = 0.11$, $t = -1.202$, $p = .243$). Consequently, uncalibrated zero-shot GPT-5 marks failed to predict official examination outcomes, leading to the retention of the null hypothesis $H03$.

Discussion

The empirical findings of this study offer practical insights into the capabilities and operational limitations of utilizing zero-shot Large Language Model prompting for automated writing evaluation in the high-stakes context of the Malaysian University English Test (MUET).

Inter-Rater Reliability and Rank Consistency (RQ1)

The analysis for Research Question 1 indicated an absence of systematic inter-rater reliability between GPT-5 and certified MUET examiners ($r = -.254$, $ICC = -.066$). This finding stands in contrast to prior automated essay scoring investigations, such as those by Hackl et al. [21] and Chen et al. [23], which reported high correlation and agreement metrics ($QWK > 0.75$) when utilizing earlier LLM iterations.

This divergence is largely attributable to methodological differences in prompt calibration and validation context. Whereas earlier AES prototypes in Malaysia, such as the LightSIDE-based AESF system [18], [17], were trained and evaluated on concurrent scoring of identical essay sets, this study evaluated zero-shot automated scoring against independent, official examination outcomes. In uncalibrated zero-shot environments, the model lacks task-anchored benchmark examples, resulting in ranking drift relative to human evaluators.

Scoring Variance and Severity Bias (RQ2)

The results for Research Question 2 demonstrated a significant upward scoring leniency by GPT-5 ($M_{diff} = 23.52$, $d = 1.37$), with overprediction occurring in 82.61% of scripts. Furthermore, GPT-5 produced a wide score spread (33.00 to 80.00) compared to the more conservative range of certified human examiners (22.00 to 56.00).

This score dispersion difference reflects key variations in rating behavior between generative AI models and human raters. LLMs evaluate text primarily by processing surface lexical richness, syntactic complexity, and global grammatical fluency. When presented with structurally organized scripts containing advanced vocabulary, GPT-5 tends to award high marks across the analytic dimensions of Task Fulfilment and Structure, Lexis, and Organisation. Conversely, it penalizes shorter or simpler texts more heavily.

In contrast, certified MUET examiners undergo rigorous standardization and benchmarking training. Human examiners adhere to centralized banding norms and focus on nuanced constructs such as argument development, relevance, and pragmatic appropriateness, which prevent extreme score distribution. This

severity mismatch corroborates the findings of Yamashita [10] and Shermis [11], who observed that zero-shot LLM evaluation exhibits considerable rater bias and score variance.

Predictive Modeling Limitations (RQ3)

Regarding Research Question 3, linear regression confirmed that uncalibrated zero-shot GPT-5 scores did not significantly predict official MUET Writing performance ($R^2 = .064$, $p = .243$). This lack of predictive capacity supports recent empirical studies on LLM rating mechanics.

As demonstrated in recent literature, zero-shot prompting without model fine-tuning or few-shot exemplar anchoring produces substantial drift from human scoring baselines. While few-shot prompting with calibrated anchor scripts has been shown to reduce rater distance and increase score consensus, uncalibrated zero-shot prompting yields unstandardized outputs that fail to track candidate performance on high-stakes criteria.

Pedagogical Utility Versus High-Stakes Deployment

Given that, the MUET serves as a gatekeeping examination for university admissions and academic placement [4], [5], accuracy in score prediction is critical. The substantial upward bias and low predictive validity observed here indicate that zero-shot GPT-5 is currently unsuited for unmoderated summative grading or high-stakes score forecasting. Presenting uncalibrated LLM outputs as definitive indicators of exam readiness could introduce misleading expectations regarding a candidate's projected band level.

Nevertheless, these findings do not negate the pedagogical utility of Large Language Models within formative learning contexts. When utilized as a formative diagnostic tool, GPT-5 can assist educators and learners by providing real-time qualitative commentary on drafting mechanics, paragraph cohesion, and vocabulary usage [16], [17]. Deploying LLMs formatively allows students to receive immediate feedback to support revision, while preserving final summative judgment for certified human examiners operating within standardized scoring frameworks.

For Malaysian English language educators, these findings continue to underscore the importance of maintaining human-centred assessment practices for high-stakes examinations. GPT-5, in its current configuration, cannot replace the nuanced judgement of experienced MUET raters.

CONCLUSION

This study evaluated whether automated writing evaluation generated by GPT-5 could reliably predict official Malaysian University English Test outcomes. The empirical evidence indicates that, in an uncalibrated configuration, the language model cannot serve as a reliable substitute for certified human examiners. The model awarded scores that were 23.52 points higher than official examiner marks, overpredicting outcomes for more than 80% of the candidates. Furthermore, the automated scores demonstrated no measurable agreement with the benchmark human ratings and accounted for only 6.4% of the variance in official examination results.

These findings must be interpreted within the context of several methodological constraints. The sample comprised 23 volunteers from a single institution, which restricts statistical power and limits the generalizability of the regression analysis. Additionally, the automated evaluation was conducted on practice essays written under non examination conditions responding to prompts that differed from the official test, potentially confounding rater effects with task conditions. The reliance on a single processing iteration over scanned handwritten text leaves potential variations and legibility issues unquantified. Finally, while the scoring rubric was reviewed by experienced examiners, it was an adaptation rather than the secure operational instrument utilized by the Malaysian Examinations Council, meaning a portion of the observed leniency might stem from this rubric adaptation.

Despite these constraints, the results provide clear implications for educational practice and research. For classroom instruction, the findings suggest that raw numerical scores generated by artificial intelligence should not be presented to students as accurate estimates of their likely examination band. Instead, a more defensible application is formative deployment, where educators use the model to generate descriptive commentary on language use, organization, and task fulfilment, which lecturers moderate before sharing with candidates. Future research should prioritize investigating whether prompt calibration through anchor scripts, few shot exemplars, or model fine tuning on human marked scripts can align automated outputs with operational standards. Such studies should be executed on larger, multi-institution samples under strict examination conditions to further validate the role of automated scoring in high stakes assessment.

REFERENCES

- [1] Malaysian Examinations Council, "Malaysian University English Test (MUET) – Test specification & regulation," https://www.mpm.edu.my/images/dokumen/calonpeperiksaan/muet/regulation/Test_Specification_Regulation.pdf, 2019.

- [2] S. M. B. B. V. K. S. V. Karnine, A. S. D. Preece, I. B. S. Ahmad, and S. S. B Muhammad, "A study on difficulties encountered and perception by English as Second Language (ESL) learners in Malaysian University Examination Test (MUET)," *International Academic Symposium of Social Science* 2022. doi: 10.3390/proceedings2022082053, 2022.
- [3] B. H. Voon, A. K. Teo, and J. E. C Voon, "Improving the performance of MUET students in report-writing: From experiment to actions," *Asian Journal of University Education*, vol. 15, no. 3, 175. doi: 10.24191/ajue.v15i3.7568, 2019.
- [4] N. Mahmud, N. Lateh, N. Mahmud, A. A. Hassan, A. Mohamed, and S. A Ahmad Tarmizi, "Washback impact of the MUET: The before and after effect of a high-stake university English test in Malaysia," *International Journal of Learning, Teaching and Educational Research*, vol. 20, no. 8, pp. 1–17, 2021.
- [5] S. A. Rethinasamy and K. M Chuah, "The Malaysian University English Test (MUET) and its use for placement purposes: A predictive validity study," *Electronic Journal of Foreign Language Teaching*, vol. 8, no. 2, pp. 234–245, 2011.
- [6] D. Ramesh and S. K Sanampudi, "An automated essay scoring system: A systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527. doi: 10.1007/s10462-021-10068-2, 2022.
- [7] W. Xu, R. Mahmud, and W. L Hoo, "A systematic literature review: Are automated essay scoring systems competent in real-life education scenarios? IEEE Access, 12, 77639–77657," doi: 10.1109/ACCESS.2024.3399163, 2024.
- [8] A. Mizumoto and M Eguchi, "Exploring the potential of using an AI language model for automated essay scoring," *Research Methods in Applied Linguistics*, vol. 2, no. 2, 100050. doi: 10.1016/j.rmal.2023.100050, 2023.
- [9] A. Pack, A. Barrett, and J Escalante, "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability," *Computers and Education: Artificial Intelligence*, 6, Article 100234. doi: 10.1016/j.caeai.2024.100234, 2024.
- [10] T Yamashita, "An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0," *Computers and Education: Artificial Intelligence*, 7. doi: 10.1016/j.rmal.2024.100133, 2024.
- [11] M. D Shermis, "Using ChatGPT to score essays and short-form constructed responses," *Assessing Writing*, 66, 100988. doi: 10.1016/j.asw.2025.100988, 2025.
- [12] Y. Liu, H. Qi, and X Lu, "Enhancing GPT-based automated essay scoring: the impact of fine-tuning and linguistic complexity measures," *Computer Assisted Language Learning*, 1–20. doi: 10.1080/09588221.2025.2518430, 2025.
- [13] R. Jin, M. Zhao, C. Niu, Y. Xia, H. Zhou, and N Liu, "Evaluating the performance of ChatGPT and Claude in automated writing scoring: Insights from the Many-facet Rasch model," *Education and Information Technologies*, vol. 30, no. 18, pp. 25881–25908. doi: 10.1007/s10639-025-13774-4, 2025.
- [14] Messer, M., Neil, Kölling, M., & Shi, M, "How consistent are humans when grading programming assignments? [Preprint]," *arXiv*. doi: 10.48550/arxiv.2409.12967, 2024.
- [15] A. O. Ugbari, C. Ndeekor, and E Wobidi, "Optimizing GPT-4 for automated short answer grading in educational assessments," *International Journal of Computer Applications*, vol. 187, no. 21, pp. 32–36. doi: 10.5120/ijca2025925255, 2025.
- [16] K. Bouziane and A Bouziane, "AI versus human effectiveness in essay evaluation," *Discover Education*, vol. 3, no. 1. doi: 10.1007/s44217-024-00320-6, 2024.
- [17] W. S. Wong and C. H Bong, "Assessing Malaysian University English Test (MUET) essay on language and semantic features using Intelligent Essay Grader (IEG)," *Pertanika Journal of Science and Technology*, vol. 29, no. 2. doi: 10.47836/pjst.29.2.12, 2021.
- [18] W. S. Wong and C. H Bong, "A study for the development of Automated Essay Scoring (AES) in Malaysian English test environment," *International Journal of Innovative Computing*, vol. 9, no. 1. doi: 10.11113/ijic.v9n1.220, 2019.
- [19] M Lundgren, "Large language models in student assessment: Comparing ChatGPT and human graders [Preprint]," *arXiv*. doi: 10.48550/arxiv.2406.16510, 2024.
- [20] H. Xu, W. Gan, Z. Qi, J. Wu, and P. S Yu, "Large language models for education: A survey [Preprint]," *arXiv*. doi: 10.48550/arxiv.2405.13001, 2024.
- [21] V. Hackl, A. E. Müller, M. Granitzer, and M Sailer, "Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings," *Frontiers in Education*, 8, Article 1272229. doi: 10.3389/feduc.2023.1272229, 2023.
- [22] A. C. Uyar and D Büyükhüska, "Artificial intelligence as an automated essay scoring tool: A focus on ChatGPT," *International Journal of Assessment Tools in Education*, vol. 12, no. 1, pp. 20–32. doi: 10.21449/ijate.1517994, 2025.
- [23] X. Chen, Z. Zhou, and M Prado, "ChatGPT-3.5 as an automatic scoring system and feedback provider in IELTS exams," *International Journal of Assessment Tools in Education*, vol. 12, no. 1, pp. 62–77. doi: 10.21449/ijate.1496193, 2025.
- [24] B. Quah, L. Zheng, T. Jie, C. W. Yong, and I Islam, "Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations," *BMC Medical Education*, vol. 24, no. 1, Article 5881. doi: 10.1186/s12909-024-05881-6, 2024.