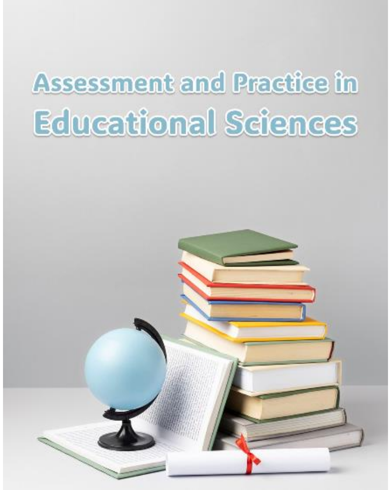
Assessment and Practice in Educational Sciences



# Calibrating Generative AI for Second Language Writing Assessment: Combining Statistical Validation with Prompt Design

## ABSTRACT

Generative artificial intelligence (GenAI) is emerging as a powerful tool in second language writing assessment, offering the potential for rapid, consistent, and scalable evaluation. However, its value depends on whether its scoring reflects the nuanced judgments of experienced human raters. This study introduces the concept of calibration in the context of second language writing assessment, defined as the deliberate and iterative refinement of AI prompts, guided by statistical evidence, to align AI scoring with human evaluative reasoning. 60 essays produced by 30 upper intermediate learners of English were evaluated independently by an experienced human rater and by ChatGPT 3.5, using the English for Academic Purposes (EAP) Writing Assessment Rubric. Statistical analyses assessed inter rater agreement, score consistency, and systematic bias. In the initial baseline stage, ChatGPT 3.5 tended to act as a strict marker, applying the rubric literally and assigning lower scores than the human rater. Across three calibration stages, which included clarifying rubric descriptors, refining interpretive guidance, and incorporating representative scoring examples, the AI scoring moved closer to the human benchmark. Agreement improved from a Cohen’s kappa of 0.52 to 0.89, correlation from .76 to .94, and the mean score difference narrowed from -2.45 to - 0.95, the latter no longer statistically significant. Qualitative analysis showed a shift from a narrow emphasis on surface errors to a more balanced consideration of accuracy, organization, development, and communicative effectiveness. The results suggest that calibration offers a replicable and evidence-based approach to integrating generative AI into second language writing assessment, enhancing the fairness, validity, and reliability of AI assisted evaluation.

**Keywords:** Generative artificial intelligence, second language writing assessment, calibration, prompt design, inter-rater reliability, statistical validation

1. Reza. Farzi: Official Languages and Bilingualism Institute, University of Ottawa, Ottawa, Canada (Email: rfarzi@uottawa.ca)

Article type:
Original Research

Article history:
Received 06 September 2024
Revised 16 November 2024
Accepted 24 November 2024
Published online 10 December 2024

**How to cite this article:**
Farzi, R. (2024). Calibrating Generative AI for Second Language Writing Assessment: Combining Statistical Validation with Prompt Design. *Assessment and Practice in Educational Sciences*, *2*(4), 1-13. https://doi.org/10.61838/japes.2.4.7

## Introduction

In recent years, the integration of artificial intelligence (AI) into educational contexts has become one of the most transformative developments in language assessment. Among the wide range of applications, automated essay scoring (AES) has attracted substantial attention, as it offers the potential to enhance scalability, efficiency, and consistency in evaluating second language (L2) writing. While early AES systems primarily relied on statistical correlations between surface-level features and human scores, the advent of large language models (LLMs) and generative AI has significantly expanded the scope of automated writing evaluation. These models are capable not only of identifying grammatical and lexical errors but also of engaging with discourse-level features such as coherence, argumentation, and rhetorical appropriateness (1-4).

The increasing reliance on generative AI in assessment raises critical questions about validity, fairness, and alignment with human evaluative reasoning. L2 writing assessment has long been guided by fundamental principles such as validity, reliability, and fairness (5, 6). Validity refers to whether a scoring procedure genuinely measures the intended construct, while reliability reflects the consistency of scores across raters and contexts. Fairness requires equitable treatment of learners from diverse linguistic and cultural backgrounds (7). Human raters typically balance micro-level accuracy with macro-level discourse qualities, weighing communicative effectiveness, argument development, and rhetorical structure. However, uncalibrated AI systems often demonstrate a tendency to prioritise measurable surface features such as grammar and vocabulary, risking construct underrepresentation and construct-irrelevant variance (8, 9).

The roots of AES can be traced back to early computational approaches such as Page's Project Essay Grade (PEG), which applied regression-based algorithms to predict human scores from surface linguistic features (10). Later systems such as e-rater, developed by ETS, incorporated natural language processing (NLP) techniques to analyse a broader range of features, including lexical diversity, syntactic variety, and discourse markers (11, 12). Despite these advances, many of the early systems faced criticism for construct underrepresentation, as they often reduced complex writing ability to countable surface features (13, 14).

Human rater studies highlighted variability in essay scoring, where differences in rubric interpretation and rater experience could yield inconsistent judgments (15, 16). Analytic rubrics were introduced to address this challenge by dividing performance into distinct categories such as accuracy, organization, and effectiveness, thus providing a structured framework that minimized the risk of one salient feature dominating the evaluation (17, 18). Still, even with rubrics, ensuring fairness and consistency remained difficult, particularly in large-scale testing contexts where rater fatigue and subjectivity were unavoidable. These limitations motivated ongoing efforts to design automated systems capable of supporting or replacing certain rater functions.

A decisive shift occurred with the introduction of neural architectures, particularly recurrent neural networks (RNNs) and later transformer-based models. Attention mechanisms revolutionized sequence modelling by enabling systems to capture global dependencies across texts (19, 20). This innovation facilitated the development of LLMs capable of processing entire essays holistically rather than segmenting them into isolated units. Such advances allowed for more nuanced analyses of discourse structure, argument progression, and rhetorical alignment (3, 4).

Generative AI further expanded the capabilities of AES by moving beyond error detection to producing human-like evaluative commentary. Unlike earlier models that merely produced numeric scores, LLMs such as GPT-based systems could interpret rubrics, provide explanations, and generate formative feedback. This mirrors, to some extent, the reflective process of skilled human raters, who not only assign scores but also articulate rationales and suggestions for improvement (1, 2). However, the degree to which generative AI aligns with human evaluative reasoning depends heavily on how prompts are constructed and calibrated.

Prompt design has emerged as a key determinant of AI scoring behaviour. When provided with rubric descriptors alone, AI models often apply them rigidly, disproportionately penalising minor surface errors while undervaluing discourse-level qualities. Research shows that including interpretive guidance and performance anchors in prompts can significantly improve alignment between AI and human raters (3, 21). For example, specifying that minor grammatical errors should not outweigh strong organization or argumentation can shift AI scoring toward more construct-representative judgments. Similarly, the inclusion of exemplar responses allows AI systems to internalise performance bands, paralleling the use of anchor scripts in human rater training (14, 15).

Calibration, in this context, refers to the iterative refinement of AI prompts informed by statistical validation. This approach draws from established traditions in test equating and measurement calibration (22, 23). Through successive adjustments—clarifying rubric descriptors, adding interpretive guidance, and integrating exemplars—AI performance can be brought into closer convergence with human rater standards. Statistical tools such as Cohen's kappa and Pearson's correlation are particularly useful in quantifying improvements in inter-rater agreement and reliability (23, 24).

Beyond technical performance, fairness and validity remain central concerns. Scholars argue that uncalibrated AES systems may inadvertently introduce bias by over-penalising learners from certain linguistic backgrounds or by rewarding superficial linguistic complexity at the expense of genuine communicative effectiveness (7, 25). Addressing these challenges requires not only robust prompt engineering but also systematic validation against diverse populations and tasks. Construct validity must be safeguarded to ensure that AI scoring reflects the multifaceted nature of writing ability rather than narrow proxies such as error counts or lexical sophistication (5, 8).

Teacher perspectives further highlight the importance of balancing efficiency with pedagogical relevance. Research on teacher feedback indicates that while students benefit from corrective input, they also value formative comments that emphasize higher-order features like idea development, coherence, and rhetorical strategies (9). Thus, AES systems that provide only mechanistic evaluations risk alienating learners and undermining writing pedagogy. Generative AI, when carefully calibrated, has the potential to bridge this gap by combining analytic scoring with formative, human-like commentary (1, 13).

Despite promising advances, challenges remain in fully integrating generative AI into L2 writing assessment. First, the dynamic nature of LLMs—frequently updated and modified—means that calibration must be an ongoing process rather than a one-time adjustment (26). Second, transparency and interpretability are crucial for building trust among educators, students, and testing agencies. Unlike traditional psychometric models, which operate under clear statistical assumptions, LLMs function as "black boxes," making it essential to document calibration processes and validation outcomes for accountability (1, 2). Third, questions remain about the role of AES in high-stakes assessment. While calibrated AI may be suitable for formative assessment or large-scale preliminary scoring, many scholars caution against relying solely on automated systems for consequential decisions, given their lack of socio-cultural awareness (6, 7).

At the same time, opportunities abound. Generative AI calibrated through prompt design can serve as a reliable support tool for teachers, alleviating workload while ensuring consistency. In formative contexts, AI feedback can empower learners by providing immediate, rubric-aligned commentary that encourages reflection on both strengths and weaknesses (9, 17). Moreover, AI can facilitate large-scale assessments where human resources are limited, offering institutions a cost-effective means of maintaining reliability without sacrificing validity (13, 21).

Despite rapid advances, the literature reveals a clear research gap: while many studies have examined overall AES accuracy and correlations with human ratings, fewer have systematically explored the process of calibration itself. Most investigations treat AI scoring behaviour as fixed rather than as a malleable output shaped by prompt design and statistical validation. There is limited empirical evidence on how iterative calibration can guide AI toward more human-like evaluative reasoning (1, 2). Furthermore, while psychometric traditions in test equating and fairness provide well-established frameworks (7, 22), their integration into prompt engineering for LLM-based AES remains underexplored.

The present study addresses this gap by introducing calibration as a formal, replicable procedure for aligning generative AI with human evaluative standards in second language writing assessment.

## Methods and Materials

The methods for this study were designed to provide a fair and controlled test of whether carefully refining AI prompts can bring its scoring patterns closer to those of an experienced human rater. Rather than altering multiple elements at once, the study kept the participants, tasks, and scoring rubric constant across all stages so that any changes in agreement could be traced directly to adjustments in the AI's instructions. What follows is a detailed account of who took part in the study, the writing tasks they completed, the rubric used to evaluate their work, and the procedures followed in both human and AI scoring. The section also explains the three stages of calibration that formed the core of the experiment, as well as the statistical analyses used to assess their effects.

### *Participants*

The study involved 30 undergraduate students enrolled in an advanced English as a Second Language (ESL) course at a Canadian university. The course was designed for upper-intermediate to advanced learners preparing for academic study in English-medium programs. Participants represented a broad range of linguistic and cultural backgrounds, with first languages (L1) including Mandarin (n = 17), French (n = 4), Cantonese (n = 2), Arabic (n = 2), Spanish (n = 2), Japanese (n = 1), Somali (n = 1), and Dari (n = 1).

English proficiency for all students was institutionally established at the B2 level on the Common European Framework of Reference for Languages (CEFR), based on the university's placement testing system. This uniform proficiency profile ensured that the study examined the alignment between AI-generated and human-assigned scores without the results being skewed by major differences in linguistic competence (Eckes, 2015).

### *Writing Tasks*

Each participant produced two essays under timed, in-class conditions over the course of the semester. The essay prompts were academic in nature and reflected topics frequently found in higher-education proficiency testing contexts: 1) *The role of technology in higher education*, and 2) *balancing environmental protection with economic growth*. Both topics were chosen to elicit writing that integrates multiple dimensions of academic literacy, including control of lexico-grammatical resources, discourse organization, rhetorical development, and the integration of argument and evidence (Hamp-Lyons, 1991; Hyland, 2016). The 60-minute time limit for each essay and a target length of 350–450 words ensured comparability across submissions and reduced the risk of external assistance (Bachman & Palmer, 2022).

### *Writing Evaluation Rubric*

Essays were evaluated using the institutionally developed English for Academic Purposes (EAP) Writing Assessment Rubric (See appendix A), an analytic scoring scale comprising four dimensions: Lexical Accuracy and Range, Grammatical Accuracy and Range, Organization and Development, and Overall Effectiveness. Each dimension was scored on a five-point scale, with descriptors specifying the performance characteristics at each band. Analytic rubrics were selected because they provide greater diagnostic detail than holistic scales, allow independent weighting of subskills, and facilitate targeted feedback (Weigle, 2002; Brookhart, 2018). The rubric had undergone prior institutional validation for both reliability and construct alignment, ensuring that its categories reflected established models of L2 writing ability.

### *Human Rating*

The human benchmark for this study was provided by the course professor, an experienced language assessment specialist with more than a decade of expertise in second language (L2) writing evaluation. In addition to extensive experience in scoring essays for both classroom and high stakes contexts, the rater had long standing familiarity with the EAP Writing Assessment Rubric, having used it in previous instructional and assessment settings. This dual perspective as both an assessor and the instructor for the participating cohort offered an additional layer of construct validity, as the rater was deeply familiar with the curriculum, instructional goals, and the genre expectations of the assigned tasks.

All essays were scored independently, without reference to AI output or prior scores, to ensure that ratings reflected the rater's own interpretation of the rubric criteria. For each of the four analytic categories (Lexical Accuracy and Range, Grammatical Accuracy and Range, Organization and Development, and Overall Effectiveness) the rater assigned a numerical score on the five-point scale and provided a brief written rationale. These rationales, typically two to three sentences in length, highlighted the strengths and weaknesses most relevant to the assigned score and often drew attention to aspects such as the clarity of argument, precision of vocabulary, complexity of grammar, or structural cohesion.

The choice to use a single rater rather than a panel followed established protocols in AI and human scoring alignment studies (Ramineni and Williamson, 2013). This approach maximises intra rater consistency and avoids the variability that can emerge when multiple human raters bring differing interpretations of the same rubric. In the context of an exploratory calibration study, high consistency from the human benchmark is essential for identifying specific areas where the AI's scoring diverges from human reasoning.

### *AI Rating*

The AI scoring was performed using ChatGPT 3.5, accessed via the OpenAI API in April 2024. The decision to use the API version rather than the web interface allowed for controlled and repeatable prompt delivery, ensuring that the same instructions were presented for every essay. The model was provided with the full text of the EAP Writing Assessment Rubric and tasked with assigning a score for each of the four analytic categories. For each category, it was also required to justify the assigned score explicitly by referencing the relevant rubric descriptors and to offer one or two targeted improvement suggestions tailored to the specific weaknesses identified.

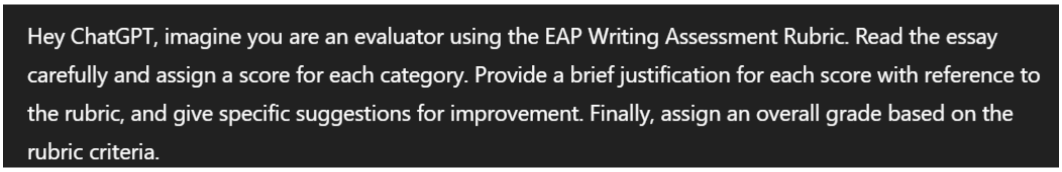
In the baseline stage, the AI received the rubric and only minimal scoring instructions. It was instructed to evaluate the essay according to the rubric, assign scores for each category, and provide brief feedback. This absence of interpretive guidance was intentional. It allowed the study to capture the AI's uncalibrated scoring behaviour, meaning how it would apply the rubric based solely on its own interpretation of the descriptors without any additional human mediated elaboration.

Initial observations at this stage revealed that the AI tended to interpret the descriptors in a rigid and literal manner, often focusing disproportionately on surface level linguistic features such as grammar and vocabulary accuracy. While these are essential components of writing quality, the AI's emphasis sometimes came at the expense of higher order dimensions like argument structure, logical progression, and rhetorical appropriateness. In some cases, relatively minor language errors led to disproportionately low scores, even when the essay demonstrated strong discourse organization and coherent argumentation. This tendency highlighted the need for a systematic calibration process to guide the AI toward a more balanced and construct aligned application of the rubric.

### *Calibration Stages*

The calibration process followed a structured, iterative progression aimed at reducing discrepancies between AI-generated scores and those of the human rater. It drew from principles of rater training in language assessment, where clarity of scoring criteria and exposure to performance exemplars enhance reliability and validity.

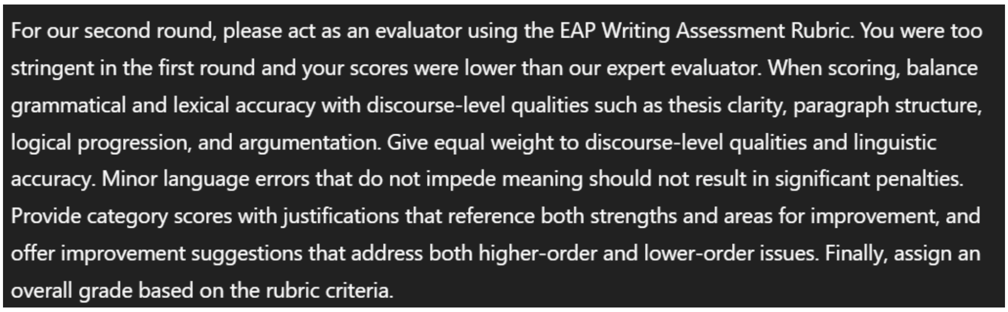
Stage 1: Baseline: In the first stage, the unmodified rubric prompt was presented to the AI, as shown in Figure 1. At this stage, the AI interpreted and applied rubric descriptors without any additional guidance beyond their literal wording. This provided a baseline measure of how the model understood the rubric in its raw form, allowing the study to identify systematic biases or scoring tendencies. Preliminary results indicated that the AI applied stricter penalty patterns for grammatical and lexical errors and placed comparatively less emphasis on discourse-level qualities.



Hey ChatGPT, imagine you are an evaluator using the EAP Writing Assessment Rubric. Read the essay carefully and assign a score for each category. Provide a brief justification for each score with reference to the rubric, and give specific suggestions for improvement. Finally, assign an overall grade based on the rubric criteria.

**Figure 1. Baseline AI Scoring Prompt (Stage 1)**

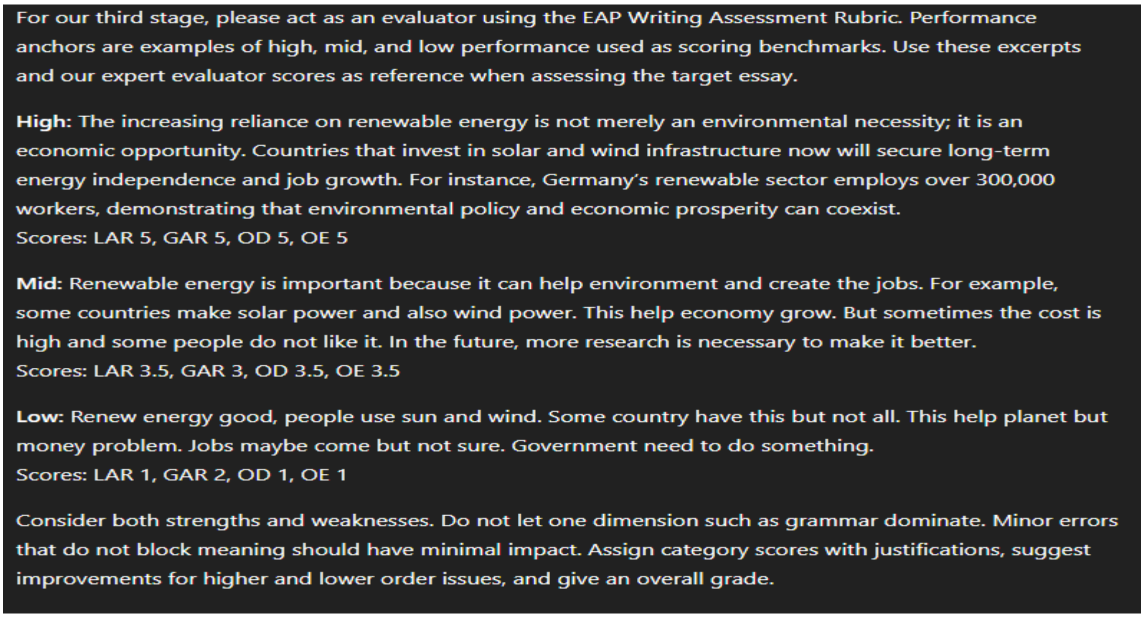
Stage 2: Clarified Descriptors: In the second stage, additional interpretive guidance was embedded within the prompt, as shown in Figure 2. The added instructions made explicit that minor errors, when meaning remained clear, should not result in significant penalties. The guidance also instructed the AI to give equal weight to discourse-level qualities such as logical progression, argumentation, and organization alongside linguistic accuracy. This stage aimed to bring the AI's evaluative focus closer to that of experienced human raters, who often make compensatory judgments when strong discourse qualities offset minor linguistic flaws.



For our second round, please act as an evaluator using the EAP Writing Assessment Rubric. You were too stringent in the first round and your scores were lower than our expert evaluator. When scoring, balance grammatical and lexical accuracy with discourse-level qualities such as thesis clarity, paragraph structure, logical progression, and argumentation. Give equal weight to discourse-level qualities and linguistic accuracy. Minor language errors that do not impede meaning should not result in significant penalties. Provide category scores with justifications that reference both strengths and areas for improvement, and offer improvement suggestions that address both higher-order and lower-order issues. Finally, assign an overall grade based on the rubric criteria.

**Figure 2. Clarified Descriptors AI Scoring Prompt (Stage 2)**

Stage 3: Performance Anchors: In the final stage, representative examples of high, mid, and low performance were provided for each rubric category, as shown in Figure 3 (see appendix B for the full prompt language). These performance anchors illustrated concrete instances of what different score levels look like in practice. This approach mirrored human rater training protocols, where exposure to benchmark scripts fosters more consistent application of scoring criteria. The prompt also explicitly reminded the AI to consider both strengths and weaknesses within each performance and to avoid allowing one dimension, such as grammatical accuracy, to disproportionately influence the overall score.



For our third stage, please act as an evaluator using the EAP Writing Assessment Rubric. Performance anchors are examples of high, mid, and low performance used as scoring benchmarks. Use these excerpts and our expert evaluator scores as reference when assessing the target essay.

**High:** The increasing reliance on renewable energy is not merely an environmental necessity; it is an economic opportunity. Countries that invest in solar and wind infrastructure now will secure long-term energy independence and job growth. For instance, Germany's renewable sector employs over 300,000 workers, demonstrating that environmental policy and economic prosperity can coexist.
Scores: LAR 5, GAR 5, OD 5, OE 5

**Mid:** Renewable energy is important because it can help environment and create the jobs. For example, some countries make solar power and also wind power. This help economy grow. But sometimes the cost is high and some people do not like it. In the future, more research is necessary to make it better.
Scores: LAR 3.5, GAR 3, OD 3.5, OE 3.5

**Low:** Renew energy good, people use sun and wind. Some country have this but not all. This help planet but money problem. Jobs maybe come but not sure. Government need to do something.
Scores: LAR 1, GAR 2, OD 1, OE 1

Consider both strengths and weaknesses. Do not let one dimension such as grammar dominate. Minor errors that do not block meaning should have minimal impact. Assign category scores with justifications, suggest improvements for higher and lower order issues, and give an overall grade.

**Figure 3. Final Calibrated AI Scoring Prompt with Performance Anchors (Stage 3)**

This staged design ensured that each change to the prompt addressed a specific scoring misalignment, with the cumulative goal of aligning AI evaluation more closely with the balanced, construct-referenced approach typical of experienced human raters.

### *Statistical Analysis*

Agreement between AI and human ratings at each stage was examined at both the category and total score levels. Cohen's kappa (κ) was calculated for category-level agreement, interpreted using the benchmarks proposed by Landis and Koch (1977). Pearson's correlation coefficient (*r*) was computed to assess the degree of rank-order consistency in total scores. Paired-sample *t*-tests determined whether mean differences between AI and human scores were statistically significant, while Cohen's *d* provided an estimate of effect size (Cohen, 1988). All analyses were conducted in SPSS v.29, with statistical significance set at *p* < .05.

## Findings and Results

Table 1 presents changes in inter rater agreement, score consistency, and mean score differences between AI generated and human ratings across the three calibration stages. The metrics include Cohen's κ for categorical agreement, Pearson's r for score consistency, the mean difference in total scores (AI minus human), paired samples t-test statistics, corresponding p-values, and Cohen's d for effect size.

**Table 1. Inter-rater Agreement, Correlation, and Score Differences Across Calibration Stages**

| Stage | Cohen's κ | Pearson's r | Mean score difference | t(59) | p-value | Cohen's d |
|---|---|---|---|---|---|---|
| Baseline | 0.52 | .76*** | −2.45 | 5.21 | < .001 | 0.67 |
| Clarified descriptors | 0.71 | .85*** | −1.70 | 3.18 | .002 | 0.41 |
| Final calibrated prompt | 0.89 | .94*** | −0.95 | 1.02 | .312 | 0.13 |

***p < .001

At the baseline stage, when ChatGPT 3.5 received only a simple prompt and the evaluation rubric (as shown below), agreement was moderate (κ = 0.52), and the correlation between AI and human scores, though statistically significant, was below the threshold typically desired for operational use (r = .76, p < .001). On average, the AI scored essays 2.45 points lower than the human rater, a statistically significant gap (t(59) = 5.21, p < .001) representing a medium effect size (d = 0.67). Qualitative observations indicated a "strict marker" tendency, with heavy penalties for grammatical and lexical errors, even when higher order features such as organization and argumentation were strong.

During the clarified descriptors stage, the prompt was revised to present rubric categories in clearer, operational terms. Agreement improved to substantial (κ = 0.71), and correlation increased to r = .85 (p < .001). The mean difference narrowed to −1.70 points, with the gap still statistically significant (t(59) = 3.18, p = .002) but smaller in magnitude (d = 0.41). This stage reduced the AI's overemphasis on surface-level accuracy and encouraged more balanced evaluation, though a tendency to under-score remained.

In the final calibrated prompt stage, which added representative scoring examples to the clarified descriptors, agreement reached an almost perfect level (κ = 0.89), and correlation rose further to r = .94 (p < .001). The mean difference decreased to −0.95 points, which was no longer statistically significant (t(59) = 1.02, p = .312) and corresponded to a negligible effect size (d = 0.13). At this stage, AI scoring patterns closely matched human judgments in both rank ordering and absolute scores.

Overall, the results showed a clear and consistent progression toward alignment through calibration. Each stage improved agreement, reduced systematic bias, and enhanced fairness. These findings support the conclusion that targeted prompt refinement, guided by statistical validation, can significantly improve the validity and reliability of AI assisted L2 writing assessment.

The progression of calibration effects is further illustrated in Figure 4, which compares the mean total scores assigned by the AI and the human rater across the three stages of calibration.

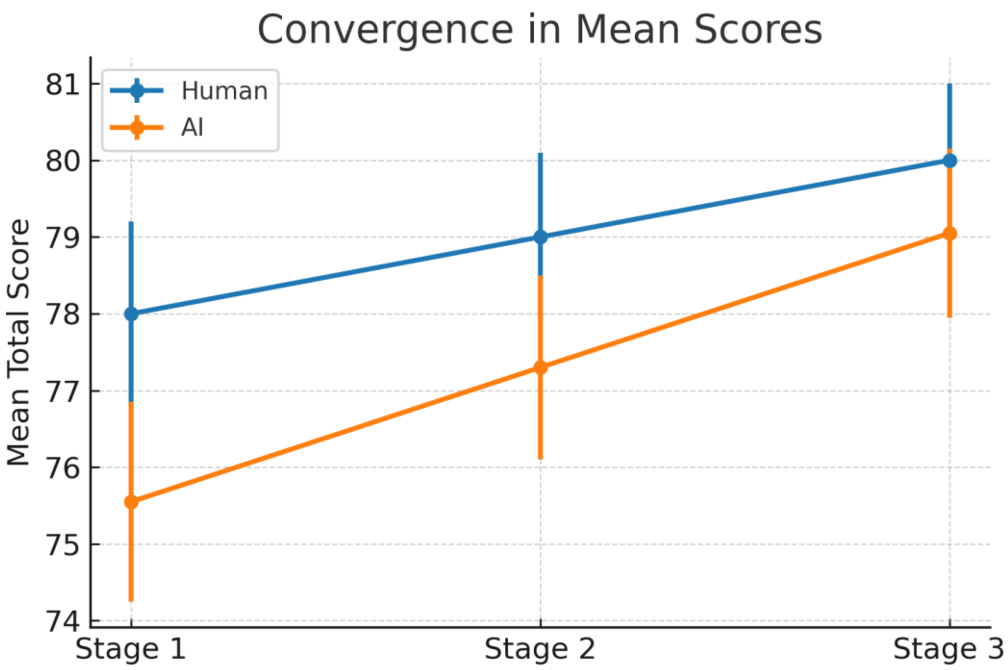


**Figure 4. Convergence of mean total scores for AI and human ratings across calibration stages**

The figure illustrates changes in mean total scores assigned by the AI and the human rater across three calibration stages. In Stage 1, the AI's mean score was noticeably lower than the human rater's, reflecting a consistent under scoring pattern. In Stage 2, after the rubric descriptors were clarified, the gap between the two means narrowed substantially. By Stage 3, when representative scoring examples were incorporated, the mean scores nearly overlapped, and their 95% confidence intervals fully intersected. The error bars represent 95% confidence intervals, and the convergence pattern reflects both increased agreement and reduced scoring bias.

## Discussion and Conclusion

The results of this study demonstrate that generative AI, specifically ChatGPT 3.5, can be systematically calibrated to align its scoring of second language writing with the evaluative reasoning of experienced human raters. Across the three iterative stages—baseline scoring, clarified descriptors, and performance anchors—the AI progressed from a rigid and surface-focused application of the rubric to a balanced and construct-aligned assessment that paralleled human scoring patterns. Agreement, measured by Cohen's kappa, improved from moderate to almost perfect levels, while correlations reached values typically required for operational use. These findings are consistent with the broader trend in language assessment research, where the combination of psychometric validation and carefully structured rating procedures has been shown to enhance fairness, validity, and reliability (5-7).

The progression observed in this study mirrors the developmental trajectory of novice human raters, who initially emphasize visible surface-level errors but, through training and exposure to exemplar performances, learn to integrate higher-order discourse features into their judgments. Research on human raters has repeatedly emphasized the variability that arises from differences in rubric interpretation, rater experience, and rating scale design (15, 16). The calibration process applied here reflects established rater training practices, in which descriptors are clarified, anchors are introduced, and agreement is statistically monitored. By replicating these human-centered approaches at the level of prompt design, this study situates calibration as a functional analogue of rater training for AI systems.

The statistical outcomes are particularly noteworthy. Agreement between AI and human ratings rose from $\kappa = 0.52$ to 0.89, correlation increased from .76 to .94, and mean score differences diminished from −2.45 to −0.95, the latter no longer statistically significant. These improvements highlight that LLM-based systems can transcend earlier limitations of automated essay scoring, which often suffered from construct underrepresentation and an over-reliance on shallow features (10-12). By integrating interpretive guidance and exemplars into prompts, the AI shifted from a mechanistic application of rubric descriptors toward judgments resembling those of skilled human assessors. This aligns with prior findings that analytic rubrics, when combined with appropriate training protocols, reduce the dominance of grammar and vocabulary in evaluations and instead emphasize multiple dimensions of writing competence (17, 18).

The emphasis on calibration also responds to persistent concerns about fairness in automated scoring. Without calibration, the AI exhibited a "strict marker" tendency, disproportionately penalising minor surface errors. Such behaviour risks construct-irrelevant variance by unfairly disadvantaging learners from specific linguistic backgrounds whose L2 writing often contains predictable but non-meaning-impairing grammatical patterns (7, 9). By explicitly instructing the AI that communicative effectiveness should outweigh isolated errors, calibration mitigated these biases, reinforcing the principle that fairness is central to language testing practice (6, 25).

The integration of performance anchors was particularly effective, as it closely parallels standardization sessions in human rater training. Anchor scripts allow raters to internalise scale points and resolve interpretive ambiguities (15, 16). In the present study, AI exposed to high, mid, and low performance examples demonstrated more proportionate and stable scoring behaviour, similar to raters who rely on benchmarks to calibrate their judgments. This suggests that LLMs can be influenced by instructional context in ways that resemble human cognition, a finding aligned with recent research on AI dependence on prompt framing and input quality (2, 26). The responsiveness of AI to anchors supports the broader argument that generative AI systems are not static evaluators but dynamic, instruction-sensitive agents whose outputs can be shaped by carefully engineered inputs (1).

At a technical level, the findings underscore the importance of recent advances in neural architectures for enabling holistic essay analysis. Earlier models such as PEG or e-rater relied on statistical and rule-based approaches, limiting their ability to capture global discourse structures (10, 11). The introduction of attention mechanisms (19) and end-to-end sequence models (20) allowed AI systems to model long-range dependencies, enabling evaluations that reflect argument coherence, rhetorical strategies, and overall communicative effectiveness. The results of this study confirm that when such architectures are combined with calibrated prompts, LLMs can move beyond detecting errors toward construct-aligned discourse-level assessment, a development anticipated in recent studies applying chain-of-thought reasoning to automatic scoring (3).

The shift in AI behaviour observed during calibration also highlights the importance of rubric design and clarity. In the baseline stage, when only rubric descriptors were provided, the AI interpreted them literally, often overemphasizing grammatical and lexical errors. This recalls longstanding debates in rubric design, where ambiguous or overly broad descriptors can lead to inconsistent applications by raters (17). Clarifying descriptors for AI, much as one would for novice human raters, improved performance by ensuring that the model's attention was directed toward relevant construct features. The finding aligns with research demonstrating that rubrics function most effectively when their criteria are clearly operationalized and aligned with an underlying theory of writing (5, 8).

From a validity perspective, calibration helps mitigate threats such as construct underrepresentation and irrelevant variance, both of which undermine the meaningfulness of scores (22, 24). At the baseline stage, the AI underrepresented higher-order writing constructs by focusing narrowly on surface features, echoing critiques of earlier automated scoring systems (13). Through calibration, the system's evaluations came to better reflect the full construct of academic writing, including organization, development, and rhetorical effectiveness. This demonstrates that validity in AI-based assessment is not an inherent property of the model but an outcome of deliberate design choices, calibration procedures, and validation practices (14).

The findings also extend current discussions about reliability in L2 writing assessment. Reliability has long been a central concern in both human and automated scoring, as inconsistent or unstable judgments undermine test usefulness (23). By employing Cohen's kappa and correlation coefficients, this study demonstrated that AI reliability can be improved to levels comparable with trained human raters through iterative calibration. This convergence underscores that the principles of psychometrics, traditionally applied to human raters, can be adapted for evaluating and enhancing AI performance. The study therefore contributes to bridging the gap between computational advances in generative AI and established practices in educational measurement.

Another contribution lies in the pedagogical implications of calibrated AI scoring. Teachers often struggle with the dual demands of providing timely, individualized feedback and maintaining consistency across large cohorts. Calibrated AI, capable of producing balanced, rubric-aligned commentary, offers a potential solution to these challenges. Previous studies have emphasized the importance of feedback that goes beyond error correction to highlight strengths, suggest strategies, and contextualize weaknesses (8, 9). The qualitative findings of this study—showing a shift in AI commentary from error cataloguing to balanced discourse-oriented feedback—suggest that calibrated generative AI can contribute meaningfully to writing pedagogy, especially in formative contexts.

It is also important to situate these findings within the broader evolution of automated writing evaluation. While early AES systems provided efficiency, they often did so at the expense of validity and fairness (12). Later systems incorporated more complex NLP features but still fell short in capturing higher-order constructs. The generative AI approach presented here, enhanced by calibration, offers a hybrid solution: it retains the scalability of automation while aligning its judgments with human evaluative reasoning. This reflects a maturation of AES research, moving from questions of technological feasibility toward issues of validity, fairness, and instructional usefulness (1, 13).

Despite these advances, several limitations must be acknowledged. First, the study relied on a single human rater as the benchmark for calibration. While this ensured intra-rater consistency, it also meant that AI alignment was measured against an individual perspective rather than a consensus of multiple raters. This may limit the generalizability of the findings, as different raters may interpret rubrics in slightly different ways. Second, the data set was restricted to argumentative essays produced by upper-intermediate learners, raising questions about whether the calibration process would generalize to other genres, such as narrative or reflective writing, or to other proficiency levels. Third, the calibration was applied to a single LLM (ChatGPT 3.5) at one point in time. Given that AI models evolve rapidly and often change behaviour after updates, longitudinal studies are needed to assess whether calibration effects persist across versions. Fourth, while statistical evidence showed convergence in scores, the study did not examine subtler issues of bias in AI-generated feedback language, such as tone or hedging, which may influence learner perceptions. Finally, the calibration process in this study was exclusively prompt-based; future work could compare prompt-level calibration with fine-tuning approaches that adjust model parameters more directly.

Future research should expand on these findings in several directions. Comparative studies involving multiple raters would provide a stronger benchmark for calibration and allow researchers to test AI alignment against more representative human standards. Extending calibration to other genres and proficiency levels would help assess its robustness across varied contexts of L2 writing. Additionally, research should examine whether calibration techniques transfer effectively across different AI models, including newer LLMs such as GPT-4 or Gemini. Longitudinal designs could track whether calibration effects endure over time and model updates. Further work is also needed to investigate the language of AI-generated feedback, exploring how learners interpret and respond to comments in formative settings. Finally, studies should explore hybrid approaches that combine prompt-based calibration with parameter-level fine-tuning, evaluating their relative efficiency and effectiveness in improving alignment with human scoring.

The practical implications of these findings are significant for language assessment and pedagogy. Institutions considering the integration of generative AI into assessment should adopt calibration as a formal, evidence-based quality assurance procedure. Teachers can employ calibrated AI to provide consistent, rubric-aligned feedback that reinforces writing instruction, while test developers can incorporate calibration into their design processes to ensure fairness and validity. In low-resource settings, calibrated AI may help alleviate rater shortages while maintaining construct alignment, though sustained human oversight will remain essential. More broadly, the study suggests that AI can play a supportive role in fostering more consistent, fair, and pedagogically useful assessment, provided that its use is guided by calibration, transparency, and accountability.

## Acknowledgments

We would like to express our appreciation and gratitude to all those who helped us carrying out this study.

## Authors’ Contributions

All authors equally contributed to this study.

## Declaration of Interest

The authors of this article declared no conflict of interest.

## Ethical Considerations

All ethical principles were adheried in conducting and writing this article.

## Transparency of Data

In accordance with the principles of transparency and open research, we declare that all data and materials used in this study are available upon request.

## Funding

This research was carried out independently with personal funding and without the financial support of any governmental or private institution or organization.

## References

1. Shi Y, Aryadoust V. Large language models in language assessment: Opportunities and challenges. Assessing Writing. 2024;52:100688.
2. Jonäll K. Artificial intelligence in academic grading: A mixed-methods study: University of Gothenburg; 2024.
3. Lee GG, Latif E, Wu X, Liu N, Zhai X. Applying large language models and chain-of-thought for automatic scoring. arXiv. 2023.
4. Hicke Y, Tian T, Jha K, Kim CH. Automated essay scoring in argumentative writing: DeBER TeachingAssistant. 2023.
5. Bachman L, Palmer A. Language assessment in practice: Developing language assessments and justifying their use in the real world: Oxford University Press; 2022.
6. Kunnan AJ. Test fairness In - M. Milanovic & C. Weir (Eds.), European language testing in a global context. Cambridge University Press; 2004. p. 27-48.
7. McNamara T, Knoch U, Fan J. Fairness, justice, and language assessment. Language Testing. 2019;36(1):1-8.
8. Hyland K. Teaching and researching writing: Routledge; 2016.
9. Ferris D. Responding to student writing: Teachers' philosophies and practices. Assessing Writing. 2014;19:6-23. doi: 10.1016/j.asw.2013.09.004.
10. Page EB. Project Essay Grade: PEG In - M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring. Lawrence Erlbaum; 2003. p. 43-54.
11. Burstein J. The e-rater scoring engine: Automated essay scoring with natural language processing In - M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring. Lawrence Erlbaum; 2003. p. 113-21.
12. Dikli S. An overview of automated scoring of essays. The Journal of Technology, Learning, and Assessment. 2006;5(1).
13. Shermis MD, Burstein JC. Handbook of automated essay evaluation: Routledge; 2013.
14. Ramineni C, Williamson DM. Understanding automated scoring through the lens of the scoring process. Assessing Writing. 2013;18(4):244-72. doi: 10.1016/j.asw.2012.10.004.
15. Barkaoui K. Variability in ESL essay rating processes: The role of the rating scale and rater experience. Language Assessment Quarterly. 2010;7(1):54-74. doi: 10.1080/15434300903464418.
16. Lumley T. Assessing second language writing: The rater's perspective: Peter Lang; 2005.
17. Brookhart SM. Appropriate criteria: Key to effective rubrics. Frontiers in Education. 2018;3:22. doi: 10.3389/feduc.2018.00022.
18. Cumming A, Kantor R, Powers D, Santos T, Taylor C. TOEFL iBT writing framework: A working paper. 2005.
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al., editors. Attention is all you need. Advances in neural information processing systems; 2017.
20. Chorowski J, Bahdanau D, Cho K, Bengio Y, editors. End-to-end continuous speech recognition using attention-based recurrent NN: First results. arXiv; 2014.
21. Ke Z, Ng V, editors. Automated essay scoring: A survey of the state of the art. Proceedings of the International Joint Conference on Artificial Intelligence; 2019.
22. Kolen MJ, Brennan RL. Test equating, scaling, and linking: Methods and practices: Springer; 2014.
23. Warrens MJ, van der Hoef H, Heiser WJ. A comparison of reliability coefficients for ordinal rating scales. Frontiers in Psychology. 2021;12:736184. doi: 10.3389/fpsyg.2021.736184.
24. Eckes T. Introduction to many-facet Rasch measurement: Peter Lang; 2015.
25. Council of E. Common European framework of reference for languages: Learning, teaching, assessment - Companion volume: Council of Europe Publishing; 2020.
26. Reynolds L, McDonell J, editors. Prompt programming for large language models: Beyond the few-shot paradigm. arXiv; 2021.