ISSN: 2278-4012, Volume:14, Issue:1, March 2025

AI-Driven Adaptive Assessment Strategies for English Language Proficiency Evaluation

Valeria R. Mendoza[1*] and Alejandro M. Quispe[2]

[1*]English Language Teaching, Universidad de Lima, Peru. E-mail: valeriarteach@outlook.com.pe

[2]English and Communication, Universidad Nacional de San Agustín de Arequipa, Peru.

Abstract: This paper introduces a unified approach to combine artificial-intelligence-based adaptive assessment with large language model (LLM) tools for the assessment and improvement of the English language proficiency of university students. Current writing instruction and proficiency assessment are mostly rigid, with fixed sequences of items and standardized feedback that are not attuned to the variable content of individual students' learning processes, their cognitive burden, or the conventions of writing in their subject matter. This research integrates empirical findings from the fields of AI-based proficiency assessment, automated essay grading, and generative-AI as a measure of writing feedback, leveraging the advancements of the computerized adaptive testing (CAT) approach and recent transformer-based language models, to create a two-level architecture comprising an Adaptive Assessment Engine and an LLM-Based Academic Writing Support Module, both linked by a learner model to promote personalized learning. The Adaptive Assessment Engine automatically adjusts item difficulty and skill coverage in real time based on ability estimation and the writing support module provides rubric-aligned, register-sensitive and iterative feedback on structure and coherence, lexical sophistication, and grammatical accuracy. The design rationale and evaluation criteria for the framework are informed by a narrative synthesis of 17 peer-reviewed studies published from 2021 to 2025. In addition to the documented improvements in scoring reliability, engagement, and revision productivity noted in the literature, there are ongoing questions about validity, algorithmic bias, over-reliance, and issues in academic integrity. The paper concludes that it is possible to integrate an adaptive-assessment and writing-support architecture design that can improve formative and summative assessment of English proficiency while still maintaining the oversight of the instructor, and it provides directions for empirical validation, cross-linguistic generalization, and longitudinal impact studies of university English-language programs.

Key Words: artificial intelligence, adaptive assessment, english language proficiency, large language models, academic writing support, computerized adaptive testing, automated essay scoring

(Received: 20 December 2024; Revised: 24 January 2025; Accepted: 24 February 2025; Published: 31 March 2025)

Introduction

With the spread of artificial intelligence (AI) into higher education, time-honored issues of how English language proficiency should be assessed and nurtured at the university level have been put on the agenda again. Traditional proficiency examinations give a standard set of items to all candidates regardless of their estimated ability level, creating sub-optimal tests for high ability students and demoralizing for low ability students. At the same time, the traditional method of academic writing instruction for English learners was based on feedback by the teacher, which was delivered slowly and with a high level of resources consumption and could not be used in large classes and in diverse student groups (Meyer et al., 2024). Two technological advances now enables to tackle both of these at the same time: computerized

adaptive testing (CAT) based on Item Response Theory (IRT), and large language models (LLMs) that create, rate and edit long academic text.

The adaptive testing platforms change the difficulty of items based on the successive responses of the test taker, and, as of writing, converge to a more accurate ability estimation with fewer items and less administration time than a fixed form test (Bock & Gibbons, 2021). With AI-generated item pools and dynamic difficulty adaptation, these platforms can then further customize content for each proficiency profile within the sub-skills of listening, reading, and writing (Abdellatif et al., 2024). Conversely, generative LLMs like GPT-4 have shown significant alignment with human raters in AES and in delivering formative feedback that is perceived as relevant, specific and motivating by students (Li & Liu, 2024; Lee, 2024). The developments point to a possible evolution from assessment and writing support as two discrete, sequential processes to a single ongoing, evolving process.

The purpose of this paper is to attempt to answer two related questions. First, it brings together the latest empirical research on AI-based English proficiency assessment and academic writing support with LLMs, and finds the convergences and differences, along with unanswered questions. Second, it suggests a holistic system – consisting of an Adaptive Assessment Engine and an LLM-Based Academic Writing Support Module, linked by a shared learner model – which could be used in university English medium and English-as-a-Foreign-Language (EFL) instruction. The framework is aimed at design contribution, based on the considered evidence and not as a report of an already completed empirical trial, and thus ends with a statement of an evaluation agenda for future studies that will validate the framework.

The rest of the paper is organized as follows: In Section 2, the literature is reviewed on adaptive assessment and LLM-based writing support. In Section 3, the proposed framework and its elements are presented. The methodological approach in deriving and structuring the framework is described in Section 4. The expected benefits and their place in the reviewed evidence are discussed in Section 5. Then, Section 6 discusses pedagogical implications, Section 7 discusses limitations and ethical considerations, and Section 8 concludes the paper.

Literature Review

The idea of using IRT models to estimate examinee ability from item response patterns and to select each subsequent item so as to maximize information at the current ability estimate is a longstanding idea in psychometrics, and is the foundation of computerized adaptive testing. This has been expanded on in recent studies for contexts of English proficiency. For the case of institutional English-proficiency screening, a university level computerized adaptive test designed on IRT and the Admissible Probability Measurement Procedure was demonstrated to be more efficient by being able to reduce test length without compromising the precision of the measurement compared with the paper-and-pencil instrument it was being used to replace (Bock & Gibbons, 2021). Likewise, adapted English proficiency systems at a local level designed for specific groups of learners have shown that test localization and IRT based item calibration can yield psychometrically sound measures of English proficiency that are freely available and adapted to a defined population of test takers.

In addition to psychometric calibration, AI can be used in adaptive assessment through dynamic difficulty adaptation (DDA) constructs, which can dynamically add or subtract items and difficulty levels from grammar and language tasks based on the current level of student performance at a particular moment. AI-enhanced assessment environments based on the concept of the portfolio have also been linked to better emotion regulation, mindfulness, and attitude toward language learning in learners, showing that adaptive

assessment impacts both assessment precision and affective engagement in the assessment process (Alazemi, 2024). In the area of listening assessment, AI-based adaptive testing has been shown to yield measurable improvements in listening proficiency for EFL students, as well as in students' self-competence and resilience, showing that adaptivity can have a positive impact on both non-reading language skills, aside from grammar and reading (Abdellatif et al., 2024). Research on the use of AI models for assessing student performance and AI systems for providing formative feedback at tertiary level also suggests better learning outcomes if assessment is flexible and responsive to learners' performance (Aijun, 2024). Together, these studies provide compelling evidence for the potential to optimize adaptive assessment to enhance measurement efficiency, engagement, and diagnostic value for skills.

Another related field of study is the impact of LLM on the learning of academic writing skills of English learners. Teng, (2024) conducted acceptance research based on the Unified Theory of Acceptance and Use of Technology (UTAUT) that revealed that performance expectancy, effort expectancy, and facilitating conditions significantly influenced EFL learners' behavioral intention to use LLMs to perform register analysis, paraphrasing, and written corrective feedback in business academic writing courses. This acceptance-based evidence is supported by intervention research that explores the direct impact of LLM feedback on writing quality. According to a randomized study, students' text revision behavior, motivation and positive emotion were higher with feedback based on LLM compared to no feedback and quasi-experimental studies indicate that AI feedback was more immediate and specific than instructor feedback, even though it was not as credible as that of instructors. Randomized studies of LLM-generated evidence-based feedback showed that it encouraged text revision behavior, motivation, and positive emotion compared with no feedback or delayed feedback and comparative studies of LLM and instructor feedback revealed that LLM feedback was more immediate and specific than instructor feedback, but that instructor feedback was still more credible.

Complementary strands of research have examined the processes and tools of self- and peer-mediated revision processes that are scaffolded by LLM. The results of structured revision research on SLP show that guided and iterative revision, whether mediated by peer or AI, is more effective than single-pass revision feedback (Li & Hebert, 2024). In addition, motivational-climate theory locates the influence of emotions and goal setting as intervening variables that drive students' reactions to generative-AI-enabled writing instruction, suggesting that the pedagogical representation of LLM feedback, in addition to its factual correctness, matters to students' learning outcomes. Conceptual models for sustainable assessment and feedback advocate embedding the LLMs into pedagogically meaningful and technologically viable higher education feedback systems, not just as a replacement for instructor feedback. In the broader context of EFL language instruction, the adoption of conversational AI in EFL classrooms has been singled out as an emerging trend, while the possibility of enhancing affordances for autonomous learning and self-directed practice has been raised, with the need for more explicit guidelines on how to integrate conversational AI pedagogically also emphasized (Klímová & Ibna Seraj, 2023; Jeon, 2024).

The reliability and validity of LLMs as automated essay scoring systems is a large volume of research that is relevant to any framework that pairs adaptive assessment with writing support. GPT-4 has been shown to produce more accurate annotations and better human rater agreement than BERT-based and locally trained language models when evaluating non-native essays on lexical richness, syntactic complexity, cohesion and grammatical accuracy, with in some cases, agreement of GPT-4 with human raters higher than agreement between two human raters (Li & Liu, 2024). A complementary study of the LLM's accuracy in marking second-language learners (SLLs) writing also finds that LLM writing scores match well with human markers' scores under specific rubric criteria (Lee, 2024). The zero-shot prompting

and fine-tuned LLM configurations for automated argumentative essay evaluation showed that the latter yielded superior scores than the former in terms of score precision, highlighting the fact that the way a LLM is configured affects the scores it can deliver (Wang et al., 2024). Previous work using GPT-3 for automated essay scoring of TOEFL-style writing showed that the accuracy of the baseline automated judgements is moderate-to-fair, and that adding explicit linguistic-complexity measures in addition to the automated judgement of the AI language model significantly improved the accuracy of such judgements (Mizumoto & Eguchi, 2023).

There is a significant body of literature that highlights the different ways in which AI tools for grading and feedback work and the varied reliability they achieve across different tasks, with LLM-based tools achieving greater reliability on holistic and rubric-anchored tasks and less on tasks that demand finer-grained content judgement and discipline-specific argumentation (Imamguluyev et al., 2024). Research that focuses on the scalability of these AI-based deep learning models for adaptive and inclusive assessment also suggests that the reliability patterns that emerge are most appropriate in high-stakes settings as an adjunct to human marking, and where appropriate, fully automated in low-stakes formative environments (Javed, 2024). This separation between formative and summative appropriateness also appears throughout the literature that was reviewed and directly influences the architecture.

The literature reviewed shows that both, IRT-based adaptive assessment and Dynamic Difficulty Adaptation (DDA) approaches have been matured and validated, while on the side of LLM academic writing support and automated essay scoring, a growing body of literature has been developed but largely isolated. There are few studies that investigate these two abilities as a tightly coupled system, in which a learner's ability estimates that changes as a result of adaptive assessment directly influences the calibration of writing feedback and in which writing performance data directly influence the evolution of the adaptive assessment engine's ability estimate. This gap will inspire the integrated framework presented in the next section, where assessment and writing support are considered complementary data streams shared with a learner model, and not as separate tools.

Proposed Integrated Framework

To support the above evidence, this paper introduces a two-level model: Adaptive Assessment Engine (AAE) and the LLM Based Academic Writing Support Module (WSM), that are orchestrated by a common Learner Model. It is designed for use in university English foreign language ability training programs as a diagnostic and supporting system, or as a supplementary layer of service on top of the existing learning management platform.

Architecture Overview

There are 4 functional layers of architecture. The Assessment Layer presents adaptively chosen items from a bank of IRT-calibrated items, along with AI-generated items whose difficulty parameters are estimated on deployment and continually calibrated. The Writing Support Layer receives a student's draft and scores it in accordance with the rubric and provides a rubric-aligned structure of feedback using an LLM trained to follow specific scoring rules, and provides iterative revision suggestions. The Learner Model Layer combines ability estimates, error pattern histories, and revision trajectories and writes them into a longitudinal learner profile which is accessed by and updated by the other layers. The Interface and Oversight Layer provides instructor-facing review tools and learner-facing dashboards while enabling instructors to review, override, or add annotations to AI-generated scores and/or feedback before the scores

ISSN: 2278-4012, Volume:14, Issue:1, March 2025

are finalized, aligning with recommendations that AI-based scoring should augment, not supplant, human judgment for consequential decisions.

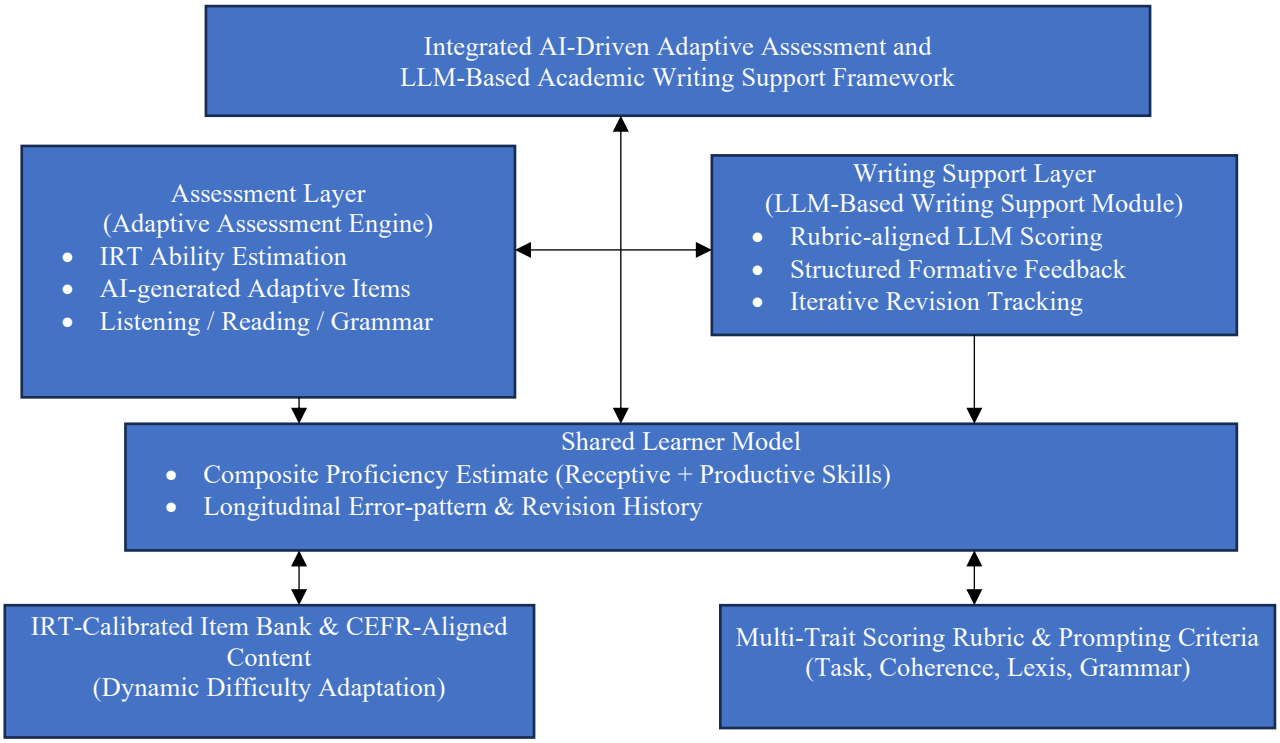

Figure 1: Conceptual Architecture of the Integrated AI-driven Adaptive Assessment and LLM-based Academic Writing Support Framework

The four layers and data flow in figure 1 shows. The two layers are parallel and symmetrical because they both provide evidence of the learner's ability: one from a collection of items chosen to assess receptive skills, and the other from a collection of writing samples that are scored and revised. The two layers contribute to the Shared Learner Model, which combines evidence of receptive and productive skills into a single composite measure of proficiency and keeps a history of the errors that students made again and again as well as of the revisions they made over time. This is a composite estimate that enables the framework to be adaptive for the skills, but not just listening, reading, grammar and writing being separate measurement tracks. The Interface and Oversight Layer is built on top of the Learner Model to ensure instructors are always in control to review, modify, or add comments to any AI-generated score and feedback before it has an impact, thereby implementing the human-in-the-loop protection. The two supporting inputs at the bottom of the diagram are the IRT-calibrated item bank and the multi-trait scoring rubric, respectively, grounding the framework's adaptivity and its writing feedback in the psychometric and prompting practices that have been validated in the reviewed literature.

Adaptive Assessment Engine

The AAE follows a standard three-parameter logistic IRT item-selection loop: First, the population mean is used as the initial ability estimate; then, the next item is chosen that maximizes Fisher information at the current ability estimate; the ability estimate is then updated after each response using maximum-likelihood or Bayesian estimation; and finally, the test is terminated if a target standard error is reached or a maximum number of items is reached. To adapt the loop to discrete-item skills, an AI-based item-generation system that generates reading passages, grammar items and short-answer prompts based on a target CEFR band is integrated into the AAE, with results from previous AI-based grammar exercise

generators validated by the dynamic-difficulty-adaptation approach (Javed, 2024). For productive skills, the AAE sends writing prompts to the WSM described below and receives back a proficiency-band estimate, which is combined with that of the receptive skill and the same overall ability is maintained for all four skills.

LLM-Based Academic Writing Support Module

There are three interrelated functions of the WSM. First, it asks students to score drafts that are prompted by an explicit, discipline-aware rubric detailing task achievement, coherence and cohesion, lexical resource, and grammatical range and accuracy, similar to a multi-trait prompting framework that has been demonstrated to enhance the reliability of scoring when compared to a holistic-only prompt. Second, it provides structured formative feedback, which provides a description of specific strengths and error patterns, instead of an overall grade, since evidence has shown that structured feedback that provides specific strengths and error patterns, rather than a general judgement, has a greater impact on revision productivity and learner motivation than general praise or criticism. Third, it implements an iterative revision process with successive drafts, demonstrating that revision through peer-mediated and self-reflective revision research that incorporates an iterative revision process is more effective in enhancing end-of-process writing quality than a single pass revision process (Li & Hebert, 2024). The WSM aims to avoid over-reliance and maintain the learner's agency by withholding direct rewrites and focusing on specific questions and localized suggestions, following the recommendations of the pedagogical-framework (Abulela et al., 2023; Agostini & Picasso, 2024) that generative-AI feedback should be used as a scaffold and not a replacement for the learner's reasoning.

Integration Layer and Learner Model

The two tiers communicate to each other via the shared Learner Model. Items that are included in the WSM as a writing prompt are also modified to both be easier and more complex to read, depending on the ability estimate, and patterns of errors found during writing feedback are incorporated into the item-selection algorithm of the AAE so that items related to the same category are emphasized. This bi-directional coupling is designed to reduce the diagnostic cycle over systems in which assessments and writing instruction take place on independent, asynchronous time frames, and reflects the recurring results from the literature cited, that adaptivity can improve the accuracy of the measurement and increase the engagement of the learner when it is tied to a current "snapshot" of learner ability.

Methodology

This paper does not follow an empirical approach, but rather a conceptual and design approach. The framework in this section was developed using a structured synthesis of the narratives of 17 peer-reviewed studies spanning the period 2021 to 2025 and obtained from targeted searches of Scopus-indexed and Web-of-Science-indexed sources related to AI-based adaptive language assessment, computer adaptive testing, automated essay scoring, and LLM-based writing feedback. Studies were eligible for inclusion if they provided empirical results (psychometric, linguistic, or perceptual) regarding the use of AI systems for assessing or supporting learners' writing performance in English or a similar second language, and if they were judged as relevant to two design questions: (a) What architectural elements of adaptive assessment systems have shown measurable improvements in efficiency or validity? and (b) What are the configurations of LLM-based writing feedback that have shown measurable improvements in scoring reliability, revision behavior, or learner motivation? The two questions were addressed through the findings and each of the four layers of an architecture was assessed for its sources in the set of reviewed findings.

ISSN: 2278-4012, Volume:14, Issue:1, March 2025

The design-science approach is similar to that of other educational-technology studies in which conceptual frameworks are first suggested and argued in light of the current empirical findings, and then pilot-tested and tested in a controlled study before being empirically validated.

Results and Discussion

The synthesis reveals three claims that fit together to support the proposed architecture. Adaptive item selection based on IRT also results in shorter tests and shorter administration than fixed-form tests without compromising or even enhancing measurement precision, as found in both institutional and localized settings of English proficiency testing. Second, with proper prompting with rubric criteria and multiple-trait scoring guidelines, LLM-based automated essay grading performs with human-like reliability for rubric-based writing tasks, with some reports achieving scores equal to or better than human-to-human agreement. Third, LLM-generated formative feedback significantly enhances revision productivity, motivation and positive affect compared to no feedback conditions, while it is perceived by the learners as a complement to, not a replacement of, instructor feedback. Table 1 describes the expected differences between traditional English proficiency testing and writing instruction, as well as the proposed framework in this paper, on the basis of the patterns that were reported in the reviewed literature.

Table 1: Comparison of Conventional English Proficiency Assessment/Writing Instruction and the Proposed AI-driven Adaptive Assessment and LLM-based Writing Support Framework

Dimension	Conventional Approach	Proposed AI-Driven Framework
Item sequencing	Fixed-form, identical for all test-takers	IRT-calibrated adaptive selection maximizing information
Writing feedback turnaround	Delayed, instructor-dependent	Immediate, rubric-aligned LLM feedback
Feedback specificity	Often holistic or generic	Multi-trait, error-pattern specific (Wang et al., 2024; Mizumoto & Eguchi, 2023)
Assessment-instruction linkage	Separate, asynchronous processes	Bidirectional learner model coupling assessment and writing support
Scoring reliability on rubric tasks	Dependent on rater training and load	Human-comparable agreement reported for GPT-4 scoring
Learner engagement	Variable; test anxiety at fixed difficulty	Improved motivation and reduced discouragement reported under adaptivity

Figure 2 illustrates that the range of Quadratic Weighted Kappa (QWK) for human–human rater agreement (0.695–0.950) is shown in comparison to human–GPT-4 agreement (0.644–0.819) across sixteen writing proficiency measures that were assessed in the non-native essay-scoring study collected. The ranges are both higher than the 0.61 value generally considered to represent a significant level of agreement, and the overlap suggests that scoring by GPT-4-based systems came close to or equaled the level of agreement seen when scores are judged independently by human raters on several individual measures. The visible difference at the high end of the human-human range, however, confirms the position taken in this paper that automated scoring should be used in tandem with human scoring on the higher stake's elements of the proposed framework.

ISSN: 2278-4012, Volume:14, Issue:1, March 2025

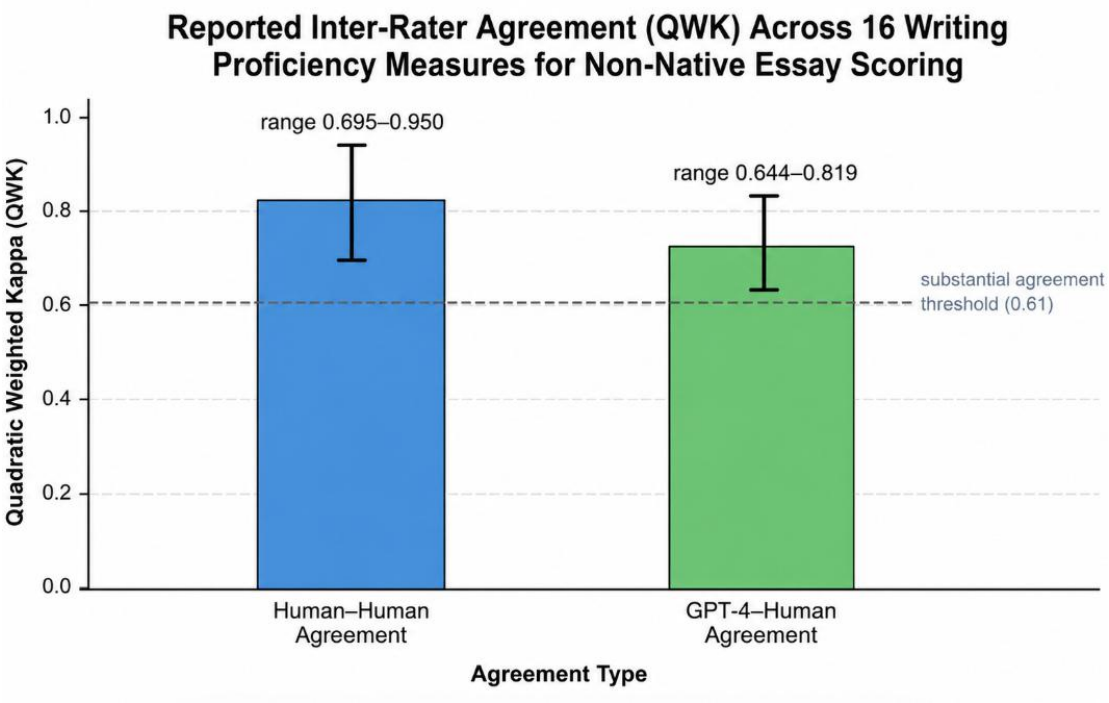

Figure 2: Reported Inter-rater Agreement (Quadratic Weighted Kappa) Between Human Raters and Between GPT-4 and Human Raters, Aggregated Across 16 Non-native Writing Proficiency Measures (Li & Liu, 2024)

The pattern of results also sheds light on where automation is not to be fully trusted. The results of reliability seem to be better for rubric-based, holistic, or multi-trait scoring, and less reliable for tasks that require content judgment and/or argumentation judgment. This distinction helps to justify the framework's design decision to include an instructor level of oversight above the WSM rather than having the WSM make the final summative grade without such oversight. Likewise, although adaptive assessment reliably shortens test length and enhances precision for receptive skills it is not as certain with productive skills, as the composite ability estimate generated by the framework would only be as reliable as its weakest scoring component. This dependency is an explicit design risk and will be quantified with future empirical validation, before high-stakes deployment.

The overall discussion suggests that the framework is feasible, albeit with a few uncertainties that may require further research, as each of its elements has some empirical basis and the proposed coupling of these elements fills a documented gap that may benefit assessment and writing instruction by giving both the ability to respond to the learner in real time.

Pedagogical Implications

The framework also provides guidance for curriculum designers that the placement and diagnostic testing of university English courses do not need to be a rigid process with long tests but can be replaced by short, adaptive tests to achieve the same level of measurement precision, thereby saving instructional time that is currently being devoted to long tests. The framework changes the role of a writing instructor from the sole feedback provider to the "feedback curator" by asking writing instructors to review and refine the feedback from AI, as research has demonstrated that students value immediacy of feedback from AI and also the credibility of their writing instructor, and that the combination of both is more beneficial than either alone (Meyer et al., 2024). The learner model is a bi-directional perspective of growth in proficiency across the receptive and productive skill areas, which helps with the historical tension between test scores

ISSN: 2278-4012, Volume:14, Issue:1, March 2025

and feedback in writing courses. The framework's use of validated, explainable rubric-based scoring, instead of opaque holistic AI scoring, ensures that it meets the needs of program administrators who must be accountable to their institutions for scoring efficiency benefits that come with the use of AI.

Challenges and Ethical Considerations

There are a number of limitations to the potential advantages of the framework. There is still some doubt over the validity of using LLM-based scores to measure overall academic writing proficiency, as they tend to favor aspects of the linguistic structure that are correlated with, but not necessarily reflective of, effective arguing. Another broad concern is algorithmic bias: scoring systems that are mainly designed to reflect rhetorical norms of dominant first-language groups, or genre norms, may systematically disadvantage students whose rhetorical patterns differ from the distribution of the scores that the algorithms were trained on, as noted in the literature on AI-based assessment in general and as discussed in this paper prior to institutional implementation and usage of such tools, and that would merit special attention during fairness audit. Another risk is over-reliance: LLM feedback is easy to access, and quick, making it easier for learners to use it as a disincentive to engaging with their own work in a critical way unless they are proactively supported to do so by the WSM, as suggested in Section 3.3, to promote thinking over direct rewriting. Academic integrity issues also emerge when the distinction between AI revision and AI authorship becomes less clear, necessitating the development of norms of disclosure regarding when it is appropriate to use an LLM in academic writing that is given grades. Lastly, any system that collects longitudinal learner ability estimates and writing samples will have data privacy and consent requirements, and institutions implementing the proposed framework will need to ensure adherence to data protection laws and to make learners aware of the ways that their data is used to inform adaptive item selection and scoring.

Conclusion

This paper has brought together the empirical evidence regarding adaptive assessment and academic writing support using LLMs and suggested a framework to integrate them using a common learner model. The reviewed literature validated the validity of IRT-based adaptive testing for measuring English proficiency efficiently and accurately, and rubric-prompted LLM scoring and feedback for providing reliable, motivating writing support, but also revealed that there is a lack of research on the two as a coordinated system. To overcome these concerns in terms of validity and over-reliance, the proposed four-layered architecture includes the following components: Adaptive Assessment Engine, LLM Based Academic Writing Support Module, shared Learner Model, and Interface and Oversight Layer, with the aim of bridging the gap between current systems and closing it. The main value of the framework is architectural, outlining the ways in which adaptive selection of items and generative text support can be mutually informed in real time, rather than used in parallel and independently. The framework presented is a design contribution that is a result of a narrative synthesis and not a validated deployment; this is the main limitation. Future research should focus on conducting controlled pilot studies that compare the outcomes of students when assessed using the integrated framework approach with the outcomes when assessed using conventional fixed-form testing and feedback from the teacher only, psychometric validation of the composite ability estimate by integrating receptive and productive ability, fairness audit for students with different first language backgrounds, and longitudinal studies to track the progress of writing achievement during the entire academic term. Replication across languages (in addition to English) would also help determine if the architecture extends to other additional-language contexts. The proposed framework, subject to this validation, provides a clear roadmap for assessment and writing instruction that

is continuous and moves toward instruction based on individual student proficiency, rather than moving from the measurement phase to the instruction phase two distinct phases of English language instruction.

References

Abdellatif, M. S., Alshehri, M. A., Alshehri, H. A., Hafez, W. E., Gafar, M. G., & Lamouchi, A. (2024). I am all ears: listening exams with AI and its traces on foreign language learners' mindsets, self-competence, resilience, and listening improvement. *Language Testing in Asia*, *14*(1), 54. https://doi.org/10.1186/s40468-024-00329-6

Abulela, M. A., Mrutu, A. P., & Ismail, N. M. (2023). Learning and study strategies as predictors of undergraduates' emotional engagement: a cross-validation approach. *Sage Open*, *13*(1), 21582440231155881. https://doi.org/10.1177/21582440231155881

Agostini, D., & Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework. *Intelligenza Artificiale*, *18*(1), 121-138. https://doi.org/10.3233/IA-240033

Aijun, Y. (2024). On the influence of artificial intelligence on foreign language learning and suggested learning strategies. *International Journal of Education and Humanities*, *4*(2), 107-120. https://doi.org/10.58557/(ijeh).v4i2.214

Alazemi, A. F. T. (2024). Formative assessment in artificial integrated instruction: delving into the effects on reading comprehension progress, online academic enjoyment, personal best goals, and academic mindfulness. *Language Testing in Asia*, *14*(1), 44. https://doi.org/10.1186/s40468-024-00319-8

Bock, R. D., & Gibbons, R. D. (2021). Computerized adaptive testing. In R. D. Bock & R. D. Gibbons (Eds.), Item response theory (Chapter 8). Wiley. https://doi.org/10.1002/9781119716723.ch8

Imamguluyev, R., Hasanova, P., Imanova, T., Mammadova, A., Hajizada, S., & Samadova, Z. (2024). Ai-powered educational tools: Transforming learning in the digital era. *International Research Journal of Modernization in Engineering Technology and Science*, *6*, 920-929. https://www.doi.org/10.56726/IRJMETS65040

Javed, F. (2024). The evolution of artificial intelligence in teaching and learning of English language in higher education: Challenges, risks, and ethical considerations. https://doi.org/10.1108/978-1-83549-486-820241015

Jeon, J. (2024). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning*, *37*(1-2), 1-26. https://doi.org/10.1080/09588221.2021.2021241

Klímová, B., & Ibna Seraj, P. M. (2023). The use of chatbots in university EFL settings: Research trends and pedagogical implications. *Frontiers in psychology*, *14*, 1131506. https://doi.org/10.3389/fpsyg.2023.1131506

Lee, O. S. (2024). Examining AI-Based Accuracy Assessment in L2 Learners' Writing. *Journal of Pan-Pacific Association of Applied Linguistics*, *28*(2). https://doi.org/10.25256/paal.28.2.3

Li, A. W., & Hebert, M. (2024). Unpacking an online peer-mediated and self-reflective revision process in second-language persuasive writing. *Reading and Writing*, *37*(6), 1545-1573. https://doi.org/10.1007/s11145-023-10466-8

Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities and Social Sciences Communications*, *11*(1), 1-15. https://doi.org/10.1057/s41599-024-03209-9

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, *6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

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

Teng, M. F. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, *7*, 100270. https://doi.org/10.1016/j.caeai.2024.100270

Wang, Y., Hu, R., & Zhao, Z. (2024, November). Beyond agreement: Diagnosing the rationale alignment of automated essay scoring methods based on linguistically-informed counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 8906-8925). https://doi.org/10.18653/v1/2024.findings-emnlp.520