algorithms

MDPI

Article

Automated Grading and Professional Accounting Education: Examining the Fairness, Reliability, and Validity of AI Grades

Nusirat Ojuolape Gold [1,2,*] and Husain Coovadia [1]

[1] Department of Commercial Accounting, College of Business and Economics, University of Johannesburg, Auckland Park, 1 Kingsway Road, Johannesburg 2006, South Africa; hcoovadia@uj.ac.za
[2] Department of Accounting and Finance, Faculty of Management and Social Sciences, Kwara State University, Malete 241103, Nigeria
* Correspondence: ngold@uj.ac.za or nusirat.gold@kwasu.edu.ng

Abstract

Automated long-essay scoring (ALES) is gradually considered as a means to enhance efficiency and consistency in large-scale assessment; however, concerns remain regarding its suitability, particularly as it relates to the reliability, validity, and fairness of ALES-assigned grades relative to human-grades in high-stakes professional contexts. This study examines these concerns using over 15,000 long essay examination scripts from a professional accounting certification examination. The study examines whether the ALES confidence index (CI) meaningfully predicts grading accuracy or points to systemic grading failures. Findings reveal fair overall agreement between human and ALES grades, with high within ± 1 grade agreement, and rare yet task-concentrated ALES grading failures, while CI shows statistically significant but practically weak predictive value and limited discrimination. The results support the use of ALES as an assistive, human oversight tool rather than an independent grader, highlighting the importance of task-based validation, stronger calibration analysis, and continuous human supervision in high-stakes professional assessment contexts. The study advances innovative assessment practices, but calls for cautious deployment of ALES and recommends integration of a hybrid human-in-the-loop approach, multi-disciplinary validation, and capacity building to strengthen ethical and responsible AI usage in accounting education and professional practice, aligning with SDGs 4 and 9.

Keywords: AI-grading; human grading; confidence index; long-essay scoring; professional accounting assessment; human-in-the-loop

Check for updates

Academic Editors: Domenico Ursino, Gianluca Bonifazi, Enrico Corradini and Michele Marchetti

Received: 13 January 2026
Revised: 15 February 2026
Accepted: 19 February 2026
Published: 1 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The shift in artificial intelligence (AI) technology, especially for assessment, has paved the way for innovation in the education sector, redefining teaching and learning boundaries through improved efficiency, scalability, and consistent grading [1]. In the past, long essay responses have relied on the traditional manual assessment, which is generally time-consuming and cognitively demanding for instructors [2]. Now, it is advancing with the use of AI technologies that allow instructors ample time to teach, engage more with students, reduce workload, provide timely feedback, and enhance grading consistency [3]. While multiple-choice questions (MCQs) can be easily automated for assessment, grading long essay responses remains difficult because it requires not only the assessment of linguistic accuracy but also the evaluation of reasoning, argument structure, and originality [4].

Therefore, within the education realm, it has become a pertinent concern to ensure grading consistency and reliability for long-essay responses [2].

Scholarly interest in AI-assisted grading has been increasing greatly, particularly due to recent advancements in transformer-based large language models (LLMs) [5–7]. These models have proven remarkable capabilities in text generation and evaluation, raising optimism for their deployment and application in educational assessment. Recorded benefits include the ability to process and handle large volumes of scripts efficiently, consistently grade them, reduce human grading burdens, mitigate human bias, and provide timely feedback [7]. Notwithstanding their apparent benefits, questions remain about their reliability, validity, fairness, transparency, and explainability [8–12]. Evidence suggests that the performance of AI grading systems often varies by task type, subject domain, and scoring dimensions [13]. Thus, empirical validation in high-stakes and professional contexts remains critical.

These concerns are even more pronounced in domain-rich fields where assessments rely on strong analytical reasoning and professional judgement. Wherein, grading goes beyond understanding surface-level grammar or vocabulary [5], to an in-depth understanding of the content relevance of the specific domain, argument quality, and critical reasoning [1]. Although prior studies have examined automated grading across various subjects, including secondary education [13] and higher education disciplines such as physics [12], chemistry [3], mathematics [11], English language [10], and dental education [5]. Yet, empirical evidence on its extension to professional education and high-stakes certification contexts remains limited. In particular, little is known about how AI grading systems perform in professional accounting assessments, how well AI confidence indexes (CI) reflect grading accuracy, and whether certain types of scripts are systematically more difficult for AI systems to grade. This represents an important gap, given the high stakes and professional implications associated with assessment decisions in accounting education.

Furthermore, the application of AI in assessment is more challenging due to technical issues such as model bias and calibration. AI systems may be overly stable or overly sensitive owing to bias in the training dataset [14], while poor calibration can lead to CIs that do not correspond to actual accuracy [15]. These issues can lead to uneven performance across subject domains and response types [1], raising questions about reliability, fairness, and credibility, and consequently undermining trust in automated grading outcomes [16]. Therefore, understanding these risks is crucial when assessing the reliability and validity of AI-assigned grades, interpreting CI, and identifying systemic patterns in AI-ungradable scripts.

To address these gaps, this study empirically examines the performance of an AI system (i.e., ALES) in grading long essay responses within a professional accounting examination context. The study focuses on the reliability and validity of AI-generated grades relative to human grading, the role of AI CI in indicating grading accuracy, and the determination of systematic patterns in AI-ungradable scripts.

This study addresses the following research questions:

1. How reliable and valid are the ALES grades relative to the human grades?
2. What is the role of CI in grading accuracy?
3. Are there systemic patterns in ALES ungradable scripts?

While this study does not propose a new scoring algorithm, it introduces a structured, multi-layer assessment framework for testing the reliability, fairness, and confidence behaviour of operational ALES systems under real examination scenarios. Hence, the authors make several significant contributions to the growing discourse in educational assessment, particularly those in high-stakes contexts requiring higher-order and domain-specific reasoning. Practical contributions include:

- The current study, unlike simulation-focused evaluations often employed in AES research, evaluates an operational ALES system deployed for use in a large-scale, real-world, high-stakes professional accounting context. This domain has received less attention in prior research. Hence, broadening the frontier of AI assessment research beyond the general classroom settings.
- To assess various aspects of validity, reliability, and calibration, the study integrates QWK as a psychometric agreement measure, ROC/AUC for discriminatory analysis, and logistic regression for inferential modeling into a unified evaluation. Hence, it extends on prior literature's single metric validation approaches.
- By examining ALES system performance in terms of gradable and ungradable scripts, the study identifies structural contributors to ALES grading failure rather than treating ungraded tasks as residual cases. Therefore, addressing potential equity and fairness concerns associated with automated grading and advancing discourse on sustainable quality education and technological innovation, aligning with SDGs 4 and 9.
- Through integrating ALES CI into performance evaluation, the study links internal model accuracy with observed grading performance, enhancing transparency in AI-based assessment.
- Through the adoption of a task-specific modeling approach, the study acknowledges that long-essay responses require construct-sensitive evaluation instead of pooled grading assumptions.
- Finally, the study findings provide practical insights to policymakers, professional bodies, and educators pertaining to the responsible and transparent deployment of AI in high-stakes educational assessment.

Theoretically, the study enriches contemporary debates by offering insights into AI reliability through the lenses of validity theory [17,18], fairness framework [19,20], and calibration theory [15], underscoring the contextual nature of AI systems' performance across domains.

2. Literature Review

2.1. Conceptual Overview of Automated Long-Essay Scoring (ALES)

The AI system evaluated in this study is an Automated Long-Essay Scoring (ALES) system. ALES refers to the use of computational or technological systems (AI), typically underpinned by natural language processing (NLP) and machine learning (ML), to evaluate extended written student responses (i.e., essays) and assign grades or rubric scores [7,10]. Within the broader field of automated essay scoring (AES), ALES specifically targets longer, domain-specific responses that require the assessment of higher-order writing components like argument structure, evidence use, task relevance, and discipline-based reasoning. Compared to multiple-choice scoring or short-answers, ALES systems have extended beyond merely evaluating surface features like vocabulary, grammar, and response length, to approximate substantive, rubric-based human judgments [5,7]. In this study, ALES operates as an embedded scoring module within the examination platform and offers both predicted grades and a grading indicator.

Changes in underlying computational approaches are reflected in the evolution of ALES. For instance, artificial features such as word counts, grammar checks, and readability indices were what early systems relied upon. This captured basic text attributes but was insensitive to discourse and semantic aspects of writing. Recent studies, especially using pretrained transformer models (e.g., BERT and its variants), have enhanced the ability to describe contextual and semantic linkages in text, allowing for better modeling of essay content beyond surface patterns [21,22]. Despite these developments, the ability of ALES

models to reflect deeper aspects of writing quality, such as domain knowledge and critical thinking, remains an area for which research is still ongoing.

ALES systems are now gaining traction in educational settings that involve a lot of writing and where human grading requires huge resources. However, when used in high-stakes or professional assessment contexts, their deployment raises concerns about reliability, validity, fairness, and explainability. These concerns reflect both the technical limitations of the existing models and broader debates on AI's role in educational assessment.

2.2. Technical Approaches to ALES

Researchers have grouped ALES into three categories comprising feature-based/shallow models, deep learning neural models, and hybrid approaches. Feature-based or shallow models use statistical learners like linear regression or decision trees and human-generated textual features such as word counts, essay length, and grammatical accuracy. Although these models are good at capturing surface textual attributes, they are typically insensitive to coherence and semantic meaning, i.e., weak on deep content [23]. Early AES study showed that while feature-based models can approximate human scores on basic text classification tasks, they struggle with complex semantic tasks that require a thorough understanding of disciplinary content and argumentative structure.

Deep learning models rely on neural networks to automatically extract scores directly from text [7,21]. Architectures like recurrent neural networks (RNNs) and, more recently, transformer-based models have become dominant. Transformer models like Bidirectional Encoder Representations from Transformers (BERT), RoBERTa, and their derivatives leverage self-attention techniques to represent contextual linkages in text, making them especially suitable for grading tasks that rely on meaning and coherence [22,24]. This model tends to perform better at capturing semantic relationships and some discourse phenomena [7]. A recent study also examines graph neural techniques to capture structural linkages within essays, like integrating transformers with graph attention networks to enhance analytic grading across dimensions such as grammar and cohesion [25].

The last is the hybrid models, also referred to as the ensemble models. It combines handcrafted linguistic features or rubric signals with deep contextual embeddings. Hybrid models seek to capitalize on the strengths of rich semantic representations and structured verbal signals. For instance, relative to models relying solely on either approach, hybrid systems that integrate pretrained transformer embeddings with attributes like discourse patterns or grammar complexity have demonstrated increased accuracy and robustness [2,26].

Recent advancements also include models designed for explainability and multi-trait scoring. For instance, rationale-driven scoring frameworks produce two things: they generate a score and produce justification/explanation for the score. This improves transparency and interpretability without undermining accuracy [27]. Prior studies explore integrating contextual enrichment approaches into transformer models to improve their ability to identify complexities in essay content, producing competitive results on standard scoring benchmarks [28].

Although, these technical developments have enhanced ALES capabilities, models may still be susceptible to training data biases and content limitations. The need for continued empirical assessment and improvement is highlighted by performances varying by essay type, domain-specific knowledge, and scoring dimensions [14,28].

2.3. Key Evaluation Criteria for ALES

As ALES systems advance through LLMs and hybrid approaches, evaluation has gone beyond mere score matching to a broader set of psychometric and ethical requirements.

Scholarly literature now examines ALES via multiple dimensions like reliability, validity, domain sensitivity, interpretability, and fairness.

Pertaining to reliability and human assessors' agreement, a common approach in ALES evaluation is gauging agreement between human and AI scores using Quadratic Weighted Kappa (QWK). QWK considers the ordinal character of scoring scales while estimating inter-rater reliability. Moderate to strong agreement, often between 0.60 and 0.85, is reported in studies like [23,26,29,30]. These findings imply that human scoring patterns can be approximated by automated systems. However, measurement research noted that agreement does not automatically translate into validity because Kappa statistics might yield contradictory results owing to their sensitivity to scale distribution, prevalence effects, and rater marginal distributions [31]. High QWK scores could therefore mask deeper inconsistencies across subject areas, skill levels, and scoring contexts.

In terms of validity and subject domain specificity, researchers also noted that a substantial agreement may be found between AI and human grades. However, exact agreement remains difficult to achieve [9,13]. One persistent issue is that automated systems often capture surface coherence and linguistic fluency better than domain-specific reasoning. LLMs, according to [13], may replicate comprehensive human essay assessments but show reduced performance when tasks require applied or domain-specific reasoning. This issue is apparent in applied and professional settings. In an AES study for undergraduate dental examinations, ref. [5] reported strong agreement in one clinical scenario but noticeably weaker agreement in another, underscoring the sensitivity of ALES performance to baseline human grading consistency and rubric clarity.

On the performance of ALES on longer essays, findings documented from studies that specifically focused on this present mixed but typically positive results. Ref. [10] evaluated multiple LLMs (GPT-3.5, GPT-4, PaLM 2, and Claude 2) on English language learners' long essays and found promising reliability across assessors. Yet, they warned that the linguistic complexity and variability of longer essays introduce significant fairness concerns, particularly across demographic groups.

Hybrid modeling approaches have reported even higher agreement levels. Ref. [32] combined handcrafted linguistic features with deep contextual embeddings (RoBERTa), reporting very high levels of agreement with human assessors (QWK of 0.94). Similarly, ref. [26] employed transformer embeddings with handcrafted linguistic features and documented improved accuracy over pure neural/shallow models. Yet both studies remain limited by training data dependence, raising concerns about generalisability to professional, high-stakes contexts. Meanwhile, ref. [33] in psychology assessment observed that ALES systems often generate "average" human scoring patterns, particularly in areas with explicit rubrics. While this alignment lends credence to reliability, it raises concerns that automated grading may mask outlier or nuanced responses, where subject-matter expertise is crucial, as it is in high-stakes exams.

The importance of CI and explainability, particularly in assessing the reliability of ALES in addition to predicted accuracy, is also highlighted by an increasing body of empirical work. A recurring question is whether CI can alternatively be used as a reliable assessment tool for decision-making in high-stakes exams or whether they only provide a surface level of transparency without addressing fundamental validity issues. In this regard, ref. [34] found that explanations and CI in automated systems only slightly boost user trust, but do not significantly increase decision quality. Additionally, ref. [2] show that CI frequently represent system's internal process certainty rather than actual grading accuracy. Similarly, ref. [12] argue that CI only offers surface-level transparency and fails to address core issues like construct validity or equity. Collectively, these submissions

reinforce that in a high-stakes domain where ALES is being applied, CI cannot be relied upon as sufficient proof of validity.

Fairness and bias have grown into a core assessment component. AES trained on skewed datasets may perform poorly for underrepresented competency levels or demographic groups, as [35] shows. This suggests that systematic subgroup disparities, which are especially important in consequential assessments, might be concealed by overall agreement metrics.

An underexplored aspect of ALES assessment is system robustness, particularly regarding the handling of ungradable scripts. Although occasional grading failures are acknowledged by [23], these cases are typically regarded as minor inconsistencies. Some researchers, such as [7,34], noted that the automated systems sometimes discard invalid responses or fail to grade all scripts, yet systemic analysis of these failures is largely absent. Whether this failure is random or concentrated in certain subjects remains unclear.

2.4. Benefits and Limitations of AES/ALES Models

Scalability and efficiency are two core advantages of ALES that scholarly work has highlighted several. According to [7], automated systems can process large volumes of essays in less time compared to human graders, reducing grading and feedback turnaround time. This scalability is especially beneficial in large-scale assessments, professional certification, and other high-stakes contexts where thousands of scripts must be reviewed within short timeframes. ALES systems are also linked to consistency in scoring. Unlike human assessors, automated models are unaffected by fatigue, mood, or assessor drift, which can introduce variability over time. Research on automated grading demonstrates that, once calibrated, systems may apply grading criteria consistently to all responses, potentially eliminating random subjectivity in grading [8,23].

The potential of formative feedback is another documented strength. Ref. [9] mentioned that automated systems can offer diagnostic comments on writing quality, organization, and language use, fostering learning that goes beyond merely calculating scores. This feedback feature may assist students in matching their answers to professional standards and discipline-specific expectations in professional education settings like accounting [26]. In addition, rather than replacing human judgment completely, ALES is increasingly being explored as a decision-support tool. Several studies also suggest hybrid workflows where automated systems grade scripts first, flagging low-confidence or odd responses for human review while flagging high-confidence cases through automated processing [13,34]. For complex or unusual cases, the model improves efficiency while maintaining human oversight.

ALES have also been noted as contributing to cost effectiveness and reproducibility [6]. Once implemented, ALES lowers reliance on a large number of trained human graders, which is especially crucial in resource-constrained settings. Additionally, because grading criteria are stable over time, consistent algorithmic grading facilitates program evaluation and longitudinal tracking [4]. Collectively, these benefits position ALES as a promising innovation to persistent challenges in large-scale assessment systems.

Despite the numerous benefits highlighted, literature notes persistent issues regarding validity, fairness, transparency, and suitability, particularly in high-stakes and professional contexts.

One issue of particular concern is construct validity. While ALES often generates strong agreement with human grades, a number of studies demonstrate that models may rely heavily on surface-level textual attributes like essay length, word count, or grammatical complexity. This raises the probability that systems will prioritize language proficiency over in-depth analytical reasoning or domain knowledge [8,13]. This concern is particularly

pronounced in fields like accounting, medicine, and law, where evaluation is meant to capture applied reasoning and professional competence rather than mere writing style [4,5].

Another critical limitation centres around fairness and bias [36,37]. ALES systems may systematically disadvantage specific demographic groups, or students from certain cultural backgrounds or proficiency bands, if training data underrepresent such groups [20,35,38]. Reviews of algorithmic bias reveal that poor model generalization outside of the populations represented in training data may give rise to performance disparities [20,35]. For example, essays written by students who are non-native English speakers or those that use unconventional writing styles may be disproportionately assessed and misclassified, notwithstanding whether their disciplinary reasoning is sound [34]. These risks question the equity of automated grading in high-stakes contexts.

Domain generalization presents another challenge for ALES systems. When models are applied to prompts, genres, or professional settings that are different from their training data, performance often reduces. Claims that general-purpose language models can be used without task-specific adaptation are undermined by this poor cross-domain transfer [7,39]. Therefore, before being used in professional exams, thorough domain-specific validation is necessary.

System transparency is another weakness of ALES. The majority of existing grading models, particularly those based on DL and LLMs, function as "black boxes," making it difficult for users and educators to comprehend how scores are generated [7,13]. Limited interpretability compounds appeal requests, accountability, and stakeholder trust.

Lastly, system robustness is another technical issue that ALES models continue to encounter. Automated systems could sometimes fail to grade scripts due to formatting errors, problems with handwriting recognition, or unusual content structures. These ungradable scripts may pose serious consequences for students in high-stakes professional assessments [6]. Yet, this issue has received little to no attention in the literature, leaving unanswered questions about how often systemic failures occur and whether they disproportionately affect specific subjects or response types.

Collectively, these limitations imply that although ALES has potential, its application to professional certification disciplines like accounting necessitates its in-depth analysis.

2.5. Synthesis of Literature and Research Gaps

ALES has recorded significant strides according to extant literature. Previously unattainable levels of agreement with human assessors are now attained by hybrid and LLM-based systems. In addition, evaluation systems have grown beyond raw accuracy to now include fairness, CI, and interpretability. Despite this growth, significant gaps still exist. Domain-specific validity is an area that is still insufficiently tested, as the majority of ALES study is conducted in general language or education assessment contexts. Professional fields like accounting and finance, which require structured professional judgement, interpretation of domain-specific standards, and applied reasoning, remain largely unexamined. Examination of fairness is also still limited. Although a number of studies acknowledged subgroup performance disparities, a comprehensive analysis across subject areas has yet to be explored, particularly for long, cognitively demanding essays. Furthermore, CIs are usually overinterpreted. Empirical study suggests that CIs do not consistently track true grading accuracy or validity in high-stakes situations, despite being presented as transparency tools. The most crucial yet almost non-existent in empirical evidence is the issue of ungradable scripts. Previous research concentrates on the degree of agreement between graded responses, ignoring the potential for automated systems to completely fail to grade some scripts. Whether these failures disproportionately affect certain students or subject areas is unclear. These gaps are especially crucial in professional accounting

examinations, because assessment ought to reflect not merely writing proficiency but also technical accuracy, standards interpretation, and professional reasoning. In this domain, demonstrating overall agreement might be insufficient. ALES system must prove fairness, domain-sensitive reliability, and robustness to failure. Therefore, addressing these unresolved issues provides the central motivation for this study.

2.6. Theoretical Background

The foundational complementary theories guiding this study are validity theory in educational measurement, fairness theory in AI decision systems, and calibration theory. Collectively, these frameworks offer a structured basis for assessing the performance of ALES in a high-stakes professional accounting examination while acknowledging its limits.

The validity theory argues that assessment scores are not inherently meaningful by themselves, but their value is dependent on supporting evidence for their interpretation and application [17,18]. This theory is core to educational assessment, especially high-stakes examinations, where students' professional progression is influenced by their results. Validity in high-stakes professional exams refers to whether score interpretations appropriately reflect the competencies the exam ought to measure and if the score implications used are justifiable. Within ALES systems, agreement with human grades offers criterion-related evidence; however does not by itself establish construct validity or justify high-stakes consequences. It is, however, important to understand that even if human and AI scores seem similar, the system may still be flawed, especially if it rewards surface-level linguistic characteristics rather than domain-specific reasoning. Hence, instead of confirming full construct or consequential validity, this study defines AI-human agreement as limited validity evidence.

Fairness is another factor that ought to be considered when automated systems are employed in grading. This theory emphasises that AI systems should not trigger systemic performance inequalities across groups or contexts unless reasonably justified [19,20]. Fairness is crucial in high-stakes professional examinations because it extends beyond demographic parity to include procedural fairness. This implies that the system operates consistently across subject areas without disproportionately failing to process any specific response forms. This study operationalises task-level fairness as it examined whether ALES system failures are clustered in certain task subjects. Hence, analysed ungradable scripts task-specific clustering using a procedural fairness perspective. This perspective aids in determining whether reliance on ALES can inadvertently disadvantage certain task subjects.

Finally, calibration theory is concerned with the relationship between CI and accuracy in decision-making [15,40,41]. It is believed that a well-calibrated system is one in which higher CI accurately align to a higher likelihood that a prediction is correct. Although most automated systems now generate CI alongside automated grades. However, the practical meaning of these indices in high-stakes professional contexts is still unclear, particularly if higher CI reliably means higher probability of accurate grades. This study, therefore, examines whether AI-generated CI meaningfully translates to minimal grading differences. Calibration theory is employed here to assist in determining if internal certainty signals offer actionable information for human oversight, and not to assume total reliability.

Overall, these frameworks position ALES as a probabilistic assessment tool whose outputs necessitate interpretation, monitoring, and contextual justification. Specifically, validity theory explains agreement as partial evidence, fairness theory emphasises the essence of auditing performance consistency across task areas, and calibration theory establishes whether ALES CI is informative or deceptive.

By incorporating these perspectives, our study moves the discussion beyond mere accuracy comparisons and provides a logical perspective to evaluate the validity, fairness, and reliability of automated grading while maintaining caution in its deployment in high-stakes professional contexts. Figure 1 summarises the integrated theoretical framework explained herein.

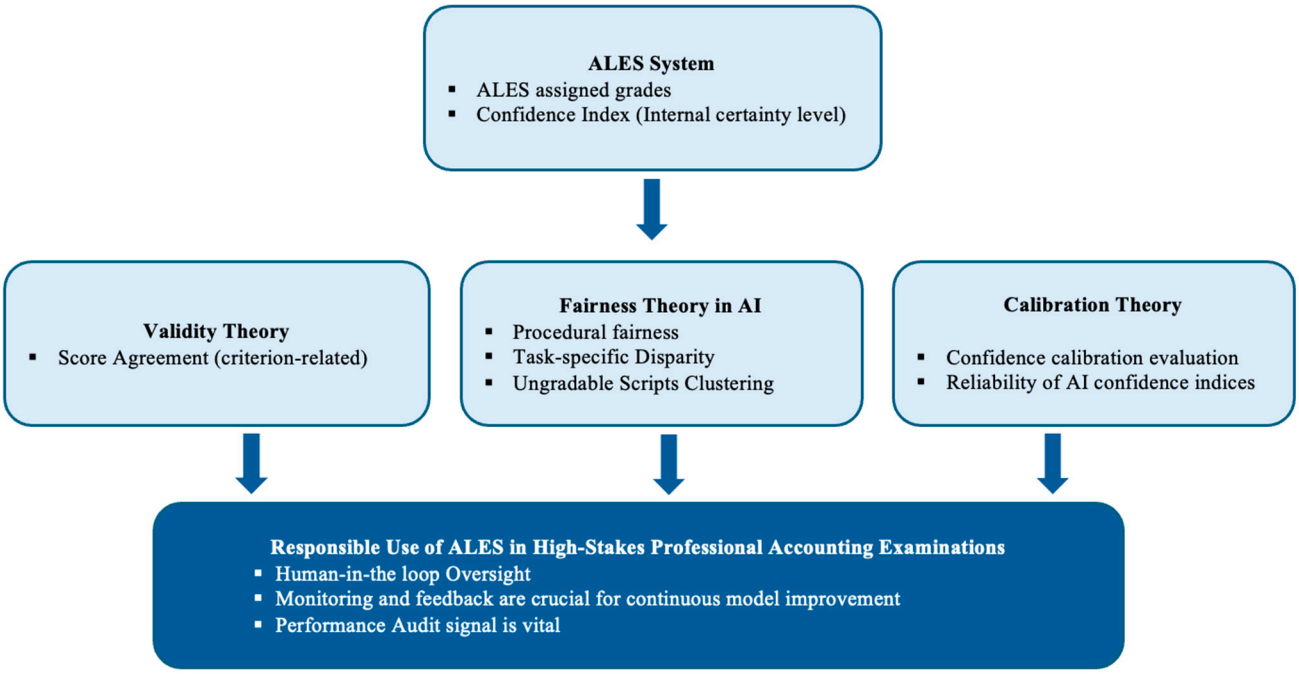


Figure 1. Theoretical framework. Source: Author's design.

3. Method

3.1. Research Design

This study employed a quantitative evaluation research approach to examine the performance of the ALES system relative to human grading in a high-stakes professional examination context. Structured as a multi-dimensional evaluation framework, this design integrates reliability analysis, fairness diagnostics, and probabilistic confidence assessments to systematically assess algorithmic grading behaviour under operational conditions instead of serving as a mere comparison.

The study scope is the assessment of the professional competence (APC) programme, a nationally approved and highly qualifying examination pathway to professional accounting certification in South Africa. The APC is a high-stakes assessment designed to examine candidates' professional prowess in terms of technical competence and integrative problem-solving skills through extended, case-study written responses. This exam is conducted online via secured platform named Secure Client, which provides a controlled MS-Word environment. Candidates must complete eight subject-specific long-essay computer-based exams at a designated venue, with the exam process spanning six days.

This context offers an operational, naturalistic test platform for assessing algorithmic grading systems under real-world assessment conditions. As a result, the research design places more emphasis on methodological assessment of algorithmic outputs than algorithm development, concentrating on the behaviour of ML-based scoring systems in complex, expert-judgment domains. Through the integration of subgroup fairness diagnostics, agreement analysis, and confidence calibration assessment into a single framework, the study offers an applied evaluation approach that aligns with emerging research on the responsible application of ML in high-stakes decision environments.

3.2. Data Description, Source, and Sample

The data used comprised 15,688 candidates' scripts drawn from the APC examination across eight tasks (subjects). In total, 1961 scripts were taken from each task, culminating in a balanced task-level distribution. Each script contained a long essay case-study written responses having an expected length of several thousand words. This required evaluative professional judgment rather than a short answer. The subjects assessed were Accounting I (IFRS selected standards), Finance I (Decision-making), Taxation, Finance II (ABC standard), Technology, Accounting II (Foreign operations), Auditing, and Ethics.

For each task, the individual script received a human-assigned grade from a trained professional examiner, an AI-generated grade produced by the ALES system, and a CI also generated by the ALES system, indicating its internal certainty in relation to the assigned grade.

The official APC ordinal scale, ranging from 0 (No attempt) to 5 (Highly competent), was used for human grades. Meanwhile, the ALES system was designed as a decision-support classifier and only assigned grades in the 2–4 range. Scripts that were predicted to fall into the extreme categories (0–1 or 5) were automatically flagged for human review. The analysis conducted in this study considers this operational constraint, which is crucial for interpreting agreement statistics. The model's internal estimate of classification certainty is represented by the CI, a probabilistic output ranging from 0 to 1. CI values were rescaled to a 0–100 metric for interpretability. The CI functions as an uncertainty signal that is assessed independently through discrimination and calibration analyses rather than directly representing grading accuracy. Ungradable scripts were those for which the ALES system failed to provide a grade. Therefore, to ascertain whether non-grading occurred randomly or clustered by task, these cases were retained in the dataset and coded as a binary outcome variable for further modeling. Table 1 summarises the variables employed in this study.

Table 1. Variable definition.

Measures	Definition
Human Grades	Ordinal score ranging from 0–5 assigned by professional APC examiners following the established grading rubrics.
AI Grade	Ordinal category ranged 2–4 predicted by ALES classifier with extreme categories (0–1 or 5) referred to human review.
Confidence Index (CI)	ALES system generated probability (0–1) which indicates internal prediction certainty. This was rescaled to 0–100 for analysis.
Ungraded Scripts	Binary indicator revealing whether ALES system didn't grade a script. Here, 1 = ungraded, and 0 = graded.
Tasks A–H	Subject area of the script from the eight APC examination tasks.

3.3. Human Graders and the Grading Process

The APC grading process takes approximately three weeks, and it involves about 100 professional examiners, all of whom hold the Chartered Accountant (South Africa) [CA(SA)] qualification. All selected graders are subject experts who have prior experience in APC marking. While most graders are returning examiners, new experts are recruited periodically, and they undergo similar training procedures.

Every examiner takes part in a standardized training and calibration process prior to the commencement of grading. This entails using a grading expectation framework that outlines performance indicators for each grade category, reviewing exemplar responses, and having organized conversations about case requirements. Each grader is required to score three pre-marked consistency scripts as part of the calibration process. Task setters and senior markers have already assessed these scripts, offering a benchmark. When discrepancies arise between their scores and the benchmark grades, graders adjust their interpretations. Only graders who successfully complete this calibration step are allowed

to proceed to operational marking. The scripts were graded using the official APC ordinal scale below:

5 = Highly competent
4 = Competent
3 = Borderline competence
2 = Limited competence
1 = Not competent
0 = No attempt

In reality, candidates often have grades concentrated in the range 2–4, with very few candidates obtaining 5. Minimal or non-substantive responses are usually represented by scores between 0 and 1. All marking is done electronically through a secure platform called AppSheet, where graders access and submit scores using structured marking grids. Senior examiners monitor the grading progress and score distributions daily to identify anomalies or drift. While operational APC marking does not require double blind-scoring of all scripts, the structured calibration process, use of benchmark scripts, and continuous monitoring and supervisory oversight act as quality control mechanisms to support grading consistency. While formal human–human inter-rater reliability statistics were not available for this dataset, the human grades were treated as the operational reference standard rather than the definitive ground truth. This limitation is acknowledged in the interpretation of AI-human agreement outcomes.

3.4. ALES System and Grading Process

ALES grading was introduced approximately five days after the human grading process began. Note that it only offered an additional, independent scoring signal that could support quality control and flag scripts needing further human review. Hence, it did not function as a replacement for human examiners.

3.4.1. Model Design and Architecture

The ALES system employed in this study implemented a task-based supervised ML pipeline instead of a generative LLM, with individual task subject modeled separately. The workflow procedure is based on the following stages:

i. Text Pre-processing: Candidates' responses were separated on a task-by-task basis, and scripts with invalid written responses or extremely few words (length less than 100 characters) were removed to avoid training on non-substantive content.
ii. Text Representation (Embedding Layer): Each task response was transformed into a dense numerical vector via a sentence-transformer embedding model hosted locally (nomic-embed-text-v1). This model converts textual data into semantic vector space representations suited for downstream classification. Local hosting ensured that sensitive examination data were not transferred to external servers, preserving data confidentiality.
iii. Predictive Target Simplification: The model was not trained to predict the entire 0–5 grade scale. Rather, it divided scripts into three categories, namely: 2 (limited competency), 3 (borderline competence), and 4 (competent). or was developed and deployed by the examination body's technology vendor prior to this research. The authors were not involved in the model development; rather technical description was obtained to support methodological transparency. Grades 0–1 and 5 were considered exceptional cases, requiring human judgment. This design reflects the operational grading distribution, which places most scripts in the range 2–4.
iv. Classifier Model: For each task, multiple logistic regression classifiers were trained using the embedded representations as input characteristics. Hyperparameter variations

(such as regularization strength) were investigated, and the best-performing model was chosen for each task.

3.4.2. Training and Evaluation Protocol

To ensure label quality, for each task, around 300 scripts previously graded by experienced professional examiners were used to train the machine. The labelled dataset for the individual task was randomly divided into training (75% and test (25%) subsets. Model performance during design was assessed exclusively on the held-out using precision, recall, and F1-score, ensuring that the model selection was not based on the same data as training. This internal evaluation step measured the classifier's ability to reproduce human-grade categories under controlled conditions. However, the core study analysis presented in this paper used the full operational dataset, reflecting real-world deployment conditions rather than a completely isolated experimental test.

3.4.3. Operational Use During Scoring

Upon completion of model training, the ALES system generated predicted grades and an associated CI for operational scripts. The CI was derived from the classifier's predicted probability for the assigned class and rescaled to a 0–1 interval, with higher values translating into stronger model certainty. Note that AI predictions were only employed in a decision-support capacity. Meanwhile, scripts that showed large discrepancies (for instance, human = 4 vs. AI = 2) were screened out for further human review by a second experienced human professional examiner, and scripts where AI and human grades were closely aligned were retained without changes.

This workflow was designed as a quality assurance technique to help eliminate oversight errors. However, because some human grades may be revised upon AI flagging, the final dataset represents an interactive human-AI process rather than a completely independent comparison. Also, this potential source of dependence is acknowledged as a limitation in the interpretation of the agreement statistics. Figure 2 presents the architectural flow of the ALES system.

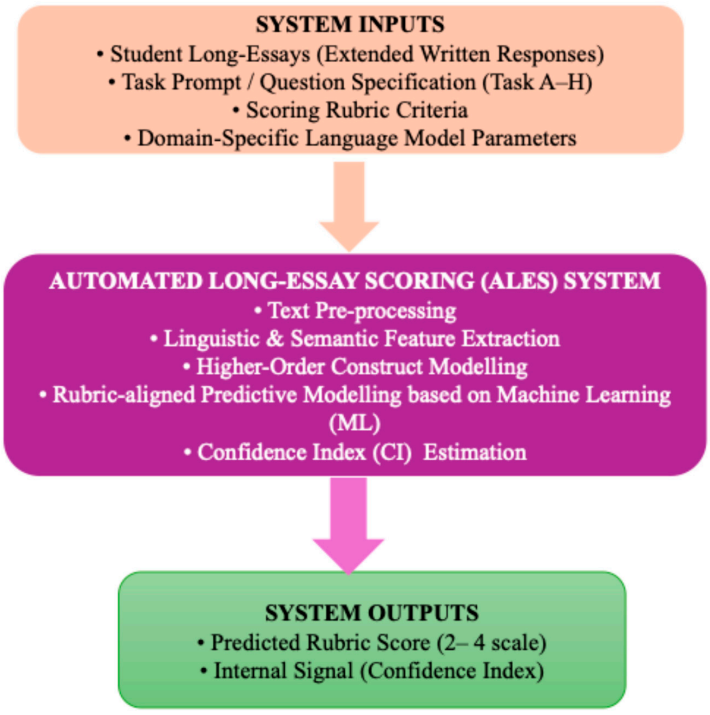


Figure 2. Evaluated ALES system architecture. Source: Author's design.

3.4.4. AI Techniques Underpinning the ALES Pipeline and Operations

Table 2 is a summary of the foundational AI techniques underpinning each of the ALES pipeline steps, along with configuration attributes, functional purposes, and known benefits and limitations.

Table 2. Foundational AI techniques of ALES pipeline.

Stages	Technique/Approach	Key Parameters/ Configuration	Purpose	Advantages	Limitations
Text Pre-processing	Task-based segmentation; filtering for minimum response length	Task partitioning and minimum character threshold (i.e., <100 characters are filtered out);	Removal of invalid or non-substantive responses and data preparation for modeling	Enhances training quality; ensure noise reduction	Very concise but valid responses may be excluded
Linguistic & Semantic Feature Extraction	Sentence-transformer embedding model (nomic-embed-text-v1)	Embedding dimensionality; locally hosted deployment	Transform essay text into contextually meaningful, dense semantic vector representations	Captures contextual information and semantic similarity	Reliant on the pre-training datasets, it might not fully capture discipline-specific nuance
Higher-Order Construct Modeling	Feature aggregation for argument structure, evidence use, and task relevance	Construct-specific feature weighting (task-specific)	Model reasoning quality and alignment with rubric structures	Extends beyond linguistic surface characteristics	Complex constructs may not be fully operationalized.
Rubric-aligned Predictive Modeling (ML)	Task-specific supervised ML classifiers	Task-level training; Model-specific hyperparameters (e.g., learning rate, regularization).	Transform semantic representations into rubric score categories	Enables consistent rubric application and task-sensitive grading	Performance is reliant on the quality of labelled data; risk of training distribution bias
Confidence Index (CI) Estimation	Threshold-based decision rule applied to model output	Confidence threshold parameter determining graded vs. ungraded scripts	Identify low-confidence predictions and failing to assign grades	Minimize erroneous scoring and introduces reliability measure	Sensitive to calibration accuracy and may increase the likelihood of grading failed attempt.

3.5. Methodological Analysis and Justification

The study employed multiple complementary analytical techniques to reflect the structure of the grading data and the distinct evaluation aims of validity, reliability, and fairness. First, because both human and AI outputs are ordinal categorical grades, inter-rater agreement statistics were employed to determine grading consistency. Specifically, Quadratic Weighted Kappa (QWK) was employed to account for the grading scale's ordered nature by penalizing larger disagreements more heavily than smaller ones. QWK is commonly used in automated scoring research to reflect performance levels rather than nominal classes [13,23,29–31].

Second, to determine how well AI scores distinguish between performance levels, Receiver Operating Characteristic (ROC) curve analysis was used. ROC analysis assesses the trade-off between sensitivity and specificity across decision thresholds and is useful for assessing a classification system's diagnostic accuracy in comparison to a reference standard. In this study, human grades were used as an operational benchmark to assess the ALES system's ability to correctly distinguish between adjacent competence categories [42,43].

Instead of relying on a single agreement statistic, this methodological framework was chosen to assess ALES performance across complementary validity aspects, namely agreement validity, discriminative validity, and structural discrepancy/systemic bias. Specifically, QWK evaluates the degree of concordance between ALES-assigned grades and human grades while accounting for the ordinal structure and partial disagreement. Meanwhile, because agreement alone does not evaluate discriminatory capacity, ROC analysis was

incorporated to assess the extent to which ALES's system generated CI distinguish between accurate and inaccurate grading.

Third, logistic regression modeling was employed to further assess structural determinants of grading differences in addition to agreement and discrimination. It is believed that this inferential approach would assist in identifying systemic bias patterns that may be associated with task-specific group/subject areas covered, which may not be detectable via agreement indices alone. For instance, specific script qualities may consistently influence the risk of the script not being graded by ALES. This approach is appropriate when the outcome variable is binary and allows for the estimation of the likelihood of an event occurring while controlling for various predictors. In short, the approach was employed to discover potential structural trends that could imply unintended bias in ALES grading outcomes, rather than to demonstrate causal correlations [44].

Combining psychometric (i.e., agreement statistic and diagnostic accuracy) with inferential (logistic regression) modeling technique allows for multidimensional evaluation and offers methodological complementarity that encompasses reliability, discrimination, and structural validity. QWK measures grading consistency, ROC analysis assesses classification effectiveness, and logistic regression investigates whether grading outcomes are related to script features. Collectively, these approaches allow for a structured evaluation of the ALES system's operational alignment with human grading, while acknowledging that human grades represent a realistic, not perfect, benchmark.

Note that evaluations based on kappa statistics alone run the risk of ignoring systemic bias patterns and discriminatory system performance. Hence, this triangulation approach employed in this study mitigates this limitation. Table 3 provides a comparative justification of the methodological approach employed in this study.

Table 3. Comparative methodological approach justification.

Method	Analytical Objective	Parameters	Benefits	Limitations	Reference
Quadratic Weighted Kappa (QWK)	QWK measure ordinal agreement between ALES assigned grades and human grades	Uses quadratic weighting scheme	Accounts for degree of disagreement on ordinal rubric scales	Sensitive to grade distribution and scale restriction	[6,23,30–32,45,46]
ROC & AUC	Assess classification discrimination (e.g., accurate vs. inaccurate grades	Thresholds, AUC statistics	Indicates discrimination independent of prevalence	Requires categorizing outcomes into binary and limits ordinal nuance	[2,42,46]
Binary Logistic Regression	Model predictors of specific outcomes (Gradable and ungradable scripts, Systemic bias patterns)	Odds ratios, coefficients, model fit indices	Multivariate analysis of systematic patterns	Assumes independence of observations	[47]

3.6. Ethical Considerations

To ensure confidentiality, all data were anonymized prior to analysis, and candidate's identifiers were eliminated. Permission was sought from the appropriate bodies, and ethical clearance under the code SAREC20230621/12 was obtained. The study acknowledges the high stakes of professional accounting assessments, hence, conform to the principles of responsible research in AI and education.

4. Results Interpretation and Discussion

4.1. How Reliable and Valid Are the AI Grades Relative to the Human Grades?

To address the first research question, the analysis was separated into two parts. The first looks at the level of agreement between the ALES system and human graders, while the second part examines whether ALES grading can be a reliable alternative to human

graders. For the first part, the task-specific agreement indices were computed, and per-task confusion matrices were examined to identify specific grading patterns. The agreement level was evaluated using exact agreement, within-one agreement (±1), Spearman rank correlation, and QWK with bootstrap confidence intervals. The second examines whether ALES grading can serve as a reliable support tool rather than a substitute for human assessment. Table 4 presents descriptive statistics for the variables used (human grades, AI grades, and agreement indicators).

Table 4. Descriptive statistics.

Variables/Predictors	N	Min	Max	Mean	Std. Dev
Human Grade	15,688	0	5	2.99	0.933
AI Grade	15,507	2	4	3.14	0.661
AI Confidence	15,507	0	1	0.58	0.160
AbsDiff	15,507	0.00	4.00	0.6631	0.74093
Exact	15,507	0.00	1.00	0.4593	0.49836
Within ±1	15,507	0.00	1.00	0.9102	0.28586
AI Ungraded Scripts	181	--	--	--	--

4.1.1. Descriptive Analysis

With a mean of 2.99 (SD = 0.93), the human-assigned grades across the full 0–5 scale (N = 15,688) indicate a moderate overall performance and reasonable spread in examiner judgements. Contrarily, the ALES system generated grades that were restricted to the 2–4 range by design for 15,507 scripts, with a mean of 3.14 (SD = 0.66). The narrower dispersion shows the system's constrained prediction scale and reduced variance instead of inherently higher prediction.

The AI-generated CI averaged 0.58 (SD = 0.16), suggesting moderate certainty in assigned grades [2]. However, caution must be taken so as not to interpret this as a direct probability of correctness, rather as a relative internal scoring signal, the predictive strength of which is evaluated separately using ROC and calibration analysis. The mean absolute difference (AbsDiff) between both ALES and human grades was 0.66 (SD = 0.74) on a scale of 0–5. Exact agreement level averaged 0.46 (SD = 0.50), implying it occurred in 46% of scripts, while 91% of AI scores fell within one point (±1) of the human grade, aligning with [6,8]. Although the within-one agreement level seems high, caution must be exercised when interpreting this result because the ALES system does not predict extreme scores, which systematically increase proximity to mid-scale human grades.

4.1.2. Ungraded Scripts

ALES system failed to grade 181 scripts, accounting for (1.15%). Although 7 of these scripts also had human-assigned grades between 0 and 1, suggesting no attempt or short written-response. However, the remaining 174 ungraded scripts were mostly between 2 and 3 human grades, with a disproportionate share clustered in task F. This pattern suggests task-specific structural bias instead of an evenly low-response quality as [12] previously opined. Further investigation pertaining to potential systemic model limitations is explored by subsequent regression modeling in the later part of the study.

4.1.3. Task-Specific Agreement Statistics

To determine whether performance varied meaningfully across tasks, agreement metrics were computed separately for tasks A–H using R Studio (version 4.3.1, 2023)

software. The results are presented in Table 5. The findings revealed that agreement varied substantially across tasks. Task F had the lowest (κ = 0.057), demonstrating weak ordinal reliability, with a bootstrap lower bound close to zero (0.007). This indicates instability and near-random ordinal alignment. Task A was also relatively weak (κ = 0.10) despite having very high within-one agreement (98.8%). The remaining tasks had (κ = 0.0.227–0.306), demonstrating a fair agreement level. Overall, these findings show that ALES grading reliability is task dependent.

Table 5. Per-task agreement statistic between human and AI grades.

Task	N	Exact Agreement	Within ±1	Spearman ρ	QWK (κ) Value	95% CI	Interpretation
A	1888	0.468	0.988	0.143	0.100	[0.069, 0.131]	weak agreement
B	1810	0.439	0.895	0.306	0.306	[0.263, 0.351]	fair agreement
C	1873	0.555	0.980	0.246	0.231	[0.189, 0.270]	fair agreement
D	1848	0.468	0.925	0.317	0.298	[0.257, 0.337]	fair agreement
E	1884	0.536	0.967	0.253	0.227	[0.186, 0.264]	fair agreement
F	1764	0.387	0.914	0.058	0.057	[0.007, 0.100]	weak agreement
G	1840	0.564	0.977	0.279	0.273	[0.233, 0.311]	fair agreement
H	1827	0.446	0.911	0.272	0.267	[0.225, 0.308]	fair agreement
	Overall inter-rater statistic (κ = 0.31)						

4.1.4. Per-Task Confusion Matrices

To further clarify where disagreements occur, per-task confusion matrices were examined, and the results are shown in Table 6. Note that because the ALES system is restricted to predicting grades 2–4, all analyses were conducted within the common grade range (2–4) to ensure easy comparability.

Table 6. Confusion matrices for tasks A–H.

Task A				Task B				Task C				Task D			
Human Grades	AI Grades			Human Grades	AI Grades			Human Grades	AI Grades			Human Grades	AI Grades		
	2	3	4		2	3	4		2	3	4		2	3	4
2	27	455	14	2	307	218	74	2	180	256	15	2	99	241	118
3	29	800	35	3	223	252	178	3	224	833	63	3	88	375	308
4	9	463	56	4	116	206	236	4	23	253	26	4	21	208	390
Task E				Task F				Task G				Task H			
Human Grades	AI Grades			Human Grades	AI Grades			Human Grades	AI Grades			Human Grades	AI Grades		
	2	3	4		2	3	4		2	3	4		2	3	4
2	0	123	63	2	77	278	77	2	9	116	37	2	335	288	89
3	0	327	462	3	139	478	176	3	10	473	285	3	204	342	198
4	0	227	682	4	75	337	127	4	6	349	555	4	74	160	137

The matrices above demonstrate that in task A, ALES almost never predicts score 2 correctly (for instance, when a human assessor assigned 2, ALES mostly graded 3). This explains the very high within ±1 (98.8%) and very low QWK (0.10), indicating systemic upward drift and not random disagreement. In task E, ALES failed to assign grade 2 completely, suggesting extreme central grade compression towards 3 and 4, and an indication

of structural classifier bias. As for task F, when the human assessor generally assigned grades 4 and 2, ALES either downgraded or upgraded to 3. This explains the extremely low QWK (0.06), suggesting heavy middle-grade compression. Contrarily, tasks B and D showed the most balanced results for all grade categories and better diagonal concentration. This explains their slightly higher QWK relative to other tasks.

Overall, these findings offer insight into why overall QWK was approximately 0.31, and why task-level QWK varied substantially (0.06–0.31) despite within ±1 agreement consistently high (≥0.89) across tasks. Predominantly suggesting boundary-level disagreement and not extreme misclassification. Hence, an indication that ALES alignment is not uniform across prompts.

The findings build on [13,18,48] and support bounded reliability interpretation, implying that ALES can function as a decision support tool for preliminary grading, consistency checks, and flagging unusual scripts. However, given task-specific variability, particularly the almost zero reliability observed in task F and fair ordinal agreement, the ALES system cannot yet be considered a standalone grading authority in high-stakes professional assessment. This aligns with contemporary validity theory, which posits that ALES grading systems can act as a decision-support tool, flagging cases that need human review rather than a substitute for an independent evaluative judgement.

4.2. What Is the Role of CI in Assessment Accuracy?

To address RQ2, the CI generated by the ALES system was assessed for whether it is a good predictor of grading accuracy relative to human grades. Accuracy was operationalized using the absolute difference between ALES and human grades (AbsDiff) and binary agreement indicators, including exact agreement and within-one grade point (within ±1). Correlation analysis (Table 7), which evaluated the association between CI and accuracy (AbsDiff), revealed a statistically significant but very weak negative association (r = −0.077 ***), suggesting that higher CI is linked to slightly smaller deviations from human grades. Although the magnitude of the linkage is negligible in practical terms, indicating that CI offers only a slight signal of grading accuracy.

Table 7. Summary of outputs on the relationship between ALES confidence and grading accuracy.

Statistic/Model	Dv	Predictor(s)	Estimate	SE	95%CI
Correlation Coeff.	AbsDiff (AI-Human)	ALES CI	−0.077 ***	-	-
Logistic_Reg	Within1 (1 = within (±1 grade point)); Odd Ratio (OR) = 2.59	ALES CI	0.953 ***	0.186	[1.80, 3.73]
Constant	-	-	1.771 ***	0.058	-
ROC Curve (AUC)	Within1 (classification accuracy)	-	0.543 **	0.008	[0.527, 0.559]
Quartile Test	Within1 (by confidence quartiles)	Confidence quartiles	$\chi^2(3) = 68.04$ ***	-	-
Logistic_FE (Task FE with Cluster SE)	Exact Agreement (1 = exact match)	ALES CI	β = 1.52 ***, (OR = 4.57)	0.166 *	-
Task Fixed Effects	-	Task indicators	Included	-	-

Significance Level *** 1%, ** 5%, * 10%. All tests are 2-tailed. Note: Dv = dependent variable; key predictors used; SE = Standard error; 95% confidence interval; Pearson and Spearman correlation coefficients; Logistic_Reg = binary logistic regression output; constant = model constant; Quartile Analysis test output.

Binary logistic regression further examined if CI significantly predicts agreement. When predicting within ±1 grade point, CI was statistically significant (OR = 2.59, 95% CI [1.80,3.73], $p < 0.001$); however, model fit was minimal (Nagelkerke $R^2 = 0.004$). Given that

within ±1 agreement is highly prevalent (94.5%), classification accuracy (91%) largely mirrored the base rate of the dominant class, instead of meaningful predictive discrimination.

4.2.1. Discrimination Analysis

As a diagnostic check on the model, an ROC analysis was conducted for exact agreement (i.e., to ascertain whether CI discriminates between accurate and inaccurate grading). While the analysis showed inherently weak discriminative power, the ROC curve (Figure 3) further reveals an area under the curve (AUC) of 0.543 (SE = 0.008, 95% confidence interval (i.e., CI [0.527, 0.559]), indicating only marginal discrimination over random classification (AUC = 0.50). The ROC curve lies close to the diagonal reference line, confirming weak discriminatory power. This demonstrates that no threshold of CI meaningfully achieves an acceptable balance of both sensitivity and specificity. This suggests CI alone cannot serve as a reliable classifier of grading accuracy.

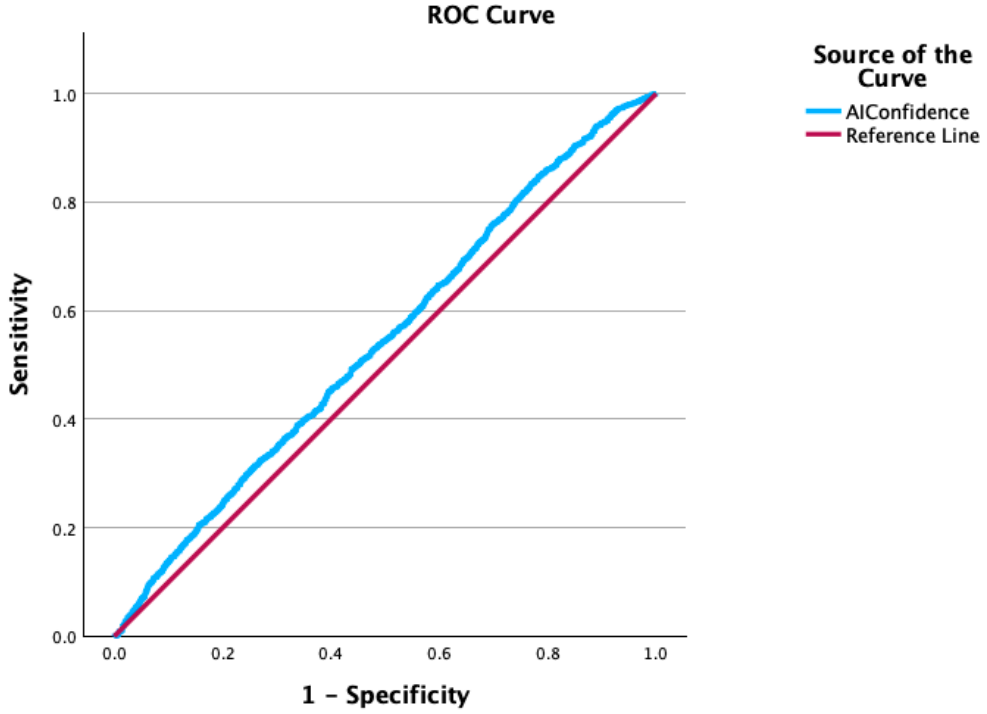


Figure 3. ROC curve showing the extent to which AI confidence accurately predicts grading (within ±1 point of human grade). Source: Author's design.

Furthermore, as a non-parametric robustness test of the relationship between the CI and agreement, a quartile-based check was conducted where CI were binned into quartiles (very low, low, medium, and high). The exact agreement moderately increased from 40.7% in the lowest quartile to 49.3% in the highest quartile. Although this trend was statistically significant (chi-square ($\chi^2(3) = 68.04$, $p < 0.001$, effect size was very small with Cramer's V ≈ 0.066). Despite that, at the highest confidence levels, agreement was still below 50%, reinforcing the limited practical value of CI as a standalone grading accuracy indicator.

4.2.2. Calibration Analysis

Given that discrimination and calibration capture different aspects of predictive performance, and CI is interpreted operationally as an indicator of grading reliability, a probabilistic calibration was also assessed using the Brier score and expected calibration error (ECE). Outputs are presented in Table 8. This is further substantiated with a reliability diagram.

Table 8. Calibration diagnostics of ALES CI.

Metric	Value	Interpretation
Brier Score	0.270	Worse than naive prevalence baseline (~0.249)
Expected Calibration Error (ECE)	0.109	Substantial miscalibration
Task-level Brier (Min–Max)	0.246–0.301	Heterogeneity across tasks
Highest Decile Confidence	0.902	Predicted
Highest Decile Observed	0.595	Empirical agreement
Confidence Gap (Decile 10)	0.307	Strong overconfidence

When evaluated for exact agreement, the overall Brier score was 0.27, exceeding the naïve prevalence baseline by approximately 0.249, implying limited probabilistic accuracy. The task-level Brier scores ranged between 0.246 and 0.301, showing meaningful heterogeneity in calibration across tasks. The ECE was 0.109, suggesting high deviation between the observed exact agreement level and the predicted confidence.

Systematic miscalibration was shown by a decile-binning-based reliability diagram (Figure 4). The model was slightly underconfident at lower confidence levels. However, from mid-range confidence levels, the system became increasingly overconfident. The observed exact agreement was 0.60, a discrepancy of roughly 30%, while the mean predicted confidence in the highest confidence decile was 0.90. Predicted confidence increased significantly (0.39 to 0.90), while observed agreement only slightly increased across bins (0.40 to 0.60). This divergence explains the higher Brier score and ECE.

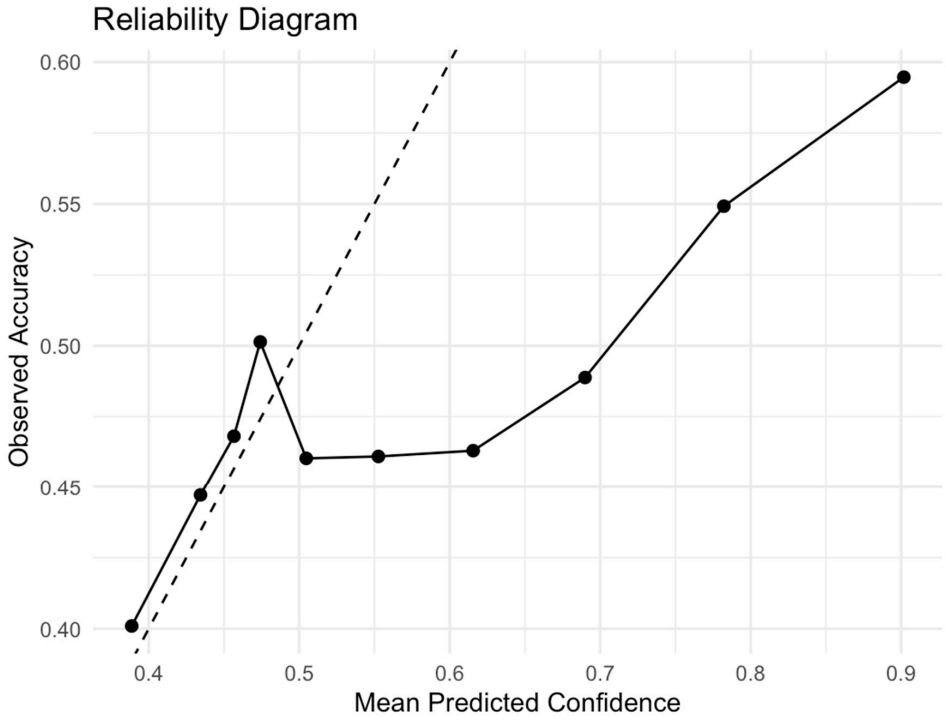


Figure 4. Reliability diagram of ALES CI. Source: Author's design.

4.2.3. Task Fixed Effects (FE) and Cluster-Robust Inference

Logistic regression models were re-estimated at the task level with task FE and cluster-robust standard errors (Table 7) to serve as a partial mitigation for potential heterogeneity. CI was still statistically associated with exact agreement ($\beta = 1.52$, SE = 0.17***; OR = 4.57).

However, when interpreted over realistic intervals, a 0.1 increase in CI translates to roughly a 16% increase in the odds of exact agreement, suggesting a modest substantive effect.

Collectively, these findings highlight that the CI's role in grading accuracy is of limited practical value. Reason being it only offers a small signal of grading accuracy, but insufficient alone as predictive utility to guide decision-making without human-in-the-loop oversight. Hence, it should be interpreted as a relative internal model signal and not as a calibrated probability of correctness. This aligns with the view of calibration theorists that CI offers some useful information but not sufficiently discriminative and calibrated for use as the single criterion for assessment measures in high-stakes professions without post-hoc calibration and human oversight. This finding builds on prior studies such as [2,4,12,13,15,34]. These studies argued that CIs have limited usefulness because they are apparently indicators of internal model reliability rather than grading accuracy.

Overall, CI operationally may support prioritizing scripts for further review, rather than serving as a standalone determinant of final grades. The moderate agreement level between the ALES system and human-graders suggests that the scalable quality assurance campaign in SDG4 is possible, where fairness and reliability are crucial. The heterogeneity observed across task types further reinforces the need for continuous improvement and human monitoring. These findings stress the role of AI as a catalyst for the design and development of a technologically sound and ethically grounded educational ALES systems, consistent with SDG 9's emphasis on resilient and responsible technological infrastructure.

4.3. Are There Systematic Patterns in AI Ungradable Scripts?

RQ3, which focuses on whether ALES system's inability to assign grades to certain scripts followed a systemic pattern, was addressed by delineating this aspect into three parts. First was the identification of the proportion of ungraded scripts by the ALES system. Second, was a comparison of their human-assigned grades distribution, and lastly was modeling of the predictors of ALES grading failure. Table 9 summarizes ALES-graded and ungraded scripts.

Table 9. Proportion of ungradable vs. graded scripts.

AI Grade Status	N/Percent	Mean Rank	Sum of Ranks
Ungraded = 0	181 = (1.2%)	7944.06	1,437,875.50
Graded = 1	15,507 = (98.8%)	7843.34	121,626,640.50
Total scripts	15,688		
Mann–Whitney U		1,385,362.50	
Z		–0.316	
p-value Asymp. Sig. (2-tailed)		0.752	

Out of the 15,688 scripts analyzed, the ALES system successfully assigned grades to 15,507 scripts (98.8%), while it failed to grade 181 scripts (1.2%). This indicates that AI failure was relatively rare but operationally vital in a high-stakes assessment context. Furthermore, to ascertain if ungraded scripts were systemically concentrated in a specific task or systematically due to performance level, a Mann–Whitney U test was conducted to compare the distribution of human grades between graded and ungraded scripts. The results were statistically non-significant with a *p*-value of 0.752, suggesting no evidence that ALES disproportionately failed on either higher or lower performing candidates. In short, this implies that grading failure was not performance-driven as there was no difference between the two groups. The average grade ranks were almost the same (mean rank of ungraded scripts = 7944.06; graded = 7843.34).

Contrarily, task-level prediction to determine if grading failure varied by task (estimated with a binary logistic regression model where 1 = ALES graded and 0 = ALES ungraded) is vital for interpreting coefficient direction. Indicators of task (A–H) and human grade were included as predictor variables (see Table 10). From the findings, task F alone was negative and statistically significant, suggesting that scripts in task F were significantly more likely to be ungraded relative to Task H ($\beta = -2.441$, $p < 0.001$). Hence, ALES grading failures were strongly concentrated in this task (see also Figure 5). All other tasks and human grade were insignificant, confirming the earlier nonparametric conclusion that failures were not related to candidate performance level. This concentration suggests task-specific constraint rather than random error or performance bias. This pattern is reflective of possible structural task attributes like domain-specific language, unusual response style, or response structure that deviate from the distribution of training data [49].

Table 10. Logistic regression predicting task-specific ALES grading status.

Variable (Predictor)	Coefficient	Wald	Exp(B) (OR)	Interpretation
Task A	0.704	1.643	2.021	Not significant
Task B	−0.257	0.369	0.774	Not significant
Task C	0.005	0.000	1.005	Not significant
Task D	−0.086	0.038	0.918	Not significant
Task E	0.019	0.002	1.020	Not significant
Task F	**−2.441** ***	**53.830**	**0.087**	**Highly significant**
Task G	−0.162	0.141	0.850	Not significant
Human Grade	−0.032	0.150	0.968	Not significant
Constant	5.360 ***	190.775	212.697	Highly significant
Model fit: $\chi^2(8)$		243.07 ***		
Nagelkerke R^2		0.13		
Classification Accuracy:		98.8%		

*** ALES grading failures were strongly concentrated in Task F. Hence, the higher likelihood of scripts here being ungradable.

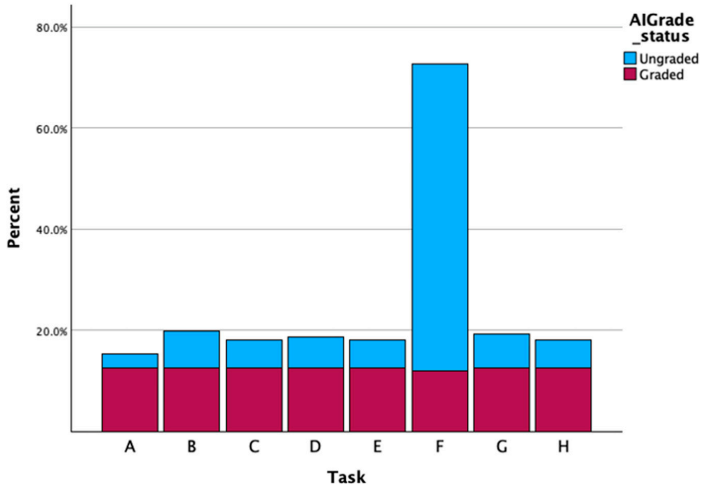


Figure 5. Bar Chart depicting task-specific ungraded scripts. Source: Author's design.

Linking the findings to fairness theory, this reflects a form of domain-specific performance difference where system effectiveness often varies across subject areas. While this issue has been raised in debates on algorithm fairness within education [20,34], and though not suggestive of demographic bias, the variation is still essential in professional assessment because it can build unequal measurement conditions across competency areas.

From the lens of calibration, the findings indicate that the AI model and its CI were not equally well-tuned across all task areas. This reinforces the importance of task-level validation and monitoring, rather than depending exclusively on aggregate performance measures.

To provide further insights into the ungraded script, task-level counts and proportion was computed, including exact binomial confidence intervals (see Table 11). Findings from this show that all tasks except task F had ungraded rates ranging between 0.25% and 0.66%, while task F alone had a higher ungraded rate of 5.61% (accounting for 110 out of 180 ungraded scripts), suggesting task F clearly drove the imbalance. The exact binomial CIs (4.63–6.72%) further confirm this is not noise, highlighting that task F failed to overlap meaningfully with the other tasks (~0.2–1.1%).

Table 11. Task-level ungraded scripts with exact binomial confidence intervals.

Task	Total Scripts	Ungraded (*n*)	Proportion (%)	95% CI
A	1961	5	0.26	[0.08, 0.59]
B	1961	13	0.66	[0.35, 1.13]
C	1961	10	0.51	[0.25, 0.94]
D	1961	11	0.56	[0.28, 1.00]
E	1961	10	0.51	[0.25, 0.94]
F	1961	110	5.61	[4.63, 6.72]
G	1961	12	0.61	[0.32, 1.07]
H	1961	10	0.51	[0.25, 0.94]
Total		181		

In addition, given the extreme imbalance, a Firth penalized likelihood model was estimated to assess rare-event bias and ensure robustness. However, the results of the standard logistic regression were stable, indicating no substantial change in estimates (Task F $\beta = 3.054$; human grade $p = 0.720$).

To further rule out potential positional bias, the mean absolute error for graded scripts was computed across the human grade (see results in Table 12). The results revealed a U-shaped pattern, with substantial increase observed in error at the extremes of the scale (Grades 0: M = 3.11; 5: M = 1.37), while it was lowest in the middle (Grades 3: M = 0.40; 2: M = 0.84; 4: M = 0.60). This pattern suggests regression toward the mean, with the ALES system exhibiting lower accuracy at performance extremes.

Table 12. Mean absolute error across human grade scale (graded scripts only).

Human Grade	Mean Absolute		
	Error	SD	N
0	3.11	0.66	433
1	1.61	0.67	207
2	0.84	0.64	3496
3	0.40	0.49	6502
4	0.60	0.61	4736
5	1.37	0.57	133
	Total		15,507

5. Conclusions

This study examined ALES systems' performance relative to human graders in a high-stakes professional accounting examination. Specifically, the study analysis explored the agreement level between AI and human-assigned grades; the role of CI as an indicator of accurate grading predictions, and the determination of whether systemic patterns exist in AI-ungradable scripts. Overall findings show a fair agreement level between AI and human-graders, with high within-one-grade agreement but low exact agreement. Rather than strong system equivalence, this pattern reflects central score clustering and the restricted ALES grading scale. The findings therefore imply that while ALES can resemble human judgment in many circumstances, particularly because it can approximate human grades, it cannot yet achieve the level of consistency required for autonomous use in high-stakes professional assessment. Its most appropriate use is as a support tool for workload management and flagging of scripts for human review.

The findings on CI showed statistically significant but practically weak relationships with grading accuracy. When class imbalance was considered, CI showed extremely limited discriminative capacity and minimal incremental predictive value. This suggests that the CI should not be viewed as a valid predictor of grading accuracy in isolation, but rather as a supplemental signal within a larger human-supervised workflow.

Further investigation of ALES ungradable scripts also revealed that ALES grading failures were not frequent overall. However, while it occurred, it was concentrated in one task area, indicating task-specific system constraint and not performance-induced, highlighting the importance of domain-specific validation, particularly stressing that performance measures may conceal unequal effectiveness across task areas.

Collectively, the study contributes to the ongoing debate on the operational behaviour of the ALES system within the professional certification context. The study contribution, therefore, lies not in proposing a novel scoring algorithm, but rather in systematically evaluating the ALES system using psychometric agreement measures, CI, and failure pattern modeling in a real-world high-stakes operational environment. This form of evaluation is vital to understand how ALES systems perform under a real assessment scenario. The findings support the intersection of SDGs 4 and 9 in shaping the responsible integration of AI in professional education. Evidence of ALES grading potential reflects the move toward an innovation-driven educational quality. Yet, preserving interpretability, fairness, and transparency remains essential for sustainable adoption. Therefore, the study holds several implications for policy, education, and accounting professional practice. For accountability, transparency, and fairness, education regulators and accrediting bodies need set explicit guidelines for automated grading systems. Additionally, professional bodies in the future can be better prepared to assess and calibrate AI tools responsibly should accounting curricula decide to fully adopt ALES grading for assessment in professional exams. AI can enhance feedback generation and grading efficiency for the profession. Although human-in-the-loop oversight is critical to ensure ethical deployment and contextual integrity. Furthermore, policymakers and educators must strengthen digital infrastructure to promote the achievement of SDG 9 while ensuring that AI deployment promotes quality learning outcomes (SDG 4). Ultimately, aligning educational innovation with sustainability imperatives will foster technological advancement that contributes significantly to inclusive and resilient academic systems.

From the lens of validity, the study provides criterion-related evidence regarding the alignment of AI and human scoring in professional accounting essays, while highlighting the need for further research on construct representation and the consequential validity of AI use in high-stakes assessment. Similarly, fairness was examined at task and performance levels; the absence of demographic data constrained conclusions about subgroup equity.

Overall, the findings lend credence to a cautious, evidence-based approach to integrating AI into professional assessment. ALES can enhance efficiency and consistency. However, existing evidence suggests ALES functions better as an assistive tool under structured human oversight instead of independent assessors.

This study is limited, firstly, in terms of a lack of human–human inter-rater reliability benchmark, which would have offered context for reporting AI-human agreement. Future research should integrate human–human reliability benchmarks, task-based performance diagnostics over multiple assessment areas, and detailed model calibration analysis. Second, the scripts could not be linked to candidate identifiers, limiting multilevel modeling and subgroup fairness analysis. Also, the CI calibration was not immediately changeable. Hence, further restricts the likelihood of post-hoc calibration enhancements. In addition, future studies should explore human moderation to ascertain whether ALES systems influence validity, reliability, and examiner workload in practice. Finally, the study is constrained by incomplete documentation by the vendor, specifically, regarding the author's inability to re-access hyperparameter values after system production and inter-assessor agreement statistics. Future studies that would be relying on third-party systems for automated grading should ensure proper and complete documentation and preservation of all prompts used during training, including statistics and values produced by the system, in accessible logs by the vendor.

Author Contributions: Conceptualization, N.O.G. and H.C.; data curation, H.C.; formal analysis, N.O.G.; investigation, N.O.G.; methodology, N.O.G. and H.C.; supervision, H.C.; validation, N.O.G. and H.C.; writing—original draft preparation, N.O.G.; writing—review and editing, N.O.G. and H.C.; visualization, N.O.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This work sought necessary permission to use the study data from the appropriate bodies, and ethical approval with code SAREC20230621/12 was obtained. This study acknowledges the high stakes of professional accounting assessments, hence, conform to the principles of responsible research in AI and education.

Informed Consent Statement: Informed consent was waived by the ethics committee due to the retrospective nature of the study and the use of anonymized data.

Data Availability Statement: Data for this study will be made available upon reasonable request from the corresponding author. However, availability is subject to approval from the authorizing body in charge of the professional examinations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ramesh, D.; Sanampudi, S.K. An Automated Essay Scoring Systems: A Systematic Literature Review. *Artif. Intell. Rev.* **2022**, *55*, 2495–2527. [CrossRef]
2. Li, Y.; Raković, M.; Srivastava, N.; Li, X.; Guan, Q.; Gašević, D.; Chen, G. Can AI Support Human Grading? Examining Machine Attention and Confidence in Short Answer Scoring. *Comput. Educ.* **2025**, *228*, 105244. [CrossRef]
3. Ade-Ibijola, A.; Chikezie, I.J.; Oyelere, S.S. Human vs. Machine Marking: A Comparative Study of Chemistry Assessments. *J. Sci. Educ. Technol.* **2025**, *34*, 1430–1440. [CrossRef]
4. Rupp, A.A.; Casabianca, J.M.; Krüger, M.; Keller, S.; Köller, O. Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. *ETS Res. Rep. Ser.* **2019**, *2019*, 1–23. [CrossRef]
5. Quah, B.; Zheng, L.; Sng, T.J.H.; Yong, C.W.; Islam, I. Reliability of ChatGPT in Automated Essay Scoring for Dental Undergraduate Examinations. *BMC Med. Educ.* **2024**, *24*, 962. [CrossRef] [PubMed]
6. Emirtekin, E. Large Language Model-Powered Automated Assessment: A Systematic Review. *Appl. Sci.* **2025**, *15*, 5683. [CrossRef]
7. Misgna, H.; On, B.W.; Lee, I.; Choi, G.S. A Survey on Deep Learning-Based Automated Essay Scoring and Feedback Generation. *Artif. Intell. Rev.* **2025**, *58*, 36. [CrossRef]

8. Hussein, M.A.; Hassan, H.; Nassef, M. Automated Language Essay Scoring Systems: A Literature Review. *PeerJ Comput. Sci.* **2019**, *5*, e208. [CrossRef]
9. Ifenthaler, D. Automated Essay Scoring Systems. In *Handbook of Open, Distance and Digital Education*; Springer Nature: Berlin/Heidelberg, Germany, 2023; pp. 1057–1071.
10. Pack, A.; Barrett, A.; Escalante, J. Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100234. [CrossRef]
11. Liu, T.; Chatain, J.; Kobel-Keller, L.; Kortemeyer, G.; Willwacher, T.; Sachan, M. AI-Assisted Automated Short Answer Grading of Handwritten University Level Mathematics Exams. *arXiv* **2024**, arXiv:2408.11728. [CrossRef]
12. Kortemeyer, G.; Nöhl, J. Assessing Confidence in AI-Assisted Grading of Physics Exams through Psychometrics: An Exploratory Study. *Phys. Rev. Phys. Educ. Res.* **2025**, *21*, 010136. [CrossRef]
13. Tate, T.P.; Steiss, J.; Bailey, D.; Graham, S.; Moon, Y.; Ritchie, D.; Tseng, W.; Warschauer, M. Can AI Provide Useful Holistic Essay Scoring? *Comput. Educ. Artif. Intell.* **2024**, *7*, 100255. [CrossRef]
14. Singla, Y.K.; Parekh, S.; Singh, S.; Li, J.J.; Shah, R.R.; Chen, C. Automatic Essay Scoring Systems Are Both Overstable and Oversensitive: Explaining Why and Proposing Defenses. *Dialogue Discourse* **2023**, *14*, 1–32. [CrossRef]
15. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, Australia, 6–11 August 2017; Volume 3, pp. 2130–2143.
16. Kizilcec, R.F.; Lee, H. Algorithmic Fairness in Education. In *The Ethics of Artificial Intelligence in Education: Practices, Challenges, and Debates*; Holmes, W., Porayska-Pomsta, K., Eds.; Taylor and Francis: New York, NY, USA, 2022; pp. 174–202.
17. Messick, S. Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *Am. Psychol.* **1995**, *50*, 741–749. [CrossRef]
18. Kane, M.T. Validating the Interpretations and Uses of Test Scores. *J. Educ. Meas.* **2013**, *50*, 1–73. [CrossRef]
19. Binns, R. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the Machine Learning Research*; PMLR: Cambridge, MA, USA, 2018; Volume 81, pp. 149–159.
20. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2021**, *54*, 115. [CrossRef]
21. Taghipour, K.; Ng, H.T. A Neural Approach to Automated Essay Scoring. In Proceedings of the EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 1882–1891.
22. Elmassry, A.M.; Zaki, N.; Alsheikh, N.; Mediani, M. A Systematic Review of Pretrained Models in Automated Essay Scoring. *IEEE Access* **2025**, *13*, 121902–121917. [CrossRef]
23. Shermis, M.; Burstein, J. Handbook of Automated Essay Evaluation: Current Applications and New Directions. In *Journal of Writing Research*; Stevenson, M., Ed.; Routledge: New York, NY, USA; London, UK, 2013; Volume 5, pp. 239–243.
24. Mayfield, E.; Black, A.W. Should You Fine-Tune BERT for Automated Essay Scoring? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 151–162.
25. Aljuaid, H.; Alhothali, A.; Alzamzami, O.; Assalahi, H.; Aldosemani, T. ET-GNN: Ensemble Transformer-Based Graph Neural Networks for Holistic Automated Essay Scoring. *IEEE Access* **2025**, *13*, 58746–58758. [CrossRef]
26. Atkinson, J.; Palma, D. An LLM-Based Hybrid Approach for Enhanced Automated Essay Scoring. *Sci. Rep.* **2025**, *15*, 14551. [CrossRef]
27. Do, H.; Ryu, S.; Lee, G.G. Teach-to-Reason with Scoring: Self-Explainable Rationale-Driven Multi-Trait Essay Scoring. *arXiv* **2025**, arXiv:2502.20748.
28. Chakravarty, A. Empirical Analysis of the Effect of Context in the Task of Automated Essay Scoring in Transformer-Based Models. Master's Thesis, University of Edinburgh, Edinburgh, UK, 2023.
29. Bond, W.F.; Zhou, J.; Bhat, S.; Park, Y.S.; Ebert-Allen, R.A.; Ruger, R.L.; Yudkowsky, R. Automated Patient Note Grading: Examining Scoring Reliability and Feasibility. *Acad. Med.* **2023**, *98*, S90–S97. [CrossRef]
30. Ndukwe, I.G.; Amadi, C.E.; Nkomo, L.M.; Daniel, B.K. *Automatic Grading System Using Sentence-BERT Network*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; Volume 12164 LNAI.
31. Kurdhi, N.A.; Saxena, A. Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. In Proceedings of the 16th International Conference on Educational Data Mining (EDM 2023), Bengaluru, India, 11–14 July 2023; pp. 103–113.
32. Faseeh, M.; Jaleel, A.; Iqbal, N.; Ghani, A.; Abdusalomov, A.; Mehmood, A.; Cho, Y.I. Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics* **2024**, *12*, 3416. [CrossRef]
33. Wetzler, E.L.; Cassidy, K.S.; Jones, M.J.; Frazier, C.R.; Korbut, N.A.; Sims, C.M.; Bowen, S.S.; Wood, M. Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teach. Psychol.* **2025**, *52*, 298–304. [CrossRef]
34. Conijn, R.; Kahr, P.; Snijders, C. The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *J. Learn. Anal.* **2023**, *10*, 37–53. [CrossRef]

35. Schaller, N.-J.; Ding, Y.; Horbach, A.; Meyer, J.; Jansen, T. Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays from Secondary Education. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 210–221.
36. Kwako, A.; Wan, Y.; Zhao, J.; Chang, K.W.; Cai, L.; Hansen, M. Does BERT Exacerbate Gender or L1 Biases in Automated English Speaking Assessment? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics; Association for Computational Linguistics (ACL)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 668–681.
37. Yancey, K.P.; LaFlair, G.T.; Verardi, A.R.; Burstein, J. Rating Short L2 Essays on the CEFR Scale with GPT-4. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 576–584.
38. Baffour, P.; Saxberg, T.; Crossley, S. Analyzing Bias in Large Language Model Solutions for Assisted Writing Feedback Tools: Lessons from the Feedback Prize Competition Series. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 242–246.
39. Yang, K.; Raković, M.; Li, Y.; Guan, Q.; Gašević, D.; Chen, G. Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 22466–22474. [CrossRef]
40. Lichtenstein, S.; Fischhoff, B.; Phillips, L.D. Calibration of Probabilities: The State of the Art to 1980. In *Judgment under Uncertainty*; Cambridge University Press: Cambridge, UK, 2013; pp. 306–334.
41. Yáñez, F.; Luo, X.; Valerio Minero, O.; Love, B.C. Confidence-Weighted Integration of Human and Machine Judgments for Superior Decision-Making. *Patterns* **2025**, *7*, 101423. [CrossRef] [PubMed]
42. Wang, D.; Keller, L.A. Using ROC Analysis to Refine Cut Scores Following a Standard Setting Process. *Educ. Psychol. Meas.* **2025**, *85*, 313–335. [CrossRef] [PubMed]
43. Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern. Med.* **2013**, *4*, 627.
44. Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E. *Multivariate Data Analysis*, 7th ed.; Prentice-Hall: Hoboken, NJ, USA, 2010.
45. Shermis, M.; Burstein, J.C. *Automated Essay Scoring: A Cross-Disciplinary Perspective*, 1st ed.; Routledge: London, UK, 2016.
46. Karabeg, M.; Petrovski, G.; Holen, K.; Sauesund, E.S.; Fosmark, D.S.; Russell, G.; Erke, M.G.; Volke, V.; Raudonis, V.; Verkauskiene, R.; et al. Comparison of Validity and Reliability of Manual Consensus Grading vs. Automated AI Grading for Diabetic Retinopathy Screening in Oslo, Norway: A Cross-Sectional Pilot Study. *J. Clin. Med.* **2025**, *14*, 4810. [CrossRef]
47. Doewes, A.; Saxena, A.; Pei, Y.; Pechenizkiy, M. Individual Fairness Evaluation for Automated Essay Scoring System. In Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022, Durham, UK, 24–27 July 2022; p. 206.
48. Lundgren, M. Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders. *arXiv* **2024**, arXiv:2406.16510. [CrossRef]
49. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.