

AI-Assisted Grading in English Language Education: A Systematic Review of Reliability, Bias, and Ethical Implications

Lodewikus Janse van Rensburg¹

¹ Department of Business Management, University of Johannesburg, Johannesburg, South Africa

Abstract. This systematic review synthesises peer-reviewed and emerging evidence (2011–2026, N = 27 studies, including a small number of preprints, working papers, and unpublished manuscripts) on the reliability, bias, and ethical implications of AI-assisted grading in English language education, applying a PRISMA-aligned thematic narrative synthesis. In contrast to reviews that treat these themes separately, it examines reliability, bias, stakeholder perceptions, and governance together, with rubric operationalisation and prompting conditions as cross-cutting explanatory factors. Findings challenge prevailing optimism surrounding AI assessment: reliability is not a fixed property of AI technology but a construct of rubric operationalisation, with studies employing prescriptive, criterion-anchored rubrics reporting good agreement (ICC = 0.814; Koraishi, 2024) and studies using holistic or domain-inferential rubrics reporting consistently poor agreement (ICC3 = 0.385; Manning et al., 2025; QWK $\approx$ 0; Gaggioli et al., 2025). Studies report systematic bias disadvantaging ELL and low-SES students (−0.10 z each; Kwako & Ormerod, 2024), with racial/ethnic disparities operating in non-intuitive directions (Yamashita, 2025). Critically, comparisons across studies are confounded by uncontrolled prompting conditions, temperature settings, and rubric design, a form of methodological heterogeneity that is rarely acknowledged as a threat to validity. Preliminary evidence suggests that open-source LLMs show comparable fairness to closed models at significantly lower cost (Oketch et al., 2025), challenging resource-barrier assumptions. The review concludes that AI grading calls for rubric transparency as a precondition, bias auditing as a regulatory standard, and human oversight as a core safeguard for equitable English language assessment. These conclusions rest on a narrative synthesis of heterogeneous studies rather than on direct empirical testing.

Keywords: AI-assisted grading; automated essay scoring; English language learners; assessment bias; rubric design; human-AI collaboration

1 Introduction

The adoption of artificial intelligence in educational assessment has accelerated dramatically since the public release of large language model (LLM) interfaces in late 2022. Within English language education, a field in which writing assessment is central to both placement

¹ Corresponding author: leonjvr@uj.ac.za

decisions and learning feedback, AI-assisted grading tools now sit at the intersection of competing pressures: institutional demands for efficiency, pedagogical commitments to feedback quality, and equity obligations toward linguistically diverse student populations. The promise is substantial: AI systems that can score essays consistently, instantly, and at scale. The risk, however, is equally substantial: deploying systems that systematically disadvantage the very populations they are meant to serve.

Research on automated essay scoring (AES) predates the LLM era by more than two decades, with rule-based and machine learning systems demonstrating moderate effectiveness in controlled benchmark settings (Shermis, 2014). Intelligent tutoring system research established that automated feedback can approach human tutoring effectiveness under certain conditions (Ma et al., 2014; VanLehn, 2011). However, the shift to LLM-based scoring represents a qualitative change: these systems were not purpose-built for assessment and bring different reliability profiles, failure modes, and fairness risks than their predecessors.

Despite growing empirical output, four fundamental questions remain insufficiently resolved for English language education practitioners. First, under what conditions does AI-assisted grading achieve acceptable reliability, and how do rubric design and prompting methodology interact with that reliability? Second, which student populations bear disproportionate scoring disadvantages, and through what mechanisms? Third, what do student and educator perceptions reveal about the acceptability conditions for AI grading, and do those perceptions align with actual reliability? Fourth, what ethical frameworks and governance structures are emerging to regulate AI assessment, and where do critical gaps remain? This review addresses these questions through systematic synthesis of recent evidence, with particular attention to methodological limitations that inflate or conflate apparent reliability gains.

This review differs from earlier syntheses in scope and purpose. Existing reviews have tended to address neighbouring topics separately: AES performance (Shermis, 2014), generative AI feedback (Lee & Moore, 2024), intelligent tutoring systems (Létourneau et al., 2025), and AI ethics in education generally (Yan et al., 2025). To the author's knowledge, none has jointly synthesised reliability, demographic bias, student and educator perceptions, and governance for English language education, or treated rubric operationalisation and prompting conditions as cross-cutting explanations for why findings diverge. The specific contribution of this review is therefore an integrative, explanatory synthesis that moves beyond summarising performance metrics to identify the conditions under which AI-assisted grading can be considered reliable and fair.

The review is organised as follows. Section 2 describes the systematic search and analytic methodology. Section 3 presents findings thematically across reliability, bias, student perceptions, and ethics. Section 4 discusses cross-cutting findings, including the rubric operationalisation problem and the prompting confound, before recommending a conditional, human-supervised deployment model for English language education contexts.

2 Method

2.1 Design

This study employs a systematic literature review following PRISMA 2020 reporting conventions (Page et al., 2021), with thematic narrative synthesis as the analytic approach. Narrative synthesis is appropriate because the included studies employ heterogeneous designs, outcome measures, and AI tools, precluding statistical meta-analysis. The review

does not seek to compute a pooled effect size but to identify patterns, contradictions, and boundary conditions across the evidence base.

2.2 Search Strategy and Corpus

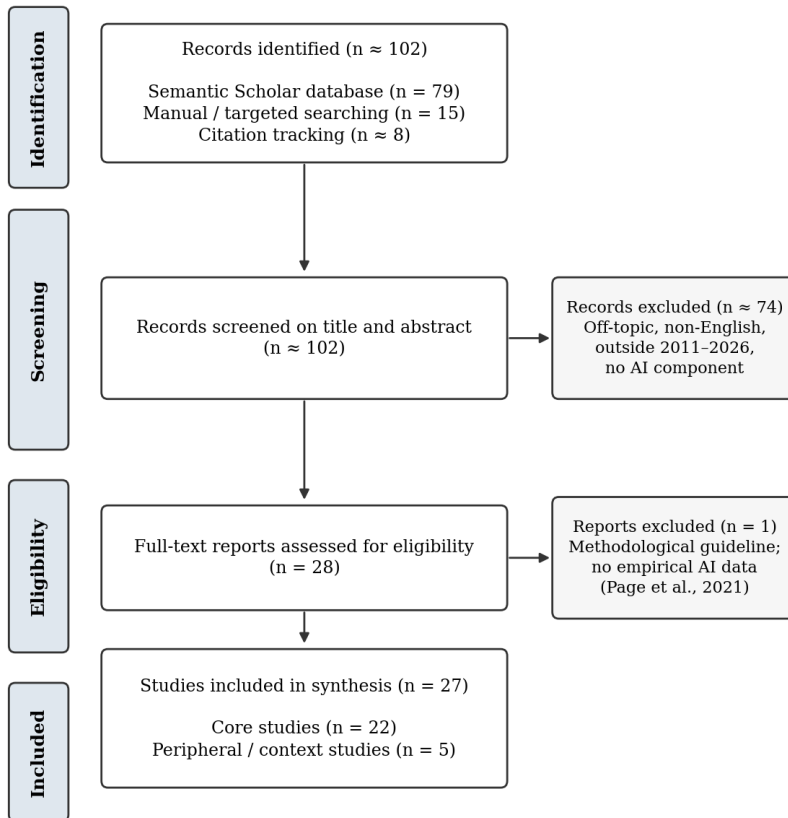


Fig 1. PRISMA 2020 flow diagram for the identification, screening, eligibility and inclusion of studies. The symbol ~ denotes counts recorded as approximate in the screening log

Four sources were searched in March 2026: Semantic Scholar (primary automated search), Google Scholar (targeted manual search), ArXiv (preprints), and forward/backward citation tracking from core papers. Search terms included: *automated essay scoring*, *AI-assisted grading*, *LLM assessment*, *ChatGPT essay*, *automated essay feedback*, *ELL writing bias*, *AES fairness*, and *intelligent tutoring writing*. The date range was 2011–2026, with the lower bound determined by VanLehn’s (2011) foundational ITS meta-review; the majority of included studies (24 of 27) were published 2023–2026.

From approximately 102 records identified, 28 were retrieved for full-text review. One was excluded (Page et al., 2021, a methodological guideline with no primary AI data), yielding 27 included studies: 22 classified as core (directly addressing AES reliability, bias, or perceptions in English-language or comparable contexts) and 5 as peripheral (providing foundational ITS or framework context). Screening followed the PRISMA 2020 flow: approximately 102 records were identified (79 via Semantic Scholar, 15 via manual or targeted searching, and approximately 8 via citation tracking), approximately 74 were excluded at title and abstract screening, 28 reports were assessed in full text, 1 was excluded

with reason, and 27 studies were included. The PRISMA flow diagram is shown in Fig 1, and the complete screening log is provided in Table A1 (Appendix A).

The included sources differ in publication status. Most are journal articles or peer-reviewed conference proceedings. Four sources cited in this review were not peer-reviewed publications at the time of writing: a preprint (Fan & Yun, 2026), a manuscript submitted for publication (Gaggioli et al., 2025), a working paper (Oketch et al., 2025), and an SSRN preprint (Xiao et al., 2023). They were retained because they address emerging questions (LLM-based scoring bias, open-source model access, and institutional policy responses) for which peer-reviewed evidence is still sparse; claims that depend on them are labelled as provisional and qualified accordingly in the Findings and Discussion. The evidence base is therefore described as peer-reviewed and emerging evidence. Findings from non-English contexts (Seßler et al., 2025, German; Gaggioli et al., 2025, Italian) are likewise treated as indicative rather than conclusive for English language education.

2.3 Inclusion and Exclusion Criteria

Studies were included if they: (a) evaluated AI-based scoring or feedback against human raters using quantitative agreement measures; (b) examined bias or fairness in AI assessment across student demographics; (c) reported student or educator perceptions of AI grading; or (d) proposed or evaluated ethical frameworks for AI assessment. Studies were excluded if they focused solely on plagiarism detection, reported no empirical data, or were published as conference abstracts without full text. Studies using non-English assessment instruments were retained when the methodology and findings were generalisable to English-language assessment contexts.

2.4 Data Extraction and Analysis

Full texts were processed using Python/PyMuPDF to extract quantitative findings, which were coded into a structured summary table ($N = 27 \times 14$ variables: author, year, design, context, N, AI tool, key metrics, and five thematic relevance ratings). Thematic analysis proceeded deductively across five pre-specified themes: (1) reliability and accuracy, (2) bias and fairness, (3) student perceptions, (4) ethics and governance, and (5) ITS and feedback systems. Cross-theme interactions were mapped in a synthesis matrix. Two studies in the corpus (Gaggioli et al., 2025; Seßler et al., 2025) used non-English essays (German and Italian, respectively); their methodological findings are incorporated as evidence about LLM scoring behaviour under rubric and context variation, with explicit notation of their linguistic contexts.

3 Results and Discussion

3.1 Results

3.1.1 *Reliability of AI-Assisted Grading*

Headline statistics mask rubric-dependency

At first glance, the reliability literature sends mixed signals. On the favourable end, Koraishi (2024) reported that ChatGPT-4 achieved ICC = 0.814 (95% CI: 0.702–0.887) with weighted

kappa = 0.811 against certified IELTS raters on Writing Task 2. These values are generally characterised as good-to-excellent agreement (Koo & Li, 2016). Seßler et al. (2025), testing five LLMs against 37 teachers across 20 student essays, found that the o1 model achieved Spearman $r = .74$ with human assessments overall and ICC = .80 for internal consistency. Y. Kim (2025) reported high intra-rater reliability for GPT-4 across eight prompting conditions (Pearson $r = .77-.91$), with the best human-alignment condition (detailed rubric + rationale request + scoring examples) yielding $r = .61$ and placement QWK = .450.

On the unfavourable end, Manning et al. (2025) found that GPT-4 achieved ICC3 = 0.385 (Poor) against human raters, compared to human-human ICC3 = 0.819 (Good); GPT-3.5 reached only ICC3 = 0.521 (Moderate). Crucially, combining a human rater with GPT-4 did not rescue reliability: human + GPT-4 ICC3 = 0.493, remaining in the Poor range. Pack et al. (2024), testing four LLMs on ELL placement writing with a holistic 1–6 rubric, found GPT-4 achieved interrater ICC = 0.843 (T1) and 0.779 (T2) relative to human raters, but with Pearson $r = 0.731$ and 0.638 , which represent moderate, not strong, correlations that raise questions about practical substitutability for human judgement. Gaggioli et al. (2025, unpublished manuscript) tested five advanced LLMs (Claude 3.5, DeepSeek v2, Gemini 2.5, GPT-4, Mistral 24B) on 67 Italian university psychology essays: all human–LLM Quadratic Weighted Kappa values approximated zero (non-significant), and within-model Kendall's W ranged 0.00–0.42 (all non-significant), indicating both poor human agreement *and* poor intra-model consistency. As this study is an unpublished manuscript that used Italian-language essays, it is treated here as indicative rather than conclusive evidence.

The apparent contradiction dissolves when rubric design is examined systematically. The studies reporting good reliability share a structural feature: their rubrics are **externally anchored**, either through the public IELTS band descriptors (Koraishi) or through language-level criteria amenable to surface evaluation (Seßler's spelling, literal speech markers). The studies reporting poor reliability share the opposite feature: rubrics requiring **domain inference** (Gaggioli's Originality and Feasibility criteria), holistic or minimally specified scoring (Manning), or criteria that can only be applied with knowledge of the writer's intent and disciplinary context.

Tang et al. (2024) provide the clearest experimental evidence for this dependence on rubric specification. In a multi-dimensional AES study, prompts that included both criteria definitions *and* sample essays produced 112% accuracy gains in QWK over dimension-only prompts, demonstrating that explicit instruction in what counts as quality at each score level drives reliability more than model capability. This finding reinforces the rubric-operationalisation thesis: the key variable is not the AI's general intelligence but the degree to which evaluative criteria are specified in terms that an LLM can operationalise without requiring inferences beyond the essay text.

Seßler et al.'s (2025) dimension-level data confirm this with independent evidence. Strong agreement was achieved for language *surface* criteria (spelling $r = 0.814$, literal speech $r = 0.805$, verbal imagery $r = 0.738$), while content-level criteria (Introduction and Main Part) showed only weak-to-moderate correlations with teacher ratings. This pattern appears despite the study using German essays, suggesting that the rubric-operationalisation effect may extend across languages, although evidence from non-English contexts should be interpreted cautiously and needs replication with English-language essays. Notably, S. Kim & Jo (2024) found that GPT-4 was sufficient for AES using a comparative judgment approach, a methodology that reduces many rubric-specification issues by asking the AI to rank paired essays rather than assign absolute scores, further illustrating that how scoring is structured matters as much as which model is deployed.

Temperature and prompting as uncontrolled confounds

A critical methodological issue under-acknowledged in this literature is the role of prompting conditions and temperature settings as confounds. LLM outputs are stochastic: at default temperature (0.7), the same input will generate variable outputs across runs. Y. Kim (2025) explicitly notes that local test administrators typically lack the technical expertise to adjust this parameter, observing that reliability varied substantially across prompting conditions even for the same model. The weakest prompting condition in Kim's study (minimal rubric, no rationale request, no examples) produced Pearson $r = .47$ and placement QWK = .004 (effectively random) while the optimal condition produced $r = .61$ and QWK = .450 from the identical model (GPT-4, gpt-4-0613).

Gaggioli et al. (2025, unpublished manuscript) used temperature = 0.25 and three replications per essay, a more methodologically careful approach, yet still observed near-zero QWK across all five LLMs. This is important: it suggests that temperature control and replication are necessary but not sufficient; the ceiling is set by how well criteria can be operationalised without domain inference. Manning et al. (2025) do not report their temperature settings or prompting protocol, which means their ICC3 = 0.385 finding cannot be attributed to AI capability alone; it may partially reflect prompting conditions that a more careful study would not test. This is a systematic threat to validity across the reliability literature: "studies that do not report, control, or vary their prompting conditions produce findings that are not comparable", yet they are routinely placed on the same reliability spectrum.

The practical implication is stark: a practitioner who reads that GPT-4 achieves ICC = 0.814 (Koraishi) versus ICC3 = 0.385 (Manning) and assumes the discrepancy reflects model characteristics is misreading the evidence. Rubric type, prompting protocol, and temperature settings appear to be major sources of variance in these reliability outcomes, rather than model architecture alone.

The holistic/analytic rubric divide

Related to the rubric issue is the distinction between holistic and analytic scoring approaches. Studies using analytic rubrics with specific, bounded criteria (Seßler's 10-category rubric; Koraishi's IELTS band descriptors; Kim's four-domain placement rubric) consistently achieve higher reliability than studies using holistic approaches (Shin & Lee, 2024; Uyar & Büyükahıska, 2025). Uyar & Büyükahıska (2025) found overall Spearman $r_s = 0.305$ ($p = .031$) between ChatGPT (4o mini) and human raters for Turkish EFL essays, a weak and unacceptable level of agreement, with human raters assigning significantly higher scores across all essay genres (Wilcoxon $p < .05$ for all; effect size $r = -0.87$). The phrase "moderate correlation" that appears in some preliminary interpretations of similar findings should be treated with caution: statistical significance does not imply practical agreement, and an effect size of $r = -0.87$ indicates strong systematic disagreement in scoring level, even if directional trends correlate.

Hallucination risk

Chelli et al. (2024), studying LLM-generated systematic review references, documented hallucination rates of 39.6% (GPT-3.5), 28.6% (GPT-4), and 91.4% (Bard/Gemini). While this study addressed reference generation rather than essay scoring, it establishes the baseline hallucination behaviour of these models. In assessment contexts, hallucination risk manifests as fabricated feedback, meaning plausible-sounding comments not grounded in the actual

essay text, which carries accountability implications for educators. Almegren et al. (2025) noted higher consistency in AI feedback compared to human feedback, but consistency in fabrication is not validity.

3.1.2 *Bias and Fairness*

The ELL scoring gap is empirically robust

The most consistent and concerning finding across the bias literature is a systematic scoring disadvantage for ELL/ESL students in AI-graded assessments. Fan & Yun (2026, preprint; not yet peer-reviewed) quantified this as a 10.3% scoring gap for high-proficiency second language writers, noting that the gap persisted even when essay quality was controlled. Using contrastive learning techniques, they reduced this gap by 39.9% while maintaining QWK = 0.756, suggesting, provisionally, that the bias may be remediable through architectural intervention.

Kwako & Ormerod (2024), analysing XLNet AES performance on ASAP 2.0 and PERSUADE 2.0 benchmarks, found that ELL students received scores -0.10 z below the mean relative to comparable writing quality, alongside low-SES students (-0.10 z) and Black/African American students (-0.07 z). Critically, ELL status was recoverable from the model's hidden states with $\kappa = 0.75$ for one prompt, demonstrating that the model encodes language background as a latent feature, making it available as a construct-irrelevant score determinant even when the rubric does not specify language background as a criterion. This is a structural problem: the model has not been instructed to penalise ELL status, but ELL status has been learned as a correlate of lower scores from training data.

Racial and ethnic bias operates in non-intuitive directions

Yamashita (2025) provides the most methodologically rigorous bias analysis in the corpus, applying many-facet Rasch measurement (MFRM) to 6,482 essays from the ELLIPSE corpus, scored by 27 human raters and GPT-4o. The findings are counterintuitive by the assumptions often brought to this literature. GPT-4o exhibited no significant bias for gender ($\chi^2 = 69.7$, $p = .10$) or socioeconomic status ($\chi^2 = 49.8$, $p = .71$). However, the rater $\times$ race/ethnicity interaction was significant ($\chi^2 = 267.0$, $p < .05$): GPT-4o was unexpectedly lenient toward Asian/Pacific Islander writers (bias = $+0.42$) and unexpectedly severe against Hispanic/Latino writers (bias = -0.1).

Two implications are important. First, the direction of racial bias is not the one most commonly assumed in the literature: GPT-4o did not uniformly disadvantage all minority groups. The differential treatment reflects the demographic composition of the ELLIPSE training corpus rather than a simple pro-majority bias. This means bias audits that test only the most-discussed demographic contrasts risk missing present but differently-directed disparities. Second, the Rasch MFRM approach used by Yamashita, which controls for essay ability, rater severity, and item difficulty simultaneously, provides a template for bias detection methodology that most other studies in this corpus do not approach. Studies that compute group mean differences without controlling for writing quality may not be measuring AI bias alone; they may be measuring AI accuracy at detecting writing quality differences that already exist.

AI reproduce human bias; this is not exculpatory

Wilson and Huang (2024) found that MI Write AES did not introduce new scoring bias against elementary ELL students beyond what was already present in human rater training data: the AI reproduced existing human biases rather than creating novel ones. This finding is sometimes cited as reassuring: the AI is no worse than humans. This reading should be resisted. If the AI has been trained on human ratings, and those ratings encode systematic bias against ELL students, then the AI is not a neutral technology; it perpetuates bias at scale. Removing human raters from the loop eliminates the oversight mechanism by which those biases might be identified and corrected.

Yang et al. (2024), examining 9 AES methods across demographic groups, found that cross-prompt generalisation specifically amplified SES bias: models trained on one writing prompt and applied to another showed greater SES-related disparities than prompt-specific models. The mechanism is one of distributional shift: essay topics that are culturally proximate to high-SES experience generate better performance from students with that background, and a model trained on one such topic will encode that advantage into its parameters when generalised. This has direct implications for high-stakes placement testing, where prompt generalisation is the norm.

Open-source models challenge the access hierarchy

A necessary corrective to assumptions about the fairness landscape is provided by Oketch et al. (2025; working paper, not peer-reviewed), who compared nine LLMs, including closed (GPT series), open (Llama 3, Qwen2.5), and open-source (OLMo 2, DeepSeek-R1), on AES tasks drawn from ASAP and FCE corpora. Their key finding: open LLMs (Llama 3, Qwen2.5) performed comparably to GPT-4 in scoring accuracy for few-shot learning-based assessment, with no significant differences in fairness by age or race. Llama 3 was up to 37 times more cost-efficient than GPT-4.

This finding has consequential policy implications for English language education, particularly in contexts where institutional budgets are constrained. The assumption that reliable and fair AES requires premium closed-model API access is not supported by the evidence currently available, and insisting on it may erect unnecessary access barriers for institutions in the Global South. These findings come from a working paper and require peer-reviewed replication. For a conference convened under the theme of sustainable futures in English education, the accessibility of open models is not a peripheral technical note but a structural affordance.

3.1.3 Student and Educator Perceptions and the Trust Paradox

Chai et al. (2024) found, across three study components (N = 172, 149, 145) with Chinese EFL university students, that students perceived AI grading as fairer than teacher grading, with transparency of AI process mediating acceptance. A plausible interpretation is that students appreciate AI's perceived objectivity. However, AI in Chai's study also assigned higher scores on average, inflating grades relative to human benchmarks. This confound matters: if students are trusting AI because it grades them more favourably, that is not evidence of legitimacy; it is evidence of grade inflation tolerance.

Conijn et al. (2023) provide a direct test of whether explanations build trust in AES: they do not. Across higher education contexts, providing explicit AI reasoning for scores did not increase student trust. Instead, trust correlated with whether the grade matched the student's own expectations. This is theoretically important because it challenges a common policy

assumption, namely that making AI decision-making transparent will increase its acceptance. The implication is that AI grading must first be *accurate* to be *trusted*, because students are calibrating to their own sense of what they deserve. Explanation alone is unlikely to rescue trust if the base grade is perceived as wrong, and given the reliability evidence reviewed above, that perception may often be correct.

On the educator side, Khlaif et al. (2024) found that teachers' adoption of generative AI for assessment was mediated by perceived usefulness, ease of use, and institutional support, but constrained by concerns about academic integrity and fairness. This finding positions teacher readiness as a separate condition from student acceptance: even where students may accept AI grading, teachers may resist deployment without adequate training and governance frameworks.

Together, these findings describe a conditional acceptance structure: students will accept AI grading when it is generous and when they already expected the score they received, while educators require institutional infrastructure and governance assurances. This is a fragile legitimacy basis that may be difficult to sustain as AI is adopted in high-stakes contexts where grade accuracy governs access to opportunities.

3.1.4 *Ethical Governance and the Policy Gap*

Yan et al. (2025) synthesised 34 peer-reviewed studies on AI ethics in education (2020–2024, PRISMA), organising findings into four solution domains: AIED Ethical Frameworks, Ethical Assessment Frameworks, AIED Literacy Frameworks, and Literacy Assessment Frameworks. Recurring concerns across the literature include algorithmic bias, privacy violations from demographic inference, transparency deficits, accountability gaps, and geographic inequity in both research coverage and policy development. Notably, Southeast Asian institutions, including the Indonesian context of the present conference, are largely absent from both the research base and the policy landscape (Xiao et al., 2023, preprint), despite rapid AI tool adoption in the region.

Perkins et al. (2024) proposed the AI Assessment Scale (AIAS) as a practical governance tool for categorising AI's role in assessment tasks, ranging from tasks where AI is prohibited through to tasks where AI generates output autonomously with minimal human review. The framework is useful precisely because it avoids the binary pro/anti-AI framing and instead asks what level of human oversight is appropriate for a given assessment purpose. Applied to the reliability evidence reviewed here, the implication is clear: AI may be appropriate for formative, low-stakes, criterion-evaluable dimensions (mechanical accuracy, structural conformity), but not for holistic, high-stakes, domain-inferential assessment without substantial human review.

Kwako and Ormerod's (2024) demonstration that LLMs can infer ELL status from essay text with $\kappa = 0.75$ carries privacy implications beyond scoring: if a system can extract protected demographic characteristics from student writing, those characteristics become available as inputs to other automated decisions. This implication receives insufficient attention in current governance frameworks, which tend to treat bias and privacy as separate domains rather than as co-constituted risks.

3.1.5 *ITS and Automated Feedback: The Longer Context*

The effectiveness of automated feedback and tutoring is well-established in foundational research. VanLehn's (2011) meta-review estimated ITS effectiveness at $d \approx 0.76$ compared to classroom instruction, approaching human tutoring. Ma et al.'s (2014) meta-analysis of 107 ITS studies found comparable figures ($d = 0.66$ vs. classroom; $d = 0.79$ vs. no tutoring).

These results establish that, under controlled conditions, automated systems can approach human effectiveness. However, conditions matter critically: these studies involved purpose-built systems evaluated on specific learning objectives, not general-purpose LLMs applied to open-ended assessment.

Lee & Moore (2024) surveyed contemporary GenAI feedback research, finding that AI-generated feedback is typically more detailed than human feedback but inconsistent across drafts and tasks. Usher (2025) found GenAI grading aligned most closely with instructor grades, while peer assessment showed the most variance, a finding that positions AI as a reasonable *complement* to instructor grading, not a replacement. Létourneau et al. (2025) noted that ITS research in K-12 shows promise, with seven of eight ITS-versus-teacher comparisons showing positive effects, but ELL-specific evidence remains limited, a gap that renders baseline comparisons difficult for English language education practitioners.

3.2 Discussion

3.2.1 *The Rubric Operationalisation Thesis*

The central empirical claim of this review is that “reliability in AI-assisted grading is not primarily a function of model capability; it is a function of rubric operationalisation”. The evidence converges on this conclusion from multiple directions: Tang et al.’s (2024) 112% accuracy gain through criteria-referenced prompts with anchor examples; Seßler et al.’s (2025) dimension-level split between language surface criteria (high agreement) and content criteria (weak agreement); Gaggioli et al.’s (2025) near-zero QWK despite five advanced models and controlled conditions; Koraishi’s (2024) good ICC on IELTS band descriptors; and Kim’s (2025) demonstration that identical model versions produce reliability ranging from near-zero to moderate depending entirely on prompting condition.

The practical implication for English language educators is specific: before deploying any AI grading tool, each rubric criterion must be operationally defined such that it can be evaluated without domain inference. This means: (a) replace adjective-heavy criterion statements (“shows understanding of”) with observable behavioural indicators (“contains at least one paragraph with a clear topic sentence supported by evidence from the text”); (b) provide anchor examples at each score level in the prompt; (c) avoid criteria that require the evaluator to infer writer intent or disciplinary knowledge. Temperature control and prompt structure are necessary conditions but are insufficient; the ceiling is set by the criterion specification.

This recommendation cuts against a common pattern in EFL writing assessment, where holistic band descriptors are preferred for their pedagogical richness and resistance to gaming. The evidence suggests this preference, however educationally justified, appears difficult to reconcile with reliable AI scoring. Educators may need to choose: AI scoring with constrained, behaviourally precise rubrics, or human scoring with holistic, pedagogically rich rubrics. The hybrid approach of applying AI to holistically framed rubrics may combine limited reliability with reduced pedagogical richness.

3.2.2 *The Prompting Confound and Comparative Claims*

The prompting heterogeneity documented across studies means that cross-study reliability comparisons cannot be treated as valid without prompting equivalence data. A practitioner choosing between AI tools on the basis of published ICC values is not comparing tools; they may be comparing prompting conditions. The field needs: (a) standardised reporting of temperature settings, prompt structure, number of replications, and whether examples were

provided; (b) multi-condition within-study designs (as Kim, 2025, provides) rather than single-condition between-study comparisons; and (c) pre-registration of prompting protocols. Until these standards exist, reliability estimates in this literature should be treated as upper and lower bounds on *what is achievable under favourable and unfavourable conditions*, not as estimates of AI tool capability per se.

3.2.3 *The Layered Bias Problem*

Bias in AI grading is not a single-source problem. Yamashita (2025) demonstrates racial/ethnic bias that operates in architecturally unexpected directions (advantaging Asian/Pacific Islander, disadvantaging Hispanic/Latino) through corpus demographic encoding. Kwako & Ormerod (2024) demonstrate that ELL status is latently learned as a construct-irrelevant feature. Yang et al. (2024) demonstrate that prompt generalisation amplifies SES bias through distributional shift. Wilson & Huang (2024) demonstrate that human-annotated training data perpetuates pre-existing rater bias. Fan & Yun (2026) demonstrate that the bias is technically remediable but requires deliberate architectural intervention.

Each mechanism requires a different mitigation strategy: racial bias → corpus rebalancing and MFRM auditing; ELL latent encoding → adversarial debiasing or contrastive learning (Fan & Yun, 2026); SES and generalisation → prompt-specific model training or ensemble approaches; human bias reproduction → auditing training data rater demographics before training. None of these is a default feature of commercially available LLM APIs. Deploying a commercial LLM for high-stakes English language assessment without conducting these audits is difficult to justify on the existing evidence base.

3.2.4 *The Trust Architecture Problem*

The trust literature reveals that current student acceptance of AI grading rests on two fragile pillars: grade generosity and expectation fulfilment (Chai et al., 2024; Conijn et al., 2023). This is a trust architecture that may prove unstable once AI grading is deployed for consequential decisions, precisely because consequential decisions require accurate, not generous, grading. Institutions building acceptance strategies around transparency features and explainability interfaces may be addressing only part of the problem. Acceptance appears more likely to follow accuracy than to precede it. This has a significant practical implication: rolling out AI grading with generous default outputs to build user acceptance, then tightening calibration for consequential assessments, risks generating a trust crisis rather than a trust foundation.

3.2.5 *The Access Equity Opening*

Oketch et al.'s (2025) preliminary finding, from a working paper, that open-source LLMs achieve comparable scoring performance and fairness to GPT-4 at 37-times lower cost is under-cited in the access equity literature. For English language education institutions in the Global South, where the majority of the world's EFL learners are located, the assumption that operational AI assessment requires premium API subscriptions is not clearly supported by the available evidence and may be harmful if it deters institutions from AES adoption that could benefit students. The conference context (Indonesia; Southeast Pacific region) makes this particularly salient: Xiao et al. (2023, preprint) reported that Southeast Asian universities were largely absent from early ChatGPT policy adoption, which may reflect resource constraints as much as institutional caution.

4 Conclusion

This review synthesised evidence across 27 heterogeneous studies to argue three primary claims. These claims rest on a systematic synthesis of published and emerging research rather than on direct empirical testing by the author. First, AI grading reliability is not stable across deployment conditions; it is mediated by rubric operationalisation and prompting methodology, with poorly specified criteria producing reliability that was close to chance in some of the reviewed studies, regardless of model sophistication. Second, systematic bias against ELL, low-SES, and racially marginalised students appears robust across the reviewed studies, operates through at least four distinct mechanisms, and is not adequately addressed by current commercial AI tools. Third, student acceptance of AI grading currently rests on grade generosity rather than trust in accuracy, placing the legitimacy of AI assessment on unstable foundations.

The practical implication for English language educators is not that AI grading should be abandoned, but that its responsible deployment requires three preconditions often absent from institutional contexts: (a) rubric operationalisation that constrains criteria to observable, inference-free indicators before AI deployment; (b) pre-deployment bias auditing across at least the dimensions of language background, SES, and race/ethnicity; and (c) human oversight retained as a core component in all summative assessment decisions. AI tools are appropriate for formative, low-stakes functions, such as providing rapid first-pass feedback on mechanical and structural dimensions, within rubric frameworks designed for the purpose. These implications should be read as evidence-informed recommendations arising from a synthesis of heterogeneous studies, some of which are preprints, working papers, or non-English contexts, and they warrant confirmation through further primary research.

The broader implication for the field is methodological. The heterogeneity of prompting conditions, temperature settings, and rubric designs across published studies has produced a reliability literature that cannot be coherently compared. Standardised reporting conventions, analogous to CONSORT in clinical trials, are needed before practitioners can make evidence-based tool selection decisions. Until then, the appropriate response to claims of high AI grading reliability is to ask: under what rubric, with what temperature, with what prompting protocol, and compared to what human benchmark?

AI-assisted grading in English education promises efficiency. Delivering equity requires substantially more work than has yet been done.

Acknowledgements

The author acknowledges no external funding for this review. The author used AI tools (GitHub Copilot) for literature extraction assistance; all analytical judgements, synthesis conclusions, and argumentative claims are the author's own.

References

- Almegren, A., Mahdi, H. S., Hazaea, A. N., Ali, J. K., & Almegren, R. M. (2025). Evaluating the Quality of AI Feedback: A Comparative Study of AI and Human Essay Grading. *Innovations in Education and Teaching International*, 62(6), 1858–1873. <https://doi.org/10.1080/14703297.2024.2437122>
- Chai, C. S., Lin, P.-Y., Jong, M. S.-Y., Chiu, T. K. F., & Qin, J. (2024). Perceptions of and Behavioral Intentions Towards Learning Analytics-Based AI-Assisted Assessment among EFL Students. *Frontiers in Psychology*, 15. <https://doi.org/https://doi.org/10.3389/fpsyg.2024.1387209>

- Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, 26, e53164. <https://doi.org/10.2196/53164>
- Conijn, R., Kahr, P., & Snijders, C. (2023). The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *Journal of Learning Analytics*, 10(1), 37–53. <https://doi.org/10.18608/jla.2023.7801>
- Fan, K., & Yun, E. (2026). Mitigating Bias in Automated Essay Scoring for ESL Learners via Contrastive Learning. *ArXiv*, 1–4.
- Gaggioli, A., Casaburi, G., Ercolani, L., Collova', F., Torre, P., & Davide, F. (2025). *Assessing the Reliability and Validity of Large Language Models for Automated Assessment of Student Essays in Higher Education*.
- Khlaif, Z. N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M. A. A., Ayyoub, A., Hattab, M. K., & Shadid, F. (2024). University Teachers' Views on the Adoption and Integration of Generative AI Tools for Student Assessment in Higher Education. *Education Sciences*, 14(10), 1–24. <https://doi.org/10.3390/educsci14101090>
- Kim, S., & Jo, M. (2024). Is GPT-4 Alone Sufficient for Automated Essay Scoring?: A Comparative Judgment Approach Based on Rater Cognition. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 315–319. <https://doi.org/10.1145/3657604.3664703>
- Kim, Y. (2025). Automated Essay Scoring With GPT-4 for a Local Placement Test: Investigating Prompting Strategies, Intra-Rater Reliability, and Alignment With Human Scores. *TESOL Quarterly*, 59(S1). <https://doi.org/10.1002/tesq.3405>
- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Koraishi, O. (2024). The Intersection of AI and Language Assessment: A Study on the Reliability of ChatGPT in Grading IELTS Writing Task 2. *Language Teaching Research Quarterly*, 43, 22–42. <https://doi.org/10.32038/ltrq.2024.43.02>
- Kwako, A., & Ormerod, C. (2024). Can Language Models Guess Your Identity? Analyzing Demographic Biases in AI Essay Scoring. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 78–86). Association for Computational Linguistics. <https://aclanthology.org/2024.bea-1.7/>
- Lee, S. S., & Moore, R. L. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. *Online Learning*, 28(3). <https://doi.org/10.24059/olj.v28i3.4593>
- Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A Systematic Review of AI-Driven Intelligent Tutoring Systems (ITS) in K-12 Education. *Npj Science of Learning*, 10, 1–13. <https://doi.org/10.1038/s41539-025-00320-7>
- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology*, 106(4), 901–918. <https://doi.org/10.1037/a0037123>
- Manning, J., Baldwin, J., & Powell, N. (2025). Human Versus Machine: The Effectiveness of ChatGPT in Automated Essay Scoring. *Innovations in Education and Teaching International*, 62(5), 1500–1513. <https://doi.org/10.1080/14703297.2025.2469089>
- Oketch, K., Lalor, J. P., Yang, Y., & Abbasi, A. (2025). Bridging the LLM Accessibility Divide? Performance, Fairness, and Cost of Closed versus Open LLMs for Automated

- Essay Scoring. In O. Arviv, M. Clinciu, K. Dhole, R. Dror, S. Gehrmann, E. Habba, I. Itzhak, S. Mille, Y. Perlitz, E. Santus, J. Sedoc, M. Shmueli Scheuer, G. Stanovsky, & O. Tafjord (Eds.), *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM)* (pp. 655–669). Association for Computational Linguistics. <https://aclanthology.org/2025.gem-1.60/>
- Pack, A., Barrett, A., & Escalante, J. (2024). Large Language Models and Automated Essay Scoring of English Language Learner Writing: Insights into Validity and Reliability. *Computers and Education: Artificial Intelligence*, 6, 100234. <https://doi.org/10.1016/j.caeai.2024.100234>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Perkins, M., Furze, L., Roe, J., & MacVaugh, J. (2024). *The AI Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment*. 21(06). <https://doi.org/10.53761/q3azde36>
- Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2025). Can AI grade your Essays? A Comparative Analysis of Large Language Models and Teacher Ratings in Multidimensional Essay Scoring. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 462–472. <https://doi.org/10.1145/3706468.3706527>
- Shermis, M. D. (2014). State-of-the-Art Automated Essay Scoring: Competition, Results, and Future Directions from a United States Demonstration. *Assessing Writing*, 20, 53–76. <https://doi.org/10.1016/j.asw.2013.04.001>
- Shin, D., & Lee, J. H. (2024). Exploratory Study on the Potential of ChatGPT as A Rater of Second Language Writing. *Education and Information Technologies*, 29(18), 24735–24757. <https://doi.org/10.1007/s10639-024-12817-6>
- Tang, X., Chen, H., Lin, D., & Li, K. (2024). Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon*, 10(14), e34262. <https://doi.org/10.1016/j.heliyon.2024.e34262>
- Usher, M. (2025). Generative AI vs. Instructor vs. Peer Assessments: A Comparison of Grading and Feedback in Higher Education. *Assessment & Evaluation in Higher Education*, 50(6), 912–927. <https://doi.org/10.1080/02602938.2025.2487495>
- Uyar, A. C., & Büyükahıska, D. (2025). Artificial Intelligence as An Automated Essay Scoring Tool: A focus on ChatGPT. *International Journal of Assessment Tools in Education*, 12(1), 20–32. <https://doi.org/10.21449/ijate.1517994>
- VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>
- Wilson, J., & Huang, Y. (2024). Validity of Automated Essay Scores for Elementary-Age English Language Learners: Evidence of Bias? *Assessing Writing*, 60, 100815. <https://doi.org/10.1016/j.asw.2024.100815>
- Xiao, P., Chen, Y., & Bao, W. (2023). Waiting, Banning, and Embracing: An Empirical Analysis of Adapting Policies for Generative AI in Higher Education. In SSRN. <https://doi.org/10.2139/ssrn.4458269>
- Yamashita, T. (2025). Exploring Potential Biases in GPT-4o's Ratings of English Language Learners' Essays. *Language Testing*, 42(3), 344–358. <https://doi.org/10.1177/02655322251329435>

- Yan, Y., Liu, H., & Chau, T. (2025). A Systematic Review of AI Ethics in Education: Challenges, Policy Gaps, and Future Directions. *Journal of Global Information Management*, 33(1), 1–50. <https://doi.org/10.4018/JGIM.386381>
- Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 22466–22474. <https://doi.org/10.1609/aaai.v38i20.30254>