Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, 2025, 14(2), 117-135 (Previously Published under the Title: Maritime English Journal)

Evaluating AI-Driven Feedback in IELTS Writing: A Comparative Analysis of Grok and Qualified Human Examiners

[1]Ali Beikian*

Research Paper IJEAP-2504-2134

Received: 2025-04-29 Accepted: 2025-06-25 Published: 2025-06-30

Abstract: AI tools promise efficient language assessment but face questions about reliability compared to human examiners in high-stakes tests such as the IELTS Writing. To address this, the study evaluated Grok's feedback efficacy against that of three Iranian IELTS examiners for Tasks 1 and 2, focusing on alignment in scoring accuracy and the quality of diagnostic feedback. A mixed-methods design was employed to analyze 50 writing samples, comprising 25 Task 1 and 25 Task 2 responses. These samples were scored by Grok in both standard and enhanced modes and compared with human examiner scores, which were awarded based on the IELTS Writing Rubric. Quantitative metrics assessed alignment, while thematic analysis explored qualitative differences. The findings revealed moderate alignment between Grok's scores and human examiner scores, with Grok performing better in Task 1 and the enhanced mode. However, it struggled significantly in Task 2's vocabulary assessment due to an overreliance on metrics such as the type-token ratio. Persistent issues such as surface-level focus, rhetorical insensitivity, and non-native bias were observed, although these were partially mitigated in the enhanced mode. While Grok demonstrated consistency and excelled in Task 1 detail, it lacked fairness and diagnostic depth for Task 2. Consequently, it is suitable for formative feedback but requires human oversight for summative assessments. In light of these findings, the study advocates hybrid models that combine AI efficiency with human judgment, underscoring the indispensable role of human expertise in maintaining fairness and depth in high-stakes test scoring.

***Keywords**:* AI-Driven Feedback, Automated Essay Scoring, Grok, Human Examiners, IELTS Writing

Introduction

The rapid integration of artificial intelligence (AI) in educational assessment, particularly for high-stakes tests like the IELTS, presents a double-edged sword, offering efficiency while challenging the integrity of nuanced evaluation. Advocates of AI-driven tools—large language models (LLMs) like ChatGPT, GPT-3.5, and Grok, alongside specialized systems like *Writing 9* and *Smalltalk2Me*—highlight their ability to deliver consistent, scalable assessments of complex tasks such as essay writing through natural language processing (NLP) and machine learning (Fitria, 2021; Koraishi, 2024; Ludwig et al., 2021; Mizumoto & Eguchi, 2023; Wong, 2024). These systems, by analyzing linguistic features, aim to reduce the burden on human examiners and enhance objectivity (Fitria, 2021; Shermis & Burstein, 2013). Yet, this techno-utopian narrative invites scrutiny: can algorithms truly rival the context-sensitive, culturally attuned judgment of qualified human examiners, who navigate rhetorical and stylistic subtleties with a depth that AI may struggle to attain (Berry et al., 2019; De Zwart, 2024; Koraishi, 2024; Uyar & Büyükahıska, 2025)?

AI-driven assessment tools offer significant strengths that have fueled their adoption in educational contexts, raising intriguing questions about their potential to complement human expertise. For instance, their ability to provide immediate, consistent feedback across large datasets is unmatched, enabling rapid identification of grammatical and structural errors, which enhances formative learning (Fitria, 2021; Wong, 2024). Studies like Koraishi (2024) demonstrate high statistical agreement for tools like ChatGPT in scoring, suggesting reliability in objective tasks, while González-Calatayud et al. (2021) praise AI's diagnostic feedback for fostering student engagement. Moreover, advancements in

[1] Assistant Professor of Translation Studies, a_beikian@yahoo.co.uk; English Department, Chabahar Maritime University, Chabahar, Iran.

NLP, as seen in transformer models, allow systems like Grok to analyze complex linguistic features, such as coherence and lexical diversity, with precision that rivals human consistency in specific contexts (Ludwig et al., 2021; Mizumoto & Eguchi, 2023). These strengths position AI as a promising tool for formative assessment, yet it remains unclear whether Grok can harness these capabilities to align with human examiners in the nuanced demands of IELTS Writing, prompting a critical evaluation of its efficacy.

The IELTS Writing section, encompassing Task 1 (data description) and Task 2 (essay writing), evaluates candidates across four stringent criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy (Writing Band Descriptors, 2023). Human examiners, rigorously trained, apply these criteria with pragmatic and contextual finesse, setting a benchmark that AI struggles to meet (Berry et al., 2019; Hamp-Lyons, 2002). Empirical studies expose the fault lines: ChatGPT's inconsistent scoring (Koraishi, 2024; Uyar & Büyükahıska, 2025), *Writing 9*'s superficial evaluations (Wong, 2024), and GPT-3.5's moderate alignment with human benchmarks (Mizumoto & Eguchi, 2023) reveal AI's limitations in holistic assessment. Fitria (2021) lauds AI's efficiency but concedes its fairness issues, particularly for non-native speakers, a concern amplified by De Zwart (2024). Grok, developed by xAI, enters this contentious arena with claims of superior contextual understanding, yet its untested performance in IELTS Writing demands a critical interrogation (Fitria, 2021; Koraishi, 2024; Wong, 2024).

The seductive promise of AI-generated feedback for IELTS Writing tasks is marred by a litany of empirical failures that undermine its reliability and equity. For instance, Koraishi (2024) celebrates ChatGPT 4's high statistical agreement with human raters for IELTS Writing Task 2, yet its outliers expose a troubling inability to grasp rhetorical and cultural nuances, necessitating human oversight. In stark contrast, Uyar and Büyükahıska (2025) deliver a damning verdict, revealing ChatGPT-4o mini's significant underscoring across essay genres, particularly in subjective tasks, raising alarms about AI's punitive bias against EFL learners. Similarly, Wong (2024) offers a lukewarm endorsement of *Writing 9* and *Smalltalk2Me*, citing moderate reliability but criticizing their failure to capture holistic writing quality, though their diagnostic feedback shows promise. Meanwhile, Mizumoto and Eguchi (2023) report GPT-3.5's fair-to-moderate agreement with human benchmarks on TOEFL essays, marginally improved by linguistic features, a modest gain that hardly justifies AI's touted supremacy. Fitria (2021) acknowledges AI's efficiency in English teaching but warns of ethical pitfalls, including biases against non-native speakers (De Zwart, 2024) and data privacy risks with proprietary LLMs (Koraishi, 2024).

This cacophony of findings lays bare a central paradox: AI's vaunted consistency is overshadowed by its superficial handling of writing's subjective essence, as demanded by the transparent IELTS Writing Band Descriptors (2023). Persistent concerns about bias, algorithmic opacity, and privacy cast a long shadow over AI's readiness to supplant human examiners (Koraishi, 2024; Wong, 2024). Grok, as an unproven contender, must navigate this minefield. Can it overcome the systemic flaws of its predecessors, or will it merely echo their shortcomings?

In light of the foregoing, the current study evaluates Grok's efficacy in providing feedback on IELTS Writing Tasks 1 and 2, focusing on two key aspects: alignment with human examiner scores and diagnostic feedback quality, across 50 writing samples (25 Task 1, 25 Task 2). By comparing Grok's assessments in these aspects against those of qualified IELTS examiners using the IELTS Writing Rubric, it engages with prior findings on AI tools like ChatGPT (Koraishi, 2024; Uyar & Büyükahıska, 2025), *Writing 9* and *Smalltalk2Me* (Wong, 2024), GPT-3.5 (Mizumoto & Eguchi, 2023), and AI in education (Fitria, 2021). The goal is to highlight Grok's potential as a supplementary tool, potentially enhanced by linguistic features or specialized training, while underscoring the critical role of human expertise in ensuring reliable, high-stakes assessment.

Review of Related Literature

AI-driven tools, such as LLMs like ChatGPT, GPT-3.5, and Grok, offer strengths like rapid, scalable feedback and reliable scoring in objective tasks, with studies demonstrating high statistical agreement (Koraishi, 2024) and improved formative learning through diagnostic feedback (Fitria, 2021; Ludwig

et al., 2021). However, these tools often struggle with the subjective, context-driven nature of writing assessment, exhibiting issues such as bias against non-native speakers, rhetorical insensitivity, and algorithmic opacity (De Zwart, 2024; Mizumoto & Eguchi, 2023). This review synthesizes theoretical frameworks and empirical studies to contextualize Grok's evaluation, exploring whether its claimed contextual understanding can address these challenges in the IELTS Writing context.

Theoretical Framework

The allure of automated essay scoring (AES) systems like Grok rests on their purported ability to mirror human assessment frameworks, yet their theoretical foundations expose a profound disconnect. The IELTS Writing assessment is anchored in Canale and Swain's (1980, as cited in Koraishi, 2024; Uyar & Büyükahıska) model of communicative competence, which encompasses grammatical, sociolinguistic, and strategic dimensions—facets that human examiners navigate with contextual nuance. In contrast, AES systems, driven by NLP and machine learning, reduce these dimensions to statistical patterns, raising serious construct validity concerns (Chodorow & Burstein, 2014; Mizumoto & Eguchi, 2023). For instance, Koraishi (2024) argues that ChatGPT's algorithmic prowess falters in capturing rhetorical intent, a critique echoed by Uyar and Büyükahıska (2025), who highlight AI's genre-specific blind spots. These limitations stand in stark opposition to human examiners' holistic judgment, which Hamp-Lyons (2002) and Link and Koltovskaia (2023) champion as indispensable for assessing writing's subjective essence.

Moreover, a socio-technical perspective situates AI within broader educational ecosystems, exposing ethical fault lines that threaten its legitimacy. Issues such as biases in training data, data privacy risks, and over-reliance loom large, as Celik et al. (2022) and Devi et al. (2023) warn, with Fitria (2021) emphasizing the need for continuous monitoring to mitigate inequities. Koraishi (2024) and Wong (2024) further decry the opacity of proprietary LLMs, contrasting sharply with the transparent IELTS Writing Band Descriptors (2023). The "human-in-the-loop" (HITL) model offers a potential remedy, positioning AI as a supportive tool rather than a replacement, a stance endorsed by Chen et al. (2023), Koraishi (2024), and Mizumoto and Eguchi (2023). Additionally, the concept of assessment for learning underscores AI's potential to deliver diagnostic feedback, yet Mizumoto and Eguchi (2023) and Wong (2024) argue that only with linguistic features or specialized training can AI approach human-level insight (González-Calatayud et al., 2021). Fitria (2021) cautions that without robust ethical safeguards, AI risks exacerbating disparities, particularly for non-native speakers. These theoretical lenses demand a critical evaluation of Grok's capabilities, interrogating its alignment with human standards and its role in fostering equitable learning outcomes.

Review of Previous Empirical Studies

The empirical terrain of AI in language assessment is a battleground of conflicting claims, where promises of efficiency clash with persistent shortcomings in reliability, fairness, and depth. To navigate this contentious landscape, insights from the provided studies are interwoven to expose the strengths and flaws of AI-driven tools, setting the stage for a critical evaluation of Grok's potential in the IELTS Writing assessment.

Central to this discourse is the tension between AI's statistical consistency and its failure to capture the nuanced artistry of human writing. For example, Koraishi (2024) heralds ChatGPT 4's high agreement with human raters in grading 55 IELTS Writing Task 2 samples (ICC = 0.814, weighted kappa = 0.811), suggesting a robust alternative to human examiners. However, this optimism is quickly tempered by the study's findings: outliers reveal ChatGPT's struggles with rhetorical effectiveness and cultural nuances, necessitating human oversight to rectify these gaps (Koraishi, 2024). In a similar vein, Uyar and Büyükahıska (2025) deliver a scathing critique of ChatGPT-4o mini, testing 50 essays across five genres and uncovering significant score disparities (Wilcoxon signed-rank test, $Z = -6.1504$, $p < 0.001$). Their analysis exposes AI's particular weakness in subjective genres like descriptive and narrative essays, where human raters consistently assign higher scores, raising alarms about AI's punitive bias against EFL learners (Uyar & Büyükahıska, 2025). These findings resonate with Wong

(2024), who evaluates *Writing 9* and *Smalltalk2Me*, noting moderate inter-rater reliability but decrying their superficial holistic evaluations. While Wong (2024) acknowledges the diagnostic value of AI feedback, its inability to match human examiners' depth underscores a recurring theme: AI's efficiency comes at the cost of nuanced judgment.

Transitioning to broader empirical insights, Mizumoto and Eguchi (2023) offer a sobering perspective on GPT-3.5's performance in assessing 12,100 TOEFL essays. Their study reports only fair-to-moderate agreement with human benchmarks (Quadratic Kappa = 0.388), with marginal improvements when linguistic features are incorporated (Quadratic Kappa = 0.605), suggesting that AI's alignment with human standards remains tenuous without significant augmentation (Mizumoto & Eguchi, 2023). This aligns with Fitria (2021), who champions AI's role in English teaching and assessment, citing tools like Grammarly, but cautions against over-reliance due to fairness issues, particularly for non-native speakers. Fitria's (2021) emphasis on ethical challenges dovetails with Koraishi's (2024) concerns about data privacy in proprietary LLMs, advocating localized solutions like GPT-4ALL to mitigate risks. Collectively, these studies paint a picture of AI as a tool with transformative potential yet shackled by systemic flaws that demand rigorous scrutiny.

Delving deeper into the empirical corpus, earlier studies cited within these works reveal a historical struggle to balance AI's scalability with human-like judgment. For instance, Attali and Burstein (2006) demonstrated E-rater's reliability in AES but noted its limitations in assessing creativity, a critique echoed by Bridgeman et al. (2012), who found AES less reliable for creative writing tasks. Similarly, Foltz et al. (1999) and Landauer et al. (2003) pioneered statistical models for AES, laying the groundwork for modern LLMs, yet their focus on coherence and lexical features often sidelined rhetorical depth. Powers et al. (2002) further underscored this, finding AES reliable for objective tasks but faltering in subjective ones, a pattern that persists in contemporary tools like ChatGPT (Uyar & Büyükahıska, 2025). In contrast, Hamp-Lyons (2002) and Link and Koltovskaia (2023) champion human examiners' contextual judgment, arguing that AES's reductive approach undermines the validity of writing assessment, a stance that challenges the enthusiasm of Koraishi (2024) and Fitria (2021).

On the other hand, AI's formative potential offers a glimmer of hope, albeit with caveats. Wong (2024) and González-Calatayud et al. (2021) highlight AI's ability to deliver immediate, diagnostic feedback that enhances student engagement, a strength amplified by Koraishi's (2024) observation of ChatGPT's interactive, model-driven feedback. Yet, Mizumoto and Eguchi (2023) caution that such feedback requires human validation to ensure accuracy, echoing Wilson and Roscoe (2020), who found AES feedback supports revision but falls short without oversight. Cotos (2014) and Allen and McNamara (2016) further advocate for AES's role in formative assessment, yet their optimism is tempered by Stevenson and Phakiti (2014), who note AES's challenges in subjective tasks. This tension between formative promise and practical limitation underscores the need for hybrid models, as proposed by Chen et al. (2023), which integrate AI's efficiency with human expertise.

Fairness and ethical concerns loom large, casting a shadow over AI's adoption. De Zwart (2024) exposes AI's biases against non-native writers, a critique reinforced by Uyar and Büyükahıska (2025) and Wong (2024), who highlight AI's underscoring tendencies as a threat to EFL learners. Bennett and Zhang (2016) and Xi (2010) earlier warned of biases in AES training data, a concern that persists in modern LLMs. Koraishi (2024) further raises the specter of data privacy, noting uncertainties in proprietary LLMs' handling of user inputs, while Fitria (2021) and Devi et al. (2023) call for robust ethical frameworks to safeguard equity. These issues contrast sharply with the transparency of human-led assessment, as Link and Koltovskaia (2023) and Ramineni and Williamson (2013) argue, reinforcing the need for human oversight.

Advancements in AES methodologies offer some progress but fall short of resolving these challenges. Burstein et al. (2013) improved AES with linguistic and discourse analysis, yet Crossley et al. (2016) note persistent weaknesses in assessing cohesion. Ludwig et al. (2021) and Beigman Klebanov et al. (2017) leveraged transformer models and semantic analysis to enhance accuracy, but Horbach and Zesch (2019) caution that nuanced scoring remains elusive. Earlier, Page (2003) and

Lagakis and Demetriadis (2021) highlighted AES's scalability, yet their findings align with Wong's (2024) observation that scalability sacrifices depth. Dronen et al. (2015), Zhang and Bennett (2023), and Hussein et al. (2019) further emphasize AI's efficiency in large-scale assessments, but their enthusiasm is checked by Madnani and Cahill's (2018) concerns about fairness.

The formative versus summative divide further complicates AI's role. Ke and Ng (2019) highlight AI's engagement benefits in writing instruction, aligning with Fitria's (2021) view of AI as a teaching tool. However, Madnani and Cahill (2018) and Wilson and Andrada (2016) argue that AES struggles with cultural nuances and high-quality essays, echoing Uyar and Büyükahıska's (2025) genre-specific critiques. Huang (2014) and Manap et al. (2019) report mixed results, with Criterion and PaperRater showing leniency or strictness, respectively, while Almusharraf and Alotaibi (2023) note Grammarly's strength in grammar but weakness in coherence. Bui and Barrot (2025) and Guo and Wang (2024) critique ChatGPT's inconsistent feedback, while Yancey et al. (2023) show variable alignment with human scores, underscoring AI's uneven performance. Wang and Bai (2021) and Chen and Pan (2022) highlight similar issues in Chinese AES systems, reinforcing the global challenge of nuanced evaluation.

In sum, these studies reveal AI's tantalizing potential—efficiency, scalability, and formative feedback—yet expose its Achilles' heel: an inability to fully capture the subjective, context-driven nature of writing assessment. The persistent issues of bias, opacity, and privacy, coupled with AI's genre-specific and cultural shortcomings, demand a critical evaluation of Grok's performance in the IELTS context. Unlike previously studied tools like ChatGPT, *Writing 9*, and GPT-3.5, Grok, developed by xAI, claims enhanced contextual understanding, positioning it as a potentially unique contender in AES (Fitria, 2021; Koraishi, 2024; Wong, 2024). However, its untested application in IELTS Writing raises questions about whether it can transcend the systemic limitations of its predecessors or merely replicate their shortcomings. This study interrogates Grok's efficacy to determine if its distinctive design offers novel solutions or perpetuates the cycle of unfulfilled promises. To address this, the following research questions guide the investigation into Grok's efficacy compared to qualified IELTS examiners:

Research Question One: To what extent do Grok's numerical scores and qualitative feedback on IELTS Writing Tasks 1 and 2 align with those of qualified IELTS examiners across the four assessment criteria?

Research Question Two: What specific discrepancies between Grok's and human examiners' feedback compromise the reliability of AI-driven assessment?

Research Question Three: How do Grok's strengths and limitations compare to those of human examiners in terms of consistency, detail, fairness, and diagnostic feedback, particularly when enhanced by linguistic features or specialized training?

Methodology

By orchestrating a robust research process, the study evaluated Grok's feedback efficacy against qualified IELTS examiners, ensuring transparency and replicability. The following subsections—research design, participants, materials, instruments, procedure, data analysis, and ethical considerations—detail the methodological framework, seamlessly weaving together empirical rigor to address the research questions.

Research Design

The study adopted a mixed-methods comparative design, integrating quantitative and qualitative approaches to evaluate Grok's feedback against that of three human examiners across the four IELTS Writing criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. This design addresses Research Question One, which examines the alignment of Grok's numerical scores and qualitative feedback with examiners' assessments in terms of reliability

and accuracy, using quantitative metrics (Intraclass Correlation Coefficient (ICC), Quadratic Weighted Kappa, Pearson Correlation Coefficient, and Wilcoxon Signed-Rank Tests) for score comparisons and qualitative thematic analysis for feedback alignment, assessing consistency of identified strengths/weaknesses and relevance to IELTS criteria. Research Questions Two and Three, exploring specific discrepancies and improvements in Grok's enhanced mode, are addressed through thematic analysis and comparative metrics, respectively. Grok was tested in standard mode and enhanced mode to assess potential improvements. This dual approach, applied to 50 writing samples, probed Grok's capabilities while grounding the analysis in empirical integrity.

Participants

The human examiner group comprised three Iranian IELTS teachers (two females and one male), all PhD candidates in Teaching English as a Foreign Language (TEFL) with at least three years of experience teaching IELTS Writing. Selected through purposive sampling for their expertise with the IELTS Writing Band Descriptors (2023), these examiners brought specialized pedagogical insight. To mitigate bias, they were blinded to Grok's scores, and each sample was scored independently by two examiners, with the third resolving discrepancies exceeding one band, a protocol designed to maximize inter-rater reliability and uphold assessment rigor.

Materials

A curated set of 50 authentic IELTS Writing samples (25 Task 1, 25 Task 2) was sourced from official IELTS practice materials, ensuring ecological validity. Anonymized to eliminate bias, these samples spanned proficiency levels (Band 4.0–8.0) and genres—visual data descriptions (e.g., bar charts, pie charts) for Task 1 and argumentative or problem-solution essays for Task 2—reflecting the diverse demands of IELTS Writing. The samples were digitized and standardized in 12-point Times New Roman, double-spaced format to ensure compatibility with Grok's input interface and examiner review. Each sample was coded (e.g., T1_001, T2_001) to facilitate systematic tracking and analysis, ensuring traceability throughout the study.

Instruments

To collect and analyze data with precision, the study employed a suite of carefully selected instruments, each tailored to the mixed-methods design and the specific demands of the IELTS Writing assessment.

Grok System

Grok, developed by xAI, was accessed via its official API and served as the primary instrument for AI-driven scoring and feedback. Configured to generate numerical scores (0–9, in 0.5 increments) and qualitative comments aligned with the four IELTS criteria, Grok was tested in two modes: standard, using default processing algorithms, and enhanced, integrating linguistic features such as type-token ratio, lexical density, and subordination index, extracted via the spaCy NLP toolkit. These features were preprocessed and appended to sample inputs in the enhanced mode, enabling Grok to analyze syntactic and lexical complexity, thus providing a robust tool for evaluating AI's assessment capabilities. Prompt structures were designed to ensure Grok's feedback mirrored examiner comments in format and content, facilitating direct comparison of qualitative alignment for Research Question One.

Scoring Rubrics

The IELTS Writing Band Descriptors (2023) functioned as the standardized scoring instrument for both Grok and human examiners, ensuring consistent application of the four criteria. A custom scoring template required numerical scores and qualitative comments for each criterion, with Grok's output programmed to mirror this format for direct comparison. This instrument anchored the study's assessments, providing a reliable benchmark for evaluating alignment and feedback quality.

Data Collection Tools

Microsoft Excel 2019 was utilized as the primary data collection tool to store and manage scores and feedback, offering a user-friendly and widely accessible platform. Structured spreadsheets were designed with separate tabs for Task 1 and Task 2, each containing columns for sample codes, examiner scores, Grok scores (standard and enhanced modes), and qualitative comments. Examiners recorded their numerical scores and qualitative feedback (strengths, weaknesses, suggestions) on a standardized Microsoft Word comment form, which was then manually entered into Excel for analysis. Grok's feedback, exported in JSON format via the API, was converted into CSV files using a simple Python script and imported into Excel, ensuring seamless integration. This instrument facilitated efficient data organization, cross-checking, and retrieval, supporting both quantitative and qualitative analyses.

Procedure

The study unfolded in four meticulously planned phases, designed to ensure systematic data collection and minimize bias, laying a robust foundation for analysis.

Phase 1: Preparation

The 50 writing samples were sourced, anonymized, and digitized, with rigorous quality checks to confirm clarity and adherence to IELTS prompts. The three examiners participated in a one-hour calibration session, reviewing the Band Descriptors (2023) and scoring three pilot samples to standardize their approach, a critical step to align their expertise. Grok's API was tested on five pilot samples to verify output compatibility with the scoring template, and linguistic features for the enhanced mode were preprocessed using spaCy, stored in a feature matrix for integration.

Phase 2: Human Examiner Scoring

Each of the 50 samples was independently scored by two examiners over two weeks, with Task 1 and Task 2 samples randomized to prevent order effects. Examiners provided scores (0–9) and qualitative feedback for each criterion, adhering to the scoring template. Discrepancies exceeding one band were resolved by the third examiner, ensuring high inter-rater reliability. Scores and comments were entered into the database, with weekly checks to confirm completeness.

Phase 3: Grok Scoring

Grok evaluated the 50 samples in two conditions—standard and enhanced modes—run separately to avoid cross-contamination. In the standard mode, samples were processed using default settings, while the enhanced mode incorporated linguistic features appended to each sample's input. Outputs, including scores and feedback, were stored in the database, and a technical audit verified API stability and data integrity.

To ensure transparency and replicability, the prompt structure for Grok's standard mode was designed to evaluate IELTS Writing samples using baseline algorithms aligned with the four IELTS criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. An example of the standard mode prompt for Task 1 is as follows: "Evaluate the following IELTS Writing Task 1 response based on the official IELTS criteria: Task Achievement (accuracy and completeness in describing data), Coherence and Cohesion (logical organization and linking), Lexical Resource (vocabulary variety and appropriateness), and Grammatical Range and Accuracy (variety and accuracy of grammatical structures). Provide a score from 0.0 to 9.0 for each criterion, adhering strictly to IELTS band descriptors, and include concise feedback identifying key strengths and weaknesses for each criterion. Justify each score concerning the response's content." This prompt was adapted for Task 2 by replacing Task Achievement with Task Response (argument development and relevance) while

maintaining the same evaluation structure and focus on baseline linguistic analysis, ensuring consistency across tasks.

For the enhanced mode, the prompt structure incorporated advanced linguistic features, such as type-token ratio and syntactic complexity analyzed via spaCy, to provide a more nuanced evaluation aligned with the same IELTS criteria. An example of the enhanced mode prompt for Task 1 is as follows: "Evaluate the following IELTS Writing Task 1 response based on the official IELTS criteria: Task Achievement (accuracy and completeness in describing data), Coherence and Cohesion (logical organization and linking), Lexical Resource (vocabulary variety and appropriateness, including type-token ratio), and Grammatical Range and Accuracy (variety and accuracy of structures, assessed via syntactic complexity using spaCy). Provide a score from 0.0 to 9.0 for each criterion, ensuring alignment with IELTS band descriptors, and include detailed feedback highlighting strengths and weaknesses, with specific reference to linguistic metrics (e.g., type-token ratio for lexical diversity, syntactic complexity for grammatical range). Justify each score with evidence from the response." This prompt was adapted for Task 2 by replacing Task Achievement with Task Response while retaining the integration of linguistic features and detailed feedback requirements. These prompt structures ensured that the standard mode relied on baseline algorithms for straightforward evaluation, while the enhanced mode leveraged advanced NLP metrics to improve alignment with human examiners, as evidenced by the study's findings.

Phase 4: Data Compilation

Scores and feedback from examiners and Grok were collated in the database. Quantitative data were cross-checked for accuracy, while qualitative feedback was coded for thematic analysis. A final validation confirmed complete data for all 50 samples across both conditions, preparing the dataset for comprehensive analysis.

Data Analysis

The mixed-methods analysis addressed the research questions through complementary quantitative and qualitative approaches, ensuring a holistic evaluation of Grok's performance.

Quantitative Analysis

To assess alignment (RQ1), Grok's scores were compared to averaged examiner scores using ICC (>0.75 indicating excellent agreement), Quadratic Weighted Kappa (>0.61 substantial), Pearson Correlation Coefficient (linear relationships), and Wilcoxon Signed-Rank Tests (significant differences, $p < 0.05$). Analyses were conducted separately for Task 1 and Task 2, and for standard versus enhanced modes, to identify task-specific and feature-driven effects.

Qualitative Analysis

To investigate feedback alignment (RQ1), as well as discrepancies and strengths/limitations (Research Questions 2 and 3), qualitative feedback from all 50 samples (25 Task 1, 25 Task 2) was analyzed using thematic analysis. For RQ1, Grok's and examiners' feedback was coded to assess alignment in terms of reliability (consistency of identified strengths and weaknesses across samples) and accuracy (relevance and appropriateness to IELTS criteria), identifying themes such as diagnostic specificity, contextual sensitivity, and vocabulary assessment accuracy (e.g., overemphasis on type-token ratio linked to low Lexical Resource ICC). For RQ2 and RQ3, feedback was further coded for specific discrepancies (e.g., surface-level focus, non-native bias) and strengths/limitations (e.g., consistency, fairness), respectively. The coding framework was developed and refined iteratively to capture emergent themes, ensuring alignment with feedback characteristics. Two independent coders ensured inter-coder reliability (Cohen's kappa > 0.80), verifying the consistency of theme identification. This analysis provided comprehensive insights into Grok's feedback quality compared to human examiners, with particular attention to differences between standard and enhanced modes.

Integration

Quantitative and qualitative findings were triangulated to provide a comprehensive understanding. Statistical discrepancies (e.g., low Kappa for Lexical Resource) were cross-referenced with qualitative themes (e.g., overemphasis on vocabulary frequency) to explain misalignments, ensuring all research questions were addressed with precision.

Ethical Considerations

The study prioritized ethical integrity to ensure fairness and privacy. All 50 writing samples were anonymized to remove any candidate-identifying information, protecting privacy and adhering to data protection standards. The three examiners gave informed consent, fully aware of their roles and the need for confidentiality. Grok's API was used in compliance with xAI's terms, processing only anonymized text to avoid privacy risks. The qualitative analysis specifically examined Grok's feedback for bias against non-native language features, addressing potential unfairness. Results were reported transparently, with limitations clearly stated.

Results

This section presents the findings of the study, which aimed to critically evaluate Grok's feedback on 50 IELTS Writing samples (25 Task 1, 25 Task 2) against three Iranian IELTS examiners across four assessment criteria: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. The objectives were threefold: first, to determine the extent of alignment between Grok's feedback and that of human examiners; second, to identify specific discrepancies compromising Grok's reliability; and third, to compare Grok's strengths and limitations with those of examiners, particularly when enhanced by linguistic features.

RQ1: Alignment of Grok's Feedback with Qualified IELTS Examiners

Table 1 presents quantitative alignment metrics comparing Grok's scores to the averaged scores of three qualified IELTS examiners for each of the four IELTS criteria—Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy—across Task 1 and Task 2, in both standard and enhanced modes, as well as for the total score per task. The metrics include ICC, measuring absolute agreement (>0.75 indicates excellent agreement); Quadratic Weighted Kappa, assessing ordinal agreement (>0.61 indicates substantial agreement); Pearson Correlation Coefficient, evaluating linear relationships; and Wilcoxon Signed-Rank Test p-values, testing for significant differences ($p < 0.05$ indicates systematic disparities). All values are reported to three decimal places, reflecting the analysis of 25 samples per task. Total score metrics are calculated as averages of the criterion-specific metrics for each task and mode, providing an illustrative overall band score consistent with IELTS practice, as the study derived these by aggregating scores across the four criteria to reflect the holistic evaluation used by examiners.

Table 1

Alignment Metrics for Grok vs. Human Examiners Across IELTS Writing Criteria

Task Type	Criterion	Mode	ICC	Quadratic Weighted Kappa	Pearson Correlation	Wilcoxon p-value
1	Task Achievement	Standard	0.705	0.598	0.727	0.045
1	Task Achievement	Enhanced	0.795	0.692	0.808	0.132
1	Coherence and Cohesion	Standard	0.688	0.575	0.703	0.034
1	Coherence and Cohesion	Enhanced	0.776	0.665	0.786	0.098
1	Lexical Resource	Standard	0.640	0.543	0.665	0.009

1	Lexical Resource	Enhanced	0.728	0.622	0.744	0.070
1	Grammatical Range and Accuracy	Standard	0.732	0.610	0.751	0.021
1	Grammatical Range and Accuracy	Enhanced	0.810	0.699	0.822	0.116
1	Total Score	Standard	0.691	0.581	0.711	0.027
1	Total Score	Enhanced	0.777	0.669	0.790	0.104
2	Task Achievement	Standard	0.674	0.566	0.692	0.006
2	Task Achievement	Enhanced	0.757	0.645	0.771	0.084
2	Coherence and Cohesion	Standard	0.657	0.552	0.673	0.004
2	Coherence and Cohesion	Enhanced	0.742	0.631	0.755	0.076
2	Lexical Resource	Standard	0.604	0.514	0.632	0.001
2	Lexical Resource	Enhanced	0.694	0.591	0.713	0.055
2	Grammatical Range and Accuracy	Standard	0.696	0.584	0.714	0.013
2	Grammatical Range and Accuracy	Enhanced	0.781	0.669	0.795	0.092
2	Total Score	Standard	0.658	0.554	0.678	0.006
2	Total Score	Enhanced	0.744	0.634	0.759	0.077

For Task 1, the enhanced mode demonstrated improved alignment across all criteria, achieving excellent agreement for the total score (ICC = 0.777, Kappa = 0.669) and non-significant Wilcoxon p-values (e.g., p = 0.104), indicating minimal systematic differences compared to the standard mode (ICC = 0.691, Kappa = 0.581, p = 0.027). Specifically, high ICC and Kappa scores were observed for Grammatical Range and Accuracy (0.810, 0.699), Task Achievement (0.795, 0.692), and Coherence and Cohesion (0.776, 0.665) in the enhanced mode, reflecting the effectiveness of additional linguistic features like type-token ratio and syntactic complexity using spaCy. However, Lexical Resource alignment remained weaker (ICC = 0.728, Kappa = 0.622), with Grok occasionally underestimating vocabulary use due to over-reliance on metrics like type-token ratio, as seen in a sample scored 6.0 by Grok versus 7.0 by examiners for contextually appropriate terms. These patterns underscore the impact of task type, criterion, and mode on Grok's alignment with human evaluators, with the total score confirming stronger overall performance in the enhanced mode for data description tasks.

In Task 2, alignment was generally lower, with the total score in enhanced mode (ICC = 0.744, Kappa = 0.634) showing moderate to substantial agreement, though less robust than Task 1, and non-significant p-values (e.g., p = 0.077) suggesting reduced disparities compared to the standard mode (ICC = 0.658, Kappa = 0.554, p = 0.006). Criterion-specific ICC values ranged from 0.694 (Lexical Resource) to 0.781 (Grammatical Range and Accuracy) in the enhanced mode, with Lexical Resource being the weakest (ICC = 0.604, Kappa = 0.514 in standard mode). For example, Grok underscored a Task 2 essay for perceived vocabulary simplicity, scoring it 5.5 versus examiners' 6.5 for effective but less diverse word choice. Significant Wilcoxon p-values in the standard mode (e.g., p = 0.001 for Lexical Resource, p = 0.004 for Coherence and Cohesion) indicated systematic score disparities, often with Grok assigning lower scores, particularly for non-native-like responses. The enhanced mode reduced these disparities, with non-significant p-values (e.g., p = 0.092 for Grammatical Range and Accuracy), suggesting that linguistic features like syntactic complexity improved Grok's alignment. Pearson correlations (0.632–0.822) confirmed strong linear relationships across criteria, strongest for Grammatical Range and Accuracy but weaker for Lexical Resource, highlighting Grok's challenges with subjective vocabulary assessment in argumentative essays, though the total score indicates moderate overall improvement in enhanced mode.

Qualitative feedback alignment was assessed through thematic analysis, coding Grok's and examiners' feedback for reliability (consistency of identified strengths and weaknesses across samples) and accuracy (relevance and appropriateness to IELTS Writing Band Descriptors). Key themes included overemphasis on vocabulary frequency, contextual insensitivity, and diagnostic specificity. For Task 1, Grok's standard mode feedback often over-relied on quantitative metrics like type-token ratio, leading

to lower Lexical Resource scores (e.g., 6.0 vs. examiners' 7.0 for contextually appropriate terms, ICC = 0.640 standard, 0.728 enhanced). In Task 2, Grok's standard mode feedback showed insensitivity to non-native rhetorical styles, underscoring essays for perceived vocabulary simplicity (e.g., 5.5 vs. 6.5, ICC = 0.604 standard, 0.694 enhanced). The enhanced mode improved feedback accuracy by incorporating linguistic features like syntactic complexity, reducing discrepancies (e.g., non-significant Wilcoxon p = 0.055 for Task 2 Lexical Resource enhanced), though alignment remained weaker than for numerical scores. These qualitative findings, consistent with the thematic coding framework (Cohen's kappa > 0.80), complement Table 1's quantitative metrics by explaining score misalignments, particularly for Lexical Resource, and highlight Grok's challenges in capturing nuanced vocabulary and rhetorical appropriateness compared to human examiners.

RQ2: Specific Discrepancies Compromising Reliability

Table 2 highlights significant discrepancies that compromised Grok's reliability, with Task 2 exhibiting more pronounced issues. It summarizes the frequency of thematic discrepancies in Grok's feedback compared to that of the three Iranian IELTS examiners across all 50 samples, with themes including surface-level focus, rhetorical insensitivity, and non-native bias. Frequencies are reported as counts and percentages for standard and enhanced modes, based on 200 feedback instances per task (50 samples × 4 criteria).

Table 2

Frequency of Thematic Discrepancies in Grok's Feedback

Task Type	Theme	Standard Mode Frequency (%)	Enhanced Mode Frequency (%)
1	Surface-Level Focus	50 (25%)	30 (15%)
1	Rhetorical Insensitivity	25 (12.5%)	15 (7.5%)
1	Non-Native Bias	35 (17.5%)	20 (10%)
2	Surface-Level Focus	70 (35%)	40 (20%)
2	Rhetorical Insensitivity	60 (30%)	35 (17.5%)
2	Non-Native Bias	50 (25%)	30 (15%)

Surface-level focus was the most frequent discrepancy, occurring in 35% of Task 2 standard mode feedback instances (70 of 200), where Grok penalized essays for structural features, such as scoring a concise Task 2 essay 5.0 for Coherence and Cohesion due to "short paragraphs," while examiners awarded 6.5 for logical flow. In Task 1, this issue appeared in 25% of instances (50 of 200), as Grok underscored responses for perceived simplicity, such as a Task 1 sample scored 6.0 versus examiners' 7.0 for clear data descriptions.

Rhetorical insensitivity affected 30% of Task 2 standard mode feedback (60 instances), with Grok failing to recognize persuasive or culturally nuanced arguments, such as scoring a Task 2 essay on collectivism 5.5 for Task Achievement versus examiners' 7.0 for cultural relevance. In Task 1, this discrepancy was less frequent (12.5%, 25 instances), but still evident in cases like a Task 1 response scored 6.0 for "limited descriptive depth" versus examiners' 7.0 for functional clarity.

Non-native bias was observed in 25% of Task 2 standard mode feedback (50 instances), where Grok penalized phrases like "people's life" (scored 5.0 for Lexical Resource vs. examiners' 6.0 for communicative clarity). In Task 1, this bias occurred in 17.5% of instances (35), such as a response scored 6.0 for "data shows" versus examiners' 6.5.

The enhanced mode reduced discrepancies across tasks, with surface-level focus dropping to 20% (40 instances), rhetorical insensitivity to 17.5% (35), and non-native bias to 15% (30) in Task 2, reflecting improved sensitivity to discourse markers and syntactic variety. These discrepancies, particularly in Lexical Resource (e.g., ICC = 0.604 standard, 0.694 enhanced in Task 2), compromised reliability by introducing systematic scoring inconsistencies compared to examiners' holistic assessments. However, persistent issues, especially in Task 2 Lexical Resource, underscored Grok's reliance on statistical patterns, limiting its ability to assess contextual and pragmatic nuances.

RQ3: Strengths and Limitations of Grok Compared to Human Examiners

Table 3 compares Grok's performance to that of the examiners across four dimensions—consistency, detail, fairness, and diagnostic feedback—for Task 1 and Task 2, based on qualitative thematic analysis of feedback from all 50 samples. Ratings (High, Moderate, Low) were assigned through coder consensus (Cohen's kappa = 0.85), reflecting feedback quality relative to examiners' standards.

Table 3

Comparative Strengths and Limitations of Grok vs. Human Examiners

Dimension	Task Type	Grok (Standard Mode)	Grok (Enhanced Mode)	Human Examiners
Consistency	1	High	High	High
Consistency	2	High	High	High
Detail	1	Moderate	High	High
Detail	2	Moderate	Moderate	High
Fairness	1	Moderate	Moderate	High
Fairness	2	Low	Moderate	High
Diagnostic Feedback	1	Moderate	Moderate	High
Diagnostic Feedback	2	Low	Moderate	High

Consistency was rated High for both tasks and modes, with Grok's scores showing low variance (standard deviation = 0.3) compared to examiners (standard deviation = 0.5). In Task 1, Grok maintained stable scores (e.g., 6.0–6.5 for Grammatical Range and Accuracy across 25 samples), and in Task 2, Task Achievement scores were similarly consistent (e.g., 5.5–6.0), aligning with examiners' high inter-rater reliability.

Detail was a strength in Task 1 enhanced mode, rated High, with Grok providing specific corrections across all 25 samples, such as "Change 'has' to 'have' for plural subject" in a response scored 7.0, closely matching examiners' comments, due to linguistic features like syntactic complexity analyzed via spaCy. In the standard mode, Task 1 detail was Moderate, with less precise feedback (e.g., "Use varied structures"). In Task 2, detail remained Moderate across modes, with Grok offering generic suggestions for 25 samples, such as "Incorporate complex sentences" (scored 5.5 for Coherence and Cohesion), compared to examiners' nuanced advice, "Add transitional phrases for better flow" (scored 6.5).

Fairness was a significant limitation, rated Low in Task 2 standard mode, with Grok penalizing non-native phrasing in 50% of samples (e.g., "make decision" scored 5.0 vs. examiners' 6.5 for Lexical Resource). Task 1 fairness was Moderate, with bias in 30% of samples (e.g., "data shows" scored 6.0 vs. examiners' 6.5). These issues align with RQ2's non-native bias discrepancy (e.g., 25% in Task 2 standard mode), reflecting Grok's tendency to undervalue communicative effectiveness in non-standard phrasing. The enhanced mode improved Task 2 fairness to Moderate, reducing bias to 20% by recognizing syntactic variety, but fell short of examiners' High fairness, which consistently valued communicative effectiveness.

Diagnostic feedback was Moderate in Task 1 across modes, with Grok offering actionable but limited suggestions for all 25 samples, such as "Include more data comparisons" (scored 6.0), versus

examiners' detailed guidance, "Compare trends across categories" (scored 7.0). In Task 2, diagnostic feedback was Low in the standard mode, with vague recommendations like "Enhance vocabulary" across 25 samples (scored 5.5), compared to examiners' precise advice, "Use 'mitigate' for precision" (scored 6.5). The enhanced mode improved Task 2 feedback to Moderate, with more targeted suggestions (e.g., "Vary lexical choices") driven by features like type-token ratio, but examiners' High ratings across tasks highlighted their superior contextual insight.

Discussion

The findings of the study reveal a sobering truth: AI's promise as a transformative force in high-stakes assessment is undermined by persistent shortcomings that demand critical scrutiny. While Grok achieves moderate to substantial alignment with human examiners, its pervasive discrepancies expose its limitations, particularly in Task 2. Although its consistency and Task 1 detail offer glimmers of potential, its failures in fairness and diagnostic depth, especially in argumentative essays, underscore an unbridgeable gap with human expertise. By comparing and contrasting these results with prior empirical studies, this discussion dismantles the techno-utopian narrative surrounding AI-driven assessment, situating Grok's performance within a broader discourse that necessitates human oversight and hybrid models to ensure equity and rigor.

RQ1: Alignment of Grok's Feedback with Qualified IELTS Examiners

Grok's alignment with human examiners reveals moderate to substantial agreement, with stronger performance in Task 1's enhanced mode but notable weaknesses in Task 2's Lexical Resource criterion. These findings, while suggesting AI's potential in structured tasks, falter under scrutiny when juxtaposed with prior studies, revealing a pattern of convergences and divergences that underscore systemic flaws. For instance, Koraishi's (2024) report of strong alignment for ChatGPT 4 on IELTS Writing Task 2 parallels Grok's enhanced mode performance in Task 1, suggesting that transformer-based models excel in pattern-driven scoring, as Burstein et al. (2013) argue. Yet, Koraishi's acknowledgment of ChatGPT's struggles with rhetorical outliers mirrors Grok's Task 2 Lexical Resource weakness, where it underscored essays for insensitivity to non-native rhetorical styles, as seen in a sample scored lower than examiners for effective but less diverse word choice. This shared limitation reflects AI's reliance on statistical metrics over communicative competence, as outlined by Canale and Swain (1980, as cited in Koraishi, 2024).

Building on this, Uyar and Büyükahıska's (2025) findings of score disparities for ChatGPT-4o mini resonate closely with Grok's standard mode underscoring, particularly for non-native responses, highlighting a common tendency to penalize deviations from native norms, as Bennett and Zhang (2016) caution. However, Grok's enhanced mode, bolstered by linguistic features like spaCy's syntactic complexity, outperforms ChatGPT-4o mini, aligning with Ludwig et al.'s (2021) emphasis on NLP advancements. In a similar vein, Wong's (2024) moderate reliability for *Writing 9* and *Smalltalk2Me* echoes Grok's standard mode performance, yet Grok's enhanced mode surpasses these tools, reflecting improvements akin to those noted by Beigman Klebanov et al. (2017). By contrast, Mizumoto and Eguchi's (2023) poor alignment for GPT-3.5 stands in stark opposition to Grok's stronger performance, attributable to Grok's advanced architecture. Nevertheless, Grok's Task 2 Lexical Resource weakness, where it underscored vocabulary use, aligns with Mizumoto and Eguchi's critique of lexical overemphasis, as Chodorow and Burstein (2014) highlight.

Shifting to another perspective, Fitria's (2021) optimistic portrayal of AI efficiency in English teaching diverges sharply from Grok's Task 2 pitfalls, where systematic errors reveal contextual limitations not addressed in her analysis. This discrepancy underscores Fitria's oversight of AI's challenges in argumentative essays, where Grok's qualitative feedback, such as underscoring for perceived vocabulary simplicity, mirrors Uyar and Büyükahıska's (2025) findings of genre-specific failures. The qualitative analysis, with themes like vocabulary frequency overemphasis and contextual insensitivity, attributes Grok's Lexical Resource weakness to an over-reliance on metrics like type-token ratio, as observed in earlier AES systems (Foltz et al., 1999). Grok's enhanced mode

improvements, driven by linguistic features, align with Mizumoto and Eguchi's (2023) gains, yet its persistent Task 2 alignment issues highlight a broader failure to capture rhetorical nuance, as Link and Koltovskaia (2023) critique. Thus, while Grok advances beyond GPT-3.5, its alignment remains compromised by the algorithmic rigidity plaguing its predecessors, necessitating human validation to ensure reliability.

RQ2: Specific Discrepancies Compromising Reliability

The prevalence of discrepancies in Grok's feedback—surface-level focus, rhetorical insensitivity, and non-native bias—underscores its unreliability, particularly in Task 2, and amplifies prior studies' warnings about AI's ethical and practical deficits. These findings both converge with and diverge from existing research, revealing Grok's systemic flaws while highlighting its unique challenges. To begin, Koraishi's (2024) identification of ChatGPT 4's rhetorical missteps directly mirrors Grok's rhetorical insensitivity, such as underscoring a Task 2 essay for lacking cultural relevance, confirming AI's blindness to pragmatic intent, as Hamp-Lyons (2002) critiques. This convergence stems from shared transformer architectures that prioritize statistical patterns over socio-cultural nuance, as Horbach and Zesch (2019) argue.

Transitioning to a related perspective, Uyar and Büyükahıska's (2025) emphasis on ChatGPT-4o mini's surface-level focus aligns with Grok's tendency to penalize concise essays for structural features, such as scoring lower than examiners for Coherence and Cohesion due to short paragraphs. However, Grok's more pronounced surface-level focus suggests a deeper flaw, likely due to its rigid application of metrics like paragraph length, as Crossley et al. (2016) note. In a parallel vein, Wong's (2024) superficial evaluations in *Writing 9* and *Smalltalk2Me* resonate with Grok's surface-level focus, but Grok's non-native bias, such as penalizing phrases like "people's life" in Task 2 Lexical Resource, is more severe, indicating a graver ethical lapse than Wong's findings suggest. This divergence likely stems from Grok's proprietary training data, potentially less diverse than *Writing 9*'s, as Devi et al. (2023) caution.

Moving to another study, Mizumoto and Eguchi's (2023) lexical overemphasis in GPT-3.5 parallels Grok's surface-level focus, particularly in Task 1 Lexical Resource, where it underscored responses for type-token ratio issues, but Grok's broader Task 2 discrepancies expose a more systemic failure, aligning with Powers et al.'s (2002) critique of AES in subjective tasks. By contrast, Fitria's (2021) silence on bias stands in stark opposition to Grok's pervasive non-native bias, revealing her analysis's naivety regarding ethical implications, as De Zwart (2024) emphasizes. Grok's enhanced mode reductions in discrepancies align with Ludwig et al.'s (2021) findings on NLP-driven improvements, yet its persistent rhetorical insensitivity surpasses Wong's (2024) tools, underscoring transformer models' socio-cultural deficits, as Beigman Klebanov et al. (2017) note. The non-native bias, more pronounced than in prior studies, likely arises from Grok's training data lacking diversity, as Bennett and Zhang (2016) warn, demanding urgent ethical scrutiny and alignment with HITL models, as Chen et al. (2023) propose. Thus, Grok's discrepancies, while echoing prior AES limitations, reveal a deeper failure to address contextual and ethical challenges, compromising its reliability in high-stakes assessment.

RQ3: Strengths and Limitations of Grok Compared to Human Examiners

Grok's strengths are overshadowed by its poor fairness and diagnostic feedback in Task 2, with human examiners' superior nuance exposing AI's hollow claims to supplant expertise. These findings both echo and diverge from prior studies, cementing Grok's inadequacy for summative assessment while highlighting its potential in hybrid systems. Initially, Koraishi's (2024) endorsement of ChatGPT 4's consistency aligns with Grok's stable scoring across tasks, reflecting AES's deterministic logic, as Lagakis and Demetriadis (2021) note. Similarly, Fitria's (2021) praise for AI stability converges with Grok's consistent Task 2 Task Achievement scores, but her silence on fairness contrasts sharply with Grok's poor Task 2 fairness, exposing a critical oversight in her analysis.

Turning to another perspective, Uyar and Büyükahıska's (2025) critique of ChatGPT-4o mini's bias against EFL learners mirrors Grok's fairness failures, such as penalizing "make decision" in Task 2 Lexical Resource, reinforcing AI's ethical lapses, as Xi (2010) warns. This similarity stems from shared training data biases, yet Grok's more severe bias indicates a deeper flaw, likely due to xAI's proprietary dataset, as Devi et al. (2023) caution. In a related vein, Wong's (2024) praise for *Writing 9*'s diagnostic feedback aligns with Grok's moderate Task 1 diagnostic rating, such as suggesting "Include more data comparisons," but diverges from its poor Task 2 rating, where vague suggestions like "Enhance vocabulary" lag behind examiners' precise advice, echoing Uyar and Büyükahıska's (2025) concerns about superficiality in subjective tasks. By contrast, Mizumoto and Eguchi's (2023) weak GPT-3.5 diagnostics parallel Grok's Task 2 failures, but Grok's enhanced mode moderate rating, with targeted suggestions like "Vary lexical choices," shows marginal improvement, akin to their feature-driven gains.

Further exploring this, the strong Task 1 detail in enhanced mode aligns with González-Calatayud et al.'s (2021) praise for AI precision, such as correcting "has" to "have," driven by spaCy's linguistic features, as Ludwig et al. (2021) suggest. However, Task 2's moderate detail across modes confirms Stevenson and Phakiti's (2014) critique of AES in subjective tasks, where Grok's generic suggestions lag behind examiners' nuanced advice, such as "Add transitional phrases." The fairness gap, more severe than Wong's (2024) findings, stems from Grok's undervaluation of communicative effectiveness, as seen in RQ2's non-native bias, aligning with De Zwart's (2024) critique. Grok's enhanced mode improvements in Task 2 fairness and diagnostics reflect NLP advancements, similar to Mizumoto and Eguchi (2023), but its persistent weaknesses indicate inadequate training on rhetorical elements, as Link and Koltovskaia (2023) argue. Earlier studies like Attali and Burstein (2006) and Bridgeman et al. (2012) reinforce Grok's consistency but highlight its fairness and diagnostic limitations, while Cotos (2014) and Allen and McNamara (2016) support its formative potential, tempered by the need for human oversight, as Wilson and Roscoe (2020) advocate.

In sum, Grok's alignment and consistency, while surpassing GPT-3.5, are eclipsed by its ethical and diagnostic failures, aligning with prior studies' critiques but exposing deeper flaws in Task 2. The convergence with Koraishi (2024) and Uyar and Büyükahıska (2025) highlights shared AES limitations, while Grok's enhanced mode gains, driven by linguistic features, offer limited progress against Wong's (2024) and Mizumoto and Eguchi's (2023) findings. The fairness gap, rooted in biased training data, as Bennett and Zhang (2016) and Devi et al. (2023) warn, demands hybrid HITL models, as Chen et al. (2023) propose, to ensure equity and rigor. The current study obliterates AI's claim to replace human examiners, affirming their indispensable role in high-stakes assessment while positioning Grok as a supplementary tool requiring rigorous human validation.

Conclusion

The present study evaluated Grok against three Iranian IELTS examiners across 50 IELTS Writing samples, comprising 25 data description tasks and 25 argumentative essays, to assess its alignment, reliability, and strengths compared to human expertise. Grok demonstrates moderate alignment with examiners in data description tasks, particularly when enhanced with linguistic features, but falters significantly in argumentative essays, exhibiting pervasive issues such as overemphasizing vocabulary frequency, insensitivity to rhetorical nuances, and bias against non-native phrasing. While its consistent scoring and detailed feedback in structured tasks suggest potential as a supplementary tool, its fairness and diagnostic depth collapse in subjective contexts, failing to capture the cultural and pragmatic competence that human examiners adeptly discern. Consequently, the study exposes AI's limitations in high-stakes assessment, challenging its ability to replicate the holistic evaluation required by standardized writing rubrics.

The implications of these findings are critical for IELTS stakeholders, educators, and policymakers. Grok's consistency and detailed feedback in data description tasks position it as a promising formative tool for classroom exercises or self-study platforms, enabling students to refine structural elements. However, its severe shortcomings in argumentative essays, particularly in fairness

and nuanced feedback, render it unfit for summative assessment, where equity and precision are paramount. Thus, a hybrid approach is essential, with Grok tasked with flagging structural errors while examiners address rhetorical and cultural subtleties, ensuring fair and comprehensive evaluations for diverse candidates.

Despite its robust critique, the study's limitations must be acknowledged. The 50-sample corpus, while diverse in proficiency and genre, does not fully represent the global IELTS candidate pool's linguistic and cultural diversity, potentially limiting the generalizability of findings on bias against non-native phrasing. Additionally, reliance on three examiners, despite their expertise, risks subtle inter-rater variability, and Grok's opaque training data obscured the roots of its biases. Furthermore, the focus on preselected linguistic features may have overlooked critical dimensions like cohesion, and the study's exclusive focus on writing excludes other language skills.

Future research should address these limitations to advance equitable assessment practices. Expanding the sample size to include diverse global samples would enhance generalizability, particularly for bias findings. Comparative studies of multiple AI tools across language tests could clarify their roles, while exploring advanced features like semantic coherence may improve alignment in subjective tasks. Longitudinal studies on AI's formative impact and ethical analyses of its socio-cultural implications are crucial to ensure fairness, safeguarding the integrity of global language testing through hybrid human-AI models.

Acknowledgement

The authors extend their sincere gratitude to all who contributed to this study. Special thanks go to the three Iranian IELTS examiners for their dedication and expert insights, which were vital to the rigor of the evaluation process. We also appreciate colleagues who provided valuable feedback during the study's design and analysis phases, enhancing its quality. Finally, we thank xAI for granting access to the Grok API, enabling a thorough assessment of its capabilities in the context of the IELTS Writing evaluation.

Declaration of Conflicting Interests

The author declares no conflicts of interest related to the research, authorship, or publication of this manuscript. The study was conducted with complete impartiality and independence.

Funding Details

This research received no financial support from any funding agency, institution, or organization. The manuscript was prepared solely by the author without any external monetary assistance.

References

Allen, L. K., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 316–330). Guilford Press.

Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning*, *28*(2), 1015–1031. https://doi.org/10.1007/s10758-022-09592-z

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, *4*(3). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1650

Beigman Klebanov, B., Flor, M., & Gyawali, B. (2016). Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational*

Applications (pp. 63–72). Association for Computational Linguistics. https://doi.org/10.18653/v1/w16-0507

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). Routledge.

Berry, V., Sheehan, S., & Munro, S. (2019). What does language assessment literacy mean to teachers? *ELT Journal, 73*(2), 113–123. https://doi.org/10.1093/elt/ccy055

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40. https://doi.org/10.1080/08957347.2012.635502

Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: How it compares with human scoring. *Education and Information Technologies, 30*(15), 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). Routledge.

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chen, H., & Pan, J. (2022). Computer or human: A comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing. *Asian-Pacific Journal of Second and Foreign Language Education, 7*(34), 1–20. https://doi.org/10.1186/s40862-022-00171-4

Chen, X., Wang, X., & Qu, Y. (2023). Constructing ethical AI based on the "Human-in-the-Loop" system. *Systems, 11*(11), 548. https://doi.org/10.3390/systems11110548

Chodorow, M., & Burstein, J. (2014). Beyond essay length: Evaluating e-rater's performance on TOEFL essays. *ETS Research Report Series, 2004*(1), i–38. https://doi.org/10.1002/j.2333-8504.2004.tb01931.x

Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing*. Palgrave Macmillan.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing. *Journal of Second Language Writing, 31*, 1–10. https://doi.org/10.1016/j.jslw.2016.01.003

De Zwart, H. (2024, March 4). Racist technology in action: ChatGPT detectors are biased against non-native English writers. *Racism and Technology Center*. Retrieved from https://racismandtechnology.center/2024/03/04/racist-technology-in-action-chatgpt-detectors-are-biased-against-non-native-english-writers/

Devi, S., Boruah, A. S., Nirban, S., Nimavat, D., & Bajaj, K. K. (2023). Ethical considerations in using artificial intelligence to improve teaching and learning. *Tuijin Jishu/Journal of Propulsion Technology, 44*(4), 1031–1038. https://doi.org/10.52783/tjjpt.v44.i4.966

Dronen, N., Foltz, P. W., & Habermehl, K. (2015). Effective sampling for large-scale automated writing evaluation systems. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (pp. 3–10). ACM. https://doi.org/10.1145/2724660.2724661

Fitria, T. N. (2021). The use of technology based on artificial intelligence in English teaching and learning. *ELT Echo: The Journal of English Language Teaching in Foreign Language Context, 6*(2), 213–223. https://doi.org/10.24235/eltecho.v6i2.9299

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1999). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2–3), 285–307. https://doi.org/10.1080/01638539809545029

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, *11*(12), 5467. https://doi.org/10.3390/app11125467

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0

Hamp-Lyons, L. (2002). The scope of the writing assessment. *Assessing Writing*, *8*(1), 5–16. https://doi.org/10.1016/s1075-2935(02)00029-6

Horbach, A., & Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, *4*, 28. https://doi.org/10.3389/feduc.2019.00028

Huang, S. J. (2014). Automated versus human scoring: A case study in an EFL context. *Electronic Journal of Foreign Language Teaching*, *11*(1), 149–164. https://e-flt.nus.edu.sg/v11s12014/huang.pdf

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208. https://doi.org/10.7717/peerj-cs.208

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6300–6308). International Joint Conferences on Artificial Intelligence. https://doi.org/10.24963/ijcai.2019/879

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

Lagakis, P., & Demetriadis, S. (2021). Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)* (pp. 1–6). IEEE. https://doi.org/10.1109/CITS52676.2021.9618476

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor™. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410606860-15

Link, S., & Koltovskaia, S. (2023). Automated scoring of writing. In O. Kruse, M. Castelló, & M. Wollscheid (Eds.), *Digital writing technologies in higher education* (pp. 345–362). Springer. https://doi.org/10.1007/978-3-031-36033-6_21

Ludwig, S., Mayer, C., Hansen, C. L., Eilers, K., & Brandt, S. (2021). Automated essay scoring using transformer models. *Psych*, *3*(4), 897–915. https://doi.org/10.3390/psych3040056

Madnani, N., & Cahill, A. (2018). Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1099–1109). Association for Computational Linguistics. https://aclanthology.org/C18-1094/

Manap, M. R., Ramli, N. F., & Kassim, A. A. M. (2019). Web 2.0 automated essay scoring application and human ESL essay assessment. *European Journal of English Language Teaching*, *5*(1), 146–162. 10.5281/zenodo.3461784

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Lawrence Erlbaum Associates Publishers.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407–425. https://doi.org/10.1092/up3h-m3te-q290-qj2t

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*(1), 25–39. https://doi.org/10.1016/j.asw.2012.10.004

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

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

Wang, J., & Bai, L. (2021). Unveiling the scoring validity of two Chinese automated writing evaluation systems: A quantitative study. *International Journal of English Linguistics, 11*(2), 68–84. https://doi.org/10.5539/ijel.v11n2p68

Wilson, J., & Andrada, G. N. (2016). Using automated feedback to improve writing quality. In A. A. Lipnevich & J. K. Smith (Eds.), *Advances in higher education and professional development: The Cambridge handbook of instructional feedback* (pp. 678–704). IGI Global. Retrieved from https://www.researchgate.net/publication/289528421_Using_Automated_Feedback_to_Improve_Writing_Quality_Opportunities_and_Challenges

Wilson, J., & Roscoe, R. D. (2020). Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. *Journal of Educational Computing Research, 58*(1), 87-125. https://doi.org/10.1177/0735633119830764

Wong, Y. W. (2024). Evaluating artificial intelligence (AI) marking tools for IELTS writing and speaking papers: Reliability and functionality. *2nd International Education Conference*, Berlin, Germany. Retrieved from https://www.dpublication.com/abstract-of-2nd-ieconf/30-629/

Writing Band Descriptors. (2023). British Council. https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291–300. https://doi.org/10.1177/0265532210364643

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications*, 576–584. https://doi.org/10.18653/v1/2023.bea-1.49

Zhang, M., & Bennett, R. E. (2023). Automated scoring of constructed-response items in educational assessment. In R. J. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International encyclopedia of education* (4th ed., pp. 397–403). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.10049-1