Journal of Information Technology Review

DLINE JOURNALS

Print ISSN: 0976-3511
Online ISSN: 0976-2922

JITR 2026: 17 (4)
https://doi.org/10.6025/jitr/2026/17/4/227-243

AI-Generated Language Assessment Validity, Reliability, Equity, Privacy, and Pedagogical Value

Sopida Tuammee
King Mongkut's University
Thailand
sopida.t@cit.kmutnb.ac.th

ABSTRACT

Integrating Artificial Intelligence (AI) into language assessment offers transformative opportunities for scalability and personalised feedback, but it also raises critical concerns about validity, equity, privacy, and pedagogical value. This study evaluates the responsible deployment of AI-generated language assessments using the Responsible AI Language Education Benchmark, an analysis-ready dataset comprising 960 learner profiles across eight countries and four CEFR proficiency levels. We assessed three anonymized generative AI models (Model_A, Model_B, and Model_C) using a multidimensional, non compensatory framework requiring a minimum threshold score of 70 across four pillars: Validity, Equity, Privacy, and Pedagogical Value. Results indicate that Model_A provides the most balanced profile, excelling in equity (92.86) and pedagogical value (79.22), yielding the highest learning revision and transfer gains. Model_B shows exceptional privacy protection (95.48) and perfect policy transparency, making it well suited for privacy-sensitive contexts. Conversely, Model_C fails the deployment criterion because of inadequate privacy (66.78) and pedagogical value (69.16), along with the highest bias incident rate and the weakest learning outcomes. The findings revealed that technical accuracy alone is insufficient for educational deployment; a non-compensatory, multi-pillar evaluation is essential to ensure AI assessments are equitable, secure, and pedagogically sound.

Keywords: AI-Generated Language Assessment, Responsible AI, Educational Equity, Automated Scoring, Non-Compensatory Evaluation, Learning Analytics, AI Ethics and Governance

Received: 29 May 2026, Revised: 30 July 2026, Accepted: 12 August 2026

Copyright: **DLINE**

1. Introduction and Background

Artificial intelligence (AI) is transforming language education and assessment by challenging conventional assessment paradigms while creating new opportunities to evaluate and support language learning [1, 2]. AI-

driven technologies have expanded the construct of language ability by enabling the assessment of linguistic features that have traditionally been difficult to measure. These features include speech prosody, interactive discourse, and editing behaviours [3, 4]. In addition, AI-enabled automated scoring and personalized feedback can facilitate the integration of assessment with instruction, thereby supporting learning oriented assessment (LoA) approaches [5].

The transformation extends beyond language assessment to higher education assessment more broadly. AI is increasingly being used to generate, interpret, and support learning evidence [6]. Contemporary AI systems incorporate adaptive feedback mechanisms, predictive analytics, and tools that support complex evaluative judgments, moving beyond the traditional focus on automated scoring [7]. These developments offer the potential to improve the efficiency, consistency, and personalization of assessment while expanding learners' access to high-quality feedback and reducing faculty workload [8, 9]. Thus, AI is increasingly positioned not simply as an automated assessment mechanism but as part of a broader assessment ecosystem that connects evidence generation, evaluation, feedback, and learning.

1.1 Validity and Reliability of AI-Based Assessment

Despite these opportunities, integrating AI into language assessment raises fundamental questions about validity and reliability. The emergence of large language models (LLMs) has substantially disrupted established approaches to higher education assessment because conventional assessment instruments may now be susceptible to AI-assisted completion. This development poses challenges to construct validity across disciplinary and institutional contexts [10]. At the same time, the application of LLMs and generative AI to assessment offers opportunities to enhance efficiency, mitigate bias, and provide more customized forms of evaluation [11].

Automated scoring is one of the most established and widely debated applications of AI in language assessment, particularly for writing and speaking. The main attraction of AI-based scoring is its ability to process large volumes of responses efficiently while potentially reducing subjectivity and increasing scoring consistency [12]. However, efficiency and consistency do not necessarily guarantee validity. The validity of AI-generated assessment must therefore be considered in relation to the linguistic construct being measured, learner characteristics, the assessment context, and the consequences of the resulting decisions. Concerns about validity in language assessment have long been established in assessment research and remain particularly important as AI assumes a greater role in the assessment process [13, 14, 15].

AI systems' characteristics similarly complicate reliability. Although automated systems may provide greater scoring consistency than human raters in some circumstances, the opacity of deep learning models can make their evaluative processes difficult to interpret. When the rationale underlying a particular score or evaluative judgment cannot be readily explained, questions arise concerning transparency, accountability, and the extent to which the assessment can be trusted. Consequently, AI-based assessment requires careful consideration of both technical reliability and the interpretability of assessment decisions.

1.2 Equity, Fairness, and Bias

Equity and fairness represent additional challenges in AI-generated language assessment. AI systems depend on the data used for development and training, and systems trained on biased, incomplete, or insufficiently representative datasets may reproduce or amplify existing inequalities. These risks are particularly significant in language assessment because learners differ substantially in linguistic backgrounds, cultural experiences, accents, varieties of English, and language use patterns. Consequently, AI systems may disproportionately disadvantage learners whose linguistic or cultural characteristics are inadequately represented in training data [16].

Fairness is especially important in automated speech assessment. Hannah et al. [17] identify potential biases in automatic speech recognition (ASR) systems that may disproportionately affect English language learners (ELLs), emphasizing the need for systematic evaluation of AI-based assessment technologies. These concerns show that technical performance alone is insufficient to establish the suitability of an AI assessment system. Evaluation must also consider whether performance is equitable across different learner populations.

Recent scholarship has therefore emphasized the development of explicit mechanisms for identifying and mitigating bias. Choi and Zu [18], for example, examine a two step framework to identify and mitigate bias in AI-generated test materials and advocate aligning AI development with equity principles. These approaches indicate that fairness should be incorporated throughout the assessment lifecycle rather than treated as an issue only after an AI system has been deployed.

1.3 Privacy, Transparency, and Ethical Governance

The increasing use of AI in assessment also raises important concerns regarding privacy, transparency, security, and ethical governance. AI-based assessment systems may process substantial amounts of learner data, including written responses, speech recordings, interactional data, performance histories, and other forms of educational evidence. Consequently, responsible deployment requires robust mechanisms to protect data and ensure learners understand how their information is collected, processed, evaluated, and potentially retained.

Ethical AI frameworks that incorporate fairness-aware algorithms, explainable AI models, and robust data-protection mechanisms can build user trust and strengthen system security [19], M]. These safeguards are particularly important in high-stakes assessment environments, where automated decisions may have significant consequences for learners. AI-based assessment should therefore be implemented within institutional and regulatory frameworks that clearly establish responsibility for assessment decisions.

The need for context-sensitive governance is also emphasized by Raza et al., [20.] who argue that although AI has transformative potential for education, its implementation requires region-specific guidelines, rigorous ethical safeguards, and collaboration among multiple stakeholders [Raza, F. A]. Such an approach recognizes that ethical AI assessment cannot be addressed solely through technical solutions. Instead, meaningful governance requires collaboration among educators, assessment specialists, AI developers, institutions, policymakers, and learners.

1.4 Pedagogical Value of AI in Language Assessment

Particular attention has been given to feedback literacy, reflective engagement, and dialogic assessment. Learners' ability to interpret and act upon automated feedback is an important determinant of whether AI-generated assessment contributes meaningfully to learning. Automated evaluation therefore should not be viewed as an end in itself; rather, it can provide evidence and feedback that learners and instructors critically interpret and use within broader learning processes.

Empirical evidence suggests that AI can support both learners and evaluators. Guo et al. [23] found that AI-aided peer feedback enhanced writing quality while also helping reviewers develop their feedback skills, suggesting pedagogical benefits for both those providing and receiving evaluation. Similarly, McLaughlin et al. [24] reported improvements in confidence, collaboration, and learning outcomes when health professionals participated in collaborative research training supported by AI tutors. These findings indicate that AI may have pedagogical value beyond automated assessment by supporting interaction, feedback, reflection, and collaborative learning.

Theoretical perspectives consequently shift attention away from the functionality of individual AI tools toward the broader relationships among AI, academic culture, institutional authority, and formative assessment practices [25]. This perspective is important because the educational value of AI cannot be determined solely by technical accuracy. Its effectiveness also depends on how teachers and learners negotiate authority, responsibility, feedback, and decision-making within AI-mediated learning environments.

1.5 Human Oversight and Hybrid Assessment

Given the opportunities and risks of AI-generated assessment, human oversight remains essential to responsible implementation. Although AI can improve efficiency, consistency, scalability, and personalization, automated systems should not necessarily replace human judgment, particularly when assessment decisions are complex

or consequential.

Sawaki et al. [26] emphasize the importance of monitoring AI–human rating alignment to address the possibility of overly strict or rigid scoring by LLMs. Their findings reinforce the need for continuous human oversight and intervention in AI-assisted assessment. Human raters can provide contextual interpretation, identify inappropriate or anomalous AI judgments, and ensure that assessment decisions remain aligned with the intended construct and educational purpose.

Accordingly, a hybrid model that combines AI capabilities with human instruction and judgment may be a more responsible approach to language assessment. Within such a model, AI can perform tasks such as large-scale analysis, preliminary scoring, feedback generation, pattern identification, and personalized support, while teachers and assessment professionals retain responsibility for interpretation, validation, moderation, and consequential decisions.

1.6 Synthesis and Implications

Overall, the literature indicates that AI-generated language assessment represents both a significant opportunity and a complex methodological and ethical challenge. AI has expanded the range of linguistic features that can potentially be assessed, increased the scalability and efficiency of scoring, enabled personalized feedback, and created new possibilities for integrating assessment with learning. At the same time, questions concerning validity, reliability, fairness, bias, privacy, transparency, and accountability remain central to its responsible adoption.

The evidence further suggests that AI's assessment value cannot be evaluated solely by technical performance. Validity must be considered in relation to the intended language construct and assessment purpose; reliability must be accompanied by transparency and interpretability; and claims of efficiency must be evaluated alongside equity and fairness. Similarly, pedagogical value depends on how well AI-generated feedback and evaluation promote learner reflection, engagement, and meaningful learning.

Consequently, responsible AI-based language assessment requires a multidimensional approach combining technical evaluation, psychometric scrutiny, fairness auditing, data protection, explainability, pedagogical alignment, and human oversight. Ethical AI frameworks, region specific guidelines, and multi stakeholder collaboration are particularly important for ensuring that technological innovation does not compromise educational equity or learner rights. A carefully designed human–AI partnership, rather than complete automation, therefore provides a more sustainable basis for integrating AI into language assessment and education.

1.7 Research Focus

The analysis addresses the research question: How valid and reliable are AI-generated language assessments relative to expert evaluation? The assessment is framed as a responsible-AI benchmarking problem rather than as a single accuracy comparison. Accordingly, we examine model performance across validity, equity, privacy, and pedagogical value, with deployment determined by a non-compensatory rule requiring each pillar to meet the specified minimum criterion.

2. Dataset and Analytical Framework

The dataset is the Responsible AI Language Education Benchmark, an analysis ready resource designed to evaluate generative artificial intelligence systems in language education while incorporating responsible-AI considerations. The benchmark contains 960 learner-profile records from Bangladesh, Brazil, Germany, India, Japan, Mexico, South Korea, and Türkiye, covering four CEFR proficiency levels from A2 through C1. It additionally contains 2,880 learner model evaluation records representing three anonymized AI conditions with repeated stochastic generations, 960 learning-outcome records, 270 privacy-audit observations, and 36 policy-governance records. SMOTE addressed class imbalance; therefore, do not interpret the resulting analytical records as population-representative observations.

The framework is organized around four non-compensatory pillars. Validity incorporates error detection, score agreement, feedback relevance, cultural sensitivity, hallucination control, bias control, and output consistency. Equity evaluates performance gaps across countries, gender, and socioeconomic status together with contextual robustness. Privacy evaluates personally identifiable information avoidance, redaction, consent clarity, data minimization, and policy transparency. Pedagogical value incorporates feedback quality, revision performance, immediate and delayed transfer, learner uptake, satisfaction, trust, and cognitive load.

The three systems are anonymized as *Model_A, Model_B*, and *Model_C* to preserve provider neutrality and avoid brand related bias. The deployment rule requires a minimum score of 70 on every pillar. This is deliberately non-compensatory: a strong score in one dimension cannot offset a failure in another. Country-level differences are interpreted as contextual differences within the dataset rather than causal national effects, and the composite weights and thresholds are recognized as study-specific design choices.

3. Overall Model Performance

The four-pillar results clearly separate the two models that satisfy the deployment criterion from the model that does not. *Model_A* and *Model_B* exceed the 70 threshold on all four pillars, whereas *Model_C* falls below the threshold for both privacy and pedagogical value. The minimum pillar score therefore becomes decisive: *Model_A* records the highest minimum score at 79.22, *Model_B* follows at 77.75, and *Model_C* reaches only 66.78.

Model	Validity	Equity	Privacy	Pedagogical Value	Minimum Deployment	Pillar Decision
Model_A	85.87	92.86	82.37	79.22	79.22	Meets threshold
Model_B	85.42	90.50	95.48	77.75	77.75	Meets threshold
Model_C	78.81	80.15	66.78	69.16	66.78	Below threshold

Table 1. Core performance metrics and deployment decisions

Validity is high for all three models, but *Model_A* has the highest score (85.87), narrowly exceeding *Model_B* (85.42) and clearly exceeding *Model_C* (78.81). The source analysis also reports a larger standard deviation for *Model_C* (4.65) than for the approximately 3.8 values reported for the other models, indicating comparatively less consistent validity performance. Thus, *Model_C* is not characterized by a complete failure of assessment validity; rather, its validity is weaker and less consistent than that of the two leading systems.

Equity scores are also favorable overall, with *Model_A* achieving the highest score of 92.86, followed by *Model_B* at 90.50 and *Model_C* at 80.15. The relatively large separation between *Model_A/Model_B* and *Model_C* indicates that equity is a relative strength of the first two systems, while *Model_C* shows greater vulnerability to contextual and demographic variation.

Privacy produces the strongest differentiation among the models. *Model_B* achieves 95.48, the highest pillar score in the comparison, while *Model_A* records 82.37. *Model_C*, in contrast, scores 66.78 and therefore fails the 70-point criterion. Pedagogical value shows a similar but less pronounced ordering: *Model_A* leads at 79.22, *Model_B* follows at 77.75, and *Model_C* falls below the threshold at 69.16.

Under the non-compensatory decision rule, only *Model_A* and *Model_B* are approved for deployment. *Model_C*

is disqualified because its privacy and pedagogical-value scores are below 70. This finding is important because *Model_C's* acceptable validity and equity scores cannot compensate for its weaknesses in privacy and pedagogical value.

4. Learning Outcomes and Pedagogical Impact

Learning gains provide an outcome-oriented assessment of whether model-generated feedback is associated with measurable improvement. Revision gain represents improvement in immediate revision performance, whereas transfer gain captures the extent to which learners apply acquired skills to new contexts. Both indicators consistently favor Model_A.

Model	Mean Revision Gain	Mean Transfer Gain
Model_	A9.14	5.08
Model_	B7.86	4.53
Model_	C5.56	3.05

Table 2. Learning outcomes and pedagogical impact

Model_A produces the largest mean revision gain, 9.14 points, compared with 7.86 for *Model_B* and 5.56 for *Model_C*. The same ranking is observed for transfer gain: *Model_A* reaches 5.08 points, *Model_B 4.53* points, and *Model_C* 3.05 points. The consistent ranking across immediate revision and transfer outcomes strengthens the interpretation that *Model_A* provides the strongest pedagogical benefit among the three systems. *Model_B* is a strong second, while *Model_C* produces the smallest observed learning improvements.

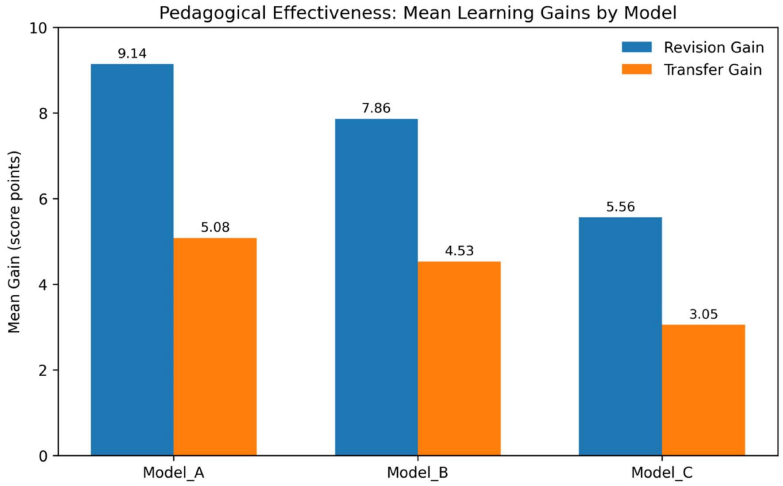

Figure 1. Comparison of mean revision and transfer learning gains across the three AI models

5. Equity and Bias Analysis

The equity audit provides a more granular examination of fairness across national and demographic contexts. (Figure 1). *Model_C* has the largest country-level performance gap (3.36), whereas *Model_A* has the smallest (0.80). This indicates that *Model_A* is comparatively stable across the national contexts represented in the benchmark, while *Model_C's* performance varies more strongly across countries.

Gender and socioeconomic-status gaps are generally small, with the source reporting values below 1.0 for the models overall. *Model_C* nevertheless records a higher SES gap of 1.01, reinforcing the evidence that its performance is less even across socioeconomic contexts. Bias incident rates provide an additional distinction: Model_B records the lowest rate at 0.011, Model_A records 0.027, and Model_C records the highest rate at 0.055. Thus, Model_B provides the strongest bias-rate profile, while Model_C combines the largest country gap and the highest reported bias incidence.

Metric	Model_A	Model_B	Model_C	Best Model
Validity	85.87	85.42	78.81	Model_A
Equity	92.86	90.50	80.15	Model_A
Privacy	82.37	95.48	66.78	Model_B
Pedagogical Value	79.22	77.75	69.16	Model_A
Min Pillar Score	79.22	77.75	66.78	Model_A
Mean Revision Gain	9.14	7.86	5.56	Model_A
Mean Transfer Gain	5.08	4.53	3.05	Model_A
Bias Incident Rate	0.027	0.011	0.055	Model_B
Policy Transparency	79.17	100.00	45.83	Model_B

Table 3. Consolidated comparison of model performance, learning, bias, and governance

Privacy and governance results provide a second major differentiator between the systems. *Model_B* achieves 100% policy transparency across the 12 audited criteria, including data collection and deletion rights. *Model_A* records 79.17%, indicating adequate but incomplete transparency, while *Model_C* records only 45.83%, indicating substantial governance deficiencies.

The interaction-level privacy results show the same ordering. *Model_B* has the highest average privacy score (93.54), followed by *Model_A* (83.74) and *Model_C* (75.76). The *Policy_Audit* also assigns *Model_C* a score of zero for retention-period transparency, along with lower scores on several other governance criteria. These results explain why privacy is a principal reason for *Model_C's* failure under the deployment rule.

Taken together, the privacy evidence positions *Model_B* as the strongest system for privacy and governance, *Model_A* as adequate, and *Model_*C as presenting the most substantial privacy and governance risks. Importantly, these conclusions concern the audited benchmark dimensions and should not be generalized beyond the study's defined criteria and observations.

7. Integrated Model Comparison

The consolidated comparison shows that *Model_A* is the most balanced system. It leads in validity, equity, pedagogical value, minimum pillar score, revision gain, and transfer gain. *Model_B* clearly leads in privacy, policy transparency, and bias incident rate. *Model_C* does not lead on any reported performance dimension

and records the weakest values for privacy, pedagogical value, minimum pillar score, learning gains, and policy transparency.

The results therefore reveal two distinct deployment profiles. *Model_A* provides the strongest all-round performance and the highest minimum pillar score, making it the most balanced candidate under the non-compensatory framework. *Model_B* offers exceptional privacy and governance performance and also satisfies every pillar threshold, but its pedagogical-value and minimum-pillar scores are slightly below those of *Model_A*. *Model_C* demonstrates some acceptable performance in validity and equity but fails the overall responsible-AI deployment test because its weakest dimensions fall below the required threshold.

8. Interpretation of the Model Performance Radar

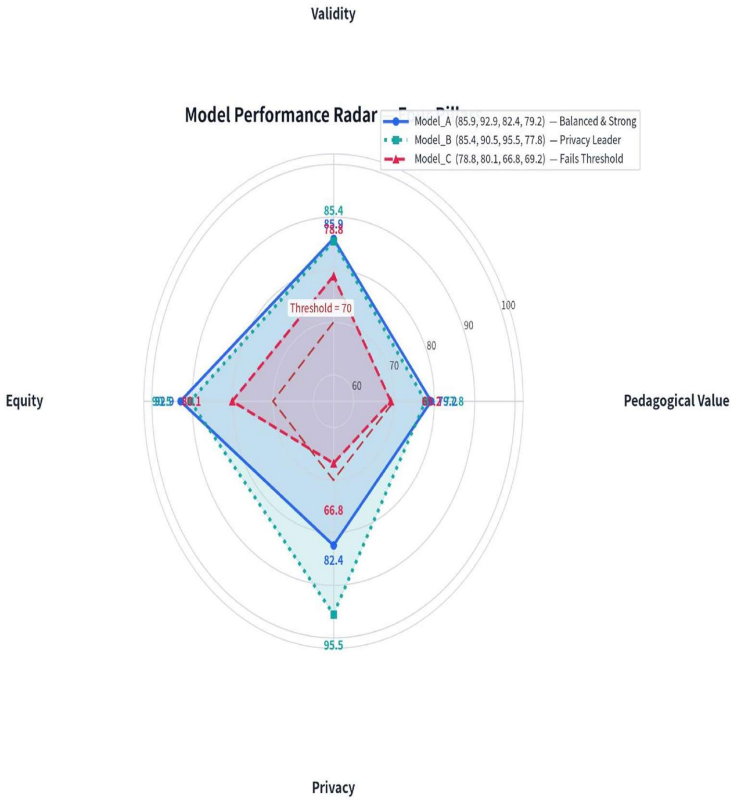

Figure 2. Model performance radar chart showing the four responsible-AI pillars and the 70-point deployment threshold

Figure 2 visually synthesises the four pillar comparison. *Model_A* has the largest and most balanced profile, exceeding the 70-point threshold on every axis and reaching its strongest value in equity (92.86). *Model_B* is similarly strong but is visibly elongated toward privacy, where it reaches 95.48, its highest score. Its pedagogical value score is slightly lower than *Model_A's*. *Model_C* is visibly constrained by its privacy and pedagogical value scores, both of which lie inside the 70-point acceptability boundary. The radar representation therefore makes the non-compensatory decision rule visually apparent: *Model_C's* strengths in validity and equity cannot offset its two pillar failures.

Figure 3 reiterates the numerical pattern using the verified values: validity of 85.9, 85.4, and 78.8; equity of 92.9, 90.5, and 80.1; privacy of 82.4, 95.5, and 66.8; and pedagogical value of 79.2, 77.8, and 69.2 for Models A, B, and C, respectively. The figure confirms that *Model_A* has the most even profile, *Model_B* has the strongest privacy-oriented profile, and *Model_C* is the only model with two dimensions below the deployment threshold.

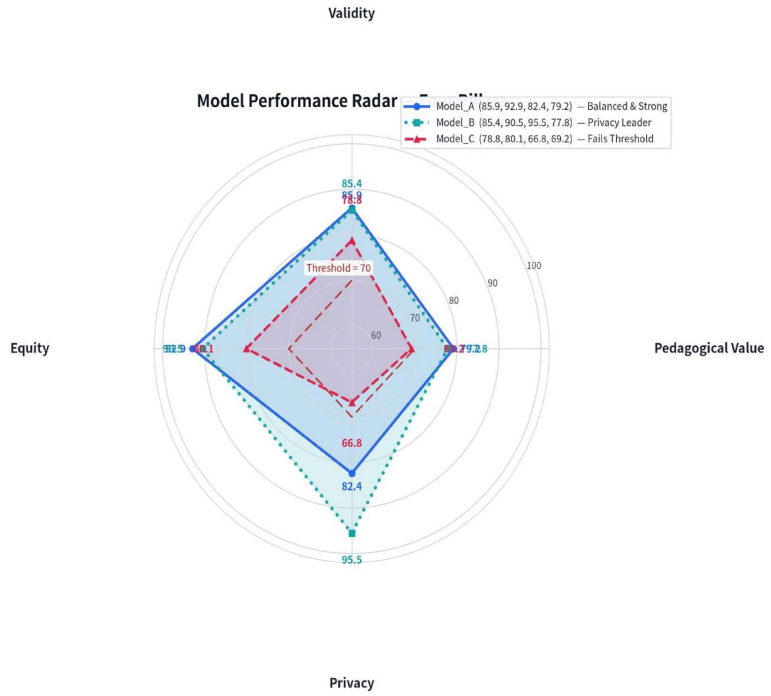

Figure 3. Verified four-pillar radar representation of *Model_A, Model_B*, and *Model_C*

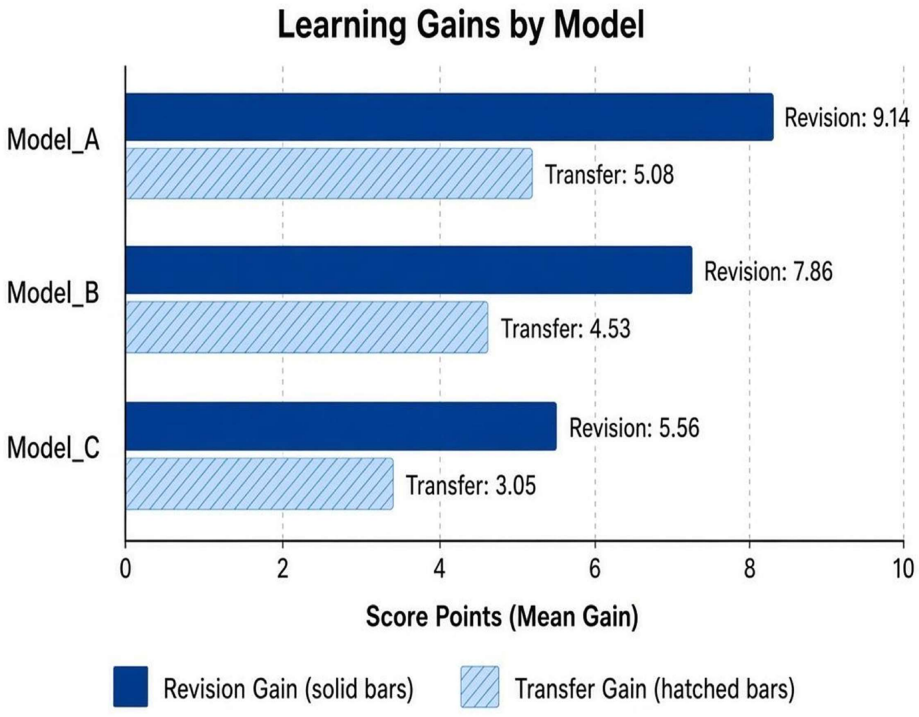

Figure 4. Learning Gains by Model

9. Interpretation of Equity–Bias Visualizations

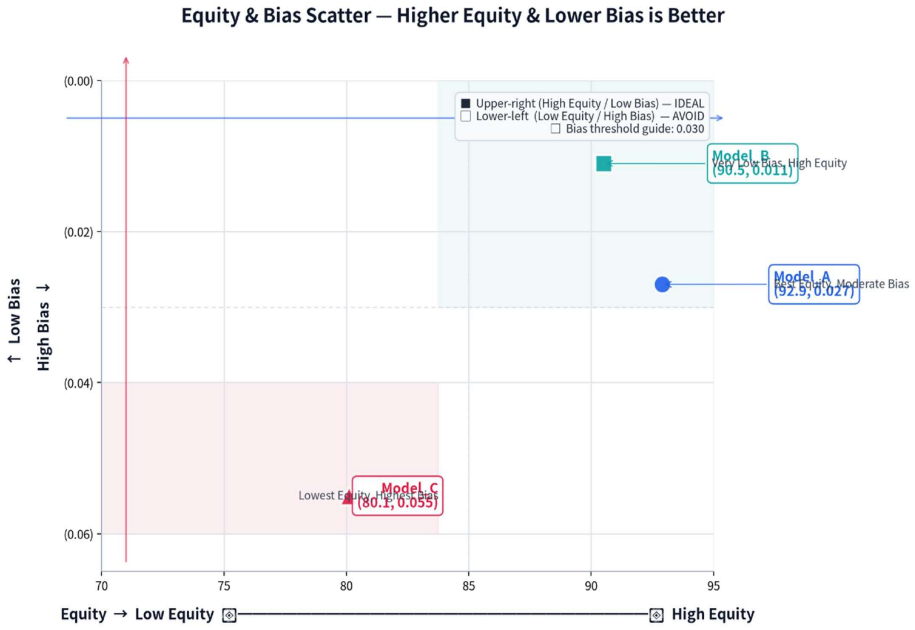

Figure 5. Equity–bias scatter visualization emphasizing higher equity and lower bias as the preferred region

Figure 5 presents equity against bias incident rate, with lower bias positioned toward the top of the display and higher equity toward the right. *Model_B* occupies the most favorable region because it combines high equity (90.5) with the lowest bias incident rate (0.011). *Model_A* is further right because it has the highest equity score (92.9), but it has a moderately higher bias rate of 0.027. *Model_C* occupies the least favorable region, combining the lowest equity score (80.1) with the highest bias rate (0.055). The figure consequently supports the conclusion that *Model_B* has the strongest equity–bias trade-off, whereas *Model_C* presents the greatest fairness concern.

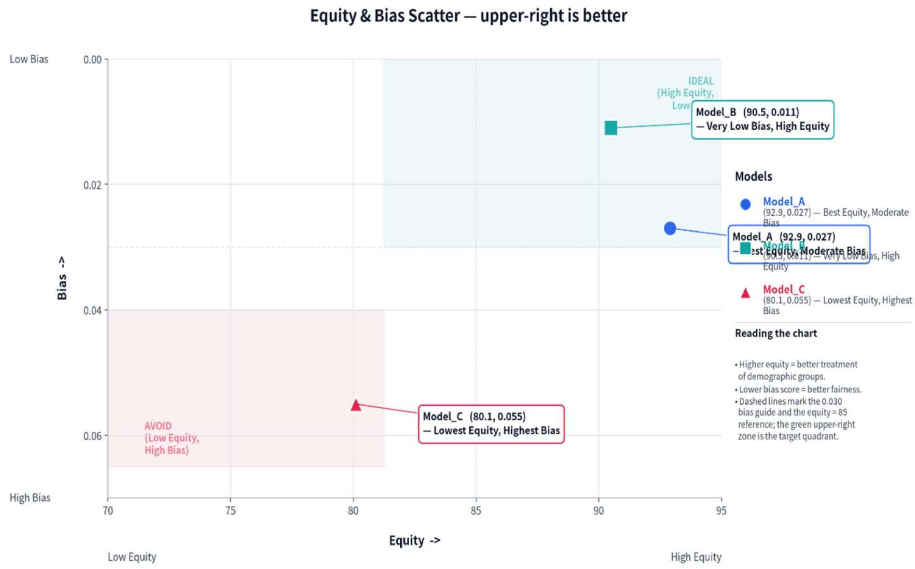

Figure 6. Equity–bias scatter plot with the preferred upper-right region and the less desirable lower-left region

Figure 6 reinforces the same trade-off using the explicitly marked equity and bias guide values. *Model_B* lies in the desirable high equity/low bias region, *Model_A* remains high in equity but with higher bias, and *Model_C* lies in the lower equity/high bias region. The spatial separation among the models is consistent with the numerical equity and bias findings.

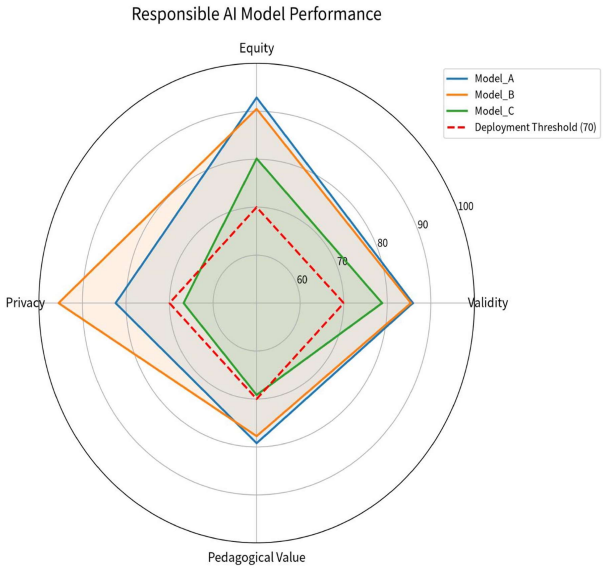

Figure 7. Responsible AI model performance radar emphasizing the relative balance of the four pillars

Figure 7 again demonstrates that *Model_A* maintains the most balanced responsible-AI profile, *Model_B* reaches the highest privacy level, and *Model_C* remains below the threshold in privacy and pedagogical value. The repeated visual representation is useful because it connects the detailed pillar scores to the final deployment classification.

10. Additional Figures and Visual Records in the Source Analysis

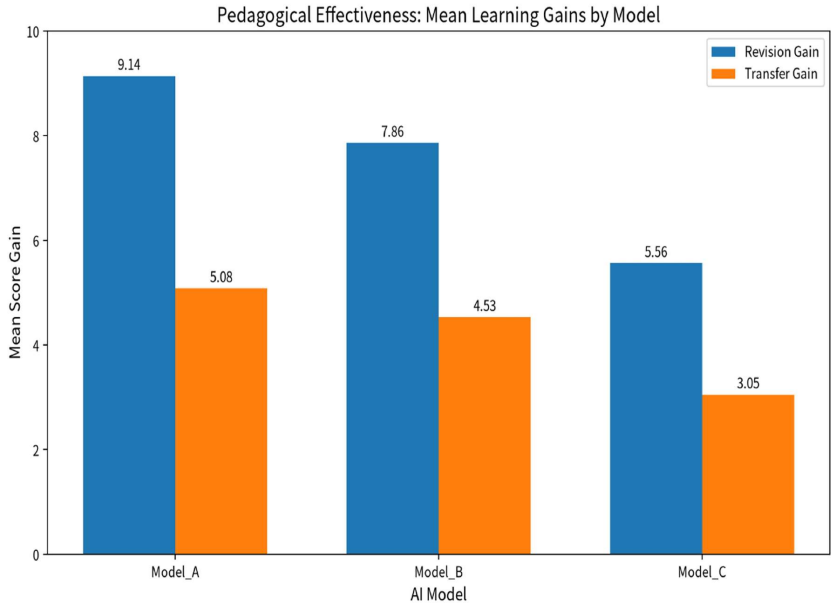

Figure 8. Learning-gain visualization retained from the source analysis; it reproduces the revision and transfer comparison reported in Table 2

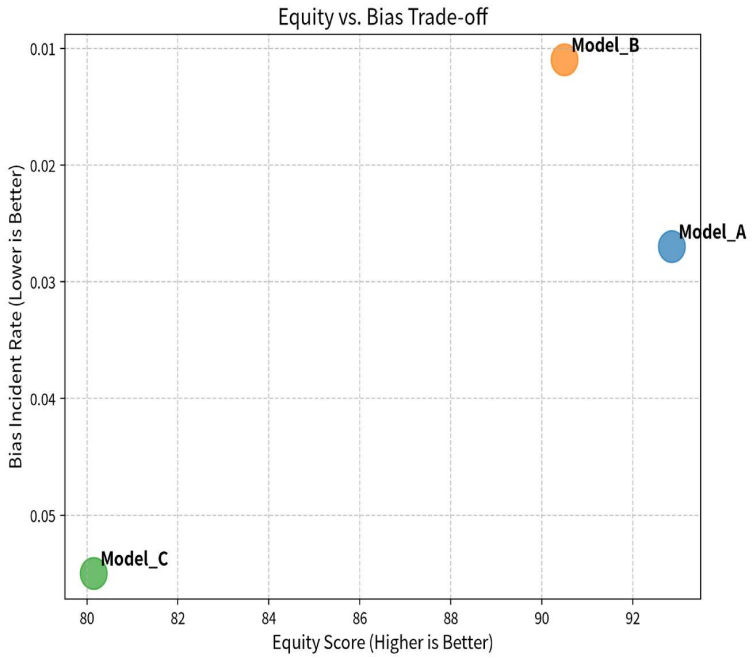

Figure 9. Equity–bias trade-off visualization retained from the source analysis; it emphasizes the relative positions of Models A, B, and C

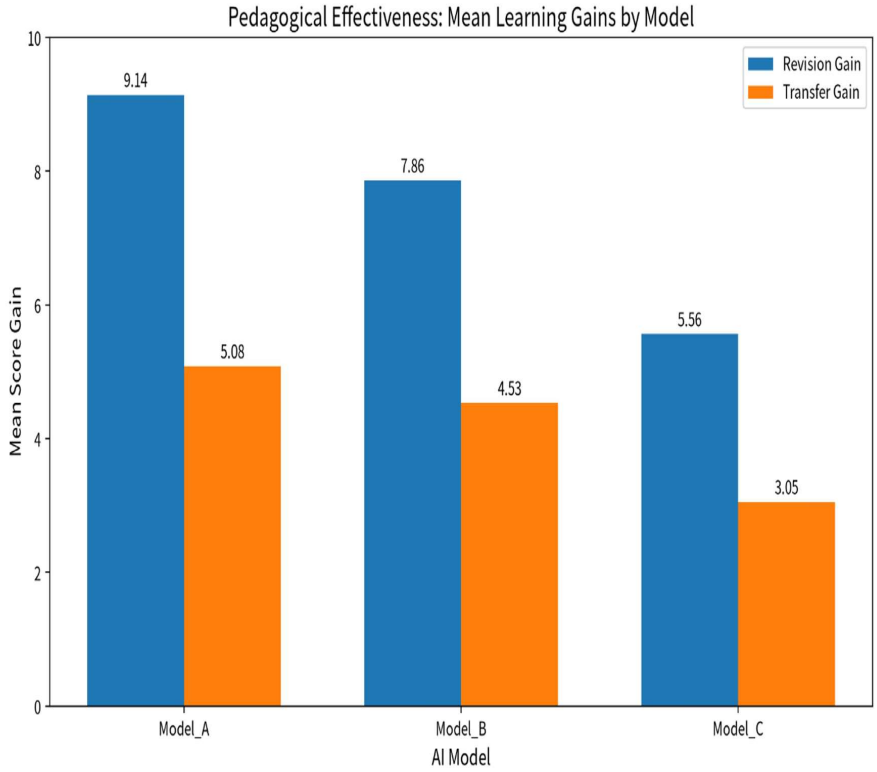

Figure 10. Additional learning-gain visualization retained from the source analysis; the numerical pattern is identical to the reported revision and transfer gains

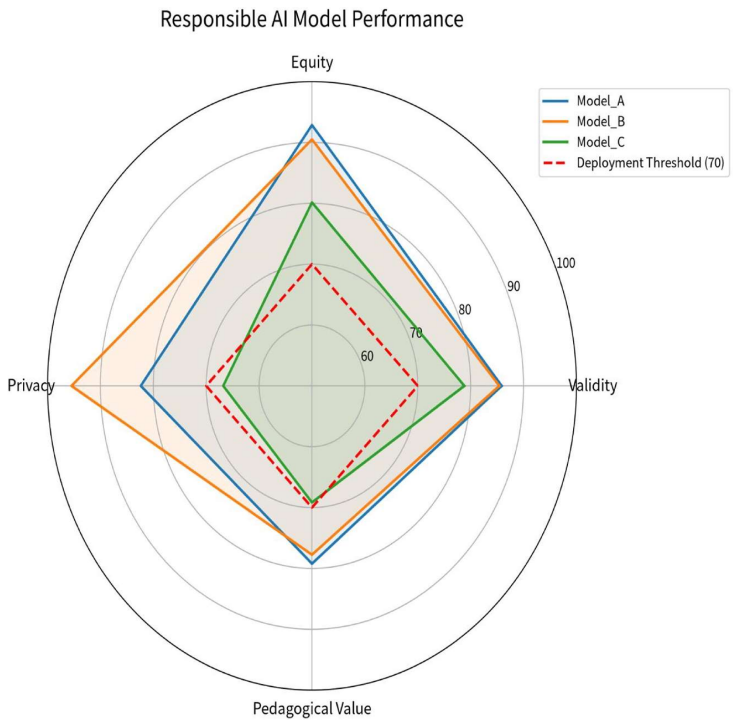

Figure 11. Additional responsible-AI radar visualization retained from the source analysis; it reproduces the four-pillar comparison

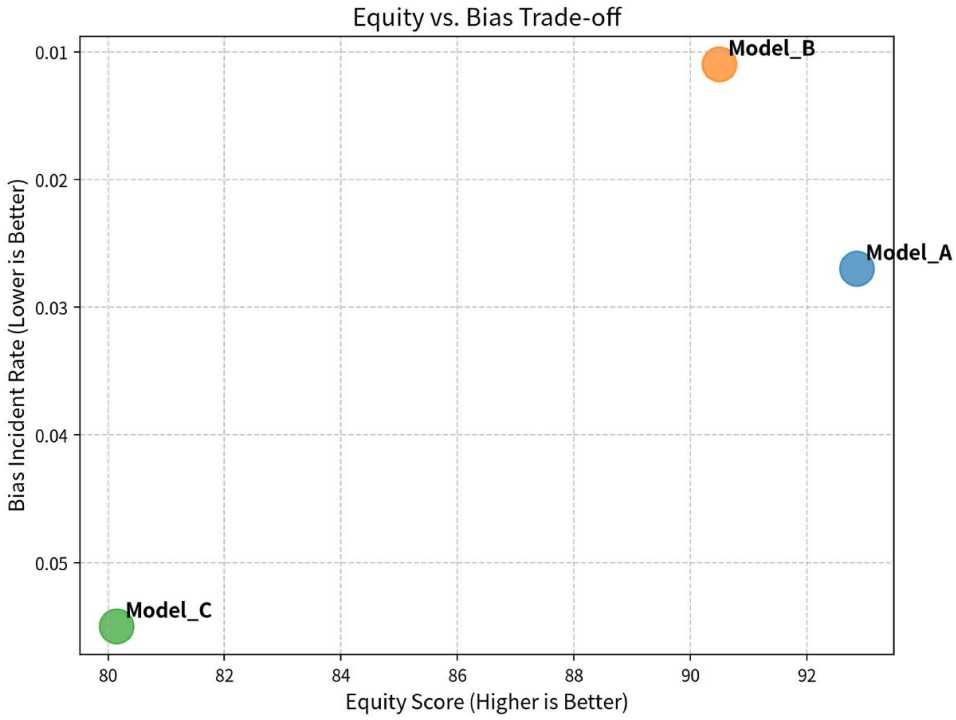

Figure 12. Additional equity–bias visualization retained from the source analysis; it provides a repeated visual record of the bias/equity trade-off

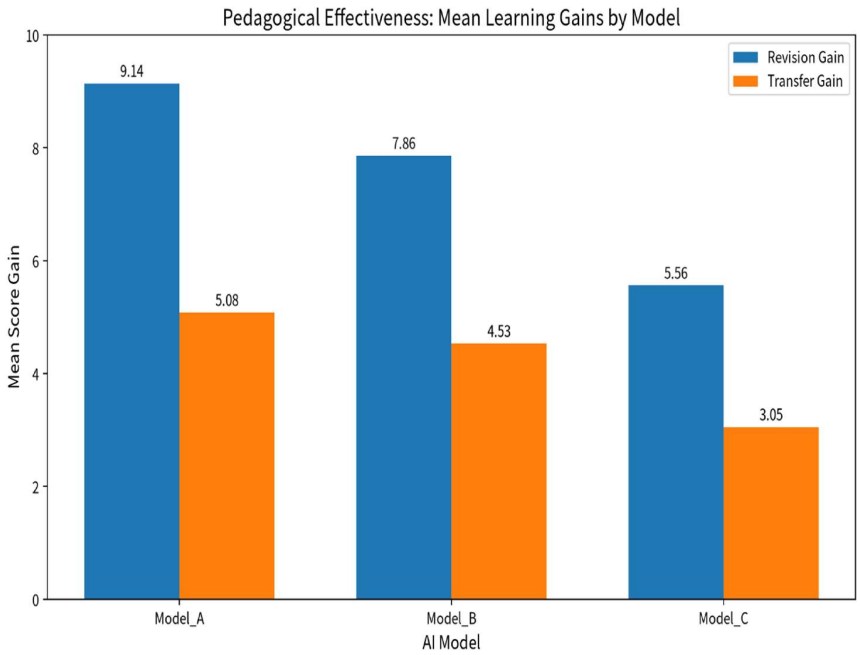

Figure 13. Additional learning-gain visualization retained from the source analysis; it confirms the ordering Model_A > Model_B > Model_C for both learning indicators

Figures 8–13 are retained because they occur as separate figure instances in the source document. The available source content indicates that these later instances repeat or reproduce the previously reported learning-gain, equity–bias, and four-pillar visualizations rather than introducing new numerical results. They therefore provide visual confirmation rather than independent statistical evidence. No additional numerical claims are inferred from these repeated visual records.

11. Overall Interpretation and Deployment Implications

The evidence supports *Model_A* as the strongest overall model when performance is evaluated across the complete responsible-AI framework. It has the highest validity, equity, and pedagogical value scores, and it produces the largest revision and transfer gains. Most importantly under the non-compensatory rule, its weakest pillar is still above the 70-point threshold, giving it the highest minimum pillar score of the three models.

Model_B represents a different but highly competitive profile. It has the highest privacy score, perfect policy transparency, the highest interaction-level privacy score, and the lowest bias incident rate. It nevertheless records slightly lower validity, equity, pedagogical value, and learning gains than *Model_A*. Because it also exceeds the threshold on all four pillars, it remains fully deployable under the study's decision rule and may be particularly attractive where privacy and governance are prioritized.

Model_C is not suitable for deployment under the specified criterion. Although its validity score of 78.81 and equity score of 80.15 remain above 70, privacy falls to 66.78 and pedagogical value to 69.16. These weaknesses coincide with the highest bias incident rate, the largest country-level performance gap, the highest reported SES gap, the lowest learning gains, and markedly weaker policy transparency. The convergence of these independent indicators strengthens the interpretation that *Model_C* requires substantial improvement before responsible educational deployment.

Overall, the analysis demonstrates that AI validity cannot be judged adequately from assessment performance alone. A model may show acceptable validity while still failing because of privacy, governance, equity, or

pedagogical limitations. The non-compensatory framework therefore produces a more stringent deployment assessment by requiring simultaneous adequacy across all four dimensions.

12. Data Interpretation and Scope of Inference

Interpret the findings within the benchmark's design. The dataset is balanced using SMOTE and is therefore an analytical benchmark rather than a population representative sample. Country comparisons represent contextual differences in the observed benchmark and should not be interpreted as causal national effects. Likewise, the 70-point threshold and any composite weighting are study specific methodological choices. These limitations do not invalidate the comparative analysis, but they define the appropriate scope of its conclusions.

13. Discussions

The Necessity of Non-Compensatory Evaluation Frameworks: The most significant finding of this study is the demonstration that high technical validity cannot compensate for deficiencies in other critical domains of responsible AI. Model_C achieved an acceptable validity score (78.81) and a passing equity score (80.15); however, it was disqualified from deployment because of its failures in privacy (66.78) and pedagogical value (69.16). This supports applying a non-compensatory decision rule in educational technology procurement. In high-stakes language assessment, a system that accurately scores grammar but violates data privacy or fails to provide actionable, learning oriented feedback is fundamentally unsuitable for the classroom. The framework proves that AI assessment value must be judged holistically, treating equity, privacy, and pedagogical impact as equally critical as psychometric validity.

Trade-offs in Model Deployment: Pedagogical vs. Governance Priorities: The comparative analysis reveals distinct deployment profiles for Model_A and Model_B, highlighting that institutional context should dictate model selection. Model_A emerges as the optimal choice for teaching and learning environments, leading in pedagogical value, equity, and tangible learning outcomes (mean revision and transfer gains). It represents an AI system deeply aligned with Learning-oriented Assessment (LoA) principles. On the other hand, Model_B represents the gold standard for data governance, achieving a 95.48 in privacy and 100% in policy transparency, alongside the lowest bias incident rate (0.011). While its pedagogical value is slightly lower than Model_A's, it still comfortably exceeds the deployment threshold. Institutions operating in regions with stringent data protection regulations (e.g., GDPR) or handling highly sensitive learner data may find Model_B the more responsible choice, illustrating the need for context-sensitive AI governance.

Equity, Bias, and the Risk of Contextual Disparities: The equity and bias audits expose the vulnerabilities of AI systems when applied across diverse linguistic and cultural backgrounds. Model_C showed the largest country-level performance gap (3.36) and the highest socioeconomic status (SES) gap (1.01), along with a bias incident rate five times higher than Model_B. This reinforces the literature's warning that AI models trained on non-representative datasets can reproduce or amplify existing educational inequalities. The spatial separation observed in the equity-bias scatter visualizations confirms that fairness cannot be assumed; it must be systematically audited. The superior performance of Model_A and Model_B across diverse demographics suggests that it is possible to engineer AI systems robust across varied cultural and linguistic contexts, provided equity principles are integrated throughout the development lifecycle.

Pedagogical Value and the Role of Human Oversight: Finally, the learning outcomes data confirm that AI's true value lies in its pedagogical utility rather than mere scoring efficiency. Model_A's superior revision and transfer gains indicate that its feedback is not only accurate but also interpretable and actionable for learners. However, as the literature emphasizes, automated evaluation is not an end in itself. The opacity of deep learning models necessitates a hybrid assessment model. Even the best performing models (Model_A and Model_B) require continuous human oversight to monitor AI-human rating alignment, provide contextual interpretation, and intervene when anomalous judgments occur. AI should be positioned as a tool that augments the assessment ecosystem, supporting feedback literacy and dialogic assessment, rather than replacing human expert judgment.

14. Conclusions

The Responsible AI Language Education Benchmark provides critical empirical evidence that integrating AI into language assessment must be guided by a multidimensional, responsible AI framework. This study concludes that Model_A is the most balanced and pedagogically effective system, offering the highest equity and learning gains, while Model_B is the best choice for environments that demand strict privacy and transparent governance. Model_C is deemed unsuitable for educational deployment, serving as a cautionary example of how failures in privacy and pedagogical design can invalidate an otherwise technically acceptable assessment tool.

Ultimately, responsible AI adoption in education requires moving beyond single metric accuracy comparisons. Institutions, policymakers, and developers must collaborate to implement non compensatory benchmarking, fairness auditing, and region specific ethical guidelines. Future research should build upon these analytical benchmarks by conducting longitudinal, population representative studies in real world classroom settings. By prioritizing a carefully designed human-AI partnership, the education sector can harness the scalability and personalization of AI while safeguarding educational equity, learner privacy, and the fundamental pedagogical goals of language assessment.

References

[1] Galaczi, E., Luckin, R. (2024). *Generative AI and language education: Opportunities, challenges and the need for critical perspectives* (Cambridge Papers in English Language Education). *Cambridge University Press*.

[2] Xi, X. (2023). Advancing language assessment with AI and ML–learning into AI is inevitable, but can theory keep up? *Language Assessment Quarterly*, 20(4–5), 357–376.

[3] Hannah, L., Jang, E. E., Shah, M., Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly*, *20*(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253.

[4] Ockey, G. J., Chukharev-Hudilainen, E., Hirch, R. R. (2023). Assessing interactional competence: ICE versus a human partner. *Language Assessment Quarterly*, 20(4–5), 377–398. https://doi.org/10.1080/15434303.2023.2237486.

[5] Jang, E. E., Sawaki, Y. (2025). Advancing language assessment for teaching and learning in the era of the artificial intelligence (AI) revolution: Promises and challenges. *Language Testing*, *42*(4), 361-368. https://doi.org/10.1177/02655322251348685.

[6] Bulut, O., Beiting-Parrish, M. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3), 3.

[7] Alfaleh, M. (2026). Sustainable AI-driven assessment in higher education: A systematic review of fairness, transparency, pedagogical innovation, and governance. *Sustainability*, 18(2), 785.

[8] Mpolomoka, D. L. (2025). Utilizing artificial intelligence for assessment in higher education. *Pedagogical Research*, 10, em0243.

[9] Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, 4.

[10] Amnie, A. G. (2026). AI-resistant and AI-resilient assessment in higher education: A systematic review of

validity grounded strategies, institutional frameworks, and equity implications. *Assessment & Evaluation in Higher Education*, 1-23.

[11] Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, 43, 16-29. https://doi.org/10.1111/emip.12602.

[12] Briggs, D. C. (2024). Strive for measurement, set new standards, and try not to be evil. *Journal of Educational and Behavioral Statistics*, 49(5), 694–701. https://doi.org/10.3102/10769986241238479.

[13] Bachman, L., Adrian, P. (2022). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

[14] Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple.... *Language Testing*, 29(1), 19–27.

[15] O'Sullivan, B. (2023). Reflections on the application and validation of technology in language testing. *Language Assessment Quarterly*, 20(4–5), 501–511. https://doi.org/10.1080/15434303.2023.2291486.

[16] Guo, S., Wang, Y., Yu, J., Wu, X., Ayik, B., Watts, F., Latif, E., Liu, L., Liu, N., Zhai, X. (2025). Artificial intelligence bias on English language learners in automatic scoring. *Proceedings of the International Conference on AI in Education*, 1–12. https://doi.org/10.48550/arXiv.2505.10643.

[17] Hannah, L., Jang, E. E., Shah, M., Gupta, V. (2023). Validity arguments for automated essay scoring of young students' writing traits. *Language Assessment Quarterly*, 20(4–5), 399–420. https://doi.org/10.1080/15434303.2023.2288253 *(Note: This is a duplicate of entry 3)*

[18] Choi, I., Zu, J. (2025). Generating language assessment content free from representational harms. *Language Testing*, 42 (4). https://doi.org/10.1177/02655322251349560.

[19] Selvam, M., Vallejo, R. G. (2025). Ethical and privacy considerations in AI-driven language learning. *LatIA*, (3), 328.

[20] Raza, F. A., Singh, A. D., Kovilpillai, J. J. S., Hamdan, A., Rajaratnam, V. (2025). Safeguarding integrity in AI-enhanced education: Stakeholder perspectives on accuracy, validity, and ethics in ASEAN. *European Journal of STEM Education*, 10(1), 22.

[21] Alwakid, W. N., Dahri, N. A., Humayun, M., & Alwakid, G. N. (2025). Exploring the role of AI and teacher competencies on instructional planning and student performance in an outcome-based education system. *Systems*, 13, 517.

[22] Mammadova, I., Aliyeva, E., Mammadova, N., Ismayilli, F. (2025). The role of artificial intelligence in shaping assessment practices in higher education. *Educational Process International Journal*, 18.

[23] Guo, K., Pan, M., Li, Y., Lai, C. (2024). Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability. *Internet and Higher Education*, 63, 100962.

[24] McLaughlin, J. E., Ponte, C. D., Lyons, K. (2025). Student perceptions of GenAI as a virtual tutor to support collaborative research training for health professionals. *BMC Medical Education*, 25, 895.

[25] Deneen, C. (2025). Editorial: Artificial intelligence and educational assessment. *Learning Letters*, 4, 38.

[26] Sawaki, Y., Ishii, Y., Yamada, H., Tokunaga, T. (2025). Examining the consistency of instructor vs. large language model ratings on summary content: Toward checklist-based feedback provision with second language writers. *Language Testing*, 42(4). https://doi.org/10.1177/02655322251349217.