CAN ARTIFICIAL INTELLIGENCE (AI) JUDGE REFLECTION? VALIDITY, RELIABILITY, AND FAIRNESS OF GENERATIVE AI-ASSISTED ASSESSMENT IN POSTGRADUATE HEALTH PROFESSIONS EDUCATION

Brekhna Jamil[1], Nowshad Asim[2]

How to cite this article

Jamil B, Asim N. Can Artificial Intelligence (AI) Judge Reflection? Validity, Reliability, and Fairness of Generative AI-Assisted Assessment in Postgraduate Health Professions Education. J Gandhara Med Dent Sci. 2026;13(1):3-11. https://doi.org/10.37762/jgmds.13-1.835

Date of Submission: 10-12-2025
Date Revised: 15-12-2025
Date Acceptance: 16-12-2025

[2]Assistant Professor, Institute of Health Profession Education and Research, Khyber Medical University, Peshawar

Correspondence

Brekhna Jamil, Professor, Institute of Medical Education and Research, Khyber Medical University, Peshawar, Honorary Professor, University of Dundee, UK

: +92-320-9591000
: brekhnajamil@kmu.edu.pk

ABSTRACT

OBJECTIVES

This study aimed to evaluate the role of GenAI in assessing reflective writing among Master of Health Professions Education (MHPE) students by comparing GenAI scores with those of human raters, examining subgroup fairness, exploring stakeholder perceptions, and proposing governance recommendations.

METHODOLOGY

A sequential mixed-methods study was conducted in an MHPE programme at Khyber Medical University, Pakistan. In Phase I, 120 Gibbs-structured reflections from 40 students were scored by three trained faculty raters and a GPT-4-level GenAI model using an eight-dimensional rubric. Inter-rater reliability, AI-human agreement, and subgroup differences by gender, discipline, and career stage were examined. In Phase II, semi-structured interviews were conducted with 10 MHPE students and the three faculty raters. Data were analysed using reflexive thematic analysis and integrated with quantitative results.

RESULTS

Human scoring demonstrated strong reliability (ICC = .82). GenAI showed high alignment with human ratings for surface-level dimensions such as clarity and language mechanics (r = .81-.84), but only modest agreement for higher-order reflective constructs including feelings, analysis, and conclusion (r = .49-.59). Exploratory subgroup analyses revealed no statistically significant differences in AI-human discrepancies. However, qualitative accounts highlighted concerns about linguistic and cultural fairness. Participants valued AI for efficient, organised feedback but consistently emphasised its inability to interpret emotional nuance, contextual meaning, or developmental trajectories. Faculty stressed the irreplaceability of human judgment and the need for transparent governance and fairness monitoring.

CONCLUSION

GenAI can effectively support the assessment of structural and linguistic aspects of reflective writing but remains limited in evaluating deeper reflective constructs central to postgraduate learning. Ethical and educationally sound integration requires hybrid human-AI approaches in which AI provides formative support while human evaluators retain primary responsibility for interpretive judgment, fairness oversight, and professional mentorship. GenAI should supplement, not replace, human assessment of reflective writing.

KEYWORDS: *Artificial Intelligence, Writing, Health Professions Education, Validity, Reliability, Gibbs' Reflective Cycle*

INTRODUCTION

Reflective practice is widely acknowledged as a cornerstone of professional development in the health professions, enabling learners to examine experiences, integrate new understanding, and enhance clinical and educational judgment.[1,2] Reflective writing operationalises this process by encouraging learners to articulate emotions, evaluate experiences, and engage in critical analysis.[1] Among the many reflective frameworks, Gibbs' Reflective Cycle is widely used because of its structured approach: description, feelings, evaluation, analysis, conclusion, and action plan, which guides learners toward deeper levels of meaning-making.[1,2,3] At Khyber Medical University (KMU), Pakistan, the Master of Health Professions Education (MHPE) programme is a two-year postgraduate degree designed to develop clinicians, educators, and academic leaders with advanced expertise in teaching, assessment, curriculum development, leadership, and

educational research. A defining feature of the KMU MHPE curriculum is its strong emphasis on reflective practice as a core component of professional development. Students are required to maintain regular reflective entries, typically structured using Gibbs' Reflective Cycle, which form an integral part of their learning portfolios and course assessments. These reflections reveal students' evolving professional identities, metacognitive development, and responses to complex educational situations. They also support faculty in understanding learners' growth trajectories and the emotional labour inherent in becoming an educator.[4] In spite of its pedagogical value, assessing reflective writing is difficult. Reflective texts are subjective, context-dependent, and often emotionally nuanced. Even with structured rubrics, human raters may diverge in their interpretation of reflective depth, emotional authenticity, and contextual meaning, leading to moderate inter-rater reliability and concerns about fairness.[5,6] This variability is particularly salient in postgraduate education, where learners draw on complex clinical, academic, and personal experiences. Advances in natural language processing and large language models (LLMs) have driven interest in using generative AI (GenAI) to support assessment. LLMs can evaluate linguistic quality, detect structural features, and generate consistent feedback at scale.[7,8] However, the suitability of AI for assessing reflective writing is uncertain. Reflection involves non-computable constructs such as emotion, identity formation, and tacit knowledge.[5,6,7,8,9] AI systems may also reproduce linguistic or cultural biases present in their training data.[10,11] Empirical work on GenAI in assessment has primarily concentrated on standard essays, short constructed-response tasks, and automated scoring of analytic writing, with limited exploration of reflective writing in postgraduate health professions education.[8,9,10,11,12] Existing studies tend to emphasise technical performance metrics, such as AI-human correlations or reliability indices, rather than examining how GenAI interprets deeper reflective constructs such as emotion, identity, and meaning-making.[10,11,12,13] Research that integrates quantitative evidence of model alignment with qualitative accounts of fairness, trust, and teacher agency remains rare, particularly in low- and middle-income settings and in South Asia.[11] Against this backdrop, the present study contributes context-specific, mixed-methods evidence on GenAI-assisted assessment of reflective writing in a postgraduate health professions education programme. Based on these considerations, the present study sought to address the following research questions: How well do GenAI-generated scores align with trained human raters using a Gibbs-aligned reflective writing rubric? And How do MHPE students and faculty perceive GenAI feedback regarding fairness, trust, and educational value?

METHODOLOGY

A sequential explanatory mixed-methods design was used. Quantitative analyses examined human inter-rater reliability, AI-human agreement, and subgroup fairness. Qualitative interviews then explored participant perceptions, interpreted alongside quantitative findings. Integration occurred during interpretation in line with mixed-methods best practice.[14] The design was selected because the use of GenAI in reflective assessment is not only a psychometric issue but also a sociotechnical phenomenon with implications for fairness, trust, and teacher agency dimensions that cannot be understood through quantitative approaches alone. The study was conducted in the Master of Health Professions Education (MHPE) programme at the Institute of Health Professions Education and Research (IHPER), Khyber Medical University (KMU), Pakistan. The MHPE programme follows a blended-learning approach, combining asynchronous online modules, synchronous virtual sessions, and face-to-face instructional blocks. Reflective writing is embedded throughout the curriculum and is typically structured using Gibbs' Reflective Cycle, enabling students to articulate descriptions, emotions, analysis, and planned actions in relation to teaching and learning experiences.[1,2,3] A total of 120 reflective entries were randomly sampled from work submitted by 40 MHPE students in the 2024-2025 academic year. Each student contributed up to three reflections (300-500 words), selected from multiple modules to ensure representation across learning contexts. All reflections were anonymised before being scored. Because multiple reflections originated from the same learners, the dataset exhibited a nested structure, rendering the scores nonstatistically independent. This limitation was acknowledged in the interpretation of quantitative findings, as nested data can influence estimates of reliability and agreement.[15,16] Three experienced faculty members from the Institute, each holding postgraduate qualifications in medical or health professions education and with prior experience assessing reflective writing, acted as human raters. These faculty members later took part in qualitative interviews to reflect on their experiences with AI-generated feedback. For the student interviews, purposive sampling was used to ensure diversity across gender, teaching/clinical background, and career stage. Ten students were recruited, and the tenth interview achieved data saturation; by the eighth interview, no substantively new codes were emerging, and the final two interviews confirmed thematic stability (Guest et al., 2006). The

three faculty raters were also interviewed to explore perceptions of GenAI, fairness, and implications for teacher agency and governance. Ethical approval was obtained from the Institutional Review Board of Khyber Medical University (Ref No: 14 IHPER/KMU/26-138). Written informed consent was secured from all participants. Students were informed that AI-generated scores would not influence their course grades. All reflections were anonymised before being shared with human raters or the AI model, and all interview data were stored securely in encrypted files. Given the personal nature of reflective writing, confidentiality and voluntary participation were emphasised throughout the research process. An eight-dimensional rubric aligned with Gibbs' Reflective Cycle was used to evaluate each reflection (Supplementary Appendix S1). The dimensions included description, feelings, evaluation, analysis, conclusion, action plan, clarity and coherence, and language mechanics. Each dimension was rated on a five-point scale (0-4), yielding a maximum total score of 32. Rubric development drew on established reflective assessment frameworks, notably the REFLECT rubric,[5] and Kember's categorisation of reflective depth.[17] To establish content validity, an initial draft of the rubric was reviewed by four health professions education scholars with experience in teaching and assessing reflection. Experts examined the relevance, clarity, and representativeness of each descriptor, leading to refinements in the definitions of the ‑feelings,‖ ‑analysis,‖ and ‑conclusion‖ dimensions. Cognitive debriefing interviews were subsequently conducted with six MHPE students to explore how they interpreted each rubric element, following recommended practices in instrument refinement.[18] Feedback revealed ambiguities in the differentiation of mid-level performance descriptors, prompting further clarification. The revised rubric was pilot-tested on 15 reflections not included in the primary dataset. Two researchers independently applied the rubric, discussed discrepancies, and made final adjustments to improve clarity and reduce overlap between adjacent levels, consistent with best practices for rubric development.[19] This iterative development supported strong content representation and clarity in scoring expectations. Three experienced faculty members with postgraduate HPE qualifications served as human raters. A three-hour calibration workshop was held to promote shared interpretive understanding. Raters reviewed exemplar reflections representing varying levels of reflective depth, discussed scoring discrepancies, and independently scored 10 anchor reflections. Pre-study inter-rater reliability was strong (ICC = 0.76), consistent with literature noting greater agreement on low-inference reflective dimensions.[5] Raters then independently scored all anonymised reflections. Automated scoring was conducted using a GPT-4-family large language model accessed through an API. To ensure scoring consistency, fixed parameters were used for all entries (temperature=0.0, top-p=1.0). A structured scoring prompt containing the full rubric, dimension-level descriptors, and formatting instructions was applied uniformly across all reflections. The complete scoring prompt is provided in Supplementary Appendix S2. As with most commercial large language models, the underlying architecture and training data are proprietary, limiting transparency into linguistic distributions and potential cultural or linguistic biases.[10,11,12,13,14,15,16,17,18,19,20] Identical parameters and prompts were applied to all reflections to ensure standardisation. Data Collection Procedures comprised of two phases:

Phase I: Human and GenAI scoring

All 120 reflections were first anonymised and then distributed to the faculty raters, who independently scored every entry using the rubric. Detailed anonymisation and rater-scoring procedures are documented in Supplementary Appendix S3. After human scoring, the de-identified reflections were submitted to the GenAI model, which generated dimension-level scores and short narrative justifications. Scores were compiled in SPSS (version 28) for analysis.

Phase II: Semi-structured interviews

After preliminary quantitative analysis, separate interview guides were developed for students and faculty. The student and faculty interview guides used to explore perceptions of GenAI feedback, fairness, validity, and teacher agency are available in Supplementary Appendix S4 (student guide) and Supplementary Appendix S5 (faculty guide). Student interviews explored experiences receiving AI feedback on reflective writing, perceptions of its fairness and usefulness, views on trust and authenticity, and preferences for human versus AI involvement in assessment. Faculty interviews addressed experiences of scoring reflections, perceptions of AI accuracy and limitations, concerns about validity and fairness, and views on assessment governance and teacher agency. Interviews, lasting approximately 30-40 minutes, were conducted via Zoom, audio-recorded with consent, and transcribed verbatim. All transcripts were anonymised; participant codes replaced real names. Quantitative data were analysed using SPSS version 28. Human inter-rater reliability was estimated using a two-way random-effects intraclass correlation coefficient for single measures [ICC (2,1)], with corresponding 95% confidence intervals. AI-human alignment was examined using Pearson correlations and absolute-agreement ICCs comparing AI-generated scores with

the mean human score for each dimension.[17,18,19,20,21] Subgroup fairness analyses compared AI-human score discrepancies across gender, clinical vs. non-clinical background, and early- vs. senior-career professionals using independent-samples t-tests. Analyses were treated as exploratory due to the nested data structure and modest sample sizes. Qualitative data were analysed using reflexive thematic analysis.[22] Two researchers (BJK, NA) independently familiarised themselves with the transcripts, noting initial impressions. Coding was then conducted inductively and deductively, with attention to experiences of AI feedback, perceptions of fairness and bias, views on teacher agency, and expectations for governance. Codes were iteratively refined and organised into candidate themes. Themes were reviewed against the dataset for coherence and distinctiveness, and refined through team discussion. To enhance trustworthiness, several strategies were employed. First, researcher reflexivity was encouraged through the use of analytic memos.[23] Second, triangulation was achieved by comparing student and faculty accounts and relating them to quantitative results. Third, concise thematic summaries were shared with a subset of participants (member checking), who confirmed that the interpretations resonated with their experiences.[24] Integration of Quantitative and Qualitative Findings occurred at the interpretation stage using a weaving approach. Quantitative findings about AI-human agreement and subgroup differences were interpreted alongside qualitative themes that contextualised these patterns, particularly in relation to fairness, authenticity, and agency. This allowed the study to move beyond psychometric performance to address consequential and ethical dimensions of GenAI-mediated assessment.[25,26]

RESULTS

The dataset consisted of 120 reflective writing samples from 40 MHPE students. In the qualitative phase, 10 students and all three faculty raters participated in interviews. No participant withdrew, and there were no missing scores across the quantitative dataset. Although multiple reflections originated from the same learners, observations were not statistically independent, a factor considered in the interpretation of psychometric findings.[15,16] Nonetheless, the dataset provided sufficient diversity to examine patterns of human-AI agreement and explore stakeholder perceptions.

Human inter-rater reliability

Inter-rater reliability among the three faculty raters was strong, supporting the stability of human scoring. The overall intraclass correlation coefficient (ICC[2,1]) for total rubric scores was 0.82 (95% CI [.77, .87]), indicating high agreement, consistent with prior research showing that structured reflective rubrics can yield acceptable reliability among trained raters.[5,6,7,8,9,10,11,12,13,14,15,16,17] Dimension-level ICCs varied according to the interpretive demands of each construct:

- **The highest reliability was observed in low-inference dimensions,** such as *clarity and coherence* (ICC = 0.89) and *language mechanics* (ICC = 0.89).
- **High reliability** was also observed for *description* (ICC = 0.84), reflecting relatively straightforward application of criteria.
- **Moderate but acceptable reliability** was found for more inferential dimensions, such as *feelings* (ICC = 0.61) and *analysis* (ICC = 0.63). These findings are consistent with the interpretive complexity associated with evaluating emotional expression and critical meaning-making in reflective writing.[9]

The level of human agreement is supported using the mean human score as the comparator for AI scoring.

AI-human agreement:

GenAI demonstrated strong alignment with human ratings for structural and linguistic dimensions. Correlations between AI and mean human scores were .81 for clarity and coherence and .84 for language mechanics. These high correlations indicate that GenAI was highly consistent with human raters in identifying surface-level qualities such as organisation, readability, grammar, and syntactic correctness (8). For reflective constructs, agreement was more modest. Correlations with human scores were 0.62 for evaluation, 0.57 for conclusion, 0.59 for action plan, 0.54 for analysis, and 0.49 for feelings. These findings suggest that while GenAI can recognise explicit evaluative statements and simple reflective patterns, it is less sensitive to nuanced emotional expression, contextual interpretation, and deeper analytical reasoning. The overall ICC between AI and human total scores was 0.71, suggesting moderate-to-strong agreement. Details of human-human and AI-human agreement across rubric dimensions are presented in Table 1. A consistent pattern across entries was a tendency for GenAI to assign slightly lower scores than human raters on the *feelings*, *analysis*, and *conclusion* dimensions. This pattern suggests partial construct underrepresentation, wherein the AI detects structural or descriptive elements but underestimates emotional depth or interpretive complexity. These trends reinforce the need for human oversight in evaluating reflective constructs that depend on contextual understanding, emotional nuance, and professional identity work.

Table 1. Inter-rater Reliability and AI-Human Agreement Across Rubric Dimensions

Dimension	Human-Human ICC	AI-Human r	AI-Human ICC
Description	.84	.68	.73
Feelings	.61	.49	.52
Evaluation	.78	.62	.67
Analysis	.63	.54	.59
Conclusion	.76	.57	.64
Action plan	.74	.59	.66
Clarity & coherence	.89	.81	.83
Language mechanics	.89	.84	.86
Overall	.82	.71	.71

ICC = Intraclass Correlation Coefficient.

Fairness audit

The fairness audit did not reveal statistically significant subgroup differences in AI-human discrepancies. Mean differences between human and AI total scores were small and non-significant for gender, clinical versus non-clinical background, and early- versus senior-career status. Early-career professionals tended to receive slightly lower AI scores than their more senior counterparts, but the differences did not reach significance. Subgroup results are summarised in Table 2. Although quantitative subgroup analyses did not demonstrate systematic disadvantages by gender, discipline, or career stage, qualitative data (see below) suggested that some students perceived GenAI to favour more polished English and certain cultural expressions, raising concerns about linguistic fairness that were not directly captured in the quantitative stratification.

Table 2: Fairness Audit: Mean Human and AI Scores Across Subgroups

Subgroup	Mean Human Score	Mean AI Score	Difference (Human-AI)	P-Value
Male	24.70	24.00	0.70	.09
Female	24.60	24.20	0.40	.18
Early-career	24.40	23.80	0.60	.11
Senior-career	24.90	24.40	0.50	.15

Scores represent totals on a 32-point rubric; no statistically significant subgroup differences were observed.

Qualitative findings

Four interrelated themes emerged from student and faculty interviews, illustrating how participants interpreted the role, limitations, and ethical implications of GenAI in reflective writing assessment. The four themes, with brief descriptions and illustrative quotations, are shown in Table 3.

Table 3: Themes, Sub-Themes, and Illustrative Quotations

Theme	Sub-themes	Illustrative quotations
1. Trust with scepticism	Usefulness for structure and clarity; doubts about depth and meaning	*—AI helps with structure, but it does not understand what I am trying to express.‖ (Int#2)* *—It gives neat feedback, but it cannot feel what I am feeling.‖ —It judged how I write, not what I learned.‖ (Int#9)*
2. Fairness concerns	Perceived linguistic disadvantage; cultural misalignment	*—My reflection was deep, but AI scored it lower because my English is not fancy.‖ (Int#5)* *—Some expressions make sense in our context, but AI treats them as errors.‖ (Int#7)* *—It rewards style more than substance.‖ (Int#3)*
3. Irreplaceability of teacher judgment	Importance of contextual interpretation; emotional and relational understanding	*—Reflection is human work. AI cannot see the growth and struggle behind the words.‖ (Int#1)* *—A teacher knows my journey. AI cannot interpret that.‖ (Int#4)* *—Human judgment is needed to see meaning beyond the text.‖ (Int#3)*
4. Expectations for transparency and governance	Need for transparency, fairness monitoring, and AI literacy for students and faculty	*—If AI is used, it must be transparent and monitored.‖ (Int#3)* *—We should have the right to request human review.‖ (Int#6)* *—There should be checks to see if AI is fair to everyone.‖ (Int#8)*

Theme 1: Trust with Scepticism

Students described GenAI feedback as immediate, structured, and helpful for improving readability. One participant noted,

- AI gives feedback instantly, and it is very organised; it highlights things I often overlook." (Int#10)

However, this trust was tempered by doubt about AI's ability to interpret meaning or emotion.

Another student explained,

- It judged how I write, not what I learned. It cannot feel what I am trying to express." (Int#2)

Faculty echoed this concern, describing AI as efficient but shallow:

- It can summarise, but it cannot understand the struggle behind the reflection." (Int#3)

Theme 2: Fairness Concerns

A recurring concern, especially among multilingual

learners, was that GenAI appeared to privilege polished English and certain rhetorical styles. A student observed,

- *My English is simple, so AI thought my reflection was simple."* (Int#7)

Faculty recognised the same pattern:

- *It rewards a certain style of English. Depth gets lost if the writing is not stylistic."* (Int#2)

Cultural nuance was also reported as being overlooked, as one participant shared,

- *Some expressions make sense in our context, but AI treated them as mistakes."* (Int#4)

Theme 3: Irreplaceability of Teacher Judgment

Faculty positioned AI as an assistant rather than an assessor. One faculty member stated,

- *Reflection is human work. You need to know the learner to understand their growth."* (Int#2)

Participants emphasised that human evaluators grasp contextual cues such as teaching challenges, emotional labour, or professional identity development that AI cannot interpret. A student added,

- *When a teacher reads my reflection, they know where I am coming from. AI cannot see that."* (Int#8)

Theme 4: Expectations for Transparency and Governance

Both groups expressed a desire for clear institutional policies governing AI use. Students wanted to know.

- *When AI is used, how it is used, and what we can do if the score is unfair."* (Int#6)

Faculty also stressed the need for safeguards, articulating concerns about accountability:

- *We need routine audits. If AI is biased, we must know."* (Int#3)

Participants requested training to support AI literacy, highlighting the need for ethical and transparent implementation.

Integrated interpretation

Quantitative and qualitative findings were mutually informative. Where AI aligned closely with human ratings on clarity and language, participants generally endorsed AI's usefulness as a feedback tool. Where divergence was most significant on depth of analysis, emotional authenticity, and action planning, participants identified these areas as inherently human and expressed reluctance to cede judgment to AI. Although quantitative subgroup analyses did not demonstrate systematic bias by gender, professional background, or career stage, qualitative accounts pointed to perceived linguistic and cultural inequities, particularly for students who felt their writing was less stylistically polished or more locally contextualised. Taken together, the findings suggest that GenAI is most defensible as a supplementary tool focused on surface-level features, within a governance framework that foregrounds human judgment, fairness monitoring, and transparency.

DISCUSSION

This study examined the performance, fairness, and perceived value of GenAI-assisted scoring of reflective writing in a postgraduate health professions education programme. Consistent with the international literature, the findings demonstrated that GenAI performs well on surface-level writing features but struggles with higher-order reflective constructs that require interpretation, emotional insight, and contextual understanding.[5,6,7,8,9] Across methods, the central conclusion is clear: GenAI can supplement but cannot replace human evaluators in reflective assessment, particularly when reflection is tied to identity formation, professional growth, and emotionally nuanced learning.[4-27]

Implications for validity

The validity of AI-assisted reflective writing assessment must be considered within a comprehensive framework that includes content representation, internal structure, response processes, and consequential aspects.[25,26] Across these dimensions, the findings highlight both potential and limitations.

Construct Representation and Under-Representation

Quantitatively, GenAI demonstrated strong alignment with human ratings on clarity, coherence, and linguistic mechanics, dimensions that rely on low-inference cues and surface textual features. This is consistent with the literature, which notes that automated scoring systems are most reliable when evaluating structural and linguistic attributes.[8-21]

However, divergent performance on feelings, analysis, and conclusion indicates construct under-representation, a known limitation of algorithmic scoring for reflective writing.[5,6,7,8,9] Reflection involves interpreting experiences, articulating emotions, and integrating personal and professional identities, demands that exceed the current interpretive capacity of large language models.[11,12,13] As such, AI-only scoring would misrepresent the construct of reflection by overemphasizing form at the expense of meaning.

Internal Structure

The pattern of inter-rater reliability among human scorers is higher for low-inference dimensions and moderate for emotion- and analysis-based dimensions, which mirrors established evidence in reflective assessment.[17] AI-human agreement followed this same pattern, but with greater attenuation on interpretive dimensions. This internal-structure profile underscores the methodological necessity of human oversight in evaluating meaning-rich constructs.

Response Processes

Qualitative findings revealed that GenAI often privileged polished, fluent English over culturally contextualised expressions and emotionally grounded

meaning-making. This aligns with literature documenting linguistic and cultural bias in LLMs.[10,11] Where students wrote in simple or regionally contextualized English, they perceived AI feedback as less favourable despite deeper reflective insight. This constitutes a form of construct-irrelevant variance tied to language proficiency rather than reflective capability.

Consequential Validity

Participants highlighted that reflective writing communicates vulnerability, struggle, growth, and emergent professional identity features central to health professions education.[6,7,8,9] Students expressed unease at the prospect of AI judging such personal narratives, while faculty emphasized the pedagogical value of human judgment. These insights reflect broader concerns in the literature that overreliance on AI may erode teacher agency, diminish mentoring relationships, and reshape assessment cultures in problematic ways.[13-28] Taken together, these validity insights strongly support hybrid, human-led approaches rather than autonomous AI scoring.

Reliability Considerations

Human scoring demonstrated strong reliability, and AI-human agreement reached moderate-to-strong levels overall. However, as scholars caution, reliability alone is insufficient to justify automated scoring in the absence of evidence of construct validity and fairness.[21,22,23,24,25,26] A scoring system can be reliably misaligned with the intended construct if it consistently rewards attributes unrelated to meaning-making, emotional depth, or professional development. The reliability patterns observed here confirm that while AI may support consistency for technical writing components, it cannot yet reliably interpret the deeper constructs necessary for summative reflective evaluation.

Fairness, Equity, and Linguistic Considerations

Although subgroup fairness analyses did not show statistically significant differences by gender, clinical background, or career stage, qualitative perceptions highlighted subtle but meaningful fairness concerns, especially linguistic fairness for students writing in non-idiomatic English. These findings align with emerging evidence that LLMs may reproduce linguistic and cultural hierarchies embedded in their training data.[10,11,20] The absence of formal language proficiency measures in the dataset limits the interpretation of fairness outcomes. Nonetheless, perceived unfairness carries significant consequences and can threaten trust, legitimacy, and students' willingness to engage with AI-mediated feedback.[13] International AI ethics frameworks emphasise fairness, transparency, and the right to contest algorithmic decisions. These guidelines support a cautious and human-led approach to AI use in assessment.[29]

Teacher Agency and Professional Judgment

A central theme across interviews was the irreplaceability of educators in interpreting reflective writing. Faculty articulated that understanding a student's reflective journey requires relational knowledge, contextual insight, and professional empathy attributes beyond the computational scope of current AI models.[4,5,6] Over-delegating reflective assessment to AI risks reducing educators' interpretive authority and diminishing the pedagogical richness of feedback.[28] Protecting teacher agency requires:

- Human raters retaining final evaluative authority
- Precise mechanisms for overriding AI outputs
- Institutional policies affirming human judgment in assessment decisions

These recommendations align with global AI governance principles promoting human oversight, accountability, and educational equity.[30]

Theoretical Contribution: What AI Can and Cannot Assess

The findings illuminate an important theoretical boundary: reflective constructs involving emotional authenticity, experiential meaning-making, or identity transformation may be fundamentally non-computable or only partially inferable from surface linguistic features. AI may approximate the structure of reflection but cannot infer the lived experience behind the text.

This distinction advances theory in AI-enabled assessment by clarifying that:

- AI is well-suited for evaluating surface-level and low-inference features.
- AI struggles to evaluate introspective, relational, and contextualised cognitive processes.
- Human interpretation remains essential for assessing metacognition, emotion, and identity work.

This aligns with broader arguments in the field that the epistemic foundations of reflection and the pedagogical relationships that sustain it cannot be mechanised.

Recommendations for Responsible Integration

The findings support several recommendations for the ethical and educationally sound use of AI in reflective writing assessment:[29]

1. Use AI as a supplementary tool, focusing on surface features and leaving Interpretive judgement to human
2. Ensure transparency, informing learners when AI is used, what it evaluates, and how scores are interpreted.
3. Enable student agency, including the right to request human review or contest AI - influenced decisions.
4. Monitor fairness continuously, considering linguistic, cultural, and stylistic dimensions that may not be detectable through quantitative analyses alone.
5. Invest in faculty and student AI literacy to enable stakeholders to engage critically with AI outputs.
6. Align institutional policy with global AI ethics frameworks, ensuring human oversight, accountability,

and equity.

LIMITATIONS

Limitations include the single-institution context, limiting generalisability; the nested structure of reflections within students; the modest sample size for subgroup analyses; and the proprietary nature of AI systems, which limits transparency regarding training data and potential sources of bias. Future studies should employ larger and more diverse datasets, incorporate formal measures of language proficiency, and use multilevel modelling to enhance analytic precision.

CONCLUSIONS

GenAI offers promising opportunities to enhance efficiency and consistency in assessing reflective writing, particularly in terms of structural and linguistic features. However, reflective writing is fundamentally a human, relational, and interpretive practice, deeply tied to learner identity, emotional insight, and contextual meaning-making. In this study, GenAI consistently underperformed on higher-order reflective constructs and raised concerns about linguistic fairness among participants. These findings underscore that GenAI should supplement, not supplant, human evaluators. Ethical and educationally sound integration requires hybrid models, transparent governance, fairness monitoring, and explicit protection of teacher agency. As AI becomes more embedded in health professions education, maintaining human judgment at the core of reflective assessment remains essential for both the integrity of learning and the development of reflective practitioners.

CONFLICT OF INTEREST: None

FUNDING SOURCES: None

REFERENCES

1. Schön DA. The reflective practitioner: How professionals think in action. London: Temple Smith; 1983.
2. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract. 2009;14(4):595–621. https://doi.org/10.1007/s10459-007-9090-2. PMID: 18274874
3. Sandars J. The use of reflection in medical education: AMEE Guide No. 44. Med Teach. 2009;31(8):685–95. https://doi.org/10.1080/01421590903050374. PMID: 19811128
4. Adeani IS, Febriani RB, Syafryadin S. Using Gibbs' reflective cycle in making reflections of literary analysis. Indonesian EFL J. 2020;6(2):139–48. https://doi.org/10.25134/ieflj.v6i2.3385.
5. Jasper M, Rosser M. Reflection and reflective practice. In: Jasper M, Rosser M, editors. Professional development, reflection, and decision-making in nursing and healthcare. Chichester: Wiley-Blackwell; 2013. p. 41–82.
6. Moon J. Using reflective learning to improve the impact of short courses and workshops. J Contin Educ Health Prof. 2004;24(1):4–11. https://doi.org/10.1002/chp.1340240103.PMID: 15069907
7. Ryan M, Ryan M. Theorising a model for teaching and assessing reflective learning in higher education. High Educ Res Dev.2013;32(2):244–57. https://doi.org/10.1080/07294360.2012.661704.
8. Wald HS, Borkan JM, Taylor JS, Anthony D, Reis SP. Fostering and evaluating reflective capacity in medical education: developing the REFLECT rubric for assessing reflective writing. Acad Med. 2012;87(1):41–50. https://doi.org/10.1097/ACM.0b013e31823b55fa.PMID:22104058
9. Williamson S, Seewoodhary R. A review and reflection on the visual rehabilitation progress of an older person following cataract surgery two years on. Int J Ther Rehabil. 2016;23(5):242–6. https://doi.org/10.12968/ijtr.2016.23.5.242.
10. Kumar P. Large language models (LLMs): survey, technical frameworks, and future challenges. Artif Intell Rev. 2024;57(10):260. https://doi.org/10.1007/s10462-024-10874-9.
11. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, et al. Bias and fairness in large language models: a survey. Comput Linguist. 2024;50(3):1097–179. https://doi.org/10.1162/coli_a_00523.
12. Liang J-C, Hwang G-J, Chen M-RA, Darmawansah D. Roles and research foci of artificial intelligence in language education: an integrated bibliographic analysis and systematic review approach. Interact Learn Environ. 2023;31(7):4270–96. https://doi.org/10.1080/10494820.2021.1982642.
13. Williamson B, Macgilchrist F, Potter J. Re-examining AI, automation and datafication in education. Abingdon: Routledge/Taylor & Francis; 2023. p. 1–5.
14. Lee D, Arnold M, Srivastava A, Plastow K, Strelan P, Ploeckl F, et al. The impact of generative AI on higher education learning and teaching: a study of educators' perspectives. Comput Educ Artif Intell. 2024;6:100221. https://doi.org/10.1016/j.caeai.2024.100221.
15. Creswell JW, Plano Clark VL. Revisiting mixed methods research designs twenty years later. Handb Mixed Methods Res Des. 2023;1(1):21–36.
16. Hox J, de Leeuw E, Klausch T. Mixed-mode research: issues in design and analysis. In: Biemer PP, de Leeuw ED, Eckman S, Kreuter F, Lyberg LE, Tucker C, West BT, editors. Total survey error in practice. Hoboken (NJ): Wiley; 2017. p. 511–30.
17. Raudenbush SW, Bryk AS. Hierarchical linear models: applications and data analysis methods. 2nd ed. Thousand Oaks (CA): Sage Publications; 2002.
18. Kember D, McKay J, Sinclair K, Wong FKY. A four-category scheme for coding and assessing the level of reflection in written work. Assess Eval High Educ. 2008;33(4):369–79. https://doi.org/10.1080/02602930701293355.
19. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30(4):459–67. https://doi.org/10.1002/nur.20199. PMID: 17654487
20. Jonsson A, Svingby G. The use of scoring rubrics: reliability, validity and educational consequences. Educ Res Rev. 2007;2(2):130-44.https:doi.org/10.1016/j.edurev.2007.05.002.
21. Chinta SV, Wang Z, Yin Z, Hoang N, Gonzalez M, Quy TL, et al. FairAIED: Navigating fairness, bias, and ethics in educational AI applications. arXiv preprint arXiv:2407.18745. 2024. Available from: https://arxiv.org/abs/2407.18745.

22. Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. Educ Meas Issues Pract. 2012;31(1):2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.
23. Braun V, Clarke V. Toward good practice in thematic analysis: avoiding common problems and becoming a knowing researcher. Int J Transgend Health. 2023;24(1):1–6. https://doi.org/10.1080/26895269.2022.2026619. PMID:36608079
24. Rankin J, McFadyen J. The role of gatekeepers in research: learning from reflexivity and reflection. J Nurs Health Care (JNHC). 2016;4(1):218–9.
25. McKim C. Meaningful member-checking: a structured approach to member-checking. Am J Qual Res. 2023;7(2):41–52. https://doi.org/10.29333/ajqr/13188.
26. Kane MT. Validation as a pragmatic, scientific activity. J Educ Meas. 2013;50(1):115–22. https://doi.org/10.1111/jedm.12007.
27. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995;50(9):741–9. https://doi.org/10.1037/0003-066X.50.9.741.
28. Gearing N. What is the ideal methodological response for the learning and teaching of critical thinking and evaluative judgement in the age of generative artificial intelligence? ATLAANZ J. 2024;7(1):1-9. Available from: https://www.atlaanzjournal.org/article/ai-critical-thinking.
29. UNESCO. Recommendation on the ethics of artificial intelligence. Paris: United Nations Educational, Scientific and Cultural Organization (UNESCO); 2021. Available from: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.
30. Corrêa NK, Galvão C, Santos JW, Del Pino C, Pinto EP, Barbosa C, et al. Worldwide AI ethics: a review of 200 guidelines and recommendations for AI governance. Patterns. 2023;4(10):100818. https://doi.org/10.1016/j.patter.2023.100818. PMID: 37827204

AUTHORS CONTRIBUTION

Brekhna Jamil *- Concept & Design; Data Acquisition; Data Analysis/Interpretation; Drafting Manuscript; Critical Revision; Supervision; Final Approval*

Nowshad Asim *- Concept & Design; Data Acquisition; Data Analysis/Interpretation; Drafting Manuscript; Critical Revision; Supervision; Final Approval*

The authors accept responsibility for all aspects of the work and will ensure that any concerns regarding the accuracy or integrity of any part are properly investigated and resolved.