Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 3055-3060 | ISSN:2251-6727

Generative AI for OSCE and Workplace-Based Assessment: Mapping Validity, Bias/Fairness, Reliability, and Defensible Governance

Mohammed Moizuddin Khan

College of Medicine, Dar Al Uloom University, Riyadh, Saudi Arabia

(Received: 25 October 2025 Revised: 27 November 2025 Accepted: 16 December 2025)

KEYWORDS

generative AI; large language models; assessment; OSCE; workplace-based assessment; validity; fairness; reliability; governance; PRISMA-ScR.

ABSTRACT:

Background: Generative artificial intelligence (GenAI), including large language models (LLMs), is increasingly used in medical education assessment as item writers, rubric-based scorers of constructed responses, and draft generators for narrative feedback. In high-stakes settings (e.g., OSCE pass/fail and programmatic workplace-based assessment [WBA]), GenAI functions as a measurement intervention and therefore requires defensible evidence for validity, fairness, reliability, and governance. [1,2,11]

Methods: We conducted a PRISMA-ScR-aligned scoping review (2020-present) across major medical education and health databases to map empirical and review evidence on GenAI in assessment, prioritizing OSCE/WBA and high-stakes contexts. Evidence was charted using Kane's argument-based validity framework (scoring, generalization, extrapolation, implications), alongside themes related to bias/fairness, reliability, and governance. [9-11]

Results: Evidence is heterogeneous and concentrated in written assessment (short answers, clinical notes, and post-OSCE documentation) and in automated analysis of narrative WBA feedback. LLM grading can be efficient, but studies show variable agreement with expert human ratings and potential shifts in grade distributions. In early OSCE-related studies, agreement between GenAI outputs and clinician judgments is inconsistent, highlighting threats to the scoring inference and the need for response-process evidence, rater calibration, and drift monitoring. [3,6,16-21]

Conclusion: For clinical skills and high-stakes decisions, GenAI use should be bounded and treated as a regulated assessment intervention. Institutions should require a complete validity argument, explicit fairness audits, program-level reliability evidence, and governance controls (versioning, audit trails, appeals, and change management). We propose a practical DEFEND-VFRG toolkit (Validity, Fairness, Reliability, Governance) for implementation and reporting. [11-15,25]

Introduction

Generative AI is rapidly reshaping assessment ecosystems in health professions education. Learners use LLMs to draft text and explanations, while educators and institutions are experimenting with LLMs for item generation, rubric-guided scoring, and narrative feedback drafting. Because these tools can alter score meaning and downstream decisions, especially in OSCE and WBA, their use should be justified with the same rigor applied to any major change in assessment design. [1,2,11]

LLMs can introduce construct-irrelevant variance through prompt sensitivity, reliance on surface linguistic features, and instability across model updates. These properties can threaten standardization and fairness, particularly in clinical skills assessment where multiple sources of error already exist across cases, raters, and occasions. Program-level reliability approaches (e.g., generalizability theory) and explicit validity arguments are therefore central when GenAI is introduced into OSCE/WBA workflows. [8,11]

This review maps the available 2020-present evidence and highlights what remains missing for defensible use in high-stakes OSCE/WBA. We address four questions: (1) what validity evidence is reported; (2) how bias/fairness is evaluated; (3) what reliability evidence

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 3055-3060 | ISSN:2251-6727

supports decisions; and (4) what governance safeguards are described.

Methods

We followed PRISMA-ScR guidance for scoping reviews, drawing on PRISMA 2020 reporting principles where applicable. [9,10] We included peer-reviewed empirical studies and evidence syntheses published from 2020 onward that examined GenAI/LLM use in assessment workflows (scoring, grading, feedback generation, OSCE/WBA support, and integrity-related applications) in medical education or closely transferable health professions education.

For each included source, we extracted the assessment context, the role of AI, and the study design. Findings were mapped to Kane's validity inferences: scoring (alignment with rubrics and the scoring process), generalization (stability across tasks, prompts, cases, cohorts, and model versions), extrapolation (relationship to clinical competence and external criteria), and implications (consequences, decision accuracy, and fairness). [11] We also charted reported fairness methods, reliability approaches (agreement and multi-facet designs), and governance controls (documentation, privacy/security, auditability, and change management).

Results

Evidence map

Most published work examines (a) LLM scoring of written outputs (short answers, clinical notes, documentation after simulations/OSCE stations) and (b) automated analysis of narrative WBA feedback. Direct GenAI scoring of observed OSCE performance is less developed and is typically exploratory. [3,4,6,16-19]

Validity evidence mapped to Kane's inferences

Scoring evidence is most common and is usually operationalized as agreement or correlation with clinician raters. However, agreement alone is insufficient for high-stakes use; response-process evidence is needed to show how the model applies rubrics, where it fails, and how human oversight corrects errors. Generalization evidence should demonstrate stability across prompts, cases, cohorts, and model versions. [11,16-19,21]

In written tasks such as short answers and post-OSCE documentation, LLM and physician scores can differ in ways that would change classifications and feedback. [16-19] Experimental OSCE-related evaluation studies also report limited agreement, reinforcing that coherent text outputs are not a proxy for a defensible scoring process when observed behaviors and station context are central to the construct. [6,21]

Bias and fairness

Fairness risks are frequently discussed but less often evaluated with explicit subgroup analyses, sensitivity testing, or mitigation plans. A pragmatic approach is to pre-specify key fairness threats and test them locally before any summative use.

Fairness threat	How it may appear	Evaluation / mitigation
Language variety / register	Lower scores for non-native or culturally different phrasing.	Subgroup analyses; rater-blind comparisons; standard prompts; bilingual checks where relevant; accommodations.
Prompt sensitivity	Different prompts yield different grades for the same response.	Locked prompt templates; robustness testing; report variance across perturbations; guardrails.
Domain shift (site / case mix)	A model tuned on one context performs poorly elsewhere.	External validation across sites; local recalibration; retrieval augmentation using local

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 3055-3060 | ISSN:2251-6727

		rubrics.
Construct-irrelevant variance	Scores track writing fluency/verbosity rather than clinical reasoning.	Feature audits; rubric redesign; separate writing vs reasoning constructs; targeted human review.
Access inequity	Unequal access to permitted tools affects preparation and performance.	Clear access policies; disclosure requirements; equitable alternatives; equity impact review.

Table 1. Common fairness threats and pragmatic evaluation/mitigation strategies for local validation.

Reliability as a program-level property

Reliability in OSCE/WBA is multifaceted (learners, cases/stations, raters, and occasions). Agreement between an LLM and a single human rater is not sufficient for high-stakes decisions. Where GenAI is used, reliability should be evaluated at the program level using multi-facet designs and decision studies. [7,8]

Approach	Use case	Key notes
Human-AI agreement (e.g., weighted kappa, ICC)	LLM scoring vs human scoring in written tasks.	Report confidence intervals; test for systematic bias; set acceptable error bounds; include subgroup checks.
Generalizability theory (G-theory)	OSCE/WBA with multiple facets.	Estimate variance components; run decision studies to plan stations/raters and evaluate any AI impact. [8]
Many-facet Rasch / IRT (where appropriate)	Model rater severity and task difficulty when design supports it.	Useful for rater effects; requires adequate data and careful assumptions.
Test-retest and drift monitoring	Model updates or prompt changes over time.	Use anchor sets; pre-specify triggers for re-validation; maintain versioned prompts and model metadata.
Reliability of feedback-quality metrics (e.g., QuAL)	Scaling evaluation of narrative feedback quality.	Some metrics have reliability evidence; downstream summative uses still require a full validity argument. [7]

Table 2. Reliability approaches that can support defensible decisions depending on assessment design.

Governance reporting gaps

Empirical reports often omit implementation-critical details such as model/versioning, prompt templates, data governance, audit logs, and appeal pathways. These controls are essential in high-stakes settings because they enable reproducibility, accountability, and learner protection. Broader health-sector guidance emphasizes transparency, accountability, human oversight, and post-deployment monitoring, which are

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 3055-3060 | ISSN:2251-6727

directly applicable to assessment programs using AI. [25]

Discussion

Across the 2020-present literature, the strongest evidence supports GenAI as an assistive tool for written assessment and feedback workflows when used with explicit human oversight. Evidence for direct OSCE/WBA scoring remains limited and does not currently justify GenAI as a sole or primary rater for high-stakes clinical skills decisions. [1,6,16-21]

Kane's framework clarifies recurring gaps: many studies report scoring agreement but provide limited evidence for generalization (across prompts, cases, and model versions), extrapolation (links to real clinical performance), and implications (classification accuracy, subgroup impact, appeals, and unintended consequences). [11] A defensible local validation plan should pre-specify acceptance criteria, archive prompts and rubrics, document model versions, and monitor drift over time using anchor sets and re-validation triggers. [14,15]

Academic integrity and consequential validity

Because learners can use GenAI to draft responses, assessment programs must define permissible use, disclosure expectations, and consequences. Detection tools have documented limitations and can generate consequential false positives, so redesigning assessment toward observation-based tasks, oral justification, and triangulation across measures is often more defensible than relying on automated detection alone. [22-24]

Operational toolkit: DEFEND-VFRG

We summarize minimum defensibility requirements in four pillars-Validity, Fairness, Reliability, and Governance (VFRG)-scaled to assessment stakes and aligned with emerging professional guidance for responsible AI integration. [12-15]

Stakes	Permitted AI role	Minimum evidence / controls
Low (formative feedback)	Draft feedback; suggestion tool with human editing.	Local calibration; basic fairness check; disclosure; privacy safeguards; logging of use.
Moderate (course grading)	Second-reader scoring with human adjudication.	Defined acceptable error; subgroup checks; locked prompts; versioning; audit trail; appeals; periodic re-validation.
High (OSCE/WBA high-stakes)	Assistive only; no autonomous pass/fail decisions.	Full validity argument (Kane); program-level reliability design; fairness audits; drift monitoring with anchors; governance board; change management; learner protections.

Table 3. Practical evidence thresholds for GenAI use by assessment stakes (VFRG).

High-stakes OSCE/WBA checklist

- State intended use, decision consequences, and what the AI will and will not do (assistive vs decisional).
- Lock and version prompts/rubrics; document model family, version/date, and settings.
- Collect response-process evidence (human review of AI failure modes; rater calibration; rubric adherence checks).
- Demonstrate generalization across stations/cases, cohorts, and repeated runs; re-test after updates using anchor sets.
- Evaluate fairness with subgroup analyses and a mitigation plan; document equity impact.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(6), 3055-3060 | ISSN:2251-6727

- Establish program-level reliability evidence (variance components / decision studies where relevant).
- Maintain auditability (immutable logs of prompts/inputs/outputs and human overrides).
- Ensure transparency and learner protections (appropriate explainability, appeal pathway, remediation).
- Implement data governance (de-identification, retention limits, security review, vendor accountability).
- Define change-management triggers and incident response (re-validation thresholds, rollback plan).

Limitations

The evidence base is evolving quickly as models, products, and institutional policies change. Studies differ in tasks, rubrics, reporting detail, and access to prompts/model settings, limiting comparability and preventing quantitative synthesis. Many reports do not include prompts, model versions, or leakage controls, which constrains reproducibility.

Future research priorities

Priority work includes pre-registered validation studies with explicit acceptance criteria, adequately powered fairness audits, longitudinal drift evaluations, multi-institutional benchmarks for clinical skills scoring, and studies that examine downstream consequences (classification consistency, appeals, learner trust, and equity impacts).

Conclusion

GenAI is already influencing assessment directly (scoring and feedback) and indirectly (learner use). Current evidence supports bounded assistive roles, particularly for written tasks and feedback analytics, but is insufficient for autonomous high-stakes OSCE/WBA scoring. A defensible pathway requires explicit validity arguments, formal fairness audits, program-level reliability designs, and transparent governance with robust learner protections.

References

1. Masters K, MacNeil H, Benjamin J, Carver T, Nemethy K, Valanci-Aroesty S, Taylor DCM, Thoma B, Thesen T. Artificial intelligence in health professions education assessment: AMEE Guide No. 178. Med Teach. 2025 Sep;47(9):1410-1424.

2. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024 Apr;46(4):446-470.

3. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, Gagarin V. Assessing the ability of a large language model to score free-text medical student clinical notes: quantitative study. JMIR Med Educ. 2024;10:e56342.

4. Thomas K, Szalacha L, Hanna K, Anibal J, Petrilli J. Evaluating the effectiveness of ChatGPT versus human proctors in grading medical students' post-OSCE notes. Fam Med. 2025;57(10):727-731.

5. Spadafore M, Yilmaz Y, Rally V, Chan TM, Russell M, Thoma B, et al. Using natural language processing to evaluate the quality of supervisor narrative comments in competency-based medical education. Acad Med. 2024;99(5):534-540.

6. Yokose M, Hirosawa T, Sakamoto T, Kawamura R, Suzuki Y, Harada Y, Shimizu T. The validity of generative artificial intelligence in evaluating medical students in objective structured clinical examination: experimental study. JMIR Form Res. 2025;9:e79465.

7. Choo EK, Woods R, Walker ME, O'Brien JM, Chan TM. The Quality of Assessment for Learning score for evaluating written feedback in anesthesiology postgraduate medical education: a generalizability and decision study. Can Med Educ J. 2023;14(6):78-85.

8. Andersen SAW, Nayahangan LJ, Park YS, Konge L. Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: a systematic review and meta-analysis. Acad Med. 2021 Nov 1;96(11):1609-1619.

9. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018 Oct 2;169(7):467-473.

11. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50(1):1-73.

12. Josiah Macy Jr. Foundation. Conference on artificial intelligence in medical education: proceedings and recommendations. Acad Med. 2025 Sep 1;100(9S Suppl 1):S4-S14.

13. Rodman A, Wachter R, Aagaard E, et al. AI and the future of medical education: recommendations from a Macy Foundation Conference. Association of American Medical Colleges (AAMC). 2025 Jun 12.

14. DiMattia A, Mulrine A, Hoover S, Schmude M. Recommendations and action steps to deploy AI in medical education: a practical guide for responsible integration using the Josiah Macy Jr. Foundation framework. Association of American Medical Colleges (AAMC). 2025.

15. Association of American Medical Colleges (AAMC). AI policy development checklist for medical education. 2025 Jul.

16. Huang TY, Hsieh PH, Chang YC. Performance comparison of junior residents and ChatGPT-4 in providing history taking and patient documentation during OSCE: exploratory study. JMIR Med Educ. 2024;10:e59902. doi:10.2196/59902.

17. Bolgova O, Ganguly P, Ikram MF, Mavrych V. Evaluating large language models as graders of medical short answer questions: a comparative analysis with expert human graders. Med Educ Online. 2025;30(1):2550751. doi:10.1080/10872981.2025.2550751.

18. Morjaria L, Burns L, Bracken K, Levinson A, Ngo Q, Lee M, Sibbald M. Examining the efficacy of ChatGPT in marking short-answer assessments in an undergraduate medical program. Int Med Educ. 2024;3(1):32. doi:10.3390/ime3010004.

19. Luordo D, Torres Arrese M, Tristan Calvo C, Shani KD, Rodriguez Cruz LM, Garcia Sanchez FJ, et al. Application of artificial intelligence as an aid for the correction of the objective structured clinical examination (OSCE). Appl Sci. 2025;15(3):1153. doi:10.3390/app15031153.

20. Kwan BYM, Zhou Z, Rogoza N, Aghaei N, de Vries I, Hanmore T, Zevin B. Leveraging large language models to evaluate the quality of narrative feedback for surgery residents in competency-based medical education. Ann Surg Open. 2025;6(4):e608. doi:10.1097/AS9.0000000000000608.

21. Johnsson V, Sondergaard MB, Kulasegaram K, et al. Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Med Educ. 2024;58(1):105-117. doi:10.1111/medu.15190.

22. Pearce J, Chiavaroli N. Rethinking assessment in response to generative artificial intelligence. Med Educ. 2023;57(10):889-891. doi:10.1111/medu.15092.

23. Dalalah D, Aldalalah OMA. The false positives and false negatives of generative AI detection tools in education and academic research: the case of ChatGPT. Int J Manag Educ. 2023;21(2):100822. doi:10.1016/j.ijme.2023.100822.

24. Ardito CG. Generative AI detection in higher education assessments. New Dir Teach Learn. 2024:1-18. doi:10.1002/tl.20624.

25. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021.