Research Article

Artificial Intelligence for Clinical Competence Assessment Compared with Human Examiners: A Systematic Review and Exploratory Meta-Analysis of Agreement, Reliability, And Validity

Abdulmohsen M. Alomair

Medical Education Department, College of Medicine, King Faisal University, Al Ahsa, Saudi Arabia. Email: amaalomair@kfu.edu.sa

Abstract

Artificial intelligence (AI) is increasingly being used to assess clinical competence, yet its measurement performance relative to human examiners remains uncertain across professions, tasks, and assessment formats. This systematic review and exploratory meta-analysis evaluated agreement, association, reliability, validity, score differences, classification performance, and human-human benchmarking for AI-based assessment of human clinical competence. Searches identified 29 eligible reports representing 27 independent studies across medicine, dentistry, nursing, pharmacy, and paramedicine. Owing to substantial heterogeneity in measurement methods, most outcomes were synthesized narratively. Two independent studies reported Spearman correlations and were included in an exploratory random-effects meta-analysis despite differing substantially in clinical assessment context using Fisher z transformation and restricted maximum likelihood estimation. The pooled AI-human association was ρ = 0.82 (95% CI, 0.55-0.93). The two correlations were numerically similar (I^2 = 0%), but between-study heterogeneity could not be estimated reliably because only two studies were available. This estimate should be interpreted cautiously as a cross-context summary of association rather than a common clinical effect. Direct agreement varied markedly across studies, and favorable correlations or reliability estimates sometimes coexisted with systematic score differences or disagreement near decision thresholds. Performance appeared strongest in structured, rubric-based tasks and more variable for communication, non-technical skills, and context-dependent judgments. Human-reference quality, input equivalence, model choice, and validation design also influenced interpretability across studies. The available measurement evidence is most compatible with human-supervised formative, supplementary, second-rater, or quality-assurance applications; autonomous high-stakes use requires stronger external validation, reproducibility, and decision-level safety evidence.

Keywords: Artificial intelligence; Clinical competence; Assessment; Reliability and agreement; Medical education.

1. Introduction

Clinical competence integrates knowledge, clinical reasoning, technical and procedural skills, communication, professionalism, and decision-making in patient care (Epstein & Hundert, 2002). Its assessment informs feedback, progression, certification, and readiness for independent practice. Contemporary competency-based and programmatic assessment approaches therefore emphasize repeated observation of performance across tasks and contexts rather than reliance on isolated examinations (van der Vleuten & Schuwirth, 2005; van der Vleuten et al., 2012; Van Melle et al., 2019). Human examiners remain central because expert judgment can integrate contextual and qualitative features of performance, but rater-mediated assessment is also affected by variability in attention, standards, frames of reference, and interpretation of competence (Gingerich et al., 2014; Prentice et al., 2020). This creates both a psychometric and operational challenge: assessment systems require sufficient sampling and expert judgment while operating within finite faculty resources.

Artificial intelligence offers a potential means of increasing scoring capacity and supporting scalable assessment. Earlier automated systems used machine learning, natural-language processing, sensor data, or structured performance indicators; more recent deep-learning, computer-vision, large-language-model, generative-AI, and multimodal systems can evaluate written clinical reasoning, transcripts, speech, images, video, movement, and combinations of these inputs (Lucas et al., 2024; Shaw et al., 2025; Tozsin et al., 2024). Primary validation studies now compare AI-generated assessments directly with trained human examiners in OSCEs, clinical documentation, simulated encounters, procedural performance, and rubric-based reasoning tasks. Reported performance, however, varies by model, task, competence domain, input modality, scoring framework, and validation design, and favorable

association can coexist with systematic differences in absolute scores (Campbell et al., 2025; Luordo et al., 2025; Quah et al., 2024; Tekin et al., 2025; Yokose et al., 2025b).

Agreement, reliability, association, and validity are related but distinct measurement properties. Agreement concerns whether AI and human examiners produce sufficiently similar measurements of the same performance; reliability concerns reproducibility or consistency; and validity concerns the evidence supporting interpretation and intended use of the resulting scores (Mokkink et al., 2010; Cook & Hatala, 2016). Pearson and Spearman correlations quantify association rather than absolute agreement: two assessors can rank candidates similarly while one is systematically more lenient or severe (Bland & Altman, 1986). Human ratings are also imperfect, so human-human reliability can provide an important contextual benchmark rather than assuming that a single examiner is an error-free gold standard.

Existing reviews describe broad applications of AI in health professions education and automated assessment of procedural skills, but the focused question of how well AI-generated assessments of human clinical competence correspond with human examiner judgments remains difficult to answer because studies report non-equivalent statistics, including ICCs, correlations, kappa coefficients, Gwet coefficients, percentage agreement, paired score differences, Bland-Altman analyses, and classification measures (Feigerlova et al., 2025; Pedrett et al., 2023; Shaw et al., 2025). The present systematic review and exploratory meta-analysis therefore evaluated AI-human association and agreement, reliability and validity evidence, systematic score or classification disagreement, factors associated with variation in measurement performance, and human-human benchmarks where directly comparable. Compatible effect measures were pooled only when sufficiently similar; non-equivalent measurement properties were synthesized separately.

2. Methods

2.1 Reporting Framework and Eligibility

This systematic review and exploratory meta-analysis was reported in accordance with PRISMA 2020 and the literature search was additionally guided by PRISMA-S (Page et al., 2021; Rethlefsen et al., 2021). Terminology and interpretation of agreement and reliability were informed by GRRAS and relevant COSMIN principles (Kottner et al., 2011; Mokkink et al., 2010). The review was not prospectively registered. Eligibility criteria, distinctions among measurement properties, and the decision not to combine conceptually different statistics into a single measure of 'AI accuracy' were established before the final quantitative synthesis.

Eligible studies evaluated students, trainees, or practicing health professionals undergoing assessment of clinically relevant competence or performance. Settings included OSCEs, simulation, procedural or technical skills, history taking, communication, clinical reasoning, documentation, workplace-oriented assessment, and related structured evaluations. Studies limited to factual knowledge or general academic performance were excluded, as were studies in which AI itself was the examination candidate.

The index assessment was an AI system that independently generated a quantitative score, ordinal rating, categorical judgment, competence classification, or other evaluative output from human clinical performance. Eligible technologies included conventional machine learning, natural-language processing, neural networks, deep learning, computer vision, speech-processing systems, large language models, generative AI, and multimodal AI. The reference assessment required at least one identifiable human examiner, trained assessor, standardized patient, clinical educator, practicing clinician, or expert panel evaluating the same underlying performance or response. Differences in information available to AI and human assessors were extracted because they could materially influence agreement.

Eligible outcomes included agreement or concordance (e.g., ICC, kappa, Gwet coefficients, percentage agreement, paired score differences, Bland-Altman results), association (Pearson or Spearman correlation), reproducibility or reliability, validity evidence, categorical or classification performance, and human-human agreement when reported alongside AI-human comparisons. Reviews, editorials, commentaries, protocols, studies without an eligible human comparator, AI used only for teaching or feedback, non-clinical academic outcomes, synthetic-only candidate performances, and reports without sufficient quantitative information were excluded. Potentially overlapping reports were linked, and duplicate participant data were not counted more than once. For the present evidence

window, eligibility was restricted to reports published or publicly available on or before 30 April 2026; reports first available from 1 May 2026 onward were excluded.

2.2 Information Sources, Selection, and Data Extraction

A comprehensive search of MEDLINE via PubMed, Embase, Scopus, Web of Science Core Collection, ERIC, and IEEE Xplore was conducted. Search concepts covered artificial intelligence, clinical competence/performance, and assessment/scoring, using controlled vocabulary where available and title, abstract, and keyword terms. Searches were supplemented by backward and forward citation searching, preprint repositories, and publication-history searches to identify eligible reports and potentially overlapping publications. No language restriction was applied during the electronic search. Full database-specific strategies are provided in Supplementary Appendix S1. Records were imported into EndNote and duplicate records were identified and removed using EndNote's duplicate-detection function. The exact database-specific retrieval counts and individual database search dates were not retained in the review records and therefore could not be reconstructed reliably.

Title and abstract screening was performed by one reviewer. Full-text eligibility assessment and data extraction were subsequently independently checked by a second reviewer. The two reviewers compared their assessments and extracted data, and no disagreements requiring resolution occurred.

Records were assessed against the prespecified eligibility criteria at title/abstract and full-text stages. Particular attention was given to separating studies in which AI assessed human clinical performance from the much larger literature in which AI systems completed medical examinations or diagnostic tasks. A structured extraction framework recorded study and participant characteristics, assessment context, AI system and input modality, human reference assessment, validation design, model configuration where available, scoring framework, and all eligible measurement outcomes. Independent test or validation results were preferred over development-set performance; multiple tasks or AI systems from the same cohort were retained for narrative interpretation but were not treated as independent studies in meta-analysis.

2.3 Outcome Classification and Methodological Appraisal

Absolute agreement was distinguished from consistency and association. Absolute-agreement ICCs, paired differences, Bland-Altman analyses, and appropriate agreement coefficients were interpreted as evidence concerning similarity of measurements; consistency ICCs were interpreted separately; Pearson and Spearman coefficients were treated only as measures of association. Reliability referred to stability of AI-generated assessment when the underlying performance was unchanged, and validity evidence was interpreted according to the inference supported rather than solely by terminology used by study authors. Human assessment was treated as a reference assessment rather than an error-free gold standard.

Methodological appraisal used an adapted domain-based framework tailored to the measurement question; it was not a standard COSMIN or QUADAS-2 assessment. COSMIN principles informed appraisal of reliability and measurement-error evidence, while QUADAS-2 principles informed consideration of categorical or classification performance (Mokkink et al., 2010; Whiting et al., 2011). Domain-level judgments were summarized as favorable, some concern/limitation, or unclear/not reported. A favorable judgment indicated that the available study reporting supported the domain without an identified material limitation; some concern/limitation indicated that the domain was only partially addressed or that an identifiable limitation affected interpretation; and unclear/not reported indicated insufficient information to support a definite judgment. These categories were used as review-specific interpretive judgments rather than as a validated numerical quality or risk-of-bias score. AI-specific features included development-evaluation separation, potential training-test leakage, external validation, adequacy and independence of the human reference assessment, assessor blinding, number of human raters, AI-human information equivalence, model/version identification, prompt or configuration reproducibility, treatment of repeated stochastic outputs, missing performances, and selective reporting. Detailed study-level characteristics and appraisal are provided in Supplementary Tables S1-S2 and Supplementary Figure S1.

The methodological appraisal of the included studies was independently checked by a second reviewer. No disagreements in appraisal judgments occurred.

2.4 Statistical Synthesis

Statistics representing different measurement properties were not collapsed into a single pooled measure. ICCs, kappa statistics, percentage agreement, paired differences, Bland-Altman results, classification measures, and validity evidence were synthesized descriptively when statistical formulation, scoring structure, or clinical construct differed materially. Quantitative synthesis was considered when at least two independent studies reported a sufficiently comparable effect measure. Because only two studies contributed to the Spearman synthesis, the pooled estimate was treated as exploratory and interpreted cautiously. Representative study-level findings were tabulated to illustrate the range of measurement properties, assessment contexts, and direction of AI–human correspondence; this illustrative table was not intended as an exhaustive outcome table.

For narrative synthesis, findings were first organized according to the measurement property assessed and were then descriptively compared across study and assessment characteristics. Assessment contexts were grouped as procedural/technical, clinical reasoning/written, OSCE-based or closely structured, and history-taking/communication/non-technical simulation. Recurring patterns in AI–human correspondence were identified by comparing findings across these groups together with AI architecture, information equivalence, scoring structure, and human-reference characteristics. These comparisons were descriptive and were not treated as formal subgroup or moderator analyses.

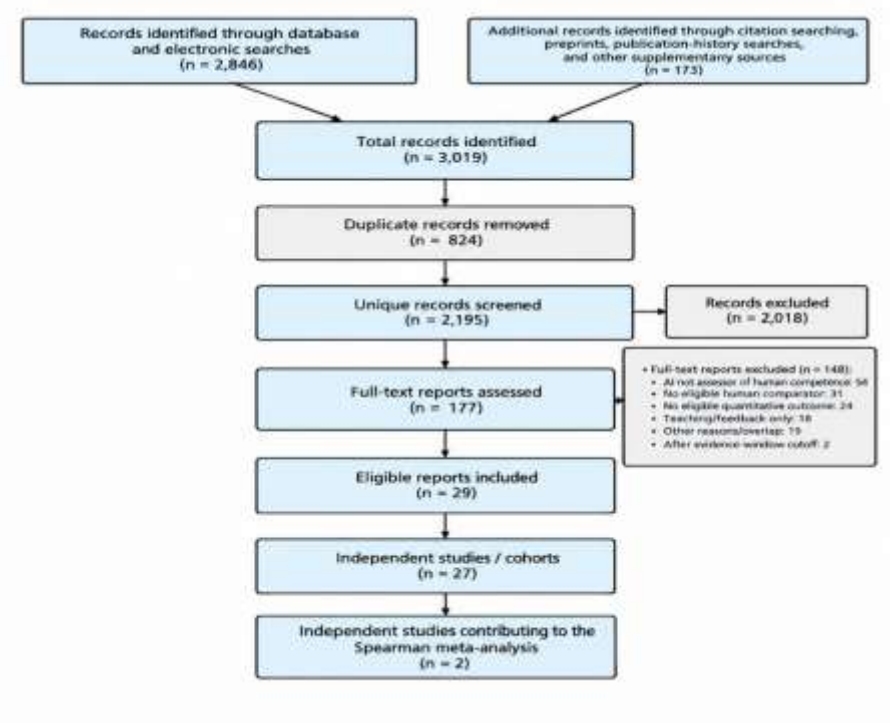
Two independent studies reported overall Spearman rank correlations between AI-generated and human-reference scores. Although the same dimensionless association metric was reported, the studies evaluated clinically distinct assessment contexts—laparoscopic surgical-skill performance and post-OSCE clinical-note grading. The quantitative synthesis was therefore undertaken as an explicitly exploratory analysis intended to summarize the direction and approximate magnitude of AI–human score association rather than to estimate a common clinical effect across equivalent constructs. Each coefficient was Fisher-z transformed and its sampling variance estimated using the Bonett-Wright approximation (Bonett & Wright, 2000). A random-effects model was fitted using restricted maximum likelihood (REML) for between-study variance. Because only two studies were available, confidence intervals were calculated using the modified Knapp–Hartung procedure with the variance safeguard described by Röver et al. (2015). For Fisher-transformed correlations y_i, random-effects weights were $w_i = 1/(v_i + \hat{\tau}^2)$ where $v_i = (1 + \rho_i^2/2)/(n_i - 3)$ was the Bonett–Wright sampling variance. The Hartung–Knapp variance factor was $q = \sum w_i (y_i - \hat{\mu})^2/(k - 1)$ with the safeguard $q^* = \max(1, q)$. The variance of the pooled estimate was therefore $q^*/\sum w_i$, and the 95% confidence interval used the student- t distribution with $k - 1$ degrees of freedom before back-transformation to Spearman's ρ. With two studies, the analysis used 1 degree of freedom. Heterogeneity was described using Cochran's Q, τ^2, and I^2, although these statistics were interpreted cautiously because only two studies were available. Meta-regression and formal tests of funnel-plot asymmetry were not performed. Analyses were conducted in Python 3.13.5 using NumPy 2.3.5 and SciPy 1.17.0.

3. Results

3.1 Study Selection and Evidence Base

The search identified 2,846 records through database and other electronic searches and 173 additional records through citation searching, preprint repositories, publication-history searches, and other supplementary sources. After removal of 824 duplicates, 2,195 unique records were screened; 177 reports underwent full-text assessment and 148 were excluded. Twenty-nine reports representing 27 independent studies or participant/performance cohorts met the eligibility criteria. Where preprint, peer-reviewed, or other companion publications represented the same underlying participant/performance cohort, they were linked and duplicate participant data or effect estimates were not counted as independent evidence. All 27 independent studies contributed quantitative evidence to at least one measurement property, and two contributed to the Spearman meta-analysis.

Figure 1. PRISMA flow diagram of study identification, screening, eligibility assessment, and inclusion.



3.2 Characteristics of Included Studies

Medicine accounted for 22 of 27 studies (81.5%), followed by dentistry (n = 2), nursing (n = 1), pharmacy (n = 1), and allied health/paramedicine (n = 1). Seventeen studies (63.0%) involved student-only populations. Assessment contexts included procedural/technical performance, written clinical reasoning or documentation, OSCE-based assessment, and history-taking, communication, or other non-technical simulated performance. AI approaches spanned conventional machine learning and NLP, deep learning and computer vision, and more recent LLM/generative and multimodal systems. Human reference assessments ranged from single clinicians or standardized patients to multiple trained examiners and expert panels.

Table 1. Aggregate characteristics of the 27 independent studies.

Characteristic	Studies, n (%)
Total independent studies	27 (100)
Medicine	22 (81.5)
Dentistry	2 (7.4)
Nursing	1 (3.7)
Pharmacy	1 (3.7)
Allied health / paramedicine	1 (3.7)
Student-only populations	17 (63.0)
Practicing professionals only	3 (11.1)
Mixed training/professional levels	5 (18.5)
Examination/licensure or unclear level	2 (7.4)
Procedural/technical assessment	8 (29.6)
Clinical reasoning/written assessment	8 (29.6)
OSCE-based/closely structured assessment	6 (22.2)
History-taking/communication/non-technical simulation	5 (18.5)

Detailed study-level characteristics are provided in Supplementary Table S1. Methodological quality varied: stronger studies used multiple trained raters, calibrated instruments, blinding, independent test sets, repeated scoring, or generalizability analyses. Common limitations were small or single-institution samples, internal rather than external validation, incomplete development-evaluation separation, limited model/configuration reporting, weak human reference standards, and mismatch in information available to AI and human assessors. Generative-AI reproducibility was often incompletely reported.

Across the 27 studies, assessor blinding with respect to the comparative AI–human assessment was unclear or not reported in all 27 (100%), and none was classified as providing independent external validation in a separate population or institution. Fourteen studies (51.9%) clearly described multiple trained, expert, or independent human reference raters or an expert panel. AI and human assessors received materially equivalent or substantially comparable information in 18 studies (66.7%), whereas 9 (33.3%) had partial equivalence or a clear information mismatch. Only one study (3.7%) explicitly evaluated repeated-run stability of AI-generated scoring.

3.3 AI-Human Agreement, Association, and Score Differences

Direct agreement varied markedly and could not be represented by one pooled coefficient because studies used different measurement models and units. Several structured assessments reported high correspondence: Latifi et al. (2016) reported 95.4% agreement between automated and physician scoring across 8,007 clinical decision-making responses; Holderried et al. (2024) reported Cohen's κ = 0.832; Liu et al. (2025) reported human-AI consistency ICCs above 0.923 across three history-taking cases. In contrast, Yokose et al. (2025b) and Mitsuhashi et al. (2025) reported poor agreement across several OSCE and non-technical-skill domains.

Table 2. Representative quantitative findings illustrating AI–human measurement performance across assessment contexts and measurement properties.

Study	Assessment context	Principal metric	Main finding
Latifi et al. (2016)	Clinical decision-making responses	Exact agreement	AI-human agreement 95.4%
Holderried et al. (2024)	Simulated history taking	Cohen's κ	κ = .832
Quah et al. (2024)	Dental clinical essays	ICC / Pearson correlation	All-rater ICC = .858 and .794; AI-manual r = .829 and .599
Luordo et al. (2025)	OSCE clinical reports	Multi-rater ICC / score difference	Single-measure ICC = .77; average-measure ICC = .91; AI mean 3.51 points lower than expert mean
Campbell et al. (2025)	Clerkship OSCE notes	Spearman / item agreement	ρ = .82; item-level agreement 83.2%
Liu et al. (2025)	History-taking assessment	ICC / item consistency	ICCs > .923; item consistency >95%
Tekin et al. (2025)	Four OSCE skills	Score difference / Bland-Altman	AI generally scored higher; largest difference for knot tying
Yokose et al. (2025b)	Medical-interview OSCE	ICC / paired scores	Poor agreement; history-taking ICC = .36
Mitsuhashi et al. (2025)	Paramedic non-technical skills	Weighted Gwet AC2	Approximately -.15 to .30 across domains
Hassanein et al. (2026)	Dental clinical responses	ICC	DeepSeek-3 = .87; ChatGPT-4 = .64; human-human = .84
Chen (2026)	Nursing empathic communication	ICC	AI-educator ICC = .81

Note. Findings were included in this table for illustrative presentation to represent the range of assessment contexts and measurement properties reported across the included studies, including agreement, association, reliability, score differences, and classification-related performance, as well as examples of both stronger and weaker AI–human correspondence. The table is not intended to provide an exhaustive listing of all study-level outcomes; full study characteristics and eligible outcome types are reported in Supplementary Table S1.

Systematic differences in absolute scoring were observed despite favorable association or reliability statistics. Luordo et al. (2025) reported a mean GPT-4 score 3.51 points below expert graders, whereas Tekin et al. (2025), Yokose et al. (2025b), and Mitsuhashi et al. (2025) identified AI leniency in several contexts. A pooled mean difference was not calculated because scoring scales, constructs, ranges, and units were not equivalent.

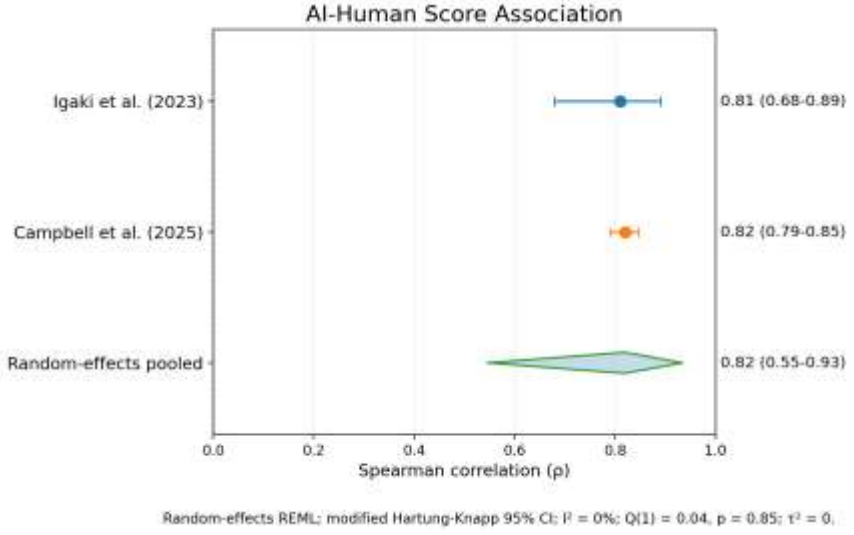
3.4 Meta-analysis of AI-Human Spearman Score Association

Two independent studies reported overall Spearman correlations suitable for an exploratory quantitative synthesis, although the assessment constructs differed substantially. Igaki et al. (2023) reported ρ = 0.81 in 60 independent validation videos of laparoscopic performance, and Campbell et al. (2025) reported ρ = 0.82 across 684 medical students' post-encounter notes. Random-effects pooling yielded a summary Spearman association of ρ = 0.82 (95% CI, 0.55-0.93). The two correlations were numerically similar (I^2 = 0%; Q (1) = 0.04, p = .85; τ^2 = 0), but between-study heterogeneity could not be estimated reliably because only two studies were available. These heterogeneity estimates should be interpreted cautiously because the synthesis contained only two studies. The analysis quantifies score association rather than absolute agreement. Accordingly, the pooled estimate should be interpreted as an exploratory cross-context summary of score association rather than as evidence of a common underlying effect across clinically equivalent assessment tasks.

Table 3. Random-effects meta-analysis of AI-human Spearman score association.

Study	Assessment	n	Spearman ρ
Igaki et al. (2023)	Laparoscopic surgical-skill assessment	60	0.81
Campbell et al. (2025)	Clerkship OSCE post-encounter notes	684	0.82
Random-effects pooled estimate	2 studies	-	0.82 (95% CI, 0.55-0.93)

Figure 2. Forest plot of Spearman correlations between AI-generated and human-reference scores.



3.5 Reliability, Validity, Classification, and Human-Human Benchmarking

Evidence on repeated-run reliability was limited but encouraging in selected studies. Liu et al. (2025) reported total-score coefficients of variation of approximately 0.87%-1.12%, ICCs above 0.923, and item-level consistency above 95% across repeated history-taking evaluations. Rajan et al. (2026) reported grading-precision ICC = 0.993 across

1,450 clinical-reasoning responses, although equivalence with faculty scoring varied by cognitive category and difficulty. Quah et al. (2024) reported high combined human/automated reliability but different AI-human correlations and absolute score similarity across two questions.

Validity evidence was strongest when the target competence was clearly defined and represented in the AI input. Cianciolo et al. (2021), Johnsson et al. (2024), and Kasa et al. (2022a) reported evidence supporting relationships with expert assessment or expected differences in competence, whereas more complex communication, non-technical, and context-dependent domains were less consistently represented. Classification findings were likewise heterogeneous: some studies reported high accuracy for structured tasks, but decision-level disagreement remained possible even when continuous-score agreement was favorable. Human-human benchmarks also varied, emphasizing that individual human ratings are not error-free; in several studies, AI-human performance approached human-human reliability, whereas in others it remained clearly weaker.

3.6 Sources of Heterogeneity and Overall Synthesis

Variation in AI–human measurement performance appeared related to the AI model or architecture, competence domain, information available to each assessor, scoring structure, decision threshold, task difficulty, and quality of the human reference assessment. Within-study comparisons showed that different models could produce materially different results on the same performances. In the grouped narrative synthesis, structured written, rubric-based, and procedural tasks tended to show the most convincing correspondence, whereas communication, non-technical skills, contextual judgment, and domains requiring nonverbal or visual information were more variable. These cross-study patterns were descriptive and were not derived from formal subgroup or moderator analyses. Studies in which humans received richer information than AI were difficult to interpret as pure comparisons of assessor quality, because disagreement in such settings may reflect construct underrepresentation in the AI input rather than solely poorer AI judgment.

The pooled Spearman analysis included only two studies and did not support meaningful meta-regression or formal publication-bias testing. Across the full evidence base, strong association or reliability did not necessarily imply absence of score bias or safe classification at consequential thresholds. Overall, the available measurement evidence is most compatible with human-supervised supplementary, formative, second-rater, or quality-assurance applications rather than autonomous replacement of trained human examiners in high-stakes clinical competence assessment.

4. Discussion

4.1 Principal Findings and Interpretation

This review found promising but highly context-dependent correspondence between AI-based assessments of clinical competence and human examiner judgments. The exploratory pooled Spearman association from two studies reporting the same association metric but evaluating substantially different clinical assessment contexts was strong ($\rho = 0.82$), although the small number of studies substantially limits the precision and generalizability of the quantitative synthesis. Although the two correlations were numerically similar and I^2 was 0%, heterogeneity cannot be estimated reliably from only two studies. The quantitative result therefore suggests that AI can preserve the ordering of performance observed by human assessors in selected contexts but should not be interpreted as a universal estimate of AI assessment performance.

The distinction between association and agreement is central. Two assessors may rank candidates similarly while one is consistently more lenient or severe; a high correlation can therefore coexist with materially different absolute scores and different progression, remediation, honors, or pass/fail decisions (Bland & Altman, 1986). This pattern was evident across the included studies: favorable ICCs or correlations sometimes accompanied systematic score differences or broad individual-level limits of agreement. For high-stakes applications, appropriately specified agreement statistics, paired-difference analyses, Bland-Altman methods, and decision-level error are more informative than correlation alone (Koo & Li, 2016).

4.2 Why AI Assessment Performance Varies

AI systems should not be treated as a homogeneous class of assessor. Within the same task, different models and configurations sometimes produced substantially different results; Hassanein et al. (2026), for example, reported stronger agreement for DeepSeek-3 than ChatGPT-4 using the same responses and rubric. Validation evidence for one architecture or model version therefore cannot automatically be transferred to another.

Methodological considerations also differed by AI technology type. Conventional automated-scoring and classical machine-learning systems were mainly limited by development-dataset dependence and external generalizability. Computer-vision and deep-learning systems additionally depended on task definition and recording conditions. LLM and generative-AI systems introduced concerns about model-version dependence, prompt sensitivity, stochastic outputs, and model updates, while multimodal systems added complexity related to modality integration and AI–human information equivalence.

The nature of the competence and the information available to the assessor were equally important. Procedural tasks, structured written responses, checklist-based assessments, and explicit information-gathering behaviors provide bounded targets for automated evaluation. Communication quality, non-technical performance, contextual reasoning, and holistic judgment may depend on tone, timing, gestures, physical technique, and other information that is not fully represented in text or limited inputs. When human examiners view a complete performance while AI receives only a transcript or note, disagreement may reflect construct underrepresentation in the AI input as well as differences in judgment. Accordingly, AI–human agreement is most interpretable when both assessors have access to materially comparable representations of the performance being judged.

Scoring structure and decision thresholds also affected apparent performance. Binary classifications can show higher apparent concordance than multi-category ratings, and strong continuous-score agreement may still permit disagreement around pass/fail or other consequential boundaries. Validation should therefore be tied to the intended use of the score rather than to an arbitrary coefficient threshold.

4.3 Human Examiners as the Reference Standard

Human assessment should be treated as a reference rather than an error-free gold standard. Assessor cognition research shows that rater variation may reflect both unwanted error and legitimate expert interpretation (Gingerich et al., 2014; Prentice et al., 2020). Human-human benchmarking was therefore an important feature of this review. Some AI systems approached human inter-rater performance in structured contexts, but human reference quality varied across studies: a single uncalibrated assessor provides a weaker criterion than multiple trained raters or expert consensus. Future validation studies should report rater expertise, training, blinding, aggregation procedures, and human-human reliability.

4.4 Implications for Clinical Competence Assessment

The available measurement evidence is most compatible with human-supervised formative, supplementary, second-rater, or quality-assurance applications. These applications may offer opportunities to reduce scoring burden and increase feedback capacity, although evidence for actual workload reduction and educational outcomes remains limited. The evidence is not yet sufficient to support autonomous AI use across high-stakes progression, certification, or licensure decisions. Before such implementation, systems require prospective external validation, explicit evaluation of absolute agreement and decision-level error, repeated-run reliability testing, robust multi-rater human reference standards, transparent model and prompt reporting, and evidence of acceptable performance across relevant populations and settings.

4.5 Strengths, Limitations, and Research Priorities

This review focused specifically on AI used to assess human clinical competence and excluded the conceptually different literature in which AI systems themselves answer medical examination questions. It also separated association, agreement, score bias, reproducibility, validity, classification performance, and human-human reliability rather than treating them as interchangeable forms of accuracy. Incompatible ICC formulations, classification

metrics, and score differences were not pooled simply to increase sample size, and overlapping publications were explicitly linked so that 29 reports represented 27 independent cohorts.

Several limitations remain. The exploratory meta-analysis included only two independent studies; consequently, the pooled estimate and heterogeneity statistics should be interpreted cautiously and cannot provide a stable estimate across clinical-assessment settings. Most other outcomes required structured narrative synthesis because measurement definitions, scales, constructs, and designs differed. The review was not prospectively registered. The evidence base is also rapidly evolving, and performance reported for commercial AI models may change after model updates. Generalizability is also limited because 22 of 27 studies were conducted in medicine and 17 involved student-only populations, leaving comparatively limited evidence from other health professions and practicing clinicians. Methodological appraisal was limited by incomplete reporting of model configuration, reference-rater procedures, development-evaluation separation, and repeated-run methods. Human-human benchmarking was available inconsistently. Search-reporting reproducibility was additionally limited because database-specific retrieval counts and exact search dates were not retained.

Future studies should prioritize prospective external validation with the AI system, prompt, rubric, and decision thresholds fixed before evaluation. Repeated-run reliability and model-version reporting are particularly important for generative AI. Statistical methods should match the intended claim: correlation for association, appropriate ICC or agreement measures for interchangeability, and paired-difference or Bland-Altman analysis where individual disagreement matters. Validation should extend beyond agreement to known-groups evidence, generalizability, consequences, fairness across candidate groups, and clinically meaningful decision-level error. Acceptable measurement error should be defined in relation to intended use, because an error that is tolerable in formative feedback may be unacceptable at a licensing boundary.

Conclusion

Artificial intelligence-based assessment of clinical competence demonstrated promising but highly context-dependent correspondence with human examiner judgment. The exploratory meta-analysis of two studies reporting the same association metric but evaluating different clinical assessment contexts showed a strong positive association between AI-generated and human-reference scores (pooled Spearman ρ = 0.82; 95% CI, 0.55-0.93). However, the very small number of pooled studies limits generalizability, and the result should not be interpreted as evidence of universal interchangeability. Evidence outside medicine and among practicing clinicians remains limited. Stronger performance was most consistently observed in structured, rubric-based, written, and procedural assessment tasks, whereas communication, non-technical skills, contextual judgment, and other complex domains were more variable. Favorable correlations or reliability estimates could coexist with systematic differences in absolute scores or disagreement near consequential decision thresholds. The available measurement evidence is therefore most compatible with human-supervised formative, supplementary, second-rater, or quality-assurance applications. Evidence remains limited regarding educational outcomes, patient safety, workload reduction, fairness, and downstream consequences. Autonomous use for high-stakes progression, certification, or licensure decisions requires stronger external validation, reproducibility, human-reference standards, and decision-level safety evidence.

Supplementary material: Full study-level characteristics, detailed methodological appraisal, and database-specific search strategies are provided separately as Supplementary Tables S1-S2, Supplementary Figure S1, and Supplementary Appendix S1.

Supplementary Table S1. Characteristics of the 27 independent studies included in the systematic review

Study	Country	Sample / performance units	Population	Assessment context / competence domain	AI system and input	Human reference assessment	Validation design	Outcomes relevant to review

Swygert et al. (2003)	United States	Human-rated patient notes from 4 Clinical Skills Examination cases	Medical examinees	Clinical skills examination; written patient notes / clinical documentation	Latent semantic analysis–based automated scoring; text input	Expert human ratings of the same patient notes	Comparative automated-scoring validation	Correlation/association between automated and human scores; score comparability
Latifi et al. (2016)	Canada	8,007 English- and French-language responses to 6 clinical decision-making questions	Medical licensure candidates	High-stakes clinical competency examination; clinical decision-making	Supervised machine-learning automated scoring; written text	Expert physician markers	Operational validation using Medical Council of Canada examination data	Human–AI agreement (95.4%); human inter-rater benchmark; classification/scoring accuracy
Brown et al. (2017)	United States	38 participants; 3 robotic peg-transfer trials per participant	Surgical trainees / surgeons	Robotic surgery simulation; technical skill	Machine learning using contact forces, robotic-arm acceleration and task-time data	3 expert raters using GEARS	Internal predictive validation	AI–human skill-rating correspondence; prediction/classification performance
Azari et al. (2019)	United States	103 video clips from 9 surgeons (6 attending, 3 resident) across 16 operations	Resident and attending surgeons	Open surgical tying and suturing; technical skill	Computer vision and markerless hand-motion/kinematic analysis	3 expert surgeons; consensus ratings of motion economy, fluidity and tissue handling	Proof-of-concept predictive validation	AI prediction of expert ratings; association; human-expert consensus
Baghdadi et al. (2019)	United States	20 pelvic lymph-node-dissection procedures; 14 development and	Practicing robotic surgeons	Robot-assisted pelvic lymph-node dissection; operative technical	Computer vision applied to console-feed operative video	Expert surgeon panel using PLACE scores	Development with holdout validation plus separate test	Classification accuracy against expert assessment (83.3% in test set)

		6 independent test procedures		performance			procedures	
Maicher et al. (2019)	United States	102 virtual-patient encounters; 20 randomly selected dialogs used for detailed validation	Medical students	Virtual standardized-patient history taking; information-gathering skill	NLP-enabled virtual standardized patient; spoken/textual dialogue	3 trained human raters	Direct agreement validation	Computer accuracy 87% versus approximately 90% for human raters; inter-rater reliability across skill categories
Cianciolo et al. (2021)	United States	414 third-year medical students; 700 diagnostic-justification essays	Medical students	Summative clinical competency examination; written clinical reasoning	NLP/machine scoring of diagnostic-justification essays	Faculty ratings	Internal validity study	Pearson correlations with faculty scores; construct/criterion-related validity; faculty–faculty benchmark
Lavanchy et al. (2021)	Switzerland	242 laparoscopic cholecystectomy videos; 949 clip-application segments used for skill modelling	Surgeons	Laparoscopic cholecystectomy; operative technical skill	Three-stage ML: CNN instrument detection, motion-feature extraction and regression	4 board-certified surgeons	Development and internal validation/test procedure	AI prediction of surgical skill; classification performance; human–human ICC benchmark
Kasa et al. (2022)	Canada	72 surgical trainees and faculty	Surgical trainees and surgeons	Simulated surgical knot tying; technical skill	ResNet image model, ResNet-LSTM kinematic model and multimodal deep-learning model	3 expert raters using OSATS Global Rating Scale	Internal model development and evaluation	AI–human ICC/MSE; human–human ICC; comparative reliability

Igaki et al. (2023)	Japan	650 source intraoperative videos; 60 model-construction and 60 independent validation videos	Practicing surgeons	Laparoscopic colorectal surgery; standardized surgical-field development / technical skill	Deep-learning surgical-field recognition producing AI Confidence Score	Human ESSQS expert skill scores	Retrospective development and validation	Spearman correlation (ρ≈0.81); ROC/classification of high- and low-skill performance
Burke et al. (2024)	United States	168 first-year medical students; 14,280 rubric scoring opportunities	Medical students	Standardized-patient encounter; history and physical note quality	ChatGPT-3.5; free-text clinical notes	Standardized patients using the same 85-element rubric	Retrospective direct comparison	AI-versus-human scoring accuracy/error; score agreement across note components
Holderried et al. (2024)	Germany	106 history-taking conversations; 1,894 question–answer pairs	Medical students	Simulated-patient history taking and communication	GPT-4-powered simulated patient and automated feedback system	Human rater evaluation of the same interactions	Direct agreement validation	Cohen's κ / AI–human agreement; automated scoring reliability
Johnsson et al. (2024)	Denmark	45 participants: 22 novices, 12 intermediates and 11 experts; 2 simulated procedures per participant	Medical learners/trainees and clinical experts	Simulated chorionic-villus sampling; complex procedural clinical skill	CNN-based video features combined with motion tracking and eye-movement information	Trained clinical experts using consensus-developed assessment instrument	Prospective comparative validity study	Validity evidence across Kane framework; reproducibility; known-groups discrimination; AI-versus-expert assessment
Quah et al. (2024)	Singapore	69 final-year dental students; 2 clinical	Undergraduate dental students	Oral and maxillofacial surgery examination; clinical	GPT-4 automated essay scoring; text	3 independent human assessors	Cross-sectional direct comparison	ICC, Cronbach's α, Pearson correlation and mean score comparison

		case essays per student		management reasoning		using common rubrics		
Tekin et al. (2025)	Türkiye	196 preclinical medical students; 4 OSCE skill stations	Medical students, Years 1–3	OSCE: intramuscular injection, square-knot tying, basic life support and urinary catheterization	ChatGPT-4o and Gemini Flash 1.5; recorded performance video	1 real-time human examiner and 2 expert video raters using standardized checklists	Cross-sectional evaluator-comparison study	Inter-rater agreement/consistency; score differences between AI and humans; task-specific performance
Luordo et al. (2025)	Spain	96 medical students; 1,824 assessed OSCE items	Medical students	OSCE written clinical reports	GPT-4 automated grading; text	Expert physician graders and an inexperienced human grader	Comparative grading validation	ICC; AI–human score differences; inter-rater comparison
Campbell et al. (2025)	United States	684 medical students across 4 academic years	Medical students in clinical clerkships	Clerkship OSCE; post-encounter clinical notes	GPT-4 Automated System Protocol; text	Legacy Standardized Patient Evaluator grading; additional physician grading comparisons	Retrospective multicohort comparison	AI–human grading agreement/performance; scoring differences; efficiency and ROI
Souria l & Hagler (2025)	United States	39 third-year pharmacy students; 5-station OSCE, with 2 interactive stations evaluated by AI	Pharmacy students	Pharmacy OSCE; clinical communication and performance checklist	Speech-to-text plus tailored transformer model	Faculty evaluators and investigator-generated validation ratings	Retrospective pilot validation	AI scoring accuracy (>96% and 94% across analyzed stations); human inter-rater variability

Liu et al. (2025)	China	31 medical students; 93 sessions across 3 clinical cases	Medical students	Simulated history taking; clinical questioning and information gathering	LLM-based medical history-taking evaluation system, including DeepSeek-V2.5 with cross-model evaluation	Human assessment of performance	Prospective multicase validation with repeated AI evaluations	Human–AI consistency; ICC; repeated-run stability/reliability; cross-model generalization
Qian et al. (2025)	United States / Australia	Pilot evaluation involving student interactions with 5 neurological clinical cases	Medical students	Case-based learning; clinical reasoning	GPT-4o scoring of written case interactions	Independent expert faculty scorers	Pilot direct-comparison and calibration study	ICC/correlation with experts; score bias; calibration; Bland–Altman-type agreement analysis
Thomas et al. (2025)	United States	127 post-OSCE SOAP notes	Medical students	OSCE post-encounter notes; history, physical examination, differential diagnosis/reasoning and management	ChatGPT-4 using the human grading rubric	Physician proctor	Retrospective comparative grading study	Paired score differences; effect sizes; pass/honors classification disagreement
Yokose et al. (2025)	Japan	11 fifth-year medical students	Medical students	Mock OSCE medical interview for abdominal pain; communication, history taking, examination, clinical reasoning and	ChatGPT-4 using encounter transcripts and patient notes	4 physicians rating videos and notes	Experimental direct-comparison study	ICC by competence domain; paired score differences; evidence of poor agreement in several domains

				management				
Mitsuhashi et al. (2025)	Japan	64 first-year paramedic students; 32 pairs completing prehospital simulations	Paramedic students	Prehospital simulation; non-technical skills including communication, information gathering, contextual actions and time management	ChatGPT-4o scoring conversation transcripts	Paramedic faculty using the same five-domain rubric	Comparative validation study	Weighted Gwet AC2; weighted agreement; paired score differences; domain-specific agreement
Kim et al. (2025)	United States / multicountry participant sites	78 participants; 111 laparoscopic performance videos across 4 procedures	Medical students, residents and attending surgeons	Laparoscopic simulation; general procedural competence	K-nearest-neighbor AI using time- and distance-based performance metrics	Human video review using Global Rating Scale competencies	Five-fold cross-validation and cross-procedure generalizability evaluation	AI–human concordance; ICC; five-class and pass/fail classification performance
Rajan et al. (2026)	United States	1,450 medical-student short-answer responses	Medical students	Case-based short-answer examination; clinical reasoning	GPT-4o; text responses	Faculty graders	Equivalence-validation study	Mean AI–human score difference; equivalence analysis; ICC = 0.993; question-level moderator analysis
Hassanein et al. (2026)	Egypt	262 undergraduate dental students; 2,358 responses to 9 clinical short-answer questions	Dental students	Clinical short-answer assessment; dental clinical reasoning	ChatGPT-4 and DeepSeek; rubric-based LLM grading	3 calibrated subject-matter experts	Cross-sectional comparative validation	AI–human ICC/correlation/agreement; human–human ICC; model-specific grading differences
Chen (2026)	Taiwan	80 nursing	Undergraduate	VR obstetric	VR-integrated	Clinical educator	Randomized	AI–educator ICC (overall strong

		students: 40 intervention and 40 control	nursing students	simulation; empathic clinical communication	AI empathic-communication assessment and automated scoring	s using the Empathic Communication Performance Rating Scale	controlled mixed-methods study with inter-rater validation	agreement, approximately 0.81); construct/intervention validity evidence

Note. Abbreviations: AI, artificial intelligence; CNN, convolutional neural network; GEARS, Global Evaluative Assessment of Robotic Skills; GRS, Global Rating Scale; ICC, intraclass correlation coefficient; LLM, large language model; MAE, mean absolute error; ML, machine learning; NLP, natural language processing; OSATS, Objective Structured Assessment of Technical Skill; OSCE, objective structured clinical examination; PLACE, Pelvic Lymphadenectomy Appropriateness and Completion Evaluation; ROI, return on investment.

Note. Each row represents one independent underlying participant or performance cohort. The evidence synthesis comprised 27 independent cohorts represented by 29 reports included in the review. Publication histories and potential companion publications were rechecked. Where preprint, peer-reviewed, or other reports represented the same underlying cohort, they were linked, the peer-reviewed publication was used as the primary source where available, and duplicate participant data or effect estimates were not counted more than once.

Supplementary Table S2. Study-level methodological appraisal of the 27 independent studies

Study	Development–evaluation separation	Human reference assessment	Blinding / assessor independence	AI–human information equivalence	Validation / generalizability	Reproducibility reporting
Swygert et al. (2003)	Separation not clearly reported	Expert human ratings	NR	Same patient-note material	Comparative/internal validation	Fixed automated-scoring approach; reporting adequate for the period
Latifi et al. (2016)	Operational validation using examination data	Expert physician markers	NR	Same written responses	Strong operational validation in a high-stakes examination setting	Fixed supervised ML system; method described
Brown et al. (2017)	Internal predictive validation	3 expert raters using GEARS	NR	Partial; AI used sensor-derived metrics whereas experts rated performance	Internal validation	Fixed ML features and model; technically reproducible
Azari et al. (2019)	Internal proof-of-concept modelling	3 expert surgeons with	NR	Partial; AI used hand-motion	Proof-of-concept/internal validation	Computer-vision and modelling

		consensus ratings		features derived from video		approach described
Baghdadi et al. (2019)	Holdout development plus separate test procedures	Expert surgeon panel using PLACE	NR	Largely equivalent underlying operative performances	Independent test set, but very small	Computer-vision pipeline described
Maicher et al. (2019)	Direct validation sample used for detailed comparison	3 trained human raters	NR	Same virtual-patient dialogues	Internal validation	NLP scoring procedure described; repeated stochastic output not applicable
Cianciolo et al. (2021)	Internal validity study	Faculty ratings	NR	Same diagnostic-justification essays	Internal validation	Fixed NLP/machine-scoring approach; methodological reporting generally adequate
Lavanchy et al. (2021)	Development with internal validation/test procedure	4 board-certified surgeons	NR	Same surgical performances, although AI extracted computational features	Internal validation/test set	Multi-stage ML pipeline clearly specified
Kasa et al. (2022)	Internal model development and evaluation	3 expert raters using OSATS	NR	Partial; AI used image, kinematic, and multimodal representations	Internal validation	Architectures and multimodal modelling described
Igaki et al. (2023)	Independent validation subset separated from model construction	Expert ESSQS reference scores	NR	Same underlying surgical performances	Independent validation sample within the same study setting	Deep-learning scoring approach described
Burke et al. (2024)	Direct comparison using a pretrained LLM; no model training on evaluation cohort reported	Standardized-patient evaluators using the same rubric	NR	High; AI and humans evaluated the same clinical notes/rubric	Single-institution retrospective validation	Model identified; prompt/configuration reporting less complete than ideal
Holderried et al. (2024)	Direct agreement validation using a pretrained LLM	Human rater assessmen	NR	High for the scored	Internal validation	GPT-4 system identified; reproducibility

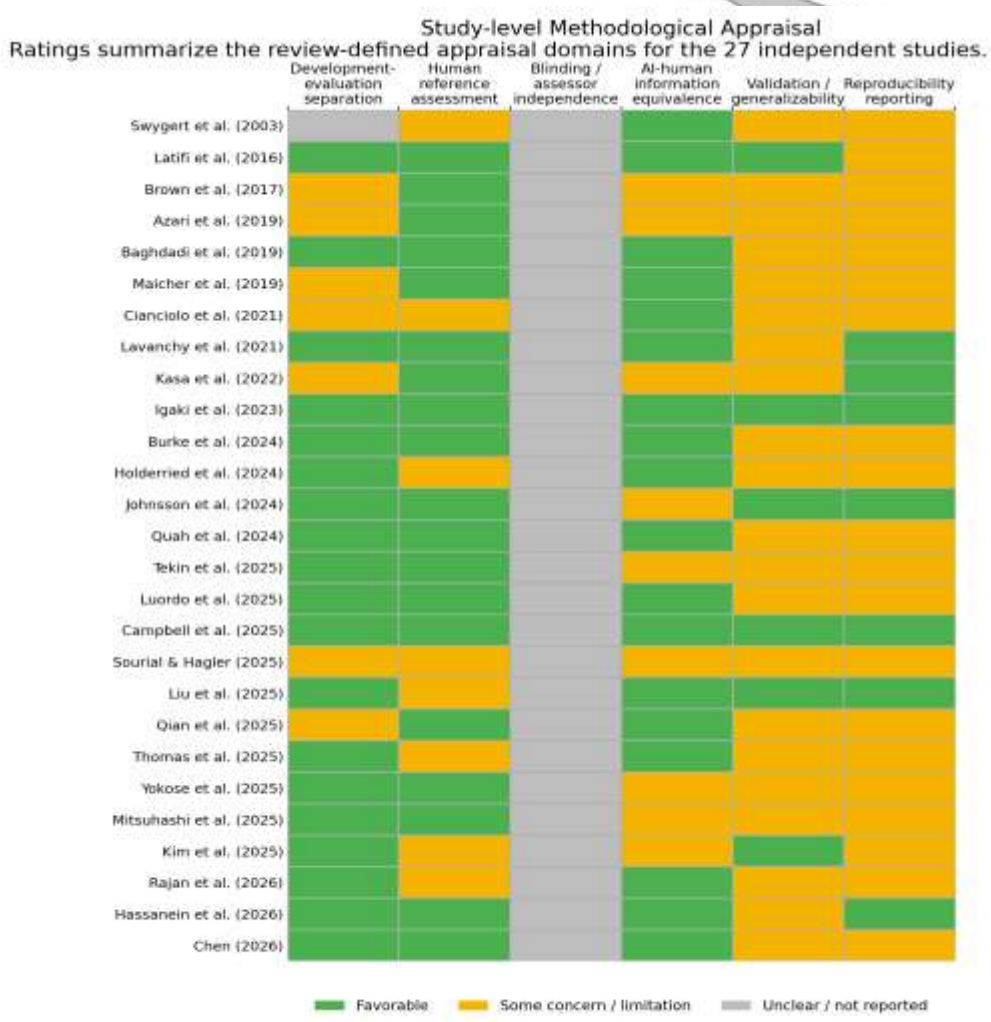
		t of the same interactions		interaction content		dependent on reported prompting/configuration
Johnsson et al. (2024)	Prospective comparative validation	Trained clinical experts using a consensus-developed instrument	NR	Partial; AI incorporated video, motion, and eye-tracking features	Prospective internal validation with known-groups evidence	Technical pipeline and reproducibility analyses comparatively strong
Quah et al. (2024)	Direct evaluation of a pretrained LLM	3 independent human assessors using common rubrics	NR	High; same essays and scoring framework	Single-setting cross-sectional validation	Model identified; generative configuration/repeated-run reporting limited
Tekin et al. (2025)	Direct evaluation of pretrained multimodal models	1 real-time examiner plus 2 expert video raters	NR	Partial; AI used recorded video, while one human assessment was real-time	Single-setting cross-sectional validation	Models identified; repeated-run/configuration reporting limited
Luordo et al. (2025)	Direct comparative grading using a pretrained LLM	Expert physician graders plus an inexperienced human grader	NR	High; same written clinical reports	Single-setting comparative validation	Model identified; repeated generative stability not comprehensively evaluated
Campbell et al. (2025)	Retrospective validation across multiple academic cohorts	Standardized-patient grading with additional physician comparisons	NR	High; same post-encounter notes and scoring task	Multicohort but institutionally limited validation	Defined automated-system protocol; relatively good implementation reporting
Sourial & Hagler (2025)	Pilot validation	Faculty evaluators and investigator-generated	NR	Partial; speech-to-text processing preceded automated scoring	Small retrospective pilot	Tailored transformer approach described; broader reproducibility limited

		validation ratings				
Liu et al. (2025)	Prospective multicase validation with cross-model testing	Human assessment of the same performances	NR	High for the assessed history-taking content	Prospective internal validation; cross-model generalization evaluated	Strong; repeated AI runs, coefficients of variation, and cross-model stability explicitly evaluated
Qian et al. (2025)	Pilot direct-comparison/calibration study	Independent expert faculty scorers	NR	High; same written case interactions	Small pilot validation	Model identified; calibration described, but repeated-run evidence limited
Thomas et al. (2025)	Direct retrospective comparison using a pretrained LLM	Physician proctor	NR	High; same post-OSCE notes and rubric	Single-setting retrospective validation	Model and rubric use described; repeated-run reliability not established
Yokose et al. (2025)	Direct experimental comparison	4 physicians rating videos and notes	NR	Clear mismatch; physicians accessed video and written information whereas AI received transcripts and notes	Very small single-setting study	Model identified; stability across repeated generations not established
Mitsuhashi et al. (2025)	Direct comparative validation	Paramedic faculty using the same five-domain rubric	NR	Partial; AI scored transcripts while human assessment included the underlying simulation context	Single-setting validation	Model identified; repeated-run/configuration evidence limited
Kim et al. (2025)	Five-fold cross-validation with cross-procedure testing	Human video review using GRS competencies	NR	Partial; AI relied on time/distance performance features whereas humans reviewed	Cross-procedure generalizability examined, but not fully external	Conventional ML pipeline reproducible in principle

				performanc e		
Rajan et al. (2026)	Direct equivalence-validation using a pretrained LLM	Faculty graders	NR	High; same short-answer responses and assessment task	Large internal validation sample	Model identified and equivalence framework described; repeated stochastic reliability limited
Hassane in et al. (2026)	Direct comparison of two pretrained LLMs	3 calibrated subject-matter experts	NR	High; same responses and scoring framework	Single-setting cross-sectional validation	Models and common rubric identified; model-specific comparisons strengthen interpretability
Chen (2026)	Inter-rater validation embedded within a randomized study	Clinical educators using the same rating scale	NR	Substantiall y comparable clinical simulation performanc e	Single-setting randomized mixed-methods study	AI/VR system described; detailed repeated-run stability limited

Note. NR indicates that blinding or assessor independence was not reported clearly enough in the available study description to support a definite judgment. 'Information equivalence' refers to whether AI and human assessors had access to materially comparable representations of the same performance; partial equivalence indicates differences in input modality or available cues. The table presents domain-specific methodological characteristics rather than a universal numerical quality score. Domain-level judgments were categorized as favorable when the available evidence supported the domain without an identified material limitation; some concern/limitation when the domain was partially addressed or an identifiable limitation affected interpretation; and unclear/not reported when available reporting was insufficient for a definite judgment. These judgments were review-specific and domain-based and did not constitute a standard COSMIN or QUADAS-2 rating or an overall numerical quality score.

Supplementary Figure S1. Study-level methodological appraisal across the review-defined domains



Note: Ratings are domain specific and do not represent an overall numerical quality score.

Supplementary Appendix S1. Electronic Search Strategies

Search strategies are reported below as numbered concept sets. The database field searched is stated in a separate column, so no bracketed field labels are used. The asterisk (*) is a literal truncation/wildcard operator used to retrieve word variants; it is not a placeholder. The original database-interface query histories were not retained; accordingly, the strategies below reproduce the documented search concepts, field specifications, and Boolean combinations rather than verbatim archived executable query logs.

The searches combined three core concepts: (1) artificial intelligence, (2) clinical competence or clinical performance, and (3) assessment or scoring. Controlled vocabulary was combined with free-text terms where the database supported controlled indexing. No language restriction was applied during searching.

MEDLINE via PubMed

Language restrictions: None

Set	Search field / indexing	Search terms or combination
1	MeSH Terms	Artificial Intelligence OR Machine Learning OR Deep Learning OR Natural Language Processing

2	Title/Abstract	"artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR "large language model*" OR LLM OR LLMs OR ChatGPT OR GPT-4 OR GPT-4o OR "generative AI" OR "generative artificial intelligence" OR "computer vision" OR "multimodal AI"
3	MeSH Terms	Clinical Competence
4	Title/Abstract	"clinical competence" OR "clinical performance" OR "clinical skill*" OR "procedural skill*" OR "technical skill*" OR "clinical reasoning" OR "diagnostic reasoning" OR "history taking" OR communication OR OSCE OR "objective structured clinical examination*" OR simulation OR "standardized patient*" OR "standardised patient*" OR "virtual patient*" OR "workplace-based assessment*" OR "clinical note*" OR "patient note*"
5	Title/Abstract	assess* OR scor* OR grad* OR rating OR ratings OR rater* OR evaluat* OR classific* OR agreement OR reliability OR validity OR concordance
6	Boolean combination	(Set 1 OR Set 2) AND (Set 3 OR Set 4) AND Set 5

Embase

Language restrictions: None

Set	Search field / indexing	Search terms or combination
1	Emtree headings; exploded	artificial intelligence OR machine learning OR deep learning OR natural language processing
2	Title/Abstract/Keywords	"artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR "large language model*" OR LLM OR LLMs OR ChatGPT OR GPT-4 OR GPT-4o OR "generative AI" OR "generative artificial intelligence" OR "computer vision" OR "multimodal AI"
3	Emtree heading; exploded	clinical competence
4	Title/Abstract/Keywords	"clinical competence" OR "clinical performance" OR "clinical skill*" OR "procedural skill*" OR "technical skill*" OR "clinical reasoning" OR "diagnostic reasoning" OR "history taking" OR communication OR OSCE OR "objective structured clinical examination*" OR simulation OR "standardized patient*" OR "standardised patient*" OR "virtual patient*" OR "workplace-based assessment*" OR "clinical note*" OR "patient note*"
5	Title/Abstract/Keywords	assess* OR scor* OR grad* OR rating OR ratings OR rater* OR evaluat* OR classific* OR agreement OR reliability OR validity OR concordance
6	Boolean combination	(Set 1 OR Set 2) AND (Set 3 OR Set 4) AND Set 5

Scopus

Language restrictions: None

Set	Search field / indexing	Search terms or combination
1	Title/Abstract/Keywords (TITLE-ABS-KEY)	"artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR "large language model*" OR LLM OR LLMs OR ChatGPT OR GPT-4 OR GPT-4o OR "generative AI" OR "generative artificial intelligence" OR "computer vision" OR "multimodal AI"
2	Title/Abstract/Keywords (TITLE-ABS-KEY)	"clinical competence" OR "clinical performance" OR "clinical skill*" OR "procedural skill*" OR "technical skill*" OR "clinical reasoning" OR "diagnostic reasoning" OR "history taking" OR communication OR OSCE OR "objective structured clinical examination*" OR simulation OR "standardized patient*" OR "standardised patient*" OR "virtual patient*" OR "workplace-based assessment*" OR "clinical note*" OR "patient note*"

3	Title/Abstract/Keywords (TITLE-ABS-KEY)	assess* OR scor* OR grad* OR rating OR ratings OR rater* OR evaluat* OR classific* OR agreement OR reliability OR validity OR concordance
4	Boolean combination	Set 1 AND Set 2 AND Set 3

Web of Science Core Collection

Language restrictions: None

Set	Search field / indexing	Search terms or combination
1	Topic field (TS)	"artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR "large language model*" OR LLM OR LLMs OR ChatGPT OR GPT-4 OR GPT-4o OR "generative AI" OR "generative artificial intelligence" OR "computer vision" OR "multimodal AI"
2	Topic field (TS)	"clinical competence" OR "clinical performance" OR "clinical skill*" OR "procedural skill*" OR "technical skill*" OR "clinical reasoning" OR "diagnostic reasoning" OR "history taking" OR communication OR OSCE OR "objective structured clinical examination*" OR simulation OR "standardized patient*" OR "standardised patient*" OR "virtual patient*" OR "workplace-based assessment*" OR "clinical note*" OR "patient note*"
3	Topic field (TS)	assess* OR scor* OR grad* OR rating OR ratings OR rater* OR evaluat* OR classific* OR agreement OR reliability OR validity OR concordance
4	Boolean combination	Set 1 AND Set 2 AND Set 3

ERIC

Language restrictions: None

Set	Search field / indexing	Search terms or combination
1	Indexed title/abstract/descriptors	"artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network*" OR "natural language processing" OR "large language model*" OR LLM OR LLMs OR ChatGPT OR GPT-4 OR GPT-4o OR "generative AI" OR "generative artificial intelligence" OR "computer vision" OR "multimodal AI"
2	Indexed title/abstract/descriptors	"clinical competence" OR "clinical performance" OR "clinical skill*" OR "procedural skill*" OR "technical skill*" OR "clinical reasoning" OR "diagnostic reasoning" OR "history taking" OR communication OR OSCE OR "objective structured clinical examination*" OR simulation OR "standardized patient*" OR "standardised patient*" OR "virtual patient*" OR "workplace-based assessment*" OR "clinical note*" OR "patient note*"
3	Indexed title/abstract/descriptors	assess* OR scor* OR grad* OR rating OR ratings OR rater* OR evaluat* OR classific* OR agreement OR reliability OR validity OR concordance
4	Boolean combination	Set 1 AND Set 2 AND Set 3

IEEE Xplore

Language restrictions: None

Set	Search field / indexing	Search terms or combination
1	All Metadata	"artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "natural language processing" OR "large language model" OR ChatGPT OR "generative AI" OR "computer vision"
2	All Metadata	"clinical competence" OR "clinical performance" OR "clinical skill" OR "procedural skill" OR "technical skill" OR "clinical reasoning" OR "history taking" OR OSCE OR "objective structured clinical examination" OR simulation
3	All Metadata	assessment OR scoring OR grading OR rating OR evaluation OR classification OR agreement OR reliability OR validity

4	Boolean combination	Set 1 AND Set 2 AND Set 3

Supplementary Searching

Electronic database searching was supplemented by the following approaches:

- Backward reference-list screening of included studies and relevant systematic or scoping reviews.
- Forward citation searching of included studies.
- Targeted searches of relevant preprint repositories.
- Targeted searches for recent publications that might not yet have been fully indexed in bibliographic databases.
- Examination of publication histories to identify earlier preprints, subsequent peer-reviewed publications, and potentially overlapping datasets.

Potentially overlapping reports were compared according to authorship, institution, participant characteristics, recruitment period, sample size, assessment task, AI system, and reported outcomes. Reports representing the same underlying cohort were linked to avoid duplicate counting.

References

1. Azari, D. P., Frasier, L. L., Quamme, S. R. P., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2019). Modeling surgical technical skill using expert assessment for automated computer rating. Annals of Surgery, 269(3), 574–581. DOI: 10.1097/SLA.0000000000002478.
2. Baghdadi, A., Hussein, A. A., Ahmed, Y., Cavuoto, L. A., & Guru, K. A. (2019). A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. International Journal of Computer Assisted Radiology and Surgery, 14(4), 697–707. DOI: 10.1007/s11548-018-1881-9.
3. Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8
4. Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. Psychometrika, 65(1), 23–28. https://doi.org/10.1007/BF02294183
5. Brown, J. D., O'Brien, C. E., Leung, S. C., Dumon, K. R., Lee, D. I., & Kuchenbecker, K. J. (2017). Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Transactions on Biomedical Engineering, 64(9), 2263–2275. DOI: 10.1109/TBME.2016.2634861.
6. Burke, H. B., Hoang, A., Lopreiato, J. O., King, H., Hemmer, P., Montgomery, M., & Gagarin, V. (2024). Assessing the ability of a large language model to score free-text medical student clinical notes: Quantitative study. JMIR Medical Education, 10, e56342. DOI: 10.2196/56342.
7. Campbell, K. K., Holcomb, M. J., Vedovato, S., Young, L., Danuser, G., Dalton, T. O., Jamieson, A. R., & Scott, D. J. (2025). Applying state-of-the-art artificial intelligence to grading in simulation-based education: Assessment, feedback, and ROI. Discover Artificial Intelligence, 5, 202. DOI: 10.1007/s44163-025-00417-3.
8. Chen, P.-J. (2026). AI-enhanced virtual reality simulation for nursing students' empathy: Automated scoring and inter-rater reliability in a randomised controlled study. Nurse Education Today, 159, 106968. DOI: 10.1016/j.nedt.2025.106968.
9. Cianciolo, A. T., LaVoie, N., & Parker, J. (2021). Machine scoring of medical students' written clinical reasoning: Initial validity evidence. Academic Medicine, 96(7), 1026–1035. DOI: 10.1097/ACM.0000000000004010.
10. Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: A primer for simulation and beyond. Advances in Simulation, 1, 31. https://doi.org/10.1186/s41077-016-0033-y
11. Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. JAMA, 287(2), 226–235. https://doi.org/10.1001/jama.287.2.226
12. Feigerlova, E., Hani, H., & Hothersall-Davies, E. (2025). A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. BMC Medical Education, 25, 129. https://doi.org/10.1186/s12909-025-06719-5
13. Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. Medical Education, 48(11), 1055–1068. https://doi.org/10.1111/medu.12546

14. Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in meta-analysis with normally distributed responses. Statistics in Medicine, 20(12), 1771–1782. https://doi.org/10.1002/sim.791
15. Hassanein, F. E. A., Hussein, R. R., Ahmed, Y., El-Guindy, J., Ahmed, D. E., & Abou-Bakr, A. (2026). Calibration of AI large language models with human subject matter experts for grading of clinical short-answer responses in dental education. BMC Oral Health, 26, 286. DOI: 10.1186/s12903-026-07665-4.
16. Holderried, F., Stegemann-Philipps, C., Herrmann-Werner, A., Festl-Wietek, T., Holderried, M., Eickhoff, C., & Mahling, M. (2024). A language model–powered simulated patient with automated feedback for history taking: Prospective study. JMIR Medical Education, 10, e59213. DOI: 10.2196/59213.
17. Igaki, T., Kitaguchi, D., Matsuzaki, H., Nakajima, K., Kojima, S., Hasegawa, H., Takeshita, N., Kinugasa, Y., & Ito, M. (2023). Automatic surgical skill assessment system based on concordance of standardized surgical field development using artificial intelligence. JAMA Surgery, 158(8), e231131. DOI: 10.1001/jamasurg.2023.1131.
18. Johnsson, V., Søndergaard, M. B., Kulasegaram, K., Sundberg, K., Tiblad, E., Herling, L., Petersen, O. B., & Tolsgaard, M. G. (2024). Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Medical Education, 58(1), 105–117. DOI: 10.1111/medu.15190.
19. Kasa, K., Burns, D., Goldenberg, M. G., Selim, O., Whyne, C., & Hardisty, M. (2022a). Multi- modal deep learning for assessing surgeon technical skill. Sensors, 22(19), 7328. DOI: 10.3390/s22197328.
20. Kasa, K., Burns, D., Goldenberg, M. G., Selim, O., Whyne, C., & Hardisty, M. (2022b). Multi- modal deep learning for assessing surgeon technical skill on a surgical knot-tying task [Preprint]. TechRxiv. DOI: 10.36227/techrxiv.20085425.v1.
21. Kim, E., Rosenthal, L. S., Ryder, C. Y., Anidi, C., Bidwell, S. S., Rooney, D. M., Yu, J., Forczmanski, P., Jeffcoach, D. R., & Kim, G. J. (2025). Generalizability of artificial intelligence assessments in laparoscopic surgery simulation. Journal of Surgical Research, 309, 249–256. DOI: 10.1016/j.jss.2025.03.030.
22. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
23. Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Journal of Clinical Epidemiology, 64(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002
24. Latifi, S., Gierl, M. J., Boulais, A.-P., & De Champlain, A. F. (2016). Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination. Evaluation & the Health Professions, 39(1), 100–113. DOI: 10.1177/0163278715605358.
25. Lavanchy, J. L., Zindel, J., Kirtac, K., Twick, I., Hosgor, E., Candinas, D., & Beldi, G. (2021). Automation of surgical skill assessment using a three-stage machine learning algorithm. Scientific Reports, 11, 5197. DOI: 10.1038/s41598-021-84295-6.
26. Liu, Y., Shi, C., Wu, L., Lin, X., Chen, X., Zhu, Y., Tan, H., & Zhang, W. (2025). Development and validation of a large language model–based system for medical history-taking training: Prospective multicase study on evaluation stability, human-AI consistency, and transparency. JMIR Medical Education, 11, e73419. DOI: 10.2196/73419.
27. Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. Medical Education, 58(11), 1276– 1285. https://doi.org/10.1111/medu.15402
28. Luordo, D., Torres Arrese, M., Tristán Calvo, C., Shani Shani, K. D., Rodríguez Cruz, L. M., García Sánchez, F. J., Lagares Gómez-Abascal, A., Rubio García, R., Delgado Jiménez, J., Pérez Carreras, M., Diez Lobato, R., Granizo Martínez, J. J., Tung-Chen, Y., & Villena Garrido, M. V. (2025). Application of artificial intelligence as an aid for the correction of the Objective Structured Clinical Examination (OSCE). Applied Sciences, 15(3), 1153. DOI: 10.3390/app15031153.
29. Maicher, K. R., Zimmerman, L., Wilcox, B., Liston, B., Cronau, H., Macerollo, A., Jin, L., Jaffe, E., White, M., Fosler-Lussier, E., Schuler, W., Way, D. P., & Danforth, D. R. (2019). Using virtual standardized patients to accurately assess information gathering skills in medical students. Medical Teacher, 41(9), 1053–1059. DOI: 10.1080/0142159X.2019.1616683.

30. Mitsuhashi, M., Akiba, Y., Saitou, M., Suzuki, K., Ogawa, S., & Masuno, T. (2025). AI-based assessment of non-technical skills in prehospital simulations: A comparative validation study. Healthcare, 13(23), 3121. DOI: 10.3390/healthcare13233121.
31. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology, 63(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006
32. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71
33. Pedrett, R., Mascagni, P., Beldi, G., Padoy, N., & Lavanchy, J. L. (2023). Technical skill assessment in minimally invasive surgery using artificial intelligence: A systematic review. Surgical Endoscopy, 37(10), 7412–7424. https://doi.org/10.1007/s00464-023-10335-z
34. Prentice, S., Benson, J., Kirkpatrick, E., & Schuwirth, L. (2020). Workplace-based assessments in postgraduate medical education: A hermeneutic review. Medical Education, 54(11), 981– 992. https://doi.org/10.1111/medu.14221
35. Qian, C., Gao, C., Park, S.-O., Gim, H., Hou, K., Cook, B., Le, J., Stretton, B., Maddison, J., McCoy, L., Goh, R., Arnold, M., Reda, H., Kaplan, T., Gheihman, G., & Bacchi, S. (2025). Use of large language models for rapid quantitative feedback in case-based learning: A pilot study. Medical Science Educator, 35(3), 1169–1171. DOI: 10.1007/s40670-025- 02343-6.
36. Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. BMC Medical Education, 24, 962. DOI: 10.1186/s12909-024-05881-6.
37. Rajan, A., Alexander, S. M., & Shenvi, C. L. (2026). Can AI grade like a professor? Comparing artificial intelligence and faculty scoring of medical student short-answer clinical reasoning exams. Advances in Health Sciences Education, 31(2), 607–617. DOI: 10.1007/s10459- 025-10462-3.
38. Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic Reviews, 10, 39. https://doi.org/10.1186/s13643-020-01542-z
39. Röver, C., Knapp, G., & Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC Medical Research Methodology, 15, 99.
40. Shaw, K., Henning, M. A., & Webster, C. S. (2025). Artificial intelligence in medical education: A scoping review of the evidence for efficacy and future directions. Medical Science Educator, 35, 1803–1816. https://doi.org/10.1007/s40670-025-02373-0
41. Sourial, M., & Hagler, J. C. (2025). A pilot study using artificial intelligence to enhance efficiency, accuracy, and objectivity in grading pharmacy objective structured clinical examinations. American Journal of Pharmaceutical Education, 89(8), 101455. DOI: 10.1016/j.ajpe.2025.101455.
42. Swygert, K. A., Margolis, M. J., King, A., Siftar, T., Clyman, S. G., Hawkins, R. E., & Clauser, B. E. (2003). Evaluation of an automated procedure for scoring patient notes as part of a clinical skills examination. Academic Medicine, 78(10 Suppl), S75–S77. DOI: 10.1097/00001888-200310001-00024.
43. Tekin, M., Yurdal, M. O., Toraman, Ç., Korkmaz, G., & Uysal, İ. (2025). Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination. BMC Medical Education, 25, 641. DOI: 10.1186/s12909-025-07241-4.
44. Thomas, K., Szalacha, L., Hanna, K., Anibal, J., & Petrilli, J. (2025). Evaluating the effectiveness of ChatGPT versus human proctors in grading medical students' post-OSCE notes. Family Medicine, 57(10), 727–731. DOI: 10.22454/FamMed.2025.954255.
45. Tozsin, A., Ucmak, H., Soyturk, S., Aydin, A., Gozen, A. S., Al Fahim, M., Güven, S., & Ahmed, K. (2024). The role of artificial intelligence in medical education: A systematic review. Surgical Innovation, 31(4), 415–423. https://doi.org/10.1177/15533506241248239
46. van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. Medical Education, 39(3), 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

47. van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. Medical Teacher, 34(3), 205–214. https://doi.org/10.3109/0142159X.2012.652239
48. Van Melle, E., Frank, J. R., Holmboe, E. S., Dagnone, D., Stockley, D., Sherbino, J., & International Competency-Based Medical Education Collaborators. (2019). A core components framework for evaluating implementation of competency-based medical education programs. Academic Medicine, 94(7), 1002–1009. https://doi.org/10.1097/ACM.0000000000002743
49. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
50. Yokose, M., Hirosawa, T., Sakamoto, T., Kawamura, R., Suzuki, Y., Harada, Y., & Shimizu, T. (2025a). The validity of generative artificial intelligence in evaluating medical students in Objective Structured Clinical Examination: A pilot study [Preprint]. JMIR Preprints. DOI: 10.2196/preprints.79465.
51. Yokose, M., Hirosawa, T., Sakamoto, T., Kawamura, R., Suzuki, Y., Harada, Y., & Shimizu, T. (2025b). The validity of generative artificial intelligence in evaluating medical students in Objective Structured Clinical Examination: Experimental study. JMIR Formative Research, 9, e79465. DOI: 10.2196/79465.