Revista Española de Educación Médica

UNIVERSIDAD DE MURCIA

Automated feedback using artificial intelligence in clinical simulation: integrative multi-source critical review of validity, learning and transfer.

Retroalimentación automatizada mediante inteligencia artificial en simulación clínica: revisión crítica integrativa multifuente de validez, aprendizaje y transferencia.

Ana Mikal Toala Mora[1]*, Adriana Denisse García Coello [2]

[1] Medical School, San Gregorio University of Portoviejo, Portoviejo, Ecuador; amtoala@sangregorio.edu.ec, https://orcid.org/0000-0002-4301-4264; [2] Adriana Denisse García Coello: adgarcia@sangregorio.edu.ec, https://orcid.org/0000-0002-7044-1894.

* Correspondence: amtoala@sangregorio.edu.ec

Received: 9/3/26; Accepted: 9/28/26; Published: 9/20/26

Summary.

Objective: To critically analyze the empirical evidence on automated feedback using artificial intelligence (AI) in clinical simulation for medical students, with an emphasis on validity, educational outcomes, and transferability. **Methodology:** A non-systematic, multi-source integrative critical review was conducted, guided by the methodological approach of Whittemore and Knafl and recommendations for integrative reviews. Structured searches were performed in PubMed/MEDLINE, ERIC, and OpenAlex, supplemented by citation tracking and comparison with recent reviews. The primary studies underwent explicit critical appraisal by domains of design, comparator, outcome independence, reproducibility, and follow-up. **Results:** Seventeen empirical studies published between 2019 and 2026 were analyzed. The evidence showed high agreement with human evaluators in some structured tasks, but also specificity errors, bias toward high scores, variability between performances, and limitations in complex communicative behaviors. The educational benefits were concentrated in immediate outcomes; The transfer to independent assessments was documented in only a subset. **Conclusions:** Automated feedback is primarily useful for formative assessment and repeated practice. The available evidence does not demonstrate general equivalence with human assessment, nor does it support its autonomous use in high-impact summative decisions. Domain-specific validations, independent outcomes, and longitudinal follow-up are required.

Keywords: artificial intelligence, automated feedback, clinical simulation, virtual patients, medical education, formative assessment

Abstract.

Objective: To critically analyze empirical evidence on artificial intelligence (AI)-generated automated feedback in clinical simulation for medical students, focusing on validity, educational outcomes, and transfer. **Methods:** A multisource integrative critical review, rather than a systematic review, was conducted following the methodological principles of Whittemore and Knafl and guidance for integrative reviews. Structured searches were performed in PubMed/MEDLINE, ERIC, and OpenAlex, supplemented by citation tracking and comparison with recent reviews. Primary studies underwent an explicit critical appraisal across design, comparator, outcome independence, reproducibility, and follow-up domains. **Results:** Seventeen empirical studies published between 2019 and 2026 were analyzed. High agreement with human

raters was reported for some structured tasks, but specificity errors, overly generous scoring, run-to-run variability, and limitations in complex communication behaviors were also observed. Educational benefits were concentrated in immediate outcomes, while transfer to independent assessments was documented only in a subset. **Conclusions:** Automated feedback currently shows its clearest utility for formative and repeated practice. Available evidence does not establish general equivalence with human assessment or justify autonomous use in high-stakes summative decisions. Domain-specific validation, independent outcomes, and longitudinal follow-up are needed.

Keywords: artificial intelligence, automated feedback, clinical simulation, virtual patients, medical education, formative assessment

1. Introduction

Clinical simulation allows for the training of clinical skills in safe and controlled environments, with the possibility of repetition, standardization, and feedback before exposure to more complex care situations. In medical education, standardized patients, physical simulators, virtual patients, and immersive environments are not equivalent modalities, but rather resources with varying degrees of realism, availability, and capacity to train specific domains. Fierro Manrique et al. recently highlighted the complementarity between high-fidelity simulation with standardized patients and virtual environments, noting distinct advantages in communication, emotional realism, accessibility, and repetition (1).

The expansion of artificial intelligence (AI), and in particular large-scale language models (LLMs), has added a new function to simulation: generating dynamic clinical responses, interpreting student dialogue, and producing automatic feedback during or after the encounter. A scoping review published in this same journal mapped AI applications in practical clinical training and found a predominance of simulated scenarios and immediate results (2). Recent international reviews also describe rapid growth in generative virtual patients and AI applications in clinical skills, but they agree on methodological heterogeneity and limited evidence of transfer to real clinical contexts (3-5). Recent work has described the expansion of generative AI, the pedagogical use of ChatGPT, case generation with LLMs, and new teaching competencies linked to AI; together, they show a rapidly evolving field with still predominantly early results (51-54).

In this context, automated feedback warrants specific analysis. Feedback is not simply post-performance information: its value depends on its ability to identify, understand, and address gaps. Classical feedback frameworks emphasize specificity, comparison to a standard, and a focus on improvement actions (6, 7). In simulation, this function is linked to deliberate practice, which requires defined objectives, repetition, and corrective feedback (8). Furthermore, from the perspective of cognitive load theory, the quantity and complexity of feedback must be adjusted to the learner's level to avoid extrinsic overload (9, 10). Therefore, a model's linguistic fluency does not, in itself, demonstrate that its judgment is valid, actionable, or pedagogically useful.

The scope review by Bonilla Mejia et al. (2) included automated assessment and feedback as one of several application domains of AI, while other recent reviews have focused more broadly on feedback generated by LLMs or on virtual patients (14-17, 44). One specific question remains less resolved: whether automated feedback produces changes that persist when the student is assessed outside the AI platform itself. Therefore, the distinctive contribution of this review is to separate three often conflated levels: agreement with human evaluators, pedagogical quality of the feedback, and transferability to independent assessments such as OSCEs or standardized patients. This focus complements, rather than duplicates, the existing general mappings.

The aim of this article is to critically analyze the empirical evidence on AI-driven automated feedback and assessment within clinical simulations for medical students. Four questions are posed: (1) What evidence exists regarding the accuracy, consistency, and stability of automated feedback?; (2) What educational outcomes have been associated with its use?; (3) To what extent do the benefits transfer to assessments independent of the AI platform?; and (4) What pedagogical and operational safeguards emerge from the available evidence?

The analysis was structured around four predefined theoretical frameworks. First, cognitive load theory guided the analysis of the quantity, sequence, and complexity of feedback (9,10). Second, criteria for good evaluation and the argumentative approach to validity allowed for a distinction between the accuracy of an output and the sufficiency of evidence for its intended evaluative use (11,12). Third, programmatic evaluation contributed the principle that high-impact decisions require multiple points of information and expert judgment, while a single piece of data can retain formative value (13). Fourth, deliberate practice underpinned the analysis of repetition, corrective feedback, and performance progression (8). These frameworks justified the domains examined: quality and actionability of feedback, validity of automated evaluation, practice conditions, cognitive load, and transfer to independent assessments. The need for a specific focus has also been reinforced by recent syntheses. Reviews of AI-powered virtual patients for communication, interviewing, and clinical reasoning have identified potential benefits, but also a lack of consistent theoretical frameworks, standardized measures, and longitudinal evidence (14-17). The present review does not aim to replace those comprehensive syntheses, but rather to critically examine the feedback/evaluation component and contrast the primary studies located with those reviews.

2. Methods

2.1. Review design and approach

An integrative critical review was conducted using a structured, multi-source search. The work was guided by the approach of Whittemore and Knafl, which organizes the integrative review into problem definition, literature search, data appraisal, analysis, and synthesis presentation, and by contemporary methodological recommendations for literature reviews (48,49). This design was chosen because it allows for the integration of experimental, quasi-experimental, validation, and mixed-methods studies around a specific educational mechanism—automated feedback within clinical simulation. The manuscript is not presented as a systematic or scoping review; therefore, it does not claim exhaustiveness or compliance with PRISMA-ScR guidelines. Inferences are limited to the primary studies identified through the described strategy and their contextualization with recent secondary syntheses.

2.2. Review and eligibility question

The question was: What empirical evidence exists regarding the validity, quality, educational outcomes, and transferability of AI-generated automated feedback or assessment within clinical simulation experiences for medical students? Empirical publications between January 2019 and September 3, 2026, were included if they: (a) targeted medical students or an explicit medical education context; (b) employed a virtual patient, digital simulated patient, recorded standardized patient interaction, conversational agent, or interactive clinical scenario; and (c) used AI to generate feedback, score performance, or compare an automated assessment to a human referent. Technology validation studies were admitted when they explicitly assessed the quality of the feedback or scoring intended for the learner. Editorials, letters, protocols, secondary reviews as primary studies, simulations without automated feedback/assessment, and AI applications used only to generate content without clinical interaction were excluded.

2.3. Bibliographic search

Identification was performed using structured and complementary searches in PubMed/MEDLINE, ERIC, and OpenAlex. In each source, the following criteria were combined: population/context (medical students, medical education), simulation (virtual patient, simulated patient, standardized patient, clinical simulation), AI (artificial intelligence, generative AI, large language model, GPT), and feedback/evaluation (feedback, automated assessment, scoring, evaluation). No restriction was applied based on the availability of full text. Publications in English or Spanish were considered; studies in other languages were eligible if they had sufficient information in English to assess the criteria. A retrospective reference search and a prospective citation search were performed. As a coverage audit, retrieved studies were compared with four recent syntheses (14–17) that included additional databases. This triangulation was used to detect potential omissions and not as a substitute for a direct systematic search of subscription databases. The original search was updated on September 3, 2026. The chains and the recovery procedure are presented in Annex 1.

Table 1. Information sources and identification strategy for integrative critical review.

Source	Function in the review	Search blocks / strategy
PubMed/ MEDLINE	Primary biomedical identification	medical education/student* + virtual/simulated/standardized patient + artificial intelligence/LLM/GPT + feedback/assessment/scoring
ERIC	Supplementary educational coverage	medical education/student* + simulation/virtual patient + artificial intelligence/LLM + feedback/assessment
OpenAlex	Multidisciplinary coverage and detection of works not retrieved in PubMed/ERIC	Combinations by title/topic of AI-generated feedback, virtual patient, standardized patient, clinical simulation and medical education
Date tracking	Retrospective and prospective research	References to included studies, related articles, and citation records
Recent multibase syntheses (14-17)	Coverage and contextualization audit	Contrast with reviews that used Scopus, Web of Science, CINAHL, PsycINFO, ERIC and other databases

2.4. Selection and duplicate control process

The records were cleaned for DOI and title matches. In the first phase, titles and abstracts were reviewed according to eligibility criteria; potentially relevant studies were then assessed for full text. Articles identified through citation tracking and OpenAlex were subjected to the same criteria. Secondary reviews were not counted among the primary studies and were used only for coverage comparison and discussion. The lead author (AMTM) performed the initial screening and data extraction using a predefined form. The second author (ADGC) independently verified the eligibility of the included studies and the critical variables extracted—design, sample, intervention, comparator, primary outcome, and DOI. Discrepancies were resolved by consensus. This procedure does not equate to independent duplicate screening from the time of identification and is acknowledged as a limitation.

The review was initially developed as a critical review, and the historical search record did not maintain an exportable count by source of all retrieved and excluded records. To avoid

reconstructing unverifiable retrospective figures, numbers that cannot be audited are not presented. Traceability is maintained for the 17 included primary studies, including their DOIs, design, sample, outcomes, and reason for inclusion (Table 2). This limitation of quantitative traceability is explicitly stated in the limitations section and distinguishes it from a systematic review.

2.5. Extraction, verification and synthesis of evidence

For each study, the following data were extracted: year, country, design, participants, simulation modality, type of AI, system function, nature of feedback or evaluation, comparator, outcomes, and relevant findings. The synthesis domains were defined before interpreting the results and were based on the frameworks presented in the introduction: (1) validity, agreement, reproducibility, and intended use; (2) specificity, actionability, and pedagogical quality of feedback; (3) educational outcomes and conditions for deliberate practice; (4) cognitive load and student experience; and (5) transferability to an independent evaluation of the platform itself. The heterogeneity of designs, competencies, scales, and comparators precluded an aggregated quantitative synthesis; therefore, a narrative analysis was conducted by domain.

2.6. Critical appraisal of the evidence

Given the heterogeneity of the designs, a critical appraisal was performed by domain, inspired by the principles of evaluating heterogeneous studies in the Mixed Methods Appraisal Tool (MMAT) (50), without converting its criteria into an overall numerical score. Five domains were considered for each study: (1) design adequacy and presence of a comparator or reference; (2) sample sufficiency and representativeness for the objective; (3) independence of the outcome from the AI platform itself; (4) reproducibility or repeatability of the evaluation; and (5) follow-up duration. The overall assessment was classified as major, intermediate, or exploratory to guide the interpretation, not as a formal level of certainty. Studies with a comparator, independent outcome, and experimental design received greater interpretive weight; small pilot studies, perceptual studies, and uncontrolled series were considered exploratory. Table 3 presents this assessment.

2.7. Transparency and reproducibility

To promote reproducibility, the sources consulted, search strings, eligibility criteria, extraction variables, and critical appraisal domains were documented. Figure 1 retains a descriptive function of the process and is not presented as a PRISMA diagram. Retrospective quantitative traceability by source is incomplete; therefore, unverifiable figures were not reconstructed. This decision prioritizes transparency over an appearance of exhaustiveness that the original design cannot sustain.

2.8. Ethical considerations

The study used exclusively information from available publications and did not involve human participants, identifiable medical records, or personal data; therefore, it did not require approval from an ethics committee. The review was guided by principles of transparency, traceability of sources, and verification of bibliographic data.

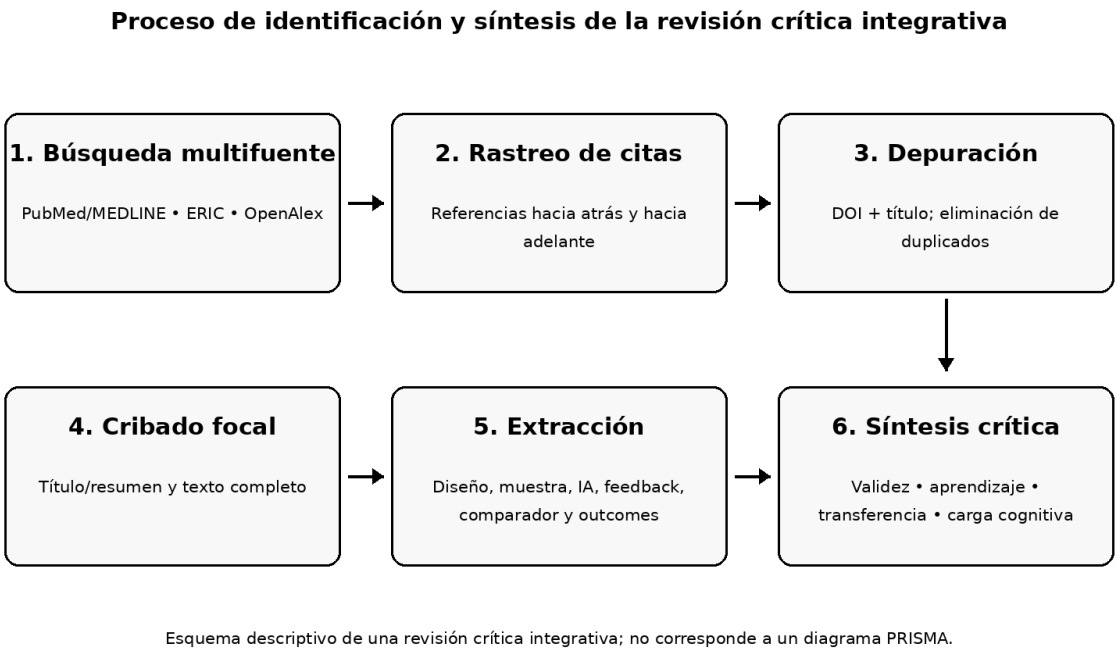

Figura 1. Proceso de identificación y síntesis de la revisión crítica integrativa.

3. Results

3.1. General characteristics

The primary synthesis included 17 empirical studies that met the focal criteria. One was published in 2019, and the remaining studies between 2024 and 2026. In addition to the previously identified studies, the multi-source search incorporated a randomized trial on simulation with LLM and structured feedback, a multimetric evaluation of ChatGPT as both standardized patient and evaluator, a pilot study of ambient listening for feedback in simulated encounters, and a study evaluating transcripts of encounters with standardized patients using various LLMs (18-21). Except for the pre-current generation LLM system described by Maicher et al. (22), recent studies predominantly used LLM or generative AI. History taking, communication, and clinical reasoning were the most frequent domains. Table 2 summarizes the included primary studies.

3.2. Critical appraisal of the evidence

The domain assessment showed that the strongest interpretive evidence came from randomized controlled trials and studies with comparators or independent outcomes, while several pilot studies, perceptual studies, and pre-post series provided exploratory evidence. No study provided long-term follow-up with real patients. Therefore, the following results are presented hierarchically by design and outcome independence, avoiding assigning the same weight to all studies (Table 3).

Table 2. Empirical studies included on automated feedback or evaluation using AI in clinical simulation.

Study	Country	Design/sample	Mode	Feedback/evaluation	Main finding	Critical reading
Maicher et al., 2019 (22)	USA	Validation; 102 encounters (20 analyzed)	AI-powered virtual 3D standardized patient	Automatic question classification and immediate feedback	Overall accuracy 87% vs. 90% human; high reliability in 19/21 categories.	Demonstrates feasibility of automated pre-LLM assessment.
Holderried et al., 2024 (23)	Germany	Prospective; 106 conversations, 1894 question-answer pairs	Simulated patient with GPT-4	Structured feedback on anamnesis	>99% of medically plausible responses; κ=0.832 with human evaluator; κ<0.6 in 8/45 categories.	Good overall agreement, but uneven specificity by category.
Yamamoto et al., 2024 (24)	Japan	Not randomized; 35 intervention/110 historical control	Simulated patient with LLM	Feedback after interviews	Preclinical OSCE: 28.1 vs 27.1; p=0.01. Limitation in non-verbal communication.	Evidence of transfer to standardized assessment, with a non-randomized design.
Scherr et al., 2025 (25)	USA	Experimental evaluation; 360 simulations	ChatGPT-4	Clinical feedback within driving simulation	100% accuracy in basic parameters; integral feedback at 78%; delayed feedback at 55%.	Adherence to advanced pedagogical rules was inconsistent.
Weisman et al., 2025 (26)	USA	Multiphase development study	Virtual patient GPT-4 for abnormal mammogram	Written feedback on communication/SPIKES	He identified strengths and areas for improvement, but occasionally misclassified adherence to the protocol.	The design of prompts and the rubric influence the quality of the feedback.
Laverde et al., 2025 (27)	Colombia/ Luxembourg	Chatbot development and evaluation	LLM agents for anamnesis	Real-time feedback and repeatable practice	High usability reported; plausible responses and flexible asynchronous practice.	Strength in scalability; requires external performance validation.
Suárez-García et al., 2025 (28)	Mexico	Pre-post; 30 students	GPT-4o, 10 virtual scenarios	Automatic feedback in natural language	Performance increased 36.7 points; 0% to 70% high performance; post-test with standardized human patients.	Strong transfer signal, but no control group.
Jacobs et al., 2025 (29)	United Kingdom	Mixed methods; 18 participants (15 students)	Virtual patient with LLM and voice	Immediate feedback	High motivation, usability and psychological safety; moderate loyalty.	The feedback was appreciated, but the realism did not match the human patient.

Luo et al., 2025 (30)	China	ECA; n=84	Digital patient with LLM and voice	Adaptive feedback	Improvement of 10.50 points in anamnesis compared to the traditional method; greater empathy.	One of the most robust designs; specialized context in ophthalmology.
Byun et al., 2026 (31)	Korea	Perception study; n=21	GPT-4, 7 virtual patients + virtual evaluator	Immediate reflective dialogue + written report	Greater appreciation of specificity/personalization and self-efficacy compared to conventional methods.	Predominantly perceptual results; requires objective validation.
Borg et al., 2026 (32)	Sweden	Quasi-experimental; n=115	Robotic virtual patient + AI	Automated post-consultation feedback	OSCE: 7.39 vs 6.68; difference 0.70; d=0.74; passed 96.7% vs 79.6%.	Favorable association with independent performance; not all subscales improved.
Sindhvananda et al., 2026 (33)	Thailand	Pre-post mixed methods; n=25	AI-simulated patient for headache	Automatic rubric-based feedback	OSCE +22.5 points (d=1.93); CRI-HT +8.8 (d=1.81); dialogue perceived as robotic.	Large effects, but small sample size and lack of parallel control.
Chan et al., 2026 (34)	Australia	Prospective pilot; 43 volunteers + 8 remediation	Virtual patient with AI, 9 emergencies	Immediate feedback	No clear improvement in volunteers; remediation subgroup improved OSCE 37 to 67 (p=0.03).	Exploratory results; potential for greater utility in specific needs.
Brügge et al., 2024 (18)	Germany	Double-blind RCTs; 21 students	LLM patient; 4 medical records	Structured feedback generated by LLM	Feedback group: 3.60 vs 3.02 in CRI-HTI at the end; F(1,18)=4.44; p=0.049; η^2p=0.198.	Direct experimental evidence of an additional feedback effect, but small sample size.
Wang et al., 2025 (19)	China	Multimetric evaluation; multiple runs per scenario	ChatGPT as a standardized patient and evaluator	Consultation scoring and feedback according to criteria	It differentiated poor consultations from moderate/good ones; prompt engineering reduced score deviation and variability.	It shows sensitivity to prompt and an initial tendency to overscore; useful for studying scoring validity.
Snedegar et al., 2026 (20)	USA	Pilot; 10 student evaluations	Simulated patient encounter + AI ambient listening	Post-meeting feedback and SMART goals	High acceptance and perceived alignment with observable behaviors; more contextualization and examples were requested.	Promising feasibility; very small n and mainly perceptual outcomes.
Nolan and	USA	Validation; 12	SP encounter	Scoring + narrative feedback	Anamnesis: κ=0.84-0.89; ICC=0.93-0.97.	High agreement on

Burke, 2026 (21)	standardized meetings	transcripts evaluated by ChatGPT-4o and Gemini	Presentation: κ=0.62-0.72; ICC=0.61-0.83.	structured elements and lower agreement on complex components; small sample size.

Table 3. Critical assessment by domains of the 17 included primary studies. "Interpretive weight" guides the narrative synthesis and does not constitute a formal certainty score.

Study	Design/ comparator	Independent outcome	Reproducibility/ Tracking	Interpretive weight
Maicher 2019 (22)	Validation with a human referent	No	Repetition by category; no tracking	Intermediate
Holderried 2024 (23)	Prospective with human evaluator	No	Multiple conversations; no follow-up	Intermediate
Yamamoto 2024 (24)	Non-randomized; historical control	Yes, OSCE	Short follow-up	Intermediate
Scherr 2025 (25)	Experimental evaluation	No	360 simulations; no tracking	Intermediate
Weisman 2025 (26)	Multiphase development	No	Repeated executions; without follow-up	Exploratory-intermediate
Laverde 2025 (27)	Uncontrolled development/evaluation	No	Limited external validation	Exploratory
Suárez-García 2025 (28)	Uncontrolled pre-post	Yes, standardized patient	Immediate follow-up	Intermediate
Jacobs 2025 (29)	Mixed methods; small sample	No	Perceptual predominance	Exploratory
Luo 2025 (30)	ECA with comparator	Partial	Short follow-up	Elderly
Byun 2026 (31)	Perceptions; small n	No	No follow-up	Exploratory
Borg 2026 (32)	Quasi-experimental	Yes, OSCE	Immediate follow-up	Intermediate-senior
Sindhvananda 2026 (33)	Pre-post; mixed methods	Yes, OSCE	Small sample; no control	Intermediate
Chan 2026 (34)	Prospective pilot	Yes, OSCE	Small subgroup	Exploratory-intermediate
Brügge 2024 (18)	Double-blind RCTs; small n	No/structured measure	Short follow-up	Elderly
Wang 2025 (19)	Multi-execution validation	Benchmark of quality	Yes; sensitivity to prompt	Intermediate
Snedegar 2026 (20)	Pilot; n=10	No	Perceptual Outcomes	Exploratory
Nolan and Burke 2026 (21)	Validation with evaluators	No	Two LLMs; n=12 meetings	Intermediate

3.3. Validity, consistency and quality of automated feedback

Validation studies show that overall quality can be high for structured tasks, but not uniform. Wang et al. (19) showed that ChatGPT could differentiate poor from moderate/good performance and that prompt engineering markedly reduced score deviation, although the model initially tended to overscore low-quality consultations. Nolan and Burke (21), evaluating transcripts of 12 encounters with standardized patients, observed high agreement for history items (κ=0.84–0.89; ICC=0.93–0.97) and lower agreement for clinical presentation components (κ=0.62–0.72; ICC=0.61–0.83). Among previous studies, Maicher et al. (22) reported automated accuracy of 87%, close to human accuracy (90%), while Holderried et al. (23) found medically plausible answers in more than 99% and near-perfect overall agreement with a human evaluator (κ=0.832), although eight of 45 categories had κ less than 0.6. Overall, the overall agreement does not exclude relevant failures by domain.

The problems are not limited to factual accuracy. Scherr et al. (25), running 360 clinical management simulations, found medical accuracy in basic parameters, but only 78% of the simulations provided comprehensive feedback and 55% adhered to the delayed feedback condition. Weisman et al. (26) described errors in judging adherence to the SPIKES protocol and variability when the same text was assessed in repeated runs. Wang et al. (19) also identified initially excessively high scores and responses that incorporated unsolicited information. These findings suggest that feedback validity should be analyzed by domain, rubric, model version, and intended use, and not inferred from linguistic fluency.

Among the primary studies included, patterns of overscoring, omission of weaknesses, and variability between performances were observed, particularly in Wang et al. (19), Weisman et al. (26), and Nolan and Burke (21). These findings do not amount to a demonstration of general invalidity; they indicate that agreement should be interpreted by domain and that a linguistically plausible output alone does not constitute sufficient evidence for a given evaluative use.

3.4. Educational outcomes and student experience

Most studies that included students reported a favorable experience in terms of availability, psychological safety, personalization, and perceived usefulness. Snedegar et al. (20), in a pilot study with simulated encounters recorded using ambient listening, found that participants perceived feedback as linked to observable behaviors and useful for generating improvement goals, although they requested more contextualization and more specific examples. Jacobs et al. (29) observed high motivation, usability, and psychological safety, despite moderate fidelity from the virtual patient, and Byun et al. (31) found that students valued the virtual evaluator's feedback as specific and personalized. These perceptions are relevant to acceptability but do not replace objective measures of learning.

Studies using objective measures show favorable signs, although with heterogeneous designs. Brügge et al. (18), in a randomized trial of 21 students, found that the group receiving LLM-generated feedback after four simulations showed greater improvement in clinical decision-making performance than the group with simulation but no feedback (partial η^2=0.198), with benefits in context creation and information assurance. Yamamoto et al. (24) reported a significant difference in the preclinical OSCE interview in favor of the group trained with AI-simulated patients. Luo et al. (30) found a 10.50-point improvement in ophthalmological history taking in a randomized trial compared to traditional methods. Sindhvananda et al. (33) observed pre-post increases in OSCE and clinical reasoning after practicing headaches with rubric-based feedback. In contrast, Chan et al. (34) did not identify a clear improvement in the volunteer group, although the small remediation subgroup showed a favorable change.

Borg et al. (32) and Yamamoto et al. (24) assessed performance using standardized stations external to the agent. Sindhvananda et al. (33) used OSCE and clinical reasoning, and Chan et al. (34) explored OSCE-like performance following virtual patient training. However, no localized study demonstrated sustained transfer to clinical outcomes with real patients or long-term retention. This gap limits the strength of any claims about clinical competence beyond the immediate educational setting.

3.5. Transfer to independent evaluations

The most demanding criterion is whether the improvement resulting from AI feedback subsequently appears in an independent evaluation of the tool itself. At least five studies used OSCE, standardized human patients, or another external evaluation as the outcome (24, 28, 32-34). Suárez-García et al. (28), for example, used standardized human patients in the post-test after virtual training, while Borg et al. (32) evaluated using an independent OSCE with a standardized patient and a blinded evaluator. These transfer signals are more informative than the scores generated by the system training the student, although they remain focused on immediate or short-term results.

Borg et al. (32) and Yamamoto et al. (24) also assessed performance using standardized stations external to the agent. Sindhvananda et al. (33) used pre-post OSCE, while Chan et al. (34) explored performance in an OSCE preparation program. These studies bring the evidence closer to educational transfer, but the time horizon remains short. No studies were identified that demonstrated retention over several months, consistent transfer to real-life patient encounters, or effects on safety/quality of care.

Table 4. Synthesis of the strength and gaps of evidence by domain.

Domain	Evidence sign	Recurring limitation	Interpretation for practice
Accuracy/concordance	High agreement on structured tasks and sensitivity to prompt engineering (19,21-23).	Variation by domain, overscoring, unequal specificity and variability between performances (19,21,23,26,35).	Validate by rubric and by domain before using the system.
Student experience	Acceptability and perceived usefulness in several studies, including feedback after recorded simulated encounters (20,29,31).	Moderate fidelity; perceived feedback does not equal valid feedback.	Useful for formative practice, not sufficient to infer competence.
Educational performance	Improvements in reasoning, anamnesis, communication or OSCE in RCTs, controlled studies and pre-post (18,24,28,30,32-33).	Heterogeneous designs, comparators, and scales; small sample sizes in several studies.	Promising as a complement; does not allow estimating an aggregate effect.

Transfer	OSCE or independent standardized patients in a subset (24,28,32-34).	Short follow-up; no results with real patients.	Prioritize external and longitudinal evaluations in future studies.
summative use	No robust evidence was identified for high-impact decisions.	Stability problems, omissions, and contextual judgment.	Maintain human supervision and avoid autonomous certification.

Figura 2. Continuo de madurez de la evidencia sobre retroalimentación automatizada mediante IA. Nota: el esquema representa una síntesis conceptual de los hallazgos. No implica que los dominios sean mutuamente excluyentes.

3.5. Triangulation with recent multibase syntheses

Triangulation with reviews published between 2024 and September 2026 showed general agreement with the patterns observed in the 17 primary studies: predominance of history-taking and communication, favorable early results, high heterogeneity, poor standardization of outcomes, and limited evidence of longitudinal transfer (14-17). The incorporation of searches in ERIC and OpenAlex allowed the identification of additional studies not present in the previous focal synthesis, reinforcing the usefulness of a multi-source strategy for a field that is published in biomedical, educational, and technological journals.

3.6. Contextualization Literature

Secondary reviews and other works not included among the 17 primary studies were used only to contextualize the findings and not to estimate the results of the primary synthesis. These sources inform the pedagogical framework, the evolution of the field, and issues of ethics, governance, implementation, and technological obsolescence (14-17, 35-47, 51-54).

4. Discussion

4.1. Main findings

The studies reviewed reveal a central tension: AI-driven automated feedback can expand learning opportunities and, in certain designs, be associated with objective improvements; however, evidence of validity and effectiveness is task-dependent. Brügge et al. (18) and Borg et al. (32) provide experimental or quasi-experimental signals of benefit, while Wang et al. (19), Holderried et al. (23), Weisman et al. (26), and Nolan and Burke (21) show that evaluative performance varies by domain, prompt, rubric, and behavioral complexity. Therefore, the most consistent finding is not general equivalence with experts, but rather potentially high utility in structured and well-defined tasks. The literature published after the search by Bonilla Mejia et al. (2) reinforces the importance of feedback as a specific subfield. The new studies not only describe virtual patients, but also progressively separate three questions: whether the AI adequately represents the patient, whether it consistently assesses performance, and whether feedback modifies an independent outcome. Brügge et al. (18), Snedegar et al. (20), Borg et al. (32), Sindhvananda et al. (33), and Nolan and Burke (21) illustrate this transition from feasibility to validation and educational outcomes, although the transfer to real clinical practice remains unresolved.

4.2. From generative feedback to pedagogically useful feedback

The quality of feedback does not depend solely on its personalization. Frameworks introduced earlier help explain why some systems produce useful results in structured tasks but are less consistent in complex competencies. Deliberate practice requires clear objectives, repeated execution, and specific feedback that allows for error correction (8). AI can reduce the time between performance and feedback and facilitate multiple attempts, but these operational advantages do not automatically translate into educational improvement if the information is imprecise, difficult to prioritize, or does not lead to a new opportunity for practice. The findings of Holderried et al. (23), Wang et al. (19), and Weisman et al. (26) illustrate why explicit rubrics and prompt control can act as calibration mechanisms. When the system must decide whether a specific behavior occurred, agreement can be high; when judgment requires integrating tone, context, empathy, or clinical priorities, consistency decreases. Technical calibration, therefore, should be considered a condition for interpreting feedback and not a guarantee of complete educational validity. Cognitive load theory provides a second interpretive lens. Immediate feedback can facilitate learning, but lengthy, non-prioritized, or contradictory feedback can increase extrinsic load. Cognitive load frameworks in medical education and simulation recommend adjusting complexity and support to the learner's level (9,10). The included studies did not sufficiently compare different doses or formats of generative feedback; therefore, limiting and prioritizing recommendations should be understood as a reasonable pedagogical hypothesis that requires specific experimental evaluation.

4.3. Transfer should be the maturity criterion

A significant gap in the localized studies is the limited evidence of transfer. Previous reviews on virtual patients have shown that outcomes are usually concentrated on knowledge, immediate skills, or perception (36,37). In the localized evidence, some studies are moving toward OSCE or independent standardized patients (24,28,32-34), but follow-up remains short. The lack of

longitudinal results with real patients prevents us from assuming that an improvement during simulation or standardized assessment will be maintained subsequently in clinical practice.

However, even an immediate OSCE remains an intermediate outcome. From a situated learning perspective, clinical competence emerges from the interaction between person, task, and context (38). Therefore, performance in a digital environment should not be assumed to be equivalent to sustained performance with real patients. Future studies could strengthen this inference through longitudinal follow-up, independent observer assessment, and supervised clinical outcomes when methodologically and ethically appropriate.

4.4. Implementation, costs and equity

Scalability is one of the most frequently cited reasons for automating feedback. Conversational agents and transcription systems can multiply practice opportunities and reduce some of the reliance on teacher observation. However, focus group studies have rarely incorporated comparative economic analyses. The literature on simulation costs shows that sustainability depends on infrastructure, licensing, maintenance, case development, and teacher time, in addition to the price of the technology (39). Therefore, the reviewed evidence allows us to consider scalability as a plausible operational hypothesis, not as proven cost-effectiveness. In institutions with limited resources, text-based systems might require less infrastructure than complex immersive experiences, but this possibility depends on connectivity, licensing, language availability, and local support. Equity should be considered as a domain to be evaluated, not as an assumed benefit. Localized studies provide little comparative evidence on feedback performance across linguistic, cultural, or socioeconomic groups.

4.5. Ethics, governance and use in evaluation

The use of AI for academic feedback raises issues of privacy, bias, opacity, and accountability. Masters (40) has proposed principles of transparency, human oversight, data protection, and attention to bias in health professions education. In this review, these principles are used as an interpretive framework for risk assessment, not as recommendations causally derived from the included studies. The strength of the safeguards should be related to the intended use: a system for low-impact formative practice places different demands on one used for progression or certification decisions. The evidence found is more compatible with formative use than with autonomous replacement of human assessment. This interpretation is supported by the observed concordance limitations and by the principles of validity and programmatic evaluation: a single data point may be useful for guiding learning, while high-impact decisions require a broader evidence base and multiple sources of information (11-13). Consequently, the review does not formulate a normative prohibition of summative use; it notes that the studies found do not yet provide sufficient evidence to justify high-impact decisions based solely on automated output.

4.6. Pedagogical quality, validity and calibration of use

The literature on feedback provides pedagogical criteria for interpreting AI-generated outputs. Hattie and Timperley noted that the effect of feedback depends on its content, level, and delivery method, while van de Ridder et al. defined it in clinical education as specific information that compares observed performance with a standard with the intention of improving it (6,7). In simulation, structured debriefing can transform the experience into learning (41,42). These frameworks, incorporated from the introduction, allow us to distinguish three questions that should not be confused: whether the generated text is correct, whether the feedback is pedagogically actionable, and whether the judgment is valid for the proposed evaluative use.

From an evaluation perspective, the usefulness of an automated system depends on the intended use of its judgments. The Ottawa criteria include validity or consistency, reproducibility, equivalence, feasibility, educational effect, and acceptability (11). Kane's argumentative approach emphasizes that it is the interpretation and use of an outcome that is validated, not the score in the abstract (12). Programmatic evaluation adds that progression decisions are strengthened by integrating multiple points of information and expert judgment (13). Consequently, agreement with a human evaluator is a source of validity evidence, but it does not exhaust the justification necessary for summative uses.

Table 4. Operational risks of automated feedback and suggested safeguards.

Risk	Observed/possible manifestation	Safeguard	Audit indicator	Basis of the safeguard
Inaccurate feedback	Omissions, factual errors, over-scoring or incorrect judgment of adherence (19,26,35).	Explicit rubric + validation with teachers before use.	% agreement with human evaluators by category.	Empirical + prudential
Variability	Different evaluations of the same text or repeated executions (19,26,35).	Model versioning, temperature control where possible, and repeatability testing.	Coefficient of agreement/reproducibility in anchor cases.	Empirical + technical
Excessive positivity	Underestimation of relevant errors or weaknesses; tendency to high scores without strict criteria (19,35).	Instructions for identifying strengths and gaps; scaling thresholds.	Proportion of critical errors omitted.	Empirical + prudential
Communication bias	Potential penalty for different linguistic/cultural styles.	Testing with diverse groups and equity review.	Systematic differences by subgroups or linguistic variant.	Prudential/ ethical
Privacy	Uploading conversations or sensitive data to external services.	Data minimization, institutional agreements, and approved platforms.	Incidents, policy compliance, and data processing traceability.	Prudential/ ethical
Premature summative use	Academic decisions based on an unvalidated output.	Human supervision and independent confirmation for high impact.	% of summative decisions confirmed by human evaluator.	Prudential/ validity

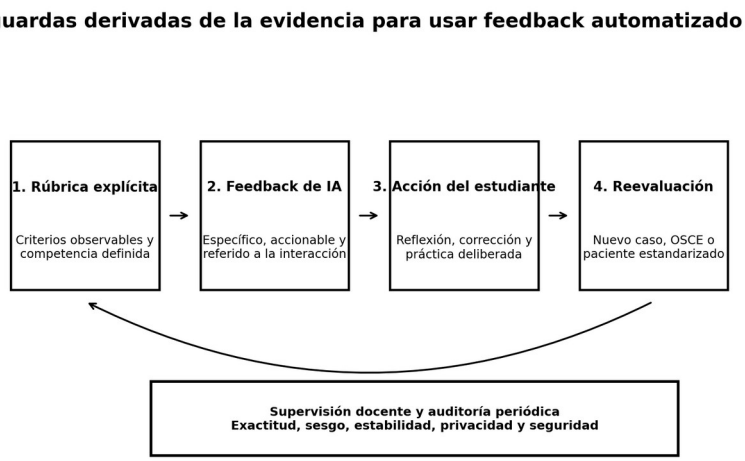

Figura 3. Salvaguardas derivadas de la evidencia para integrar retroalimentación automatizada en simulación clínica. La figura sintetiza principios de uso formativo derivados de los estudios revisados y de marcos de práctica deliberada, carga cognitiva y gobernanza ética; no representa un instrumento validado.

This interpretation aligns with contemporary syntheses. Feigerlova et al. found heterogeneous evidence on the educational outcomes of AI interventions and limited assessment of transfer to the clinical setting (43). A scope-specific review of LLM-generated feedback at the undergraduate level identified a predominance of intermediate results and variable accuracy (44). Furthermore, recent reviews of AI-assisted virtual patients converge on the finding that potential benefits are concentrated in structured practice, interviewing, communication, and reasoning, while uncertainties persist regarding authenticity, standardization of outcomes, and transferability (14-17). Additional experimental studies show that ChatGPT feedback can approximate expert feedback in some tasks, but not uniformly, and that simple warnings about errors do not necessarily modify student behavior (45-47).

4.7. Limitations of the review

This review has significant limitations. First, although the identification included PubMed/MEDLINE, ERIC, and OpenAlex, and coverage was compared with multi-database reviews, a direct systematic search of all subscription databases was not performed; the risk of omission remains. Second, the historical search log did not retain a complete exportable count by source, duplicates, and exclusions, making it impossible to retrospectively reconstruct an auditable quantitative flow without introducing unverifiable figures. Third, although the initial selection and extraction were performed by the lead author and verified by a second author, the process did not amount to independent, duplicate screening from the identification stage. Fourth, the heterogeneity of expertise, designs, and metrics precluded a quantitative synthesis. Finally, rapid technological obsolescence limits the temporal portability of the findings: performance may depend on the specific model version, vendor changes, prompts, inference parameters, and post-publication updates. Consequently, results obtained with one specific version should not be automatically extrapolated to future versions or other systems.

For these reasons, the manuscript should be interpreted as a multi-source integrative critical review and not as a comprehensive assessment of the state of the evidence. Its conclusions are limited to patterns observed in the identified studies and compared with recent syntheses. A future systematic review of effectiveness should incorporate direct searches of multiple subscription databases, pre-protocol testing, independent duplicate screening, and, where sufficient homogeneity exists, meta-analysis. It would also be useful to agree on a minimum set of outcomes that includes accuracy by experts, reproducibility, change in performance, transferability to independent assessment, retention, cognitive load, equity, and costs.

5. Conclusions

- Among the 17 studies located, favorable results were observed in some structured training and clinical tasks, but the magnitude and consistency of those results varied according to design, domain and comparator.
- Performance was not uniform. Overscoring, specificity errors, omissions, variability between runs, and lower agreement in complex domains were described; therefore, linguistic fluency and evaluative validity should not be considered equivalent.
- The most interpretive evidence comes from randomized trials and studies with OSCE or independent standardized patients; however, the samples are usually small or moderate, the follow-up is short, and there is insufficient evidence on retention or performance with real patients.
- The frameworks of feedback, deliberate practice, cognitive load, validity, and programmatic evaluation help explain why the usefulness of AI depends on the task and its intended use. Current evidence more clearly supports formative applications than high-impact, autonomous, summative decisions.

Funding: There has been no funding.

Declaration of conflict of interest: The authors declare that they have no conflict of interest.

Authors' contributions: AMTM: conceptualization, methodology, research, initial literature search and selection, data extraction and analysis, visualization, drafting of the original version, revision and editing, and approval of the final version. ADGC: methodology, validation, independent verification of the eligibility of included studies and of data extraction, critical analysis, revision and editing of the manuscript, and approval of the final version.

Declaration on the use of artificial intelligence: A generative artificial intelligence tool was used as an auxiliary support for the initial structuring of the manuscript, organization of search terms, and linguistic review. The authors verified the bibliographic sources, critically reviewed the syntheses, and assume full responsibility for the content, interpretation, and final version of the manuscript. No confidential participant data or identifiable medical records were used.

6. References

1. Fierro Manrique YV, Arteaga Losada MS, Anduquia Manzano C, Escalante Ochoa AM, Novoa Álvarez RA, Charry Cuellar JD. High-fidelity clinical simulation with acting and virtual simulation in a metaverse as strategies for the training of future physicians. *Rev Esp Edu Med*. **2026,** 5, 721571. https://revistas.um.es/edumed/article/view/721571 . https://doi.org/10.6018/edumed.721571

2. Bonilla Mejia JJ, Cortés Fuenzalida TD, Polanco Aliaga DH, Martínez Carrillo C, Herrera Alcaíno AA. Artificial intelligence in the practical clinical training of undergraduate medical students: a review of the scope of applications, results and gaps. *Rev Esp Edu Med.* **2026,** 7(3), 710071. https://doi.org/10.6018/edumed.710071

3. Thind BS, Javidi D, Schwartz LM. Artificial intelligence in undergraduate medical education clinical skills curricula: a scoping review of implementations since 2022. *Front Digit Health.* **2026,** 8, 1830254. https://doi.org/10.3389/fdgth.2026.1830254

4. Jiang J, Ye MZ, Kwok TTO, Wong JYH. GenAI-Supported Virtual Patients in Health Care Education: Systematic Review. *J Med Internet Res.* **2026,** 28, e82756. https://doi.org/10.2196/82756
5. Loubbairi S, El Moussaoui Y, Lahlou L, Chakri I, Nassik H. The impact of artificial intelligence-driven simulation on the development of non-technical skills in medical education: a systematic review. *J Educ Eval Health Prof.* **2025,** 22, 37. https://doi.org/10.3352/jeehp.2025.22.37
6. Hattie J, Timperley H. The Power of Feedback. *Rev Educ Res.* **2007,** 77(1), 81-112. https://doi.org/10.3102/003465430298487
7. van de Ridder JMM, Stokking KM, McGaghie WC, ten Cate OTJ. *What is feedback in clinical education? Med Educ.* **2008,** 42(2), 189-197. https://doi.org/10.1111/j.1365-2923.2007.02973.x
8. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med.* **2011,** 86(6), 706-711. https://doi.org/10.1097/ACM.0b013e318217e119
9. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* **2014,** 36(5), 371-384. https://doi.org/10.3109/0142159X.2014.889290
10. Fraser KL, Ayres P, Sweller J. Cognitive Load Theory for the Design of Medical Simulations. *Simul Healthc.* **2015,** 10(5), 295-307. https://doi.org/10.1097/SIH.0000000000000097
11. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *MedTeach.* **2011,** 33(3), 206-214. https://doi.org/10.3109/0142159X.2011.551559
12. Kane MT. Validating the Interpretations and Uses of Test Scores. *J Educ Meas.* **2013,** 50(1), 1-73. https://doi.org/10.1111/jedm.12000
13. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, van Tartwijk J. A model for programmatic assessment fit for purpose. *MedTeach.* **2012,** 34(3), 205-214. https://doi.org/10.3109/0142159X.2012.652239
14. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review. *Australas J Educ Technol.* **2024,** 40(3), 39-57. https://doi.org/10.14742/ajet.9307
15. Gilbert V, Philip P, Llorca PM, Samalin L. Use of artificial intelligence-enhanced virtual patients in educational approaches to medical interview training: a systematic review. *BMC Med Educ.* **2026,** 26, 588. https://doi.org/10.1186/s12909-026-08804-9
16. Fang EKM, Boon Y, Ho JY, Tey N, Low MJW, et al. Artificial intelligence generated virtual patients in health professions education: a scoping review. BMC Med Educ. 2026. https://doi.org/10.1186/s12909-026-10028-w
17. Mofidi SA, Khoshgoftar Z. Effectiveness of AI-enhanced virtual patients in clinical reasoning training of healthcare professionals and students: a systematic review. BMC Med Educ. 2026. https://doi.org/10.1186/s12909-026-10242-6
18. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, Holling M, Lu MH, Darici D. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med Educ.* **2024,** 24, 1391. https://doi.org/10.1186/s12909-024-06399-7
19. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, Liu D, Tan *J Med Internet Res.* **2025,** 27, e59435. https://doi.org/10.2196/59435
20. Snedegar R, Unger K, Craig JF, Kozlowski L, Carpenter D, Pyles E, Williamson J, Bodkins E, Williams D. Ambient Listening Devices as a Feedback Tool in the Simulated Learning Environment. *Cureus.* **2026,** 18(2), e103222. https://doi.org/10.7759/cureus.103222
21. Nolan EJ, Burke HB. Large Language Model Evaluation and Feedback of Transcribed Simulated Physician-Patient Verbal Interactions. *Front Artif Intell.* **2026,** 9, 1936749. https://doi.org/10.3389/frai.2026.1936749

22. Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. *MedTeach.* **2019,** 41(9), 1053-1059. https://doi.org/10.1080/0142159X.2019.1616683
23. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, Mahling M. A Language Model-Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR Med Educ.* **2024,** 10, e59213. https://doi.org/10.2196/59213
24. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, Ino H. Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial. *JMIR Med Educ.* **2024,** 10, e58753. https://doi.org/10.2196/58753
25. Scherr R, Spina A, Dao A, Andalib S, Halaseh FF, Blair S, Wiechmann W, Rivera R. Novel Evaluation Metric and Quantified Performance of ChatGPT-4 Patient Management Simulations for Early Clinical Education: Experimental Study. *JMIR Form Res.* **2025,** 9, e66478. https://doi.org/10.2196/66478
26. Weisman D, Sugarman A, Huang YM, Gelberg L, Ganz PA, Comulada WS. Development of a GPT-4-Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study. *JMIR Form Res.* **2025,** 9, e65670. https://doi.org/10.2196/65670
27. Laverde N, Grévisse C, Jaramillo S, Manrique R. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. *Comput Struct Biotechnol J.* **2025,** 27, 2481-2491. https://doi.org/10.1016/j.csbj.2025.05.025
28. Suárez-García RX, Chavez-Castañeda Q, Orrico-Pérez R, et al. DIALOGUE: A Generative AI-Based Pre-Post Simulation Study to Enhance Diagnostic Communication in Medical Students Through Virtual Type 2 Diabetes Scenarios. *Eur J Investig Health Psychol Educ.* **2025,** 15(8), 152. https://doi.org/10.3390/ejihpe15080152
29. Jacobs C, Johnson H, Tan N, Brownlie K, Joiner R, Thompson T. Application of AI Communication Training Tools in Medical Undergraduate Education: Mixed Methods Feasibility Study Within a Primary Care Context. *JMIR Med Educ.* **2025,** 11, e70766. https://doi.org/10.2196/70766
30. Luo MJ, Bi S, Pang J, et al. A large language model digital patient system enhances ophthalmology history taking skills. *NPJ Digit Med.* **2025,** 8(1), 502. https://doi.org/10.1038/s41746-025-01841-6
31. Byun J, Kim H, Lim J, Choi J, Ahn S. Enhancing history-taking education through GPT-4-based virtual patients and automated assessment: a study of medical student perceptions. *Korean J Med Educ.* **2026,** 38(1), 64-73. https://doi.org/10.3946/kjme.2025.108
32. Borg A, Schiött J, Ivegren W, et al. AI-generated Feedback Following Social Robotic Virtual Patient Interactions and Medical Student Performance: Nonrandomized Quasi-Experimental Study. *JMIR Med Educ.* **2026,** 12, e90368. https://doi.org/10.2196/90368
33. Sindhvananda K, Ruengwattanachot K, Leksuwankun S, et al. AI-powered simulated patients with automated feedback for enhancing headache history-taking skills: a convergent mixed-methods study. *BMC Med Educ.* **2026,** 26, 1151. https://doi.org/10.1186/s12909-026-09511-1
34. Chan BS, Marjot J, Dodds T, et al. Enhancing objective structured clinical examination performance through an artificial intelligence virtual patient: a proof-of-concept study. *Proc (Bayl Univ Med Cent).* **2026,** 1-4. https://doi.org/10.1080/08998280.2026.2699604
35. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback. *J Med Internet Res.* **2025,** 27, e68486. https://doi.org/10.2196/68486
36. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. *Med Educ.* **2020,** 54(9), 786-795. https://doi.org/10.1111/medu.14152

37. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual Patient Simulations in Health Professions Education: Systematic Review and Meta-Analysis by the Digital Health Education Collaboration. *J Med Internet Res.* **2019,** 21(7), e14676. https://doi.org/10.2196/14676
38. Durning SJ, Artino AR Jr. Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide No. 52. *Med Teach.* **2011,** 33(3), 188-199. https://doi.org/10.3109/0142159X.2011.550965
39. Hippe DS, Umoren RA, McGee A, Bucher SL, Bresnahan BW. A targeted systematic review of cost analyzes for implementation of simulation-based education in healthcare. *SAGE Open Med.* **2020,** 8, 2050312120913451. https://doi.org/10.1177/2050312120913451
40. Masters K. Ethical use of Artificial Intelligence in Health Professions Education: AMEE Guide No. 158. *Med Teach.* **2023,** 45(6), 574-584. https://doi.org/10.1080/0142159X.2023.2186203
41. Cantillon P, Sargeant J. Giving feedback in clinical settings. *BMJ.* **2008,** 337, a1961. https://doi.org/10.1136/bmj.a1961
42. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. *Med Educ.* **2014,** 48(7), 657-666. https://doi.org/10.1111/medu.12432
43. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ.* **2025,** 25, 129. https://doi.org/10.1186/s12909-025-06719-5
44. Kıyak YS, İş-Kara T, Emekli E. Applications and Outcomes of Large-Language-Model-Generated Feedback in Undergraduate Medical Education: A Scoping Review. *Med Sci Educ.* **2026,** 36(1), 81-99. https://doi.org/10.1007/s40670-025-02621-3
45. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. *Postgrad Med J.* **2025,** 101(1199), 458-463. https://doi.org/10.1093/postmj/qgae170
46. Kıyak YS, Emekli E, İş-Kara T, Coşkun Ö, Budakoğlu İİ. AI Teaches Surgical Diagnostic Reasoning to Medical Students: Evidence from an Experiment Using a Fully Automated, Low-Cost Feedback System. *J Surg Educ.* **2025,** 82(10), 103639. https://doi.org/10.1016/j.jsurg.2025.103639
47. Kıyak YS, Coşkun Ö, Budakoğlu İİ. 'ChatGPT can make mistakes' warnings fail: A randomized controlled trial. *Med Educ.* **2026,** 60(2), 138-142. https://doi.org/10.1111/medu.70056
54. Elizondo-García J. Vibe-coding as a teaching competence in the use of Artificial Intelligence for Medical Education: a scope review. *Rev Esp Edu Med.* **2026,** 7(2). https://revistas.um.es/edumed/article/view/705121, https://doi.org/10.6018/edumed.705121
53. Kaya AB, Emekli E, Kiyak YS. Mapping applications and case outcomes generated by wide language models in health professions education: a scoping review. *Rev Esp Edu Med.* **2026,** 7(1). https://revistas.um.es/edumed/article/view/691691, https://doi.org/10.6018/edumed.691691
52. Jiménez Salcedo SC, Granda Cruz CA. Use of ChatGPT as a pedagogical tool in medical education: a systematic literature review (2020-2025). *Rev Esp Edu Med.* **2026,** 7(4). https://doi.org/10.6018/edumed.716681
51. Canabal Berlanga A, Sánchez Giralt JA, Abad Santamaría B. Current trends in medical education with generative artificial intelligence. *Rev Esp Edu Med.* **2025,** 6(6). https://doi.org/10.6018/edumed.684461
50. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ Inf.* **2018,** 34(4), 285-291. https://doi.org/10.3233/EFI-180221
49. Snyder H. Literature review as a research methodology: An overview and guidelines. *J Bus Res.* **2019,** 104, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039
48. Whittemore R, Knafl K. The integrative review: updated methodology. *J Adv Nurs.* **2005,** 52(5), 546-553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

7. Annex

Annex 1. Reproducible search strategies used in the integrative multi-source critical review

PubMed/MEDLINE (updated September 3, 2026): (("medical student"[Title/Abstract] OR "medical students"[Title/Abstract] OR "medical education"[Title/Abstract]) AND ("virtual patient"[Title/Abstract] OR "simulated patient"[Title/Abstract] OR "standardized patient"[Title/Abstract] OR "clinical simulation"[Title/Abstract] OR simulation[Title/Abstract]) AND ("artificial intelligence"[Title/Abstract] OR "generative AI"[Title/Abstract] OR GPT[Title/Abstract] OR "large language model"[Title/Abstract]) AND (feedback[Title/Abstract] OR "automated assessment"[Title/Abstract] OR scoring[Title/Abstract] OR evaluation[Title/Abstract])).

ERIC (updated September 3, 2026): ("medical students" OR "medical education") AND ("virtual patient" OR "simulated patient" OR "standardized patient" OR "clinical simulation") AND ("artificial intelligence" OR "large language model" OR GPT OR "generative AI") AND (feedback OR assessment OR scoring OR evaluation). The search was performed on the bibliographic fields available in the interface, without a full-text filter. OpenAlex (updated September 3, 2026): search by terms in title/abstract/concepts using combinations of "AI-generated feedback", "automated feedback", "medical education", "virtual patient", "standardized patient", "clinical simulation", "large language model", GPT, assessment, and scoring; results were verified by title and DOI, and backward and forward citation relationships were tracked. Publications in English or Spanish were considered, and no full-text availability filter was applied. The procedure was complemented by reference tracing and comparison with recent multi-base reviews (14-17).