Supplementary Information

Contents

List of Supplementary Tables

List of Supplementary Figures

S1 Models and conversational platforms reported in the included studies

Model use was strongly concentrated in the OpenAI ecosystem (Table S1). OpenAI models were reported in 86 articles, compared with Anthropic Claude in seven, DeepSeek in six, Meta Llama in five, and Google Gemini in four. Zhipu GLM appeared in three articles, while Google Gemma, Alibaba Qwen, Baichuan, and Baidu ERNIE each appeared in two. Microsoft Phi, Mistral, Naver HyperCLOVA, and xAI Grok were each represented in one article. A further 12 articles used proprietary conversational platforms for which the underlying foundation model was not consistently identifiable, and six described an LLM- or AI-based system without reporting a specific model family. Family-level counts were non-mutually exclusive because some articles evaluated or incorporated multiple models.

Considerable variation was also present within model families. The most frequently reported OpenAI variants were GPT-4o in 27 articles, GPT-4 in 21, GPT-3.5 in 13, GPT-3.5 Turbo in 11, and GPT-4 Turbo in six. Other studies used smaller, multimodal, real-time, reasoning, or newer-generation variants, including GPT-4o mini, GPT-4 Vision, GPT Realtime, OpenAI o1 Pro, and GPT-5 variants. Claude 3.5 Sonnet was the most frequently identified Anthropic model. DeepSeek use included R1, V2.5, V3, and unspecified versions, while the open-weight or locally deployable model landscape included multiple Llama, Gemma, Qwen, Baichuan, Phi, and Mistral variants. Most of these alternatives were represented by only one or two articles, indicating that evidence for individual non-OpenAI models remained sparse.

The precision of model reporting varied substantially. Some articles reported an exact model snapshot, parameter scale, or fine-tuned version, whereas others identified only a broad product or platform name, such as ChatGPT, Claude Sonnet, Dialogflow, or LLM. This distinction is consequential because model behavior may differ across versions, access routes, hosting environments, and subsequent provider updates. Future studies should therefore report the provider, model family, exact variant or snapshot, access and hosting configuration, any retrieval or fine-tuning procedures, and the model's role within the system, such as simulated patient, assessor, feedback generator, or benchmark. Table S1 preserves the specific variants reported in the included corpus while consolidating obvious aliases of the same platform.

Supplementary Table S1 | Model families, conversational platforms, and specific variants reported in generative AI-powered simulated-patient studies.

Model family or platform	Specific variant or reported label	Unique articles	Representative papers
OpenAI models	*All reported variants*	**86**	
	GPT-4o	27	[1–3]
	GPT-4	21	[4–6]
	GPT-3.5	13	[7–9]
	GPT-3.5 Turbo	11	[10–12]
	ChatGPT, version not reported	7	[13–15]
	GPT-4 Turbo	6	[16–18]
	GPT-4o mini	5	[19–21]
	GPT-5	2	[22,23]
	OpenAI GPT, version not reported	2	[24,25]
	Custom GPT or GPTs, base version not reported	1	[26]
	GPT Realtime	1	[23]
	GPT Realtime Mini	1	[23]
	GPT-4 (0613)	1	[27]
	GPT-4 Vision	1	[28]
	GPT-4o (2024-08-06)	1	[29]
	GPT-4o (2024-11-20)	1	[30]
	GPT-4o mini (2024-07-18)	1	[30]
	GPT-5 Pro	1	[31]
	GPT-5.2	1	[32]
	OpenAI o1 Pro	1	[31]
Anthropic Claude	*All reported variants*	**7**	*Family-level total*
	Claude 3.5 Sonnet	4	[33–35]
	Claude 3 Sonnet	2	[35,36]
	Claude Sonnet, version not reported	2	[28,35]
	Claude 3 Haiku	1	[35]
	Claude 3 Opus	1	[36]

Continued on the next page.

Table S1 continued from the previous page.

Model family or platform	Specific variant or reported label	Unique articles	Representative papers
	Claude 3.7 Sonnet	1	[37]
	Claude 4 Opus	1	[35]
	Claude 4 Sonnet	1	[35]
DeepSeek	*All reported variants*	**6**	*Family-level total*
	DeepSeek, version not reported	2	[38,39]
	DeepSeek R1	2	[40,41]
	DeepSeek V2.5	1	[42]
	DeepSeek V3 671B	1	[35]
Meta Llama	*All reported variants*	**5**	*Family-level total*
	Llama, version not reported	1	[43]
	Llama 2, size not reported	1	[43]
	Llama 2 7B Chat	1	[44]
	Llama 3 70B	1	[35]
	Llama 3.1 8B	1	[45]
	TinyLlama 1.1B	1	[46]
Google Gemini	*All reported variants*	**4**	*Family-level total*
	Gemini, version not reported	1	[47]
	Gemini 2.5 Flash	1	[45]
	Gemini 2.5 Pro	1	[34]
	Gemini 3	1	[39]
Google Gemma	*All reported variants*	**2**	*Family-level total*
	Gemma 3 1B	1	[46]
	Gemma 3 27B	1	[48]
	Gemma 3 27B, fine-tuned	1	[48]
Alibaba Qwen	*All reported variants*	**2**	*Family-level total*
	Qwen-Max (2025-04-09)	1	[42]
	Qwen3 32B	1	[35]
Zhipu GLM	*All reported variants*	**3**	*Family-level total*
	ChatGLM, version not reported	2	[43,49]

Continued on the next page.

Table S1 continued from the previous page.

Model family or platform	Specific variant or reported label	Unique articles	Representative papers
	GLM-4	1	[50]
Baichuan	*All reported variants*	**2**	*Family-level total*
	Baichuan, version not reported	1	[43]
	Baichuan 13B Chat	1	[51]
Baidu ERNIE	*All reported variants*	**2**	*Family-level total*
	ERNIE 3.5-8K	1	[52]
	ERNIE 4.0	1	[53]
Microsoft Phi	*All reported variants*	**1**	*Family-level total*
	Phi-3 Mini 3.8B	1	[45]
Mistral	*All reported variants*	**1**	*Family-level total*
	Mistral, version not reported	1	[25]
Naver HyperCLOVA	*All reported variants*	**1**	*Family-level total*
	HyperCLOVA X	1	[54]
xAI Grok	*All reported variants*	**1**	*Family-level total*
	Grok 3	1	[36]
Other proprietary conversational platforms	*All reported platforms*	**12**	*Group-level total*
	Google Dialogflow, version not reported	4	[55–57]
	Dialogflow CX	1	[58]
	Character.AI	1	[59]
	Convai	1	[60]
	Danbee	1	[61]
	IBM Watson Assistant	1	[62]
	Microsoft Power Virtual Agents	1	[63]
	NotebookLM	1	[64]
	SimConverse	1	[65]
Unspecified LLM or AI model	*All incompletely specified systems*	**6**	*Group-level total*
	LLM, model not specified	3	[66–68]

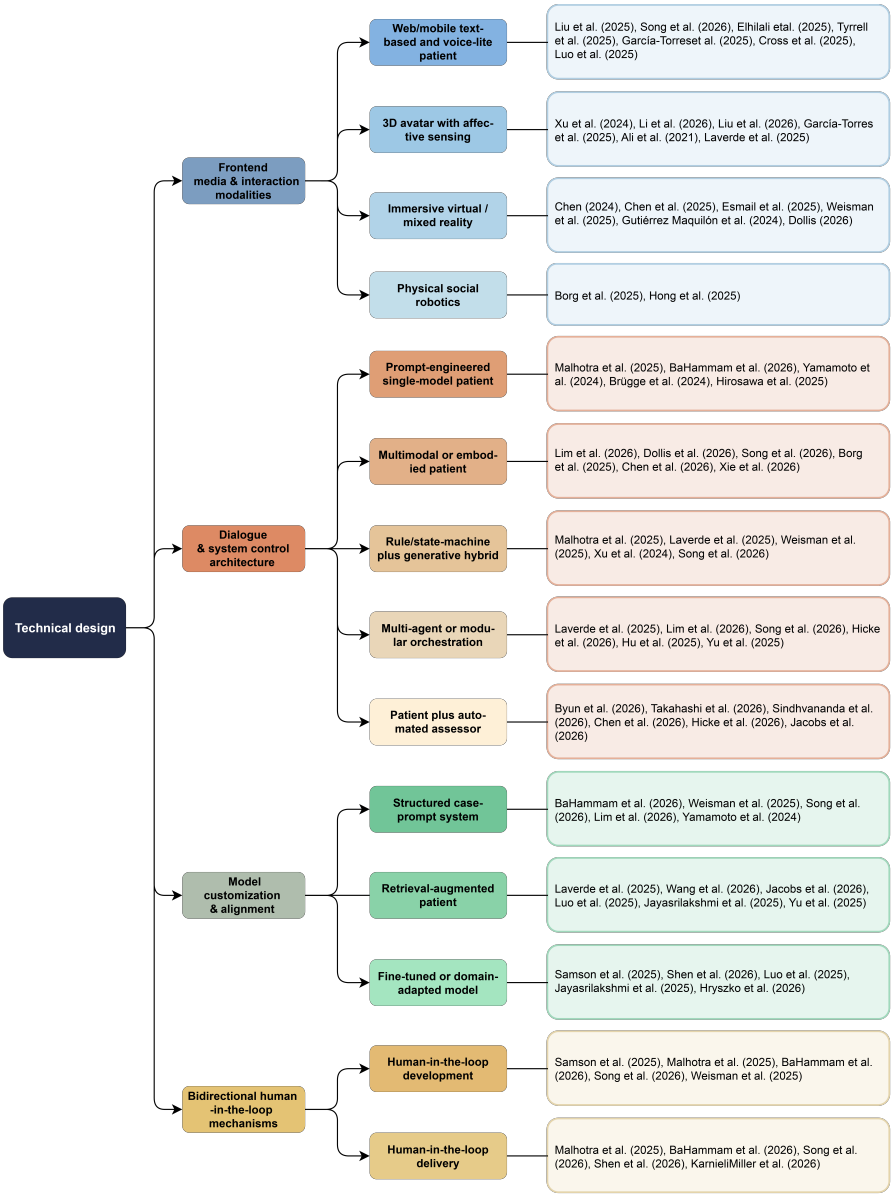
Continued on the next page.

Table S1 continued from the previous page.

Model family or platform	Specific variant or reported label	Unique articles	Representative papers
	AI-based virtual standardized-patient system	1	[69]
	LLMDP	1	[51]
	AI-based medical interview-assistance system	1	[70]

S2 Technical-design archetypes and representative studies

This subsection provides the supporting framework used to interpret the technical-design taxonomy in the main manuscript and to trace the taxonomy to its underlying evidence base.



Supplementary Fig. S1 | Taxonomy of technical design archetypes in generative AI-powered simulated patients. This hierarchical diagram maps the structural components of modern simulated-patient systems across four primary functional planes: Frontend Media & Interaction Modalities, Dialogue & System Control Architecture, Model Customization & Alignment, and Bidirectional Human-in-the-Loop Mechanisms. Each overarching plane is subdivided into specific technical and pedagogical archetypes, illustrating the diverse engineering strategies used to decouple system responsibilities. The leaf nodes link each archetype to representative empirical studies, demonstrating their practical implementation in clinical education literature.

Supplementary Table S2 | Technical and pedagogical archetypes displayed in the technical-design taxonomy, including their purpose, typical implementation, and representative studies.

Architectural plane	Technical archetype	Purpose	Typical implementation	Representative studies shown in the figure
Frontend media and interaction modalities	**Web/mobile text-based and voice-lite patient**	Provides accessible and scalable platforms for asynchronous or semi-synchronous history-taking and communication practice.	Learners interact through browser- or mobile-based text chat, speech recognition, text-to-speech, or lightweight combinations of these modalities. Systems commonly use standard web frontends with server-side dialogue generation. Brief textual descriptions may be appended to convey simple non-verbal behaviors.	[23,25,36,42,51,71,72].
	3D avatar with affective sensing	Visualizes patient personas through digital humans that display facial expressions, non-verbal cues, and emotional states.	A conversational model is connected to a screen-based 3D avatar. Lip synchronization coordinates speech with mouth movements, while gaze, facial expression, posture, or sentiment-adaptive layers represent pain, sadness, frustration, or other patient states. Some systems also analyze learner facial or behavioral signals.	[25,53,62,73–75].
	Immersive virtual/mixed reality	Recreates spatial clinical environments and supports interaction with situational, procedural, or time-sensitive scenarios.	Clinical encounters are delivered through virtual-reality headsets or mixed-reality environments with spatial audio, hand or gesture tracking, and perspective-dependent visual cues. Some implementations align virtual information, wounds, or physiological changes with physical mannequins to combine visual immersion with tactile interaction.	[10,41,47,76,77].
	Physical social robotics	Embodies the simulated patient in a shared physical environment and supports the study of gaze, presence, and mechanical-social behaviors.	A generative dialogue system is connected to a physical social robot. Facial textures, speech, gaze direction, head movements, and other behaviors are coordinated to create a physically present patient interaction. Onboard cameras or tracking systems may support eye contact and orientation toward the learner.	[11,78,79].
Dialogue and system-control architectures	**Prompt-engineered single-model patient**	Converts a general-purpose language model into a simulated patient without modifying its model weights.	A single model receives instructions defining the patient's identity, history, symptoms, emotional state, communication style, disclosure rules, and knowledge boundaries. Implementations may include few-shot examples, conversation history, response-length constraints, and instructions to remain in character. The same model may also generate post-encounter feedback.	[9,16,26,30,80].

Continued on the next page.

Table S2 continued from the previous page.

Architectural plane	Technical archetype	Purpose	Typical implementation	Representative studies shown in the figure
	Multimodal or embodied patient	Supports training that combines verbal dialogue with visual, non-verbal, spatial, or physically embodied communication.	The dialogue system is connected to speech recognition, text-to-speech, avatars, facial-expression models, virtual or augmented reality, social robots, clinical images, video, or haptic interfaces. Some systems coordinate speech with gaze, gesture, personality, emotional expression, or environmental changes.	[11,23,28,38,77,78,81].
	Rule/state-machine plus generative hybrid	Combines scripted control and reproducibility with flexible generative conversation.	Rules, branching logic, or explicit encounter states govern case progression, information disclosure, scoring, turn limits, stage transitions, and termination. The model generates dialogue only within the permitted state. Rule-based modules may also identify required questions, update the patient state, calculate checklist scores, or redirect off-case responses.	[23,30,47,62,75].
	Multi-agent or modular orchestration	Separates dialogue, memory, retrieval, safety, assessment, and interface functions into specialized components.	Role-specific agents or modules perform patient dialogue, planning, memory retrieval, emotional-state generation, consistency checking, moderation, assessment, feedback, speech processing, or avatar animation. Components may operate sequentially, share a transcript or structured patient state, or be coordinated through a supervisor, router, or fixed workflow.	[23,28,35,75,82,83].
	Patient plus automated assessor	Connects simulated-patient practice with scalable scoring, feedback, or debriefing.	The encounter transcript is evaluated by the patient model, a separate assessor model, a rule-based checklist, or a hybrid pipeline. Systems may assess history taking, communication, empathy, reasoning, diagnosis, management, counseling, or safety-netting and return rubric scores, missed questions, transcript-linked examples, or narrative recommendations.	[22,27,31,82,84,85].
Model customization and alignment	**Structured case-prompt system**	Standardizes patient content, behavior, and information disclosure across learners and sessions.	A reusable patient-behavior prompt is combined with a case-specific template or structured case record. Fields may include demographics, presenting complaint, medical and psychosocial history, medications, examination findings, emotional state, information to volunteer or withhold, and encounter-completion criteria.	[16,23,28,47,80].

Continued on the next page.

Table S2 continued from the previous page.

Architectural plane	Technical archetype	Purpose	Typical implementation	Representative studies shown in the figure
	Retrieval-augmented patient	Grounds patient responses in case-specific or clinical evidence to improve factual and contextual consistency.	Case records, standardized-patient scripts, guidelines, textbooks, knowledge bases, electronic health records, or stored dialogue memories are divided into retrievable units. Relevant content is identified through embedding-based, lexical, vector-database, or knowledge-graph retrieval and inserted into the model context before response generation.	[22,35,46,51,75,86].
	Fine-tuned or domain-adapted model	Adapts patient behavior, clinical language, or specialty knowledge to a defined educational context.	General-purpose or smaller models are trained using clinical conversations, specialty-specific cases, counseling transcripts, assessment examples, or locally authored instructional data. Approaches include supervised instruction tuning and parameter-efficient adaptation such as LoRA, sometimes combined with structured prompts or retrieval.	[43,44,46,51,68].
Bidirectional human-in-the-loop mechanisms	**Human-in-the-loop development**	Improves clinical accuracy, educational alignment, realism, and safety before learner deployment.	Clinicians, educators, standardized-patient specialists, learners, and technical developers contribute to objective setting, case authoring, prompt design, content review, model selection, prototyping, transcript review, usability testing, and clinical validation. Review may identify factual errors, inappropriate disclosure, role-breaking, bias, unrealistic communication, or assessment problems.	[23,30,47,68,80].
	Human-in-the-loop delivery	Preserves educational supervision, contextual interpretation, and opportunities for correction during curricular use.	Faculty may orient learners, facilitate sessions, monitor interactions, review transcripts, provide supplementary feedback, conduct structured debriefing, or override inappropriate outputs. Human involvement may occur synchronously during the encounter or asynchronously through post-session review.	[23,30,34,43,80].

S3 Evaluation measures and outcome-assessment approaches

This subsection establishes a common reference for interpreting the evaluation approaches used across the evidence base and for comparing how outcomes were operationalized.

Supplementary Table S3 | Metrics and instruments used to evaluate generative AI-powered simulated-patient systems.

Metric family	Definition	Common metrics or instruments	Outcome domains	Representative studies
Knowledge and diagnostic testing	These measures assess whether learners identify correct information, diagnoses, classifications, or clinical answers against a predefined reference.	Theoretical examinations; multiple-choice question scores; percentage correct; diagnostic accuracy; agreement with a predefined case diagnosis; specialty examination pass rate; classification accuracy against expert consensus.	Knowledge; diagnostic performance; clinical reasoning; OSCE or assessment performance.	[87,88]
Structured clinical-performance assessment	These measures use human or standardized-patient ratings, checklists, or rubrics to assess observable performance during or after a simulated encounter.	OSCE total and domain scores; Mini-CEX; SEGUE total and domain scores; Clinical Reasoning Indicator–History Taking Inventory; history-taking checklists; communication rubrics; Empathic Communication Performance Rating Scale; case-specific clinical-performance scores.	History taking; communication; empathy and professionalism; clinical reasoning; management planning; counseling; OSCE or assessment performance.	[9,26,38,51,89]
Automated scoring and transcript analytics	These measures use a language model, classifier, rule-based system, or other algorithm to score the encounter or derive features from its transcript.	Automated medical history-taking assessment scores; feedback-category coverage; LLM-as-a-judge rubric scores; proportions of open questions and reflections; accuracy and F1-score; cosine similarity; checklist completeness; speaking rate; turn-taking balance; question proportions; sentiment and dialogue trajectories.	History taking; communication; clinical reasoning; diagnostic performance; OSCE or assessment performance; system fidelity and response quality.	[4,35,51,90]
Learner-reported educational outcomes	These measures capture learners' perceived competence, preparedness, learning, confidence, or attitudes rather than directly observed performance.	Five-, six-, or ten-point Likert ratings; self-efficacy scales; Self-Efficacy in Patient–Physician Communication scale; Student Satisfaction and Self-Confidence in Learning questionnaire; Perceived Preparedness Inventory; retrospective pre–post comparative self-assessment.	Confidence and self-efficacy; perceived learning; communication; knowledge; empathy; satisfaction and acceptability.	[16,28,38]

Continued on the next page

Table S3 continued

Metric family	Definition	Common metrics or instruments	Outcome domains	Representative studies
Usability and technology acceptance	These measures assess ease of use, learnability, interface quality, perceived usefulness, and willingness to adopt the system.	System Usability Scale; Chatbot Usability Questionnaire; User Experience Questionnaire; Immersive Technology Evaluation Measure; Technology Acceptance Model; adapted ease-of-use and perceived-usefulness scales.	Usability; satisfaction and acceptability; engagement and adoption; system fidelity.	[3,38,91]
Qualitative experience evaluation	These methods examine how learners, faculty members, or other users interpret the educational encounter and its implementation.	Semi-structured interviews; focus groups; open-ended reflections; thematic analysis; qualitative content analysis; discourse analysis; facilitator observation and narrative feedback.	Satisfaction and acceptability; perceived learning; communication; confidence; usability; system fidelity; safety and implementation experience.	[28,51,57]
Behavioral engagement and usage	These measures use interaction records to quantify how learners engaged with the platform rather than how they rated it.	Number of sessions; completed training cycles; conversation duration; total use time; message or dialogue turns; questions asked; revisions; scenario completion; content-completion rate; frequency of practice.	Engagement and adoption; history taking; perceived learning; cost and scalability; implementation.	[4,28,87]
System fidelity and response quality	These measures assess whether the simulated patient produces clinically appropriate, coherent, relevant, realistic, and role-consistent responses.	Expert ratings of authenticity, correctness, relevance, and script coverage; response accuracy; completeness; F1-score; semantic similarity; readability indices; medical plausibility; hallucination or implausibility counts; robustness across prompts; stability across personas.	System fidelity and response quality; validity and reliability; safety and quality assurance.	[3,4,35]
Safety and psychological safety	These measures assess harmful or inappropriate system behavior and the learner's psychological response to the encounter.	State-Trait Anxiety Inventory; Test Anxiety Inventory; psychological-safety items; embarrassment or anxiety ratings; harmful-statement reports; refusal rate; forbidden-information leakage; biased-content reports; simulator sickness.	Safety and quality assurance; satisfaction and acceptability; learner experience.	[16,20,23,33]
Reliability and validity	These measures assess agreement, consistency, reproducibility, or correspondence between automated scores, human ratings, repeated assessments, or reference standards.	Cohen's kappa; intraclass correlation coefficient; Pearson correlation; Cronbach's alpha; agreement rate; item- and scale-level content-validity indices; model–human agreement; internal consistency; repeated-score reliability.	Validity and reliability; history taking; communication; OSCE or assessment performance; system fidelity.	[3,4,31,51]

Continued on the next page

Table S3 continued

Metric family	Definition	Common metrics or instruments	Outcome domains	Representative studies
Efficiency and resource use	These measures quantify the time, workload, financial cost, or computational resources required to deliver or evaluate the intervention.	Response latency; automated and human scoring time; faculty grading time; session duration; token and API costs; direct cost per learner; annual cost savings; return on investment; post-class review time.	Faculty workload; cost and scalability; system performance; implementation.	[31,71,87]
Retention and transfer	These measures assess whether learning persists over time or generalizes beyond the original training encounter.	Delayed knowledge tests; repeated performance assessments; performance after feedback removal; responses to standardized new cases; transfer-success rate; performance in practical examinations or clinical rotations.	Retention; transfer to new cases; transfer to clinical practice; knowledge; communication; clinical reasoning; assessment performance.	[87,90]
Statistical comparison and effect estimation	These analyses quantify group differences, within-person change, associations, uncertainty, or formal equivalence after an outcome has been measured.	Mean or median differences; gain scores; regression coefficients; Pearson or Spearman correlations; paired and independent-sample t-tests; Mann–Whitney and Wilcoxon tests; ANOVA; chi-square and Fisher exact tests; confidence intervals; Cohen's d; noninferiority or equivalence tests.	These analyses were used across educational outcomes, learner perceptions, system outcomes, and validity evaluations.	[9,16,51,89]

S4 Quantitative synthesis of standardized educational outcomes

Purpose and interpretation of the synthesis. The quantitative synthesis was conducted as an exploratory extension of the scoping review. It summarized standardized educational effects within four prespecified domains: history taking and information gathering, communication and empathy, OSCE global assessment and overall competence, and clinical reasoning and diagnostic performance. These domains were selected because they represented the most frequently evaluated educational-performance constructs and contained at least four estimates that could be placed on a common standardized scale.

The synthesis intentionally retained randomized, nonrandomized comparative, historical-control, and single-group pre–post evaluations. It should therefore be interpreted as a broad direction-and-magnitude summary of a methodologically heterogeneous evidence base rather than as a single causal treatment effect. Comparative and uncontrolled estimates were distinguished in the forest plot and examined separately in sensitivity analyses.

Effect-level eligibility and outcome selection. An estimate was eligible for standardization when the article reported sufficient information to derive an effect estimate and its sampling variance for a learner-level educational outcome. Technical performance, system accuracy, usability, satisfaction, qualitative findings, correlations, agreement coefficients, and post-only ratings without a comparator or reference distribution were retained in the wider evidence map but were not placed on the Hedges' g axis.

No more than one estimate was selected for each unique combination of participant cohort, outcome domain, comparator, and assessment time point. The selection hierarchy was:

1. the prespecified primary educational outcome;
2. a validated domain-specific total score;
3. an independently or human-rated performance measure;
4. an objective test or examination score;
5. a validated automated performance score;
6. a learner-reported competence, confidence, self-efficacy, or attitude measure.

Total scores were preferred to overlapping subscales. Estimates were not selected according to their magnitude or statistical significance. When an article reported both immediate and delayed assessments, the immediate post-intervention outcome was used in the principal synthesis, while delayed outcomes were retained separately. Distinct outcomes from the same cohort could enter different domain-specific models, as occurred for the OSCE and clinical reasoning outcomes reported by Shen et al., but no cohort contributed more than one estimate to the same domain model.

All effects were oriented so that positive values indicated better performance with the generative-AI-powered simulated-patient intervention or improvement after exposure. Scores for which lower values indicated better performance were multiplied by -1 before synthesis.

Small-sample correction. Continuous effects were expressed as Hedges' g. Cohen's d was multiplied by the small-sample correction

$$J(\nu) = 1 - \frac{3}{4\nu - 1}, \tag{1}$$

where ν was the corresponding error degrees of freedom. Thus,

$$g = J(\nu)d. \tag{2}$$

This approximation was used consistently in the effect-construction code.

Independent-group post-intervention outcomes. For parallel-group and historical-control comparisons reporting post-intervention means and standard deviations, the pooled within-group standard deviation was calculated as

$$s_{\mathrm{p}} = \sqrt{\frac{(n_{\mathrm{I}} - 1)s_{\mathrm{I}}^2 + (n_{\mathrm{C}} - 1)s_{\mathrm{C}}^2}{n_{\mathrm{I}} + n_{\mathrm{C}} - 2}}, \tag{3}$$

where the subscripts I and C denote the intervention and comparator groups. Cohen's standardized mean difference was

$$d = \frac{\bar{X}_{\mathrm{I}} - \bar{X}_{\mathrm{C}}}{s_{\mathrm{p}}}, \tag{4}$$

with $\nu = n_{\mathrm{I}} + n_{\mathrm{C}} - 2$. Its approximate sampling variance after small-sample correction was

$$v(g) = \frac{n_{\mathrm{I}} + n_{\mathrm{C}}}{n_{\mathrm{I}} n_{\mathrm{C}}} + \frac{g^2}{2\nu}. \tag{5}$$

Comparative change-score outcomes. For studies reporting mean changes, the change in each group was defined as

$$\Delta_j = \bar{X}_{\mathrm{post},j} - \bar{X}_{\mathrm{pre},j}, \tag{6}$$

and the independent-group standardized mean difference was calculated using the group-specific standard deviations of individual change scores. The pooled change-score standard deviation was obtained by replacing s_{I} and s_{C} in the preceding equations with $s_{\Delta,\mathrm{I}}$ and $s_{\Delta,\mathrm{C}}$.

When a standard deviation of change was not reported, it can in principle be calculated as

$$s_{\Delta,j} = \sqrt{s^2_{\mathrm{pre},j} + s^2_{\mathrm{post},j} - 2r_j s_{\mathrm{pre},j} s_{\mathrm{post},j}}, \tag{7}$$

where r_j is the within-participant pre–post correlation. No common pre–post correlation was imputed in this review. A change-score estimate was included only when the article reported the standard deviation of change, a paired test statistic, a confidence interval for change, or another quantity from which the paired variance could be recovered. The comparative change-score estimates for Shen et al. were calculated directly from the reported group-specific mean changes and standard deviations of change.

Single-group standardized mean changes. For a single-group pre–post evaluation, the standardized mean change was defined using the standard deviation of the individual change scores:

$$d_z = \frac{\bar{D}}{s_D}, \tag{8}$$

where $D = X_{\mathrm{post}} - X_{\mathrm{pre}}$. When a paired t statistic was reported,

$$d_z = \frac{t}{\sqrt{n}}. \tag{9}$$

The small-sample-corrected estimate was

$$g_z = J(n-1)d_z. \tag{10}$$

When the sampling variance was not reported directly, it was approximated as

$$v(g_z) = J(n-1)^2 \left(\frac{1}{n} + \frac{d_z^2}{2n}\right). \tag{11}$$

No pre–post correlation was required when the standard deviation of the individual changes, a paired test statistic, or a paired standardized effect was reported. Studies reporting only separate pre- and post-intervention standard deviations without a change-score variance or paired statistic were not standardized by imposing an assumed correlation.

Individual standardized effects were displayed using uniform square markers with horizontal lines indicating 95% confidence intervals. Diamonds represent the domain-level pooled estimates and their 95% confidence intervals. Evaluation design is reported in Supplementary Tables S4 and S5 for methodological transparency but is not visually encoded by marker shape or fill in the forest plot.

Effects reconstructed from test statistics or nonstandard summary statistics. When group means and standard deviations were unavailable but a two-sided p value from an independent two-group comparison was reported, the corresponding absolute t statistic was reconstructed using the reported degrees of freedom. The standardized mean difference was then calculated as

$$d = t\sqrt{\frac{1}{n_{\mathrm{I}}} + \frac{1}{n_{\mathrm{C}}}}, \tag{12}$$

with its sign assigned according to the direction of the reported group difference. This procedure was used for Aljedaani et al. and was classified as an approximate reconstruction.

For a paired comparison reported only through a two-sided p value and sample size, t was reconstructed with

$n - 1$ degrees of freedom and converted using $d_z = t/\sqrt{n}$. This procedure was used for the learner-perceived competence outcome reported by Byun et al.

Medians and interquartile ranges were not converted to means and standard deviations. When an article reported a rank-based effect r, it was converted approximately to the standardized mean-difference scale as

$$d = \frac{2r}{\sqrt{1 - r^2}}, \tag{13}$$

followed by the Hedges correction. The Xie et al. estimate was derived in this manner from the reported nonparametric effect $r = 0.61$. It was therefore identified as a rank-derived standardized effect rather than a conventional mean-based effect.

For two-group outcomes reported through partial eta squared, the approximate conversion

$$d = 2\sqrt{\frac{\eta_p^2}{1 - \eta_p^2}} \tag{14}$$

was applied, with the sign determined from the reported direction of the effect, followed by the small-sample correction. Effects reconstructed from partial eta squared, rank statistics, or p values were examined in a sensitivity analysis that excluded approximately reconstructed effects.

No medians or interquartile ranges were transformed using distributional assumptions when a compatible effect statistic was unavailable. Such findings were retained in their native metric outside the pooled analysis.

Adjusted and unadjusted estimates. For randomized parallel-group studies, arm-level post-intervention or change-score statistics were used when they corresponded to the prespecified educational outcome and no directly standardized adjusted effect was reported. For nonrandomized, matched, or longitudinal analyses, an adjusted treatment contrast was preferred when the article reported the estimate, its standard error or confidence interval, and a defensible standardizing standard deviation.

The propensity-score-matched estimate from Liu et al. was reconstructed from the reported average treatment effect on the treated and its confidence interval. The repeated-measures estimate from Hong et al. used the reported generalized-estimating-equation marginal contrast standardized by the reported control-group standard deviation. These model-derived standardizations were flagged in the effect audit and excluded in a sensitivity analysis of effects derived directly from group means, standard deviations, or

reported standardized effects. Adjusted and unadjusted estimates from the same comparison were never entered simultaneously.

Cluster, crossover, repeated-measures, and multi-arm designs. For cluster-randomized evaluations, a cluster-adjusted estimate and standard error were preferred. When only unadjusted participant-level statistics were available, an effective sample size would be calculated using

$$\mathrm{DE} = 1 + (\bar{m} - 1)\mathrm{ICC}, \tag{15}$$

where $\bar{m}$ is the mean cluster size and ICC is the intracluster correlation coefficient. No intracluster correlation was imputed.

McCarrick et al. did not report sufficient information to reconstruct a cluster-adjusted comparative standardized effect. Its plotted estimate therefore represents the reported standardized pre–post change within the intervention arm. It was retained only in the inclusive synthesis and was excluded from the comparative-design sensitivity analysis.

Crossover studies were eligible only when a paired treatment contrast and its paired sampling variance could be recovered. Crossover periods were not treated as independent parallel groups. No crossover-derived estimate entered the final pooled forest plot because the eligible crossover reports did not provide sufficient paired variance information. Their findings remained in the native-metric evidence map.

For repeated-measures studies, a reported model-based marginal contrast or a paired contrast preserving within-participant dependence was used. Repeated observations were not entered as independent effects.

For a multi-arm study with two substantively equivalent eligible intervention arms sharing one comparator, the intervention arms were combined before calculating g. Combined means and standard deviations incorporated both within-arm and between-arm variation. When intervention arms represented materially different educational strategies, the arm most closely matching the prespecified review contrast was selected and the other arm was retained narratively. A shared comparator was not duplicated across two estimates in the same domain.

Sampling variances supplied to the pooling model. Each audited effect entered the final plotting dataset as Hedges' g_i and a two-sided 95% confidence interval $[L_i, U_i]$. The pooling script reconstructed the standard

error and sampling variance as

$$\mathrm{SE}_i = \frac{U_i - L_i}{2z_{0.975}}, \qquad v_i = \mathrm{SE}_i^2, \tag{16}$$

where $z_{0.975} = 1.9599639845$. This reconstruction was used because the entered confidence intervals had been generated on the Hedges' g scale using symmetric normal-approximation intervals. The revised analysis archive stores g_i, SE_i, v_i, the source statistics, the transformation used, and all assumptions for each estimate.

Random-effects model. A separate random-effects model was fitted within each outcome domain. Between-evaluation variance τ^2 was estimated by restricted maximum likelihood. For a candidate value of τ^2,

$$w_i(\tau^2) = \frac{1}{v_i + \tau^2}, \tag{17}$$

and the pooled estimate was

$$\widehat{\mu} = \frac{\sum_{i=1}^{k} w_i g_i}{\sum_{i=1}^{k} w_i}. \tag{18}$$

The Python implementation minimized the restricted negative log-likelihood, up to an additive constant,

$$\mathcal{L}_{\mathrm{REML}}(\tau^2) = \frac{1}{2}\left[\sum_{i=1}^{k} \log(v_i + \tau^2) + \log\left(\sum_{i=1}^{k} w_i\right) + \sum_{i=1}^{k} w_i (g_i - \widehat{\mu})^2\right], \tag{19}$$

over $\tau^2 \geq 0$.

Modified Hartung–Knapp confidence intervals. Uncertainty in the pooled estimate was quantified using the ad hoc modified Hartung–Knapp method with a lower bound of one on the Hartung–Knapp variance scale. This corresponds to the modified Knapp–Hartung approach described for meta-analyses with few studies[92,93].

The conventional Hartung–Knapp scale was

$$q = \frac{\sum_{i=1}^{k} w_i (g_i - \widehat{\mu})^2}{k - 1}. \tag{20}$$

The modification used in the analysis was exactly

$$q^* = \max(1, q). \tag{21}$$

The variance and standard error of the pooled estimate were

$$\widehat{\mathrm{Var}}_{\mathrm{mHK}}(\widehat{\mu}) = \frac{q^*}{\sum_{i=1}^{k} w_i}, \qquad \mathrm{SE}_{\mathrm{mHK}} = \sqrt{\widehat{\mathrm{Var}}_{\mathrm{mHK}}(\widehat{\mu})}. \tag{22}$$

The two-sided 95% confidence interval was

$$\widehat{\mu} \pm t_{k-1,0.975}\mathrm{SE}_{\mathrm{mHK}}. \tag{23}$$

Thus, "modified Hartung–Knapp" in this review specifically denotes REML estimation of τ^2, use of $k - 1$ degrees of freedom, and the ad hoc variance-scale constraint $q^* = \max(1, q)$. It does not denote a Sidik–Jonkman estimator of τ^2.

S5 Studies and effect derivations contributing to the meta-analysis

Table S4 identifies the evaluations contributing to each outcome domain and summarizes how each standardized effect was derived. Detailed formulas and decision rules are provided in Supplementary Methods Section S4. Evaluation design is reported as supplementary methodological information but is not encoded by marker shape or fill in the forest plot.

Supplementary Table S4 | Studies and standardized-effect derivations contributing to the meta-analysis of educational outcomes.

Outcome domain	Study	Evaluation design	Effect derivation and analytical note	Reference
History taking and information gathering	Gao et al. (2026)	Randomized parallel	Post-intervention independent-group standardized mean difference calculated from 93.07 ± 3.93 (n = 20) and 88.98 ± 4.05 (n = 20). Hedges' g = 1.005.	32
	Borg et al. (2026)	Nonrandomized comparative	Post-intervention independent-group standardized mean difference calculated from 7.39 ± 0.86 (n = 61) and 6.68 ± 1.04 (n = 54). Hedges' g = 0.743.	19
	Luo et al. (2025)	Randomized parallel	Post-intervention independent-group standardized mean difference calculated from 64.62 ± 9.52 (n = 42) and 54.12 ± 8.80 (n = 42). Hedges' g = 1.135.	51
	Xu et al. (2026)	Historical control	Post-intervention standardized mean difference calculated from 86.95 ± 9.97 (n = 20) and 80.15 ± 4.04 (n = 18). Hedges' g = 0.858. The comparator was historical rather than concurrent.	94
	Aljedaani et al. (2026)	Nonrandomized comparative	Approximate standardized effect reconstructed from the reported two-sided p = 0.002 and group sizes n = 157 and n = 156. Hedges' g = 0.351. Group-specific standard deviations were unavailable.	95
	Sindhvananda et al. (2026)	Single-group pre–post	Reported paired Cohen's $d \approx 1.93$, corrected for small-sample bias. Hedges' g = 1.869. This estimate represents within-participant change.	84
	McCarrick et al. (2025)	Cluster-randomized study	Reported intervention-arm pre–post effect d = 0.37, corrected to Hedges' g = 0.364. A cluster-adjusted comparative standardized effect could not be reconstructed from the available statistics.	14
	Domain summary	**REML random effects**	g = 0.85**, 95% CI 0.38–1.32,** k = 7**;** τ^2 = 0.198**,** I^2 = 80.1%**;** $Q(6)$ = 30.13**,** $p < 0.001$**; approximate 95% PI** −0.40 **to** 2.10**.**	–
Communication and empathy	Sun et al. (2026)	Randomized parallel	Post-intervention independent-group standardized mean difference calculated from 18.53 ± 2.51 (n = 40) and 17.33 ± 2.13 (n = 42). Hedges' g = 0.512.	40
	Xie et al. (2026)	Randomized parallel with rank-based analysis	The reported rank-based effect r = 0.61 was converted using $d = 2r/\sqrt{1-r^2}$ and corrected for small-sample bias. Hedges' g = 1.507. This is an approximate rank-derived standardized effect.	38

Continued on the next page.

Table S4 continued from the previous page.

Outcome domain	Study	Evaluation design	Effect derivation and analytical note	Reference
	Chen et al. (2026)	Randomized parallel	The reported immediate post-intervention Cohen's d = 0.96 was corrected for small-sample bias using the total sample size n = 80. Hedges' g = 0.951.	81
	Karnieli-Miller et al. (2026)	Single-group pre–post	The reported paired effect d = 0.95 (n = 99) was corrected to Hedges' g = 0.943. The selected outcome measured learner-reported self-efficacy rather than directly observed communication performance.	34
	Suárez-García et al. (2025)	Single-group pre–post	A paired standardized change was reconstructed from the reported mean improvement, its confidence interval, and n = 30. Hedges' g = 2.495. Pre- and post-intervention assessments used different simulation modalities.	21
	Sanz et al. (2025)	Single-group pre–post	The reported partial η_p^2 = 0.001 was converted to a standardized mean-difference scale and corrected for small-sample bias. Hedges' g = 0.063. The outcome assessed communication attitudes.	96
	Domain summary	**REML random effects**	g = 1.03**, 95% CI 0.15–1.91,** k = 6**;** τ^2 = 0.611**,** I^2 = 89.6%**;** $Q(5)$ = 48.22**,** p < 0.001**; approximate 95% PI** −1.34 **to** 3.40**.**	–
OSCE global assessment and overall competence	Shen et al. (2026)	Randomized comparative change score	Independent-group standardized difference in change calculated from 11.9 ± 5.2 (n = 40) and 3.2 ± 5.0 (n = 40). Hedges' g = 1.689. Reported change-score standard deviations were used, so no pre–post correlation was assumed.	43
	Liu et al. (2026)	Propensity-score matched	The matched average treatment effect on the treated and its confidence interval were standardized to obtain Hedges' g = 0.246. This was a model-derived effect rather than a direct arm-level standardized mean difference.	39
	Hong et al. (2025)	Randomized repeated measures	The reported generalized-estimating-equation marginal contrast was standardized using the reported control-group standard deviation. Hedges' g = 0.656. A model-scale residual standard deviation was unavailable.	79
	Yamamoto et al. (2024)	Historical control	Post-intervention standardized mean difference calculated from approximately 28.1 ± 1.6 (n = 35) and 27.1 ± 2.2 (n = 110). Hedges' g = 0.480. The comparator was historical rather than concurrent.	16
	Byun et al. (2026)	Single-group pre–post	A paired test statistic was reconstructed from the reported two-sided p = 0.0002 and n = 21, then converted to Hedges' g = 0.953. The outcome measured perceived competence rather than objective OSCE performance.	27
	Domain summary	**REML random effects**	g = 0.77**, 95% CI 0.09–1.46,** k = 5**;** τ^2 = 0.259**,** I^2 = 88.3%**;** $Q(4)$ = 34.10**,** p < 0.001**; approximate 95% PI** −1.03 **to** 2.57**.**	–

Continued on the next page.

Table S4 continued from the previous page.

Outcome domain	Study	Evaluation design	Effect derivation and analytical note	Reference
Clinical reasoning and diagnostic performance	Esmail & Concannon (2025)	Nonrandomized controlled	The reported clinical-assessment Cohen's $d = 0.23$ was corrected for small-sample bias to obtain Hedges' $g = 0.227$. The educational-performance outcome was selected instead of the reported anxiety effect.	[76]
	Shen et al. (2026)	Randomized comparative change score	Independent-group standardized difference in change calculated from 1.3 ± 0.7 ($n = 40$) and 0.3 ± 0.5 ($n = 40$). Hedges' $g = 1.628$. Reported change-score standard deviations were used.	[43]
	Brügge et al. (2024)	Randomized feedback comparison	The reported partial $\eta_p^2 = 0.198$ was converted to a standardized mean-difference scale and corrected for $n = 21$. Hedges' $g = 0.954$. The comparison assessed the incremental effect of AI-generated feedback.	[9]
	Psychiatric CR trial (2026)	Single-group pre–post	The reported paired effect $d = 0.96$ ($n = 32$) was corrected for small-sample bias to obtain Hedges' $g = 0.937$. This estimate represents within-participant change.	[73]
	Domain summary	**REML random effects**	$g = 0.93$**, 95% CI** -0.03**–1.90,** $k = 4$**;** $\tau^2 = 0.282$**,** $I^2 = 79.9\%$**;** $Q(3) = 14.91$**,** $p = 0.002$**; approximate 95% PI** -1.69 **to** 3.56**.**	–

Note: Positive effects indicate better educational performance associated with exposure to, or improvement following, the generative-AI-powered simulated-patient intervention. Domain-level estimates were calculated using restricted maximum-likelihood random-effects models with modified Hartung–Knapp 95% confidence intervals, using the variance-scale constraint $q^* = \max(1, q)$. Prediction intervals were calculated as $\widehat{\mu} \pm t_{k-2,0.975}\sqrt{\widehat{\tau}^2 + \mathrm{SE}^2_{\mathrm{mHK}}}$ and should be interpreted cautiously because each domain contained only four to seven estimates. Evaluation design is reported for methodological transparency but is not visually encoded in the forest plot. CI, confidence interval. PI, prediction interval. REML, restricted maximum likelihood.

S6 Study-level evidence included in the educational-effect meta-analysis

This subsection documents the study-level basis of the quantitative synthesis and supports transparent interpretation and verification of the forest plot.

Supplementary Table S5 | Studies displayed in the forest plot of standardized educational effects.

Outcome domain	Study shown in the figure	Design shown in the figure	Reference
History taking and	Gao et al. (2026)	Randomized parallel	[32]
	Borg et al. (2026)	Nonrandomized comparative	[19]
	Luo et al. (2025)	randomized parallel	[51]
	Xu et al. (2026)	Historical control	[94]
	Aljedaani et al. (2026)	Nonrandomized comparative	[95]
	Sindhvananda et al. (2026)	Single-group pre–post	[84]
	McCarrick et al. (2025)	Cluster-randomized pre–post	[14]
Communication and	Sun et al. (2026)	Randomized parallel	[40]
	Xie et al. (2026)	Randomized rank-based	[38]
	Chen et al. (2026)	Randomized parallel	[81]
	Karnieli-Miller et al. (2026)	Single-group pre–post	[34]
	Suárez-García et al. (2025)	Single-group pre–post	[21]
	Sanz et al. (2025)	Single-group pre–post	[96]
OSCE global assessment and	Shen et al. (2026)	Randomized change-score	[43]
	Liu et al. (2026)	Propensity-score matched	[39]
	Hong et al. (2025)	Randomized repeated measures	[79]
	Yamamoto et al. (2024)	Historical control	[16]
	Byun et al. (2026)	Single-group pre–post perception	[27]
Clinical reasoning and diagnostic performance	Esmail & Concannon (2025)	Nonrandomized controlled	[76]

Continued on the next page.

Table S5 continued from the previous page.

Outcome domain	Study shown in the figure	Design shown in the figure	Reference
	Shen et al. (2026)	Randomized change-score	[43]
	Brügge et al. (2024)	Randomized feedback comparison	[9]
	Psychiatric CR trial (2026)	Single-group pre–post	[73]

S7 Treatment of safety and governance issues in the included literature

Safety and governance issues were most often discussed as risks or limitations, rather than documented as observed incidents or subjected to empirical evaluation (Supplementary Fig. S2 and Supplementary Table S6). Model or version instability was the most frequently represented issue and included the largest number of observed incidents, while hallucination or factual inaccuracy, bias and representativeness, and transparency or disclosure were also commonly discussed. Implemented mitigations were particularly prominent for privacy and data governance and transparency or disclosure. By contrast, harmful or inappropriate advice, external vendor or API exposure, and accessibility were reported less frequently. These patterns indicate growing recognition of safety and governance concerns, but also show that systematic empirical evaluation and incident reporting remain less common than narrative risk discussion.

Supplementary Table S6 | Treatment of safety, ethical, equity, and governance issues in the included literature.

Issue	Incident observed	Empirically evaluated	Implemented mitigation	Discussed as risk or limitation	Future recommendation only
Direct potential harm					
Harmful or inappropriate advice	0	0	1	0	1
Invalid automated feedback or scoring	1	0	8	16	0
Clinical and educational integrity					
Hallucination / factual inaccuracy	2	9	6	28	0
Clinical inconsistency	0	1	1	7	0
Model / version instability	12	4	8	54	3
Human, ethical, and equity risks					
Learner overreliance	0	0	0	21	1
Bias and representativeness	1	2	3	25	4
Cultural / linguistic appropriateness	0	1	1	14	0
Accessibility	0	0	1	9	0

Continued on the next page.

Table S6 continued from the previous page.

Issue	Incident observed	Empirically evaluated	Implemented mitigation	Discussed as risk or limitation	Future recommendation only
Transparency / disclosure	1	1	19	30	4
Data, infrastructure, and governance					
Privacy / data governance	0	0	25	14	2
External vendor / API exposure	0	0	2	2	0
Security	0	1	5	14	1
Other	3	2	2	16	1

Note: Cells report the number of unique study-unit–issue–treatment assignments. A study unit could contribute to more than one issue and to more than one treatment category for the same issue. Counts should therefore not be summed as numbers of unique studies and do not represent the incidence or severity of harm.

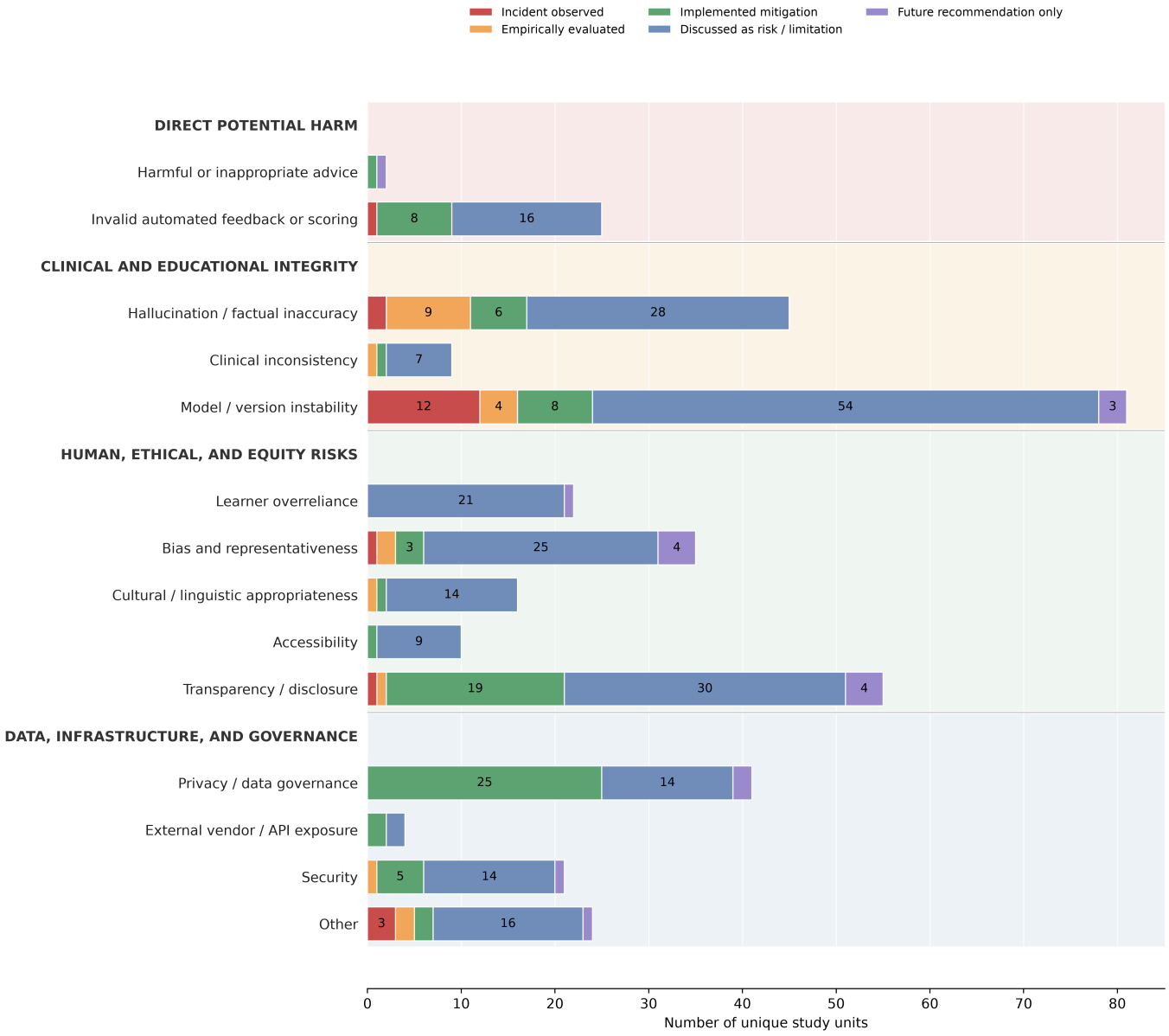
Supplementary Table S7 | Representative works illustrating the safety, ethical, equity, and governance subdomains identified in generative AI-powered simulated-patient research.

Major risk domain	Subdomain	Representative works
Direct potential harm	Emergency recognition and escalation	[84,97]
	Learner distress and psychological harm	[43,98,99]
	Invalid automated feedback or scoring	[3,34,84]
Clinical and educational integrity	Hallucination and factual inaccuracy	[3,8,35]
	Clinical inconsistency	[15,100,101]
	Patient-role fidelity and misrepresentation	[24,80,102]
	Technical reliability and availability	[4,103,104]
	Model and version instability	[34,105,106]
	Assessment integrity and gaming	[22,28,107]

Continued on the next page.

Table S7 continued from the previous page.

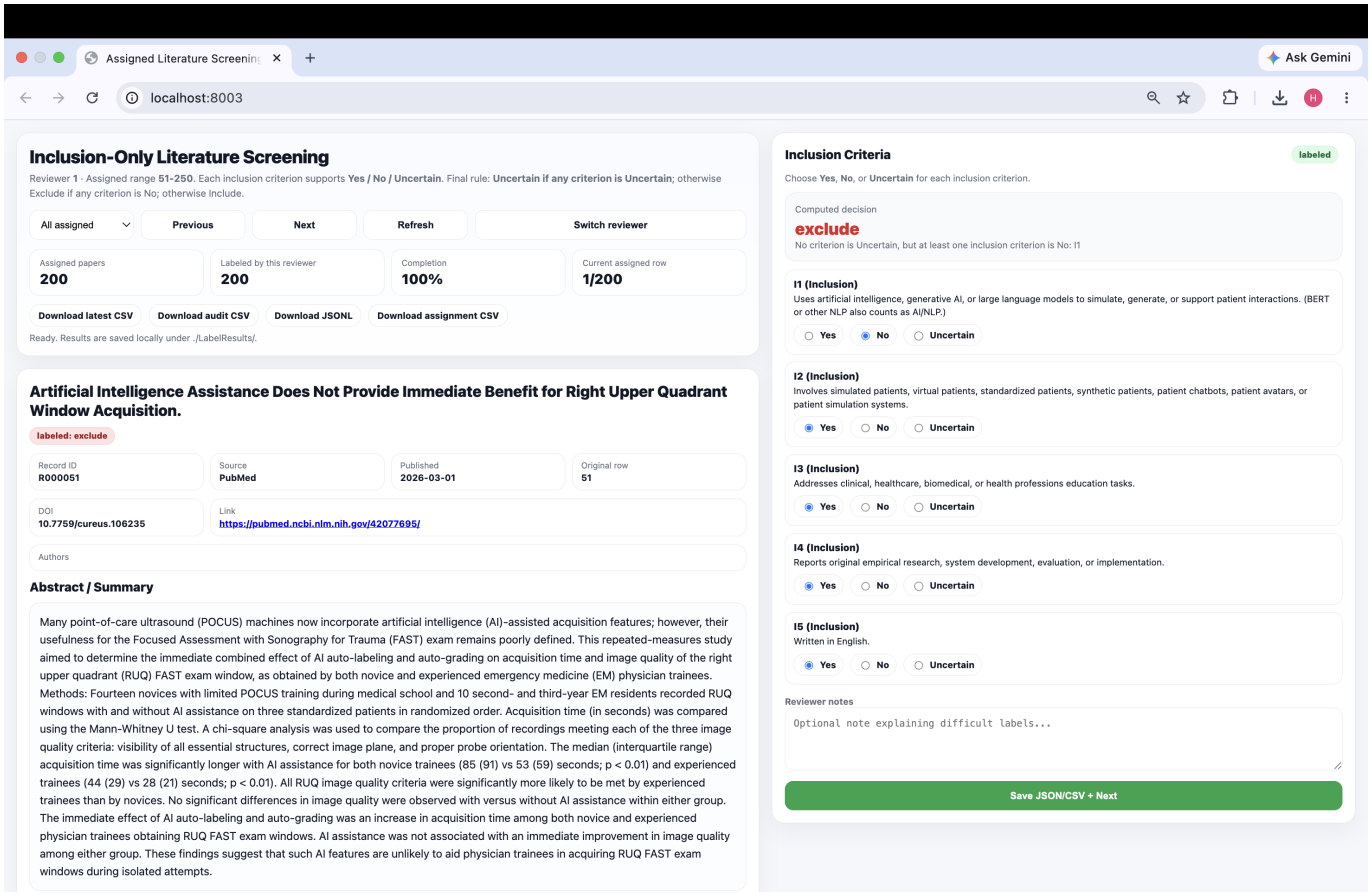
Major risk domain	Subdomain	Representative works
Human, ethical, and equity risks	Learner overreliance and automation bias	22,28,43
	Bias and representativeness	28,34,108
	Cultural and linguistic appropriateness	33,84,109
	Accessibility and the digital divide	81,110,111
	AI transparency and disclosure	5,87,112
	Professionalism and boundary risks	7,80,84
Data, infrastructure, and	Privacy and data governance	31,43,87
	Security and access control	1,56,85
	External vendor and API exposure	6,58,80
	Consent and participant rights	31,85,97
	Intellectual property and data provenance	22
	Other cross-cutting risks	38,57,113



Supplementary Fig. S2 | Treatment of safety, ethical, equity, and governance issues in the included literature. Horizontal stacked bars show the number of unique study-unit–issue–treatment assignments for each issue. Colors indicate whether the issue was reported as an incident observed, empirically evaluated, addressed through an implemented mitigation, discussed as a risk or limitation, or mentioned only as a future recommendation. Background shading groups issues into direct potential harm, clinical and educational integrity, human, ethical, and equity risks, and data, infrastructure, and governance.

S8 Title and abstract screening interface

A browser-based interface was developed to support structured title and abstract screening. For each record, reviewers examined the bibliographic information and independently assessed five inclusion criteria covering the use of artificial intelligence, the presence of a simulated-patient system, relevance to health professions education, reporting of original research or system development, and English-language publication. Each criterion could be labelled as *yes*, *no*, or *uncertain*. Records were classified as included only when all criteria were marked *yes*, as excluded when at least one criterion was marked *no*, and as uncertain when at least one criterion remained *uncertain*. The interface also displayed reviewer progress and retained record-level screening decisions for subsequent reconciliation and adjudication.



Supplementary Fig. S3 | Title and abstract screening interface. The browser-based interface displayed the title, abstract, and available bibliographic information for each record and enabled reviewers to assess the five prespecified inclusion criteria using *yes*, *no*, or *uncertain* responses. The overall screening decision was generated from the criterion-level responses, while reviewer assignments and completion progress were tracked within the interface.

S9 Human annotation and adjudication protocol

S9.1 Purpose and scope

This manual describes how trained human reviewers should independently extract and classify evidence from the included articles. The resulting adjudicated dataset will serve as the human reference standard for validation and reliability analyses.

Reviewers should use the predefined controlled vocabulary and annotation structure described in this manual. All decisions must be based on information explicitly reported in the article, tables, figures, appendices, or supplementary materials.

S9.2 Reviewer roles and source-document hierarchy

Each article should be independently annotated by at least two trained reviewers with relevant experience in health professions education, clinical research, evidence synthesis, or related fields. Questions concerning technical terminology or system architecture may be referred to a reviewer with relevant technical expertise.

Reviewers should examine the complete article and all available supplementary materials. Evidence should generally be prioritized in the following order:

1. methods, intervention, system-development, and implementation sections;
2. results, tables, figures, and supplementary materials;
3. protocols or technical appendices included with the article;
4. discussion statements describing implemented features or observed limitations; and
5. the abstract when the information is unavailable elsewhere.

Reviewers must not infer information from author affiliations, publication year, institutional reputation, clinician coauthorship, or familiarity with a named product or model.

S9.3 Units and levels of annotation

This subsection defines the unit of analysis for human coding so that reviewers apply the annotation framework consistently across articles and evaluation units.

Supplementary Table S8 | Levels and units of human annotation.

Level	Definition	Examples
Article level	Information applying to the publication as a whole.	Bibliographic information, article-wide reporting items, eligibility, overall confidence, and article-level notes.
Study-unit level	A distinct study, system configuration, learner evaluation, technical evaluation, or implementation episode.	Separate learner populations, different simulated-patient systems, distinct training and assessment uses, or separate technical and learner-facing evaluations.
Nested-record level	Repeated records within a study unit that should not be merged into a single text field.	Model components, outcomes, validity assessments, safety issues, evidence gaps, and future-research recommendations.

A new study unit should be created only when there is a meaningful difference in the study population, system, intervention, evaluation, or implementation. Multiple outcomes, subgroup analyses, or repeated reporting of the same participants should not automatically create additional study units.

S9.4 Review-question annotation modules

This subsection organizes the annotation workflow by review question and provides a consistent structure for data extraction.

Supplementary Table S9 | Human annotation modules and principal fields.

Module	Focus	Principal fields
RQ1	Research landscape	Country, profession, learner stage, specialty, educational setting, participant numbers, and implementation context.
RQ2	Technical construction and educational implementation	Case provenance, model components, architecture, grounding, patient persona, memory, interaction modalities, feedback, guardrails, validation, curriculum integration, faculty roles, and reproducibility.
RQ3	Educational activities and competencies	Educational activities, targeted competencies, measured competencies, and assessment functions.

Module	Focus	Principal fields
RQ4	Evaluation methods and educational effectiveness	Study design, participant flow, outcome measures, comparators, numerical findings, validity and reliability, effectiveness, evidence maturity, and methodological limitations.
RQ5	Safety and governance	Safety issues, evaluated risks, implemented mitigations, ethics review, privacy, oversight, disclosure, audit processes, and governance practices.
RQ6	Evidence gaps and future research	Reporting gaps, methodological limitations, translation barriers, unresolved evidence needs, and author-proposed future-research priorities.

S9.5 General annotation decision rules

These cross-cutting decision rules are intended to reduce unsupported inference and improve consistency between independent reviewers.

Supplementary Table S10 | General rules applied across all annotation modules.

Rule	Application
Use explicit evidence	Code only information that is stated or directly demonstrated. Do not reconstruct undocumented technical or educational processes.
Preserve reported terminology	Transcribe model names, versions, instruments, study-design descriptions, and numerical results as reported.
Allow multiple categories	Use multi-label coding when several professions, activities, competencies, modalities, limitations, or governance practices are supported.
Count unique people	Participant counts refer to unique human participants, not prompts, conversations, sessions, cases, ratings, or system outputs.
Separate concepts	Do not combine educational activities, intended competencies, measured competencies, learner experience, and system-performance outcomes.
Record supporting evidence	For important fields, record a concise supporting quotation and the corresponding PDF page number.
Document uncertainty	Record the disputed field, the reason for uncertainty, and the plausible alternatives. Do not force an unsupported classification.
Respect module boundaries	For example, RQ2 records whether a guardrail was implemented, whereas RQ5 records its safety or governance implications.

S9.6 Field-specific coding instructions by review question

This subsection translates the review framework into operational coding guidance for field-level annotation and adjudication.

Supplementary Table S11 | Field-specific annotation guidance by review question.

RQ	Field group	What to record	Key decision rule
RQ1	Study country	Location of system development, educational implementation, or evaluation.	Do not infer the country from author affiliations alone.
	Profession and learner stage	Profession and training stage as separate fields.	Residents are postgraduate trainees. Students without a specific stage may be classified as prelicensure, stage unspecified.
	Specialty and setting	Clinical domain and educational context.	Use the cases, curriculum, or intervention rather than the authors' departments. "Research only" is reserved for technical testing without a genuine learner intervention.
	Model components	Each materially distinct technical component and its function.	Create separate records for dialogue, feedback, scoring, retrieval, moderation, speech, translation, state management, or avatar generation.
	Model reporting	Model family, name, version, provider, and component role.	Preserve the reported terminology. Do not infer a model version from publication date.
	Grounding and retrieval	How information was supplied to the model.	A fixed case included in the prompt is not retrieval-augmented generation. Retrieval-augmented generation requires inference-time retrieval from an external collection.
	Fine-tuning	Any reported adaptation of model weights.	Prompt engineering is not fine-tuning. Fine-tuning requires weight adaptation, instruction tuning, adapters, or an equivalent process.
	Persona, state, and memory	Stable patient characteristics, changing interaction states, and mechanisms used to preserve earlier information.	Multi-turn dialogue alone does not establish a separate memory mechanism.
	Feedback and assessment	Source, timing, format, and purpose of learner feedback or scoring.	Separate patient dialogue from feedback. Identify whether feedback was produced through a model, rubric, faculty member, peer, or hybrid process.
	Guardrails and validation	Implemented safeguards, human review, pre-use testing, monitoring, and logging.	Code implemented controls only. Proposed safeguards belong under future recommendations. Clinician coauthorship alone is not formal validation.
	Educational activity	What learners did during the intervention.	Examples include history taking, communication, diagnosis, management planning, counseling, or OSCE practice.
	Targeted competency	What the intervention was intended to develop.	Use stated learning objectives or intervention descriptions.
	Measured competency	What was directly assessed.	A measured competency requires an observed task, test, rubric, rating, instrument, or reported subscale. Satisfaction, confidence, realism, and usability are not competencies.
	Study design	Definitive design classification used for evidence synthesis.	Record randomization, allocation, blinding, and power calculations only when explicitly reported.
	Participant flow	Numbers enrolled, allocated, completed, and analyzed.	Keep interactions, cases, and sessions separate from participant counts. Document inconsistent denominators.

Continued on the next page.

Table S11 continued from the previous page.

RQ	Field group	What to record	Key decision rule
	Outcome records	Construct, instrument, evaluator, timepoint, comparator, numerical result, direction, significance, and interpretation.	Create separate records for different constructs, instruments, timepoints, or comparisons.
	Educational effectiveness	Evidence of changes in learner knowledge or performance.	Satisfaction, confidence, realism, and usability alone do not demonstrate educational effectiveness. Immediate post-intervention findings do not demonstrate retention.
	Transfer and retention	Retention, performance on new cases, and performance in clinical practice.	Transfer to clinical practice requires assessment in a real or authentic workplace context.
	Validity and reliability	Formally examined measurement properties.	General expert approval or perceived realism is not automatically validity evidence.
RQ4	Safety issues	Observed incidents, evaluated risks, implemented mitigations, or specifically discussed risks.	The absence of reported harm does not demonstrate safety. When no systematic assessment was conducted, classify safety as not evaluated.
	Governance	Ethics review, consent, disclosure, privacy, oversight, audit logging, incident review, change control, and appeal mechanisms.	Do not infer governance practices from institutional affiliation or author credentials.
	Evidence gaps	Observable limitations in reporting, design, validation, safety, equity, implementation, or reproducibility.	Distinguish author-identified gaps from reviewer-identified gaps. Reviewer judgments must remain evidence-based.
	Future research	Recommendations explicitly proposed by the article authors.	Do not create recommendations on the authors' behalf.

S9.7 Independent annotation, reconciliation, and adjudication

Reviewers should complete their annotations independently and without consulting one another's records. After both reviewers have completed their records, disagreements should be compared field by field.

Disagreements may concern:

- missed information;
- study-unit boundaries;
- controlled-vocabulary mapping;
- evidence interpretation;
- numerical transcription;
- missingness or applicability; or
- ambiguity in the annotation manual.

The two reviewers should first attempt to reach consensus by jointly reviewing the relevant source passage. Unresolved disagreements should be adjudicated by a third reviewer. The final dataset should preserve the original reviewer values, the adjudicated value, and the reason for the adjudication decision.

Supplementary References

[1] Sangwon, K. *et al.* A multi-AI agent framework for interactive neurosurgical education and evaluation: From vignettes to virtual conversations. *medRxiv* (2025).

[2] Thesen, T., O'Brien, W. N., Stone, S. & Pinto-Powell, R. Generative AI as the first patient: Practice, feedback, and confidence. *Med. Sci. Educ.* **35**, 2555–2560 (2025).

[3] Bentegeac, R. *et al.* ECOSBot: a multicenter validation pilot study of a generative AI tool for OSCE-based nephrology training. *Clin. Kidney J.* **18**, sfaf308 (2025).

[4] Holderried, F. *et al.* A language model-powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Med. Educ.* **10**, e59213 (2024).

[5] Comulada, W. S. *et al.* A pilot test of an AI voice-driven simulation with feedback for medical students to practice discussing diagnostic mammogram results with patients. *Cureus* **17**, e95606 (2025).

[6] Vilanti, T. *et al.* Contraception-related topics in chat dialogues between healthcare students and generative AI patients: a natural language processing analysis. *BMC Med. Educ.* **25**, 1458 (2025).

[7] Webb, J. J. Proof of concept: Using ChatGPT to teach emergency physicians how to break bad news. *Cureus* **15**, e38755 (2023).

[8] Scherr, R., Halaseh, F. F., Spina, A., Andalib, S. & Rivera, R. ChatGPT interactive medical simulations for early clinical education: Case study. *JMIR Med. Educ.* **9**, e49877 (2023).

[9] Brügge, E. *et al.* Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Med. Educ.* **24**, 1391 (2024).

[10] Gutiérrez Maquilón, R., Uhl, J., Schrom-Feiertag, H. & Tscheligi, M. Integrating GPT-based AI into virtual patients to facilitate communication training among medical first responders: Usability study of mixed reality simulation. *JMIR Form. Res.* **8**, e58623 (2024).

[11] Borg, A. *et al.* AI-enhanced social robotic versus computer-based virtual patients for clinical reasoning training in medical education: Observational crossover cohort study. *J. Med. Internet Res.* **27**, e82541 (2025).

[12] Jacobs, C. *et al.* Application of AI communication training tools in medical undergraduate education: Mixed methods feasibility study within a primary care context. *JMIR Med. Educ.* **11**, e70766 (2025).

[13] Ji, Y.-A., Kim, G. & Seo, J.-H. Transforming medical communication education: The effectiveness of generative AI education using ChatGPT. *J. Health Inform. Stat.* **49**, 332–338 (2024).

[14] McCarrick, C. A. *et al.* A randomized controlled trial of a deep language learning model-based simulation tool for undergraduate medical students in surgery. *J. Surg. Educ.* **82**, 103629 (2025).

[15] Wang, C. *et al.* Application of large language models in medical training evaluation-using ChatGPT as a standardized patient: Multimetric assessment. *J. Med. Internet Res.* **27**, e59435 (2025).

[16] Yamamoto, A. *et al.* Enhancing medical interview skills through AI-simulated patient interactions: Nonrandomized controlled trial. *JMIR Med. Educ.* **10**, e58753 (2024).

[17] Prinz, M., Schäfer, E., Bürklein, S. & Donnermeyer, D. Endodontic diagnostics training in undergraduate dental education: An observational pilot study on AI-driven virtual patient e-learning. *Int. Endod. J.* **59**, 1145–1159 (2026).

[18] Or, A. *et al.* Enhancing dental students' history-taking skills with a generative artificial intelligence chatbot. *J. Dent. Educ.* **90**, 384–390 (2026).

[19] Borg, A. *et al.* AI-generated feedback following social robotic virtual patient interactions and medical student performance: Nonrandomized quasi-experimental study. *JMIR Med. Educ.* **12**, e90368 (2026).

[20] Harada, Y. Robustness of a large language model (LLM)-based virtual patient for japanese history-taking training under direct and indirect instructional contamination. *Cureus* **18**, e109161 (2026).

[21] Suárez-García, R. X. *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.* **15**, 152 (2025).

[22] Jacobs, B. A., Cano, L. & Bradley, J. M. Integrating AI into undergraduate medical education: an exploration of learner-centered approaches through AI. *Front. Med. (Lausanne)* **13**, 1824069 (2026).

[23] Song, J. W. *et al.* Development and validation of CPX-MATE: An end-to-end medical education platform integrating voice-based virtual patient simulation and automated real-time evaluation. *medRxiv* (2026).

[24] Jacobs, C., Johnson, H., Brownlie, K., Joiner, R. & Thompson, T. Scaling multimodal agentic AI in medical education: Multisite cross-sectional study of simulation effectiveness in primary care. *JMIR Form. Res.* **10**, e88905 (2026).

[25] García-Torres, D., Fernández, C., Mira, J. J., Morales, A. & Vicente, M. A. Using AI-based virtual simulated patients for training in psychopathological interviewing: Cross-sectional observational study. *JMIR Med. Educ.* **11**, e78857 (2025).

[26] Hirosawa, T. *et al.* Utility of generative artificial intelligence for japanese medical interview training: Randomized crossover pilot study. *JMIR Med. Educ.* **11**, e77332 (2025).

[27] 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.* **38**, 64–73 (2026).

[28] Lim, A. *et al.* Using an educator-guided generative artificial intelligence (GenAI) tool for developing communication skills in undergraduate pharmacy students. *Curr. Pharm. Teach. Learn.* **18**, 102675 (2026).

[29] Herschbach, L. *et al.* Evaluation of an AI-based chatbot providing real-time feedback in communication training for mental health care professionals: Proof-of-concept observational study. *J. Med. Internet Res.* **27**, e82818 (2025).

[30] Malhotra, A., Buller, M., Modi, K., Pajazetovic, K. & Wijesinghe, D. S. Blueprint for constructing an AI-based patient simulation to enhance the integration of foundational and clinical sciences in didactic immunology in a US doctor of pharmacy program: A step-by-step prompt engineering and coding toolkit. *Pharmacy (Basel)* **13**, 36 (2025).

[31] Takahashi, H. *et al.* AI- vs human-based assessment of medical interview transcripts in a generative AI-simulated patient system: Cross-sectional validation study. *JMIR Med. Educ.* **12**, e81673 (2026).

[32] Gao, L. *et al.* Utilizing generative artificial intelligence to create simulated patient for history taking in gynecology. *BMC Med. Educ.* **26** (2026).

[33] Elyoseph, Z. *et al.* The effectiveness of multilingual AI-based simulator for suicide risk assessment training in improving self-efficacy among young psychiatrists: a pilot study across twenty languages. *BMC Psychiatry* **26**, 98 (2026).

[34] Karnieli-Miller, O. & Elyoseph, Z. Breaking bad news with reflective AI virtual experience (BRAVE): development and evaluation of a generative AI-based simulation tool. *Adv. Simul. (Lond.)* **11** (2026).

[35] Yu, H. *et al.* Simulated patient systems powered by large language model-based AI agents offer potential for transforming medical education. *Commun. Med. (Lond.)* **6**, 27 (2025).

[36] Elhilali, A., Ngo, A. S.-H., Reichenpfader, D. & Denecke, K. Large language model-based patient simulation to foster communication skills in health care professionals: User-centered development and usability study. *JMIR Med. Educ.* **11**, e81271 (2025).

[37] Sorrick, A. *et al.* AI-based simulated patient encounters for clinical reasoning: A needs assessment. *Graduate Medical Education Research Journal* **8**, 27 (2026).

[38] Xie, Y. *et al.* An exploratory study of the use of artificial intelligence-based virtual patients to enhance dentist-patient communication training. *BMC Med. Educ.* (2026).

[39] Liu, Y. *et al.* Real-world impact and educational effectiveness of an AI-powered medical history-taking system: Retrospective propensity score-matched cohort study. *JMIR Med. Educ.* **12**, e89367 (2026).

[40] Sun, N. *et al.* Application of AI-based virtual standardized patients in physician-patient communication training: a study based on the SEGUE framework. *Front. Public Health* **14**, 1768518 (2026).

[41] Chen, G. *et al.* Virtual case reasoning and AI-assisted diagnostic instruction: an empirical study based on body interact and large language models. *BMC Med. Educ.* **25**, 1493 (2025).

[42] Liu, Y. *et al.* 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 Med. Educ.* **11**, e73419 (2025).

[43] Shen, T., Zhang, R., Wang, H., Li, D. & Han, J. Virtual patients and standardized patients combined training is associated with improved clinical reasoning among medical students. *Front. Med. (Lausanne)* **13**, 1825465 (2026).

[44] Hryszko, J., Michałek, A. & Roman, A. Application of large language models in medical interview training: a study with medical students. *BMC Med. Educ.* **26** (2026).

[45] Shindo, N. & Uto, M. Leveraging self-refinement in large language models to suppress excessive responses for virtual simulated patients. *IEICE Trans. Inf. Syst.* (2026).

[46] Jayasrilakshmi, S., Saseendran, N., Reddy, P. D., Kusuma, E. & Mahapatra, A. Virtual patient chatbot for medical training using small language models. In *2025 5th International Conference on Artificial Intelligence and Signal Processing (AISP)*, 1–6 (IEEE, 2025).

[47] Weisman, D. *et al.* 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.* **9**, e65670 (2025).

[48] Huh, C. Y. *et al.* Using large language models to simulate history taking: Implications for symptom-based medical education. *Information (Basel)* **16**, 653 (2025).

[49] Han, W. *et al.* Can large language models replace standardised patients? *Med. Educ.* **59**, 552–553 (2025).

[50] Yuan, Y. *et al.* AI agent as a simulated patient for history-taking training in clinical clerkship: an example in stomatology. *Global Medical Education* **2**, 171–177 (2025).

[51] Luo, M.-J. *et al.* A large language model digital patient system enhances ophthalmology history taking skills. *NPJ Digit. Med.* **8**, 502 (2025).

[52] Li, Y. *et al.* Development and preliminary evaluation of a virtual standardized patient system for psychiatric interview training. *BMC Psychiatry* **26** (2026).

[53] Liu, H. *et al.* Revolutionizing clinical skills training with generative AI: the intelligent inquiry training system framework. *Front. Med. (Lausanne)* **13**, 1781314 (2026).

[54] Yi, Y. & Kim, K.-J. The feasibility of using generative artificial intelligence for history taking in virtual patients. *BMC Res. Notes* **18**, 80 (2025).

[55] Suárez, A., Adanero, A., Díaz-Flores García, V., Freire, Y. & Algar, J. Using a virtual patient via an artificial intelligence chatbot to develop dental students' diagnostic skills. *Int. J. Environ. Res. Public Health* **19**, 8735 (2022).

[56] Elío, I. *et al.* Innovative application of chatbots in clinical nutrition education: The E+DIEting_Lab experience in university students. *Nutrients* **18**, 257 (2026).

[57] Shorey, S. *et al.* A virtual counseling application using artificial intelligence for communication skills training in nursing education: Development study. *J. Med. Internet Res.* **21**, e14658 (2019).

[58] Lee, J. *et al.* Development and evaluation of an artificial intelligence-based virtual patient chatbot for diagnostic education in korean medicine. *J. Korean Med.* **47**, 76–98 (2026).

[59] Sooful, P., Zablon, P. & Thornton, M. Interaction with an AI chatbot in audiology education as a critical learning agent. *Hear. J.* **78**, 1–5 (2025).

[60] Sardesai, N., Russo, P., Martin, J. & Sardesai, A. Utilizing generative conversational artificial intelligence to create simulated patient encounters: a pilot study for anaesthesia training. *Postgraduate Medical Journal* **100**, 237–241 (2024).

[61] Han, Y. Usability and educational effectiveness of AI-based patient chatbot for clinical skills training in korean medicine. *Korean J. Acupunct.* **41**, 27–32 (2024).

[62] Xu, J., Yang, L. & Guo, M. Designing and evaluating an emotionally responsive virtual patient simulation. *Simul. Healthc.* **19**, 196–203 (2024).

[63] Hall, S. *et al.* Evaluation of the use of chatbots as simulated patients in capstone clinical pharmacy education: A pilot study. *Curr. Pharm. Teach. Learn.* **18**, 102588 (2026).

[64] Nolan, E. J. & Burke, H. B. Accuracy of large language model transcription of simulated physician-patient verbal interactions. *BMC Med. Inform. Decis. Mak.* **26** (2026).

[65] Ting, P. W. & Wolffsohn, J. S. Artificial intelligence-driven patient history and symptoms combined with slit-lamp eye simulation for enhancing the clinical training of students. *Clin. Exp. Optom.* **109**, 434–443 (2026).

[66] Fung, T. C. J. *et al.* Effects of generative artificial intelligence (GenAI) patient simulation on perceived clinical competency among global nursing undergraduates: a cross-over randomised controlled trial. *BMC Nurs.* **24**, 934 (2025).

[67] Huynh, L. *et al.* A feasibility study to assess virtual patients authenticity on building confidence in students' clinical decision-making. *J. Undergrad. Res. (Gainesv.)* **27** (2025).

[68] Samson, J. *et al.* Human-centred artificial intelligence: Pilot design of a virtual patient. In *2025 IEEE Conference on Serious Games and Applications for Health (SeGAH)*, 01–08 (IEEE, 2025).

[69] Liu, H. *et al.* AI-based virtual standardized patients improve oral mucosal disease training in fourth-year dental students. *J. Dent. Educ.* (2026).

[70] Kanazawa, A. *et al.* Evaluation of a medical interview-assistance system using artificial intelligence for resident physicians interviewing simulated patients: A crossover, randomized, controlled trial. *Int. J. Environ. Res. Public Health* **20**, 6176 (2023).

[71] Tyrrell, E. G. *et al.* Web-based AI-driven virtual patient simulator versus actor-based simulation for teaching consultation skills: Multicenter randomized crossover study. *JMIR Form. Res.* **9**, e71667 (2025).

[72] Cross, J. *et al.* Assessing ChatGPT's capability as a new age standardized patient: Qualitative study. *JMIR Med. Educ.* **11**, e63353 (2025).

[73] Li, X. *et al.* AI-driven multimodal simulation for psychiatric clinical education: conceptual framework and preliminary validation of the cognitive core. *BMC Med. Educ.* (2026).

[74] Ali, M. R. *et al.* Novel computational linguistic measures, dialogue system and the development of SOPHIE: Standardized online patient for healthcare interaction education. *IEEE Transactions on Affective Computing* (2021).

[75] 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.* **27**, 2481–2491 (2025).

[76] Esmail, S. & Concannon, B. Immersive virtual reality and AI (generative pretrained transformer) to enhance student preparedness for objective structured clinical examinations: Mixed methods study. *JMIR Serious Games* **13**, e69428 (2025).

[77] Dollis, J. S. *et al.* When avatars have personality: Effects on engagement and communication in immersive medical training. In *2026 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 965–974 (IEEE, 2026).

[78] Borg, A. *et al.* Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: Mixed methods study. *J. Med. Internet Res.* **27**, e63312 (2025).

[79] Hong, H., Huang, J., Yuan, Y., Huang, X. & Huang, X. Research on the application of standardized patient robot teaching in the training of medical students' reception skills: GEE analysis 2.0 based on generalized estimation equation. *erd* **7**, 264–272 (2025).

[80] BaHammam, F. Feasibility of large language model-based standardized virtual patients to support clinical decision-making training in operative dentistry: Mixed methods study. *JMIR Form. Res.* **10**, e91021 (2026).

[81] Chen, P.-J. AI-enhanced virtual reality simulation for nursing students' empathy: Automated scoring and inter-rater reliability in a randomised controlled study. *Nurse Educ. Today* **159**, 106968 (2026).

[82] Hicke, Y. *et al.* MedSimAI: Simulation and formative feedback generation to enhance deliberate practice in medical education. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, vol. 1, 415–424 (ACM, New York, NY, USA, 2026).

[83] Hu, Y., Xiong, Q., Yi, L. & Yoon, I. Nurse town: An LLM-powered simulation game for nursing education. In *2025 IEEE Conference on Artificial Intelligence (CAI)*, 215–222 (IEEE, 2025).

[84] Sindhvananda, K. *et al.* AI-powered simulated patients with automated feedback for enhancing headache history-taking skills: a convergent mixed-methods study. *BMC Med. Educ.* **26** (2026).

[85] Chen, N. *et al.* Developing and validating a coding scheme for clinical reasoning in history taking using generative AI-based virtual patients: Systematic text condensation approach. *JMIR Med. Educ.* **12**, e84347 (2026).

[86] Wang, Q., Hu, C., Li, Y., Zhang, T. & Zhang, S. AI-assisted case-based learning and flipped classroom to improve clinical decision-making: a randomized controlled trial in reproductive medicine. *Med. Educ. Online* **31**, 2670047 (2026).

[87] Ji, F., Xiao, W. & Li, X. AI-driven intelligent training enhances clinical competence in oncology residency: a randomized controlled trial. *Front. Med. (Lausanne)* **13**, 1768388 (2026).

[88] Han, Y., Park, Y. & Lee, J. Developing AI-powered virtual patient chatbots for diagnostic reasoning training. *Acad. Med.* **101**, 799–804 (2026).

[89] Lippitsch, A. *et al.* Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars. *Comput. Methods Programs Biomed.* **244**, 107964 (2024).

[90] Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C. & Imel, Z. E. Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *J. Med. Internet Res.* **21**, e12529 (2019).

[91] Aster, A. *et al.* Development and evaluation of an emergency department serious game for undergraduate medical students. *BMC Med. Educ.* **24**, 1061 (2024).

[92] Röver, C., Knapp, G. & Friede, T. Hartung-knapp-sidik-jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC medical research methodology* **15**, 99 (2015).

[93] Jackson, D., Law, M., Rücker, G. & Schwarzer, G. The hartung-knapp modification for random-effects meta-analysis: A useful refinement but are there any residual concerns? *Statistics in medicine* **36**, 3923–3934 (2017).

[94] Xu, J. *et al.* Application and evaluation of a virtual patient system based on a multi-branch dynamic decision model in orthopedic graduate medical education. *BMC Med. Educ.* **26** (2026).

[95] Aljedaani, S., Sultan, I. & El-Tarhouny, S. Knowledge-based chatbots in clinical teaching: AI-powered virtual patients to improve history-taking during internal medicine rotations. *International Medical Education* **5**, 51 (2026).

[96] Sanz, A., Tapia, J. L., García-Carpintero, E., Rocabado, J. F. & Pedrajas, L. M. ChatGPT simulated patient: Use in clinical training in psychology. *Psicothema* **37**, 23–32 (2025).

[97] Haber, Y., Levi-Belz, Y., Elbak, Y. S., Elyoseph, Z. & Levkovich, I. Validating GenAI feedback in suicide prevention training: a mixed-methods study of QPR skill assessment. *Front. Med. (Lausanne)* **12**, 1709743 (2025).

[98] Levkovich, I., Haber, Y., Levi-Belz, Y. & Elyoseph, Z. A step toward the future? evaluating GenAI QPR simulation training for mental health gatekeepers. *Front. Med. (Lausanne)* **12**, 1599900 (2025).

[99] Holderried, F. *et al.* Using AI to train future clinicians in depression assessment: Feasibility study. *JMIR Med. Educ.* **12**, e87102 (2026).

[100] Szydlak, R., Kiyak, Y. S., Hege, I., Torre, D. & Kononowicz, A. A. Comparison of human and GPT-generated concept maps in a clinical reasoning collection of educational virtual patients. *Stud. Health Technol. Inform.* **327**, 1024–1028 (2025).

[101] Kim, J., Won, J. & Lee, Y. Use of a generative pre-trained transformer-based virtual patient for health assessment and communication training in nursing education: A mixed-methods study. *Nurse Educ. Pract.* **88**, 104536 (2025).

[102] Fąferek, J. *et al.* Applying ChatGPT to plan and create a realistic collection of virtual patients for clinical reasoning training. *BMC Med. Educ.* **25**, 1277 (2025).

[103] Öncü, S., Torun, F. & Ülkü, H. H. AI-powered standardised patients: evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med. Educ.* **25**, 278 (2025).

[104] Cui, S. *et al.* AI virtual standardized patient training for perinatal dental communication: A pilot study. *J. Dent. Educ.* (2026).

[105] Szydlak, R. *et al.* ChatGPT versus human authors: A comparative study of concept maps for clinical reasoning training with virtual patients. *Med. Teach.* **48**, 712–720 (2026).

[106] Low, M. Y. H. *et al.* Implementation of a virtual patient chatbot for physiotherapy students training. In *2023 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 1285–1289 (IEEE, 2023).

[107] Damaševičius, R., Maskeliūnas, R. & Blažauskas, T. Enhancing virtual medical history taking: Effects of customized guidelines in two serious games for medical education. *Ann Pharm Educ Saf Public Health Advocacy Spec* **5**, 39–49 (2025).

[108] Tsao, S., Clancy, C., Bennett, N., Rose, S. & Kassutto, S. Medical education in the metaverse: Integration of virtual learning experiences into core clerkships for medical students. *Acad. Med.* **98**, S189–S190 (2023).

[109] Chinwong, D., Penthinapong, T. & Chinwong, S. Integrating ChatGPT for smoking cessation counseling practice in pharmacy education: A single group quasi-experimental study. *Tob. Induc. Dis.* **23**, 1–12 (2025).

[110] Ahmad, A. *et al.* Voices of innovation: Reflective report on integrating artificial intelligence-simulated mental health patient scenarios into undergraduate nursing education in the united arab emirates. *JMIR Med. Educ.* **12**, e78161 (2026).

[111] Arzumanova, R. A. & Bokizhanova, G. K. Developing skills for clear and empathetic physician communication in medical students using artificial intelligence tools within the "russian language and culture of speech" course. *Philology. Theory & Practice* **19**, 504–512 (2026).

[112] Saggar, A., Dimitrova, V., Sarikaya, D., Hogg, D. & Darling, J. C. AI-simulated clinical consultations: Assessing the potential of ChatGPT to support medical training. *Arch. Dis. Child.* **111**, 638–644 (2026).

[113] Majmundar, P. R. A game-based simulation builder for adaptive learning, design, and nursing care. *Int. J. Drug Deliv. Technol.* **16**, 737–748 (2026).