Revista Española
de Educación Médica

Artificial intelligence in the practical clinical training of undergraduate medical students: a review of application scope, results and gaps.

Artificial intelligence in the practical clinical training of undergraduate medical students: a scoping review of applications, outcomes, and gaps.

Juan Bonilla[1,] Tomás Cortés[2,] Diego Polanco[3,] Carlos Martínez[4,] Álvaro Herrera[5]

1, School of Medicine, Faculty of Medicine, University of Chile. Santiago, Chile; juan.mejia@ug.uchile.cl, ORCID-ID 0009-0007-4086-083X; 2, tomas.cortes.f@ug.uchile.cl, ORCID-ID 0009-0008-4768-2746; 3, diego.polanco@ug.uchile.cl, ORCID-ID 0009-0008-9502-465X; 4, carlosmartinez@ug.uchile.cl ORCID-ID 0009-0001-1136-1026; 5, Department of Health Sciences Education, Faculty of Medicine, University of Chile. Santiago, Chile; levarito@uchile.cl, ORCID-ID 0009-0007-4861-2144
* Correspondence: juan.mejia@ug.uchile.cl

Received: 8/4/26; Accepted: 4/5/26; Published: 6/5/26

Summary.

Objective: To map and synthesize the available evidence on the use of artificial intelligence (AI), including large-scale language models (LLMs) and generative tools, in the practical clinical training of undergraduate medical students. **Methodology:** A scoping review was conducted following the methodological framework of Arksey and O'Malley and reported according to PRISMA-ScR. The literature search was carried out on January 28, 2026, in PubMed/MEDLINE, Scopus, and Web of Science Core Collection. Empirical studies published from 2021 onward, in English, Spanish, or Portuguese, that evaluated AI-based educational interventions in undergraduate medical students in supervised clinical practice training contexts in real and/or simulated settings were included. **Results:** 2112 records were identified, of which 789 were excluded due to duplication. After screening and full-text appraisal, 24 studies were included. The evidence focused on simulated or structured scenarios and domains such as clinical interviewing/communication, clinical reasoning, and technical or procedural skills. LLMs and generative tools were the most frequently studied technologies. A subset of comparative studies reported comparable results or, in certain domains and specific contexts, results favorable to AI-based interventions; however, the methodological heterogeneity of the comparators, outcomes, and designs prevents drawing aggregate conclusions about effectiveness, and the evidence focused mainly on immediate educational outcomes. **Conclusions:** AI shows potential as a complementary tool to expand deliberate practice, standardize feedback, and support more accessible and, in some cases, more personalized learning experiences. Nevertheless, limitations persist related to methodological heterogeneity, the scarcity of evaluation in real clinical settings, and the lack of longitudinal follow-up. Therefore, more robust studies and clear ethical and pedagogical frameworks are needed to guide its responsible integration into undergraduate medical education.

Keywords: Artificial intelligence, Large-scale language models, Medical education, Clinical competence, Medical students, Clinical simulation, Scope review.

Resumen.

Objetivo: Mapear y sintetizar la evidencia disponible sobre el uso de inteligencia artificial (IA), incluidos los modelos de lenguaje de gran escala (LLM) y las herramientas generativas, en la formación clínica práctica de estudiantes de medicina de pregrado. **Metodología:** Se realizó una scoping review siguiendo el marco metodológico de Arksey y O'Malley, y se reportó de acuerdo con PRISMA-ScR. La búsqueda bibliográfica se llevó a cabo el 28 de enero de 2026 en PubMed/MEDLINE, Scopus y Web of Science Core Collection. Se incluyeron estudios empíricos publicados desde 2021 en adelante, en inglés, español o portugués, que evaluaran intervenciones educativas basadas en IA en estudiantes de medicina de pregrado, en contextos de formación práctica clínica supervisada en escenarios reales y/o simulados. **Resultados:** Se identificaron 2112 registros, de los cuales 789 fueron eliminados por duplicación. Tras el cribado y la evaluación de texto completo, se incluyeron 24 estudios. La evidencia se concentró en escenarios simulados o estructurados y en dominios como entrevista clínica/comunicación, razonamiento clínico y habilidades técnicas o procedimentales. Los LLM y las herramientas generativas fueron las tecnologías más frecuentemente estudiadas. Un subconjunto de estudios comparativos reportó resultados comparables o, en determinados dominios y contextos específicos, favorables para intervenciones basadas en IA; sin embargo, la heterogeneidad metodológica de los comparadores, outcomes y diseños impide extraer conclusiones agregadas sobre efectividad, y la evidencia se concentró principalmente en desenlaces educativos inmediatos. **Conclusiones:** La IA muestra potencial como herramienta complementaria para ampliar la práctica deliberada, estandarizar la retroalimentación y apoyar experiencias de aprendizaje más accesibles y, en algunos casos, más personalizadas. No obstante, persisten limitaciones relacionadas con la heterogeneidad metodológica, la escasa evaluación en contextos clínicos reales y la falta de seguimiento longitudinal, por lo que se requieren estudios más robustos y marcos éticos y pedagógicos claros que orienten su integración responsable en la educación médica de pregrado.

Keywords: Inteligencia artificial, Modelos de Lenguaje a Gran Escala, Educación médica, Competencia clínica, Estudiantes de medicina, Simulación clínica, Revisión de alcance.

---

1. Introduction

Practical clinical training in medicine aims to develop observable competencies in students in real or simulated contexts—for example, clinical interviewing, physical examination, communication, applied clinical reasoning, and performance in OSCE-type scenarios or skills stations. However, teaching and assessing these competencies is often resource-intensive (teaching time, evaluators, standardized patients, infrastructure, etc.) and, therefore, difficult to consistently scale to growing cohorts (1). In parallel, competency-based curricula have increased the demand for deliberate practice and frequent, standardized feedback throughout undergraduate studies.

In recent years, artificial intelligence (AI)—including machine learning, adaptive systems, and, more recently, generative models and language models (LLMs)—has begun to be incorporated into medical education for purposes ranging from learning and assessment support to simulation and skills training. Previous scoping reviews have broadly mapped AI applications, highlighting surgical skills training, automated/objective assessment, and tools to support diagnostic reasoning, among others (e.g., admissions, teaching, laboratory training), emphasizing both opportunities and the need for ethical and educational governance frameworks (2-3). However, these syntheses often cover multiple domains and training levels, and therefore, specific evidence on AI applied to the purely practical training of medical students remains scattered.

In particular, the emergence of LLM-based tools has spurred new proposals for scalable practical training, such as conversational simulated patients and systems that seek to automate

components of feedback and evaluation (4-5). For example, agent-driven simulated patient systems based on LLM have been described that aim to replicate clinical-communicative interactions, although their effectiveness, reliability, and implementation conditions remain key challenges (4, 6). Similarly, AI applications for training/evaluation in OSCE-type formats have emerged (e.g., chatbots for clinical interviews; automated performance evaluation) (3). Despite this growth, it is still unclear what types of tools are being used, in what practical scenarios, with what assessment designs, what competency outcomes are being measured, and what methodological and ethical gaps persist, especially considering the variability in definitions of "AI" and the limited robust evidence on educational/clinical outcomes noted by previous syntheses. In this context, systematically understanding how AI is being used in practical clinical training is especially relevant, given that these competencies directly impact the quality and safety of future healthcare. Furthermore, the pressure to scale practical training in resource-constrained settings makes AI-based solutions attractive, but their adoption requires a clear characterization of their scope, reported performance, and potential ethical and pedagogical implications.

For the purposes of this review, practical clinical training is broadly understood as a continuum of scenarios aimed at developing and assessing applied clinical competencies. This ranges from supervised real-world clinical practice to structured simulation contexts, including OSCE, Mini-CEX, skills labs, virtual or digital patients, and interactive simulated cases. However, these scenarios are not considered equivalent, as they differ in their degree of pedagogical authenticity, contextual complexity, and potential for transfer to clinical practice with real patients. This distinction is relevant for interpreting the available evidence, particularly in a field where AI applications tend to focus on more structured and simulable environments.

Given that this is an emerging field, heterogeneous in terms of technologies, educational contexts, and evaluation metrics, the objective of this study is to map, explore, and describe the available evidence on the use of AI, including LLM and generative tools, for the practical clinical training of undergraduate medical students. This review does not seek to estimate effect sizes or formally compare interventions, but rather to characterize the landscape of applications, contexts, reported outcomes, and knowledge gaps, which is consistent with the objective and utility of a *scoping review* .

2. Methodology

This review was conducted following the methodological framework of Arksey and O'Malley (7). The report was prepared in accordance with the PRISMA-ScR guidelines to ensure transparency and reproducibility.

Stage 1: Identifying the research question

As a first stage of this review, the following research question was defined: "How has artificial intelligence been used in the practical clinical training of undergraduate medical students, in what contexts and competency domains, with what reported results and what methodological and ethical gaps persist?"

Stage 2: Identification of relevant literature

The literature search was conducted on January 28, 2026, in PubMed/MEDLINE, Scopus, and Web of Science (Core Collection). Boolean operators and a combination of controlled vocabulary and free-text terms related to artificial intelligence, medical education, and clinical skills training were used. In PubMed/MEDLINE, MeSH terms such as "Artificial Intelligence," "Machine Learning," "Deep Learning," "Students, Medical," "Education, Medical, Undergraduate," and "Clinical Competence" were used; in addition, free-text terms such as *"large language model," "generative AI," "ChatGPT," *"chatbot," *"medical student," "clinical reasoning," *"clinical skill,"

"simulation," "objective structured clinical examination," and "mini-CEX" were included in all three databases. The complete search strategy for each database is presented in Appendix 1. In addition, exclusion terms were used to reduce studies focused on pure algorithmic performance, such as "model performance", "algorithm performance", "predictive model" or "diagnostic model".

For the selection of studies, inclusion and exclusion criteria were defined and summarized in Table 1 of the appendix. Empirical studies (quantitative, qualitative, or mixed methods) published from 2021 onwards, in English, Spanish, or Portuguese, were included. These studies had to evaluate educational interventions based on explicitly defined AI (e.g., ML/DL, LLM, generative AI, NLP, computer vision, or trained algorithms) and used for teaching, training, assessment, or feedback in clinical competencies. Studies had to involve undergraduate medical students (MD/MBBS or equivalent); in studies with mixed populations, only those reporting separable outcomes for undergraduate students were included.

For the purposes of this review, studies were considered that ranged from supervised real-world clinical practice to various structured simulation contexts, such as OSCE, Mini-CEX, skills labs, virtual or digital patients, and interactive simulated cases. These contexts were included because they share a focus on the performance of observable clinical competencies and the application of skills in clinical situations. However, it is important to note that they are not considered equivalent, as they differ in their degree of pedagogical authenticity, contextual complexity, and potential for transfer to performance with real patients. This distinction was incorporated from the review design stage and maintained throughout the analysis, allowing for the interpretation of the evidence while considering this heterogeneity.

To avoid terminological ambiguities, in this review the term "practical clinical training" was used broadly to refer to the continuum of scenarios aimed at developing applied clinical competencies, which includes both supervised real-world clinical practice and structured simulation contexts, such as OSCE, skills labs, virtual or digital patients, and interactive simulated cases. The term "simulation" was reserved for scenarios that do not involve real patients, while "structured contexts" was used to refer to scenarios with explicit performance criteria and standardized assessment conditions, regardless of whether they occur in simulation or real-world clinical practice. Studies focused exclusively on residents, specialists, or other healthcare professionals were excluded, as were those describing educational interventions without an AI component or where AI was used for non-educational purposes. Also excluded were studies that assessed only satisfaction, perceptions, or attitudes toward AI without learning/performance/implementation outcomes, and non-empirical publications, including editorials, commentaries, letters to the editor, opinion essays, protocols without results, and secondary reviews.

Stage 3: Selection of appropriate studies (screening)

The identified records were imported into Rayyan for duplicate detection and manual selection. Eligibility was assessed in two phases (title/abstract and full text) using predefined inclusion and exclusion criteria (Table 1 in the appendix). The process was carried out by four reviewers. Discrepancies were resolved through team discussion and consensus; when necessary, further review was conducted for adjudication.

Stage 4: Data Extraction, Mapping, and Graphing

charting form was designed to record, in a standardized manner, the methodological characteristics and main findings of the included studies. The variables extracted included: author/year, country, study design, type and number of participants (including educational level when available), educational context (supervised clinical practice and/or simulation), AI technology

(e.g., LLM, ML/DL, NLP/vision), educational purpose (teaching, training, assessment, or *feedback*), *outcomes* assessed, and main results. The extracted data were synthesized using two complementary products. First, a descriptive table (Table 2 in the appendix) was created summarizing, for each included study, the country, design, participants, AI technology, and main results, with the aim of mapping the scope of the evidence and the heterogeneity of approaches, contexts, and applications. Second, a comparative table (Table 3 in the appendix) was constructed, focusing on the subset of studies that reported direct comparisons between an AI-based intervention and a comparator (e.g., traditional method, human instruction, historical cohort, or other condition), recording the competency assessed, the comparator, and the main outcome reported. This table allowed us to identify which competencies and study designs offer comparative evidence.

Additionally, a derived analytical variable ("competency domain") was added to Table 2 to classify each study according to the predominant practical clinical competency addressed by the intervention. A closed taxonomy of six domains was used for this purpose: (1) clinical interviewing, medical history taking, and communication; (2) clinical reasoning and diagnostic decision-making; (3) imaging and diagnostic interpretation; (4) technical/procedural skills; (5) assessment-centric performance evaluation and feedback; and (6) integrated clinical competence (e.g., OSCE/Mini-CEX). This taxonomy was constructed as a derived analytical variable, for the purpose of synthesis, based on the review of the included studies and the review's interest in characterizing the predominant practical clinical competency addressed by each intervention. Its application was carried out through a consensus-based review among the reviewers . When discrepancies arose in the assignment of the primary domain, these were resolved through team discussion until agreement was reached. The assignment was mutually exclusive, defining one primary domain per study. To resolve ambiguities, the following rule was applied: if the main outcome corresponded to consistency, stability, or validity of the assessment (e.g., human-AI agreement or inter-rater agreement), the study was classified in domain 5; if the main outcome corresponded to improvement in student performance (pre/post or comparison with control), it was classified in domains 1–4 or 6, depending on the predominant competency.

Based on this classification, a bar chart was created showing the number of studies per domain of competence (Figure 2), as well as a geographical map showing the distribution of countries with publications on the topic, using greater color intensity to represent those with a higher frequency of appearance in the sample (Figure 3). Finally, a heat map is also shown representing the distribution of studies according to domain of competence and type of AI (Figure 4).

Stage 5: Summary and presentation of results

The results were synthesized using a descriptive and narrative approach. Frequencies and proportions were reported to characterize the distribution of studies by country, design, type of participants, educational context (real clinical setting vs. simulation), type of AI technology, and educational purpose. The evidence was presented grouped by the previously defined "domain of competence" and, where relevant, by type of AI and implementation context. The study selection process was documented using a PRISMA-ScR flowchart, including the number of records identified, duplicates removed, studies assessed in full text, and reasons for exclusion. Findings were presented in a descriptive table (Table 2, Appendix) to map the scope and heterogeneity of the evidence, and in a comparative table (Table 3, Appendix) for the subset of studies with direct comparisons between AI-based interventions and a comparator. Given the heterogeneity of designs, interventions, and outcomes, no quantitative synthesis or meta-analysis was performed, nor were the results interpreted as an overall estimate of effectiveness. Comparative findings were reported descriptively, highlighting patterns, methodological gaps, and priority areas for future research.

On the other hand, in accordance with the PRISMA-ScR guidelines, it is explicitly stated that the decision not to conduct a formal methodological quality assessment is consistent with the

mapping purpose of a scoping review, but implies that the findings do not allow for ranking the strength of the evidence or establishing recommendations based on the quality of the studies. This limitation directly affects the applicability of the results: the reported comparative patterns should be understood as descriptive indicators of the field, and not as a basis for recommendations for clinical or curricular implementation. Future systematic reviews should incorporate risk of bias assessments to move toward conclusions about effectiveness.

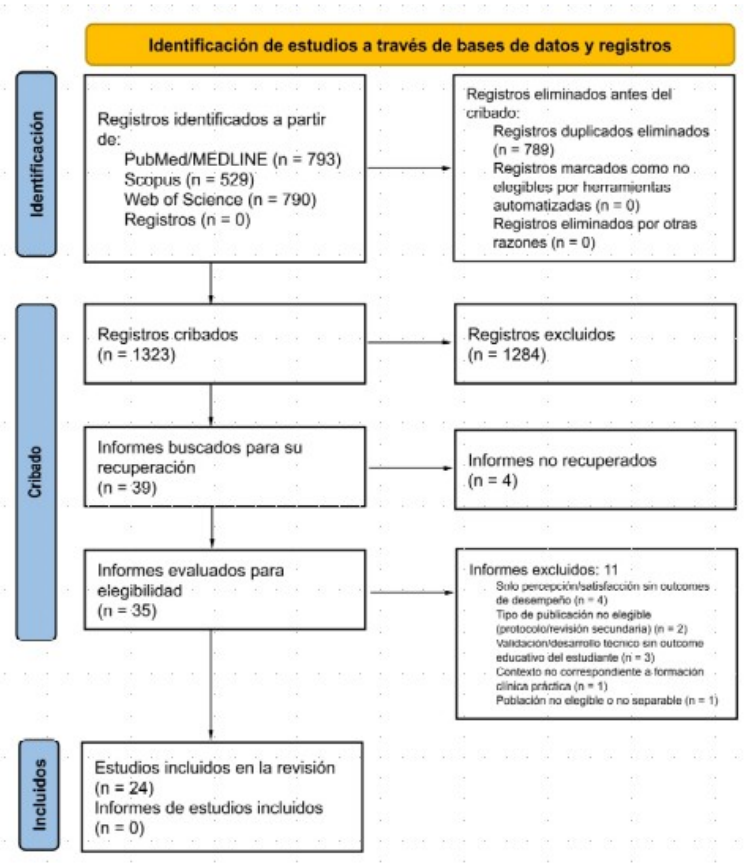


Figure 1. PRISMA flowchart of the process.

3. Results

3.1 Results selection process

The study selection process is presented in Figure 1, according to the PRISMA diagram. A total of 2112 records were identified, of which 789 were duplicates, resulting in 1323 records being included in the screening stage. These were evaluated by reviewing the title and abstract, applying the inclusion and exclusion criteria previously defined in the methodology section. After this stage, 1284 records were excluded, leaving 39 articles for full-text review. During the retrieval process, 4 articles could not be obtained due to paid access restrictions. Consequently, 35 full-text articles were evaluated. Of these, 11 were excluded for the following reasons: evaluating only perception/satisfaction (n = 4), corresponding to an ineligible publication type (n = 2), constituting validation or technical development without an educational outcome for the student (n = 3), presenting a context not corresponding to practical clinical training (n = 1), or including an ineligible or inseparable population (n = 1). Finally, 24 studies were included in the review.

3.2 General Characteristics of the Selected Items

In total, 24 studies were included that examined the use of different AI technologies in the practical clinical training of medical students. The articles were published between 2022 and 2026,

with a concentration in 2024 and 2025, reflecting a recent increase in publications on educational applications of AI, particularly in simulated clinical contexts and in interventions based on generative models and large-scale language models (8-31).

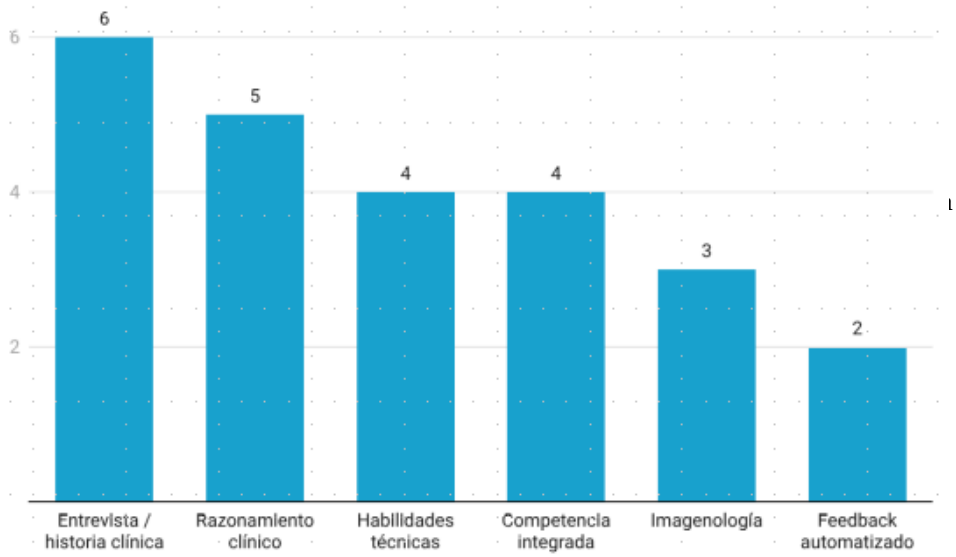


Figure 2. Distribution of articles based on main clinical competence.

From a geographical perspective, the studies were mainly concentrated in China, Germany, and Canada, with additional representation from Japan, Switzerland, Turkey, Hong Kong, Singapore, and the United States, as shown in Figure 3 and Table 2 in the appendix. This distribution suggests that the available evidence comes mostly from contexts with high technological capacity and digital infrastructure for integrating AI into medical training environments (10, 13-18, 24, 28-30).

Regarding the methodological design, a combination of randomized clinical trials, non-randomized controlled studies, prospective parallel designs, pilot or feasibility studies, mixed studies, and studies focused on the validation of automated assessment systems were identified. In general, a significant proportion of the sample corresponded to comparative or quasi-experimental studies, especially in clinical simulation, medical interviewing, Mini-CEX, and procedural training (8, 9, 11-13, 15, 17, 28, 31).

The sample size was heterogeneous, ranging from studies with small samples typical of pilot interventions or simulation trials to larger cohorts implemented in curricular or institutional contexts (8, 9, 11, 12, 15, 27, 28). Most participants were undergraduate medical students, including students in early, intermediate, or advanced stages of training, and some studies also incorporated clinical experts as a reference to validate evaluation criteria, agreement, or quality of automated feedback (10, 18, 31).

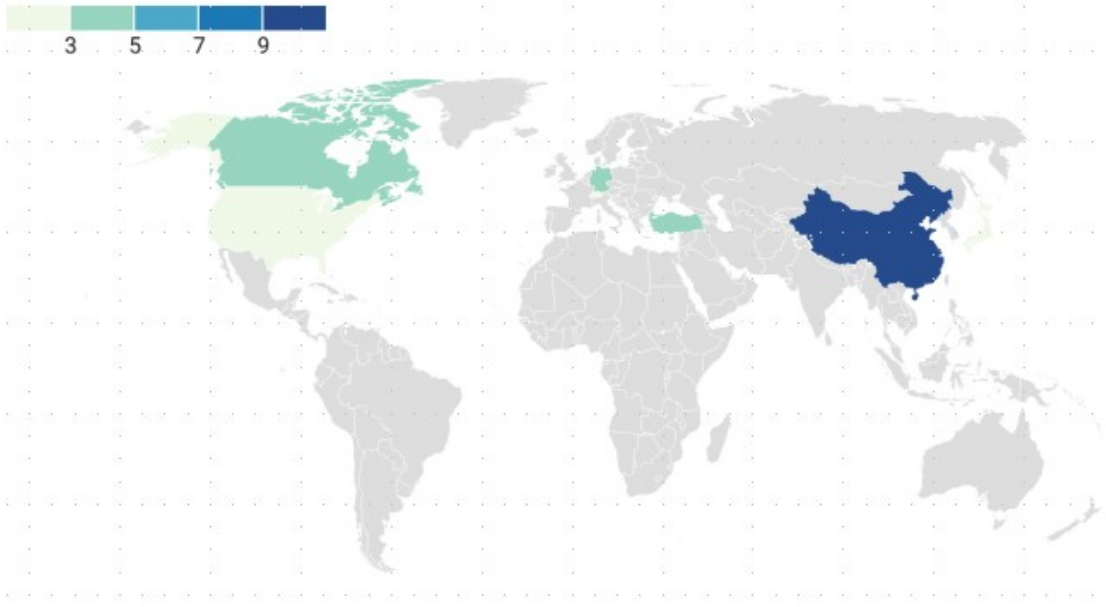


Figure 3. Geographical map of the selected articles.

Regarding the educational context, the development of interventions in simulated or structured environments predominated, such as simulated or virtual patients, OSCE-type stations, Mini-CEX assessments, virtual tutoring platforms, skills laboratories, and surgical simulation (8, 9, 11-13, 15, 17, 20, 23, 25). To a lesser extent, some studies were integrated into curricular contexts closer to clinical practice or real diagnostic workflows, such as clinical courses or AI-supported PACS systems (29-30).

Regarding the technologies used, LLMs and generative tools constituted the most frequent category, especially in interventions focused on clinical interviewing, diagnostic reasoning, conversational simulated patients, and automated feedback (9, 10, 12, 16-18, 22-25, 27). AI systems for technical/procedural training and surgical simulation were also identified, including computer vision models, multimodal analysis, and adaptive tutors (8, 20, 28, 31), as well as tools associated with diagnostic interpretation and imaging (21, 26, 30).

Finally, the reported outcomes were mainly concentrated on performance measures in simulated scenarios, structured scores of clinical competence (e.g., OSCE, Mini-CEX), measures of technical performance and agreement between AI and human evaluators, while studies with longitudinal follow-up or with evaluation of transfer to the real clinical environment were less frequent (10, 12, 13, 15, 17, 18, 20, 31).

3.3 Distribution according to domain of competence

To characterize the practical clinical competencies in which AI is being used, the studies were classified into a taxonomy of six domains, the distribution of which is shown in Figures 1 and 2. The most frequent domains were clinical interviewing, medical history taking, and communication (n = 6) and clinical reasoning and diagnostic decision-making (n = 5). The first group included studies focused on medical history taking, clinician-patient interaction, and interview practice using conversational simulated patients, chatbots, or digital patients (9, 11, 12, 16, 22, 25). The second group included interventions in which AI was used to guide clinical reasoning, case resolution, or diagnostic decision-making (14, 19, 24, 27, 29).

The domains of technical/procedural skills (n = 4) and integrated clinical competence (n = 4) were located at an intermediate frequency. The former grouped studies on technical or surgical training with simulation, adaptive tutoring, and automated feedback of psychomotor performance (8, 20, 28, 31). The latter included interventions evaluated using global measures of clinical competence, such as Mini-CEX or integrated performance in structured clinical scenarios (13, 15, 17, 23).

The least represented domains were imaging and diagnostic interpretation (n = 3) and automated performance evaluation and feedback (n = 2). In imaging, studies were identified that integrated AI into diagnostic interpretation or image workflows, including ultrasound training, keratitis, and AI integration in PACS (21, 26, 30). Meanwhile, the automated feedback domain included studies focused primarily on consistency, stability, or agreement between AI and human evaluators in feedback or scoring tasks (10, 18).

Taken together, this distribution shows that the evidence is mainly concentrated in domains that can be simulated, structured, and scaled relatively easily using AI, especially clinical interviewing, case-based reasoning, and technical simulation training (9, 14, 16, 20, 28).

3.4 Relationship between domain of competence and type of AI

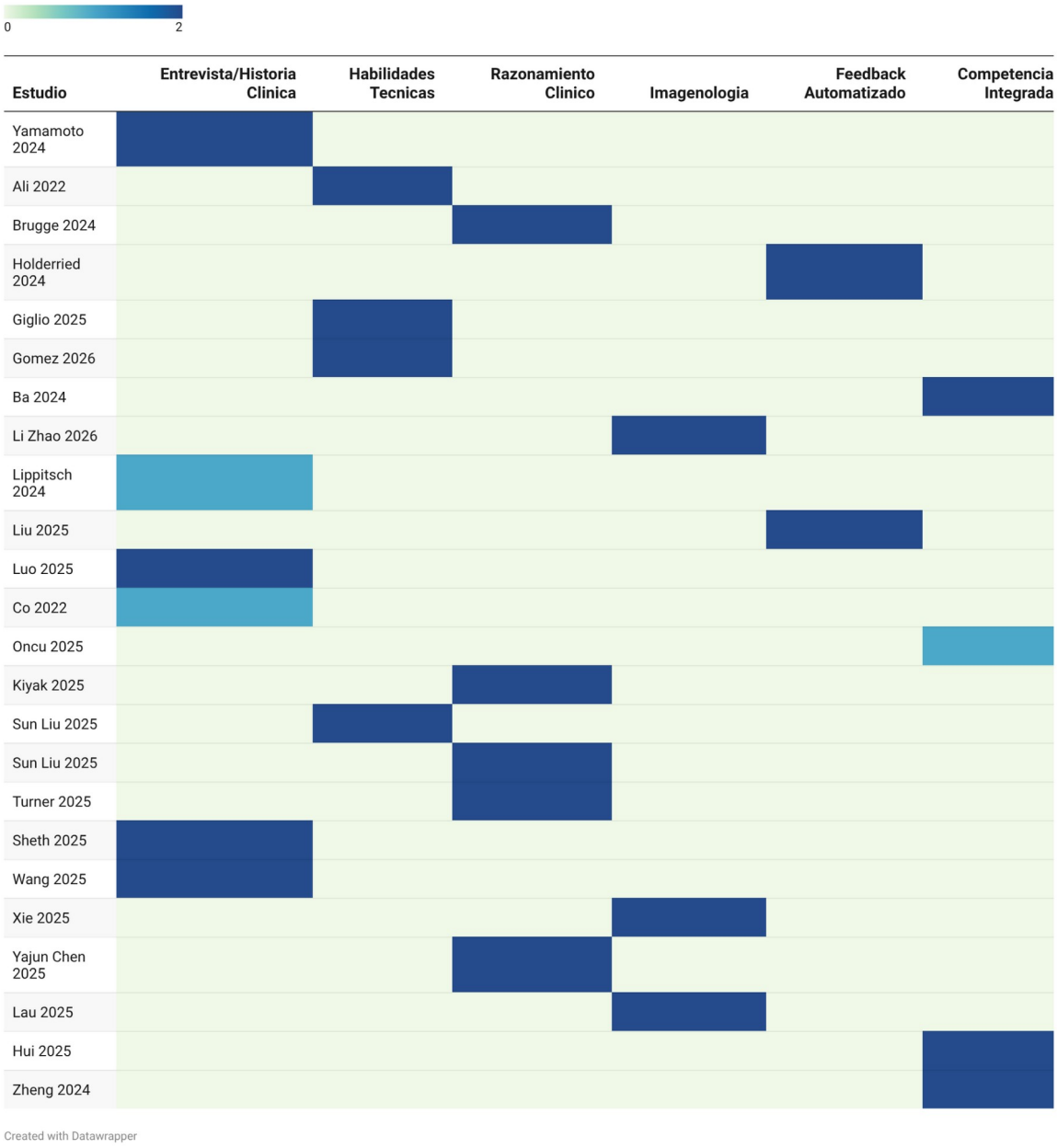
The relationship between competence domain and AI type is represented in Figure 2. In general terms, LLMs were the predominant technology in the clinical interview/communication and clinical reasoning domains, where they were used to build conversational simulated patients, support anamnesis, guide clinical discussions, and provide feedback or decision support (9, 10, 12, 14, 16, 19, 22, 24, 25, 28).

In contrast, in the domain of technical/procedural skills, AI systems combined with simulation, computer vision, motion analysis or adaptive tutoring were more frequently observed, aimed at monitoring performance, identifying errors and providing structured feedback in real time (8, 20, 28, 31).

In imaging and diagnostic interpretation, interventions were primarily associated with tools integrated into diagnostic environments or AI-assisted training materials, including AI-supported PACS systems, image generation, and support for specific interpretive tasks (21, 26, 30). Furthermore, in the automated feedback domain, AI was mainly used to standardize assessment and feedback, with an emphasis on consistency or agreement with human evaluators (10, 18).

Overall, the general pattern suggests that LLMs dominate applications focused on verbal interaction, reasoning, and simulated patients, while other approaches such as ML/DL, computer vision, or multimodal systems appear more frequently in technical or structured assessment domains (8, 10, 18, 20, 31).

Figura 4. Mapa de calor de la distribución de los estudios por dominio de competencia clínica.



3.5 Subset of studies with direct comparison

Table 3 in the appendix summarizes the subset of studies that included a direct comparison between an AI-based intervention and a comparator. This subset consisted of studies that compared AI with traditional teaching methods, human instruction, historical cohorts, or other intervention conditions, and constituted a significant part of the total corpus (8, 9, 11-13, 15-17, 20, 25, 28, 31).

In the clinical interviewing and communication domain, direct comparisons focused on simulated patients, chatbots, or digital patients for history-taking training, showing results that ranged from comparable performance to improvements in some specific measures of performance or interview structure (9, 11, 12, 16, 25). In integrated clinical competence, comparisons were observed in contexts assessed using Mini-CEX or other structured formats, comparing traditional modalities with AI or LLM support (13, 15, 17).

In procedural skills, studies compared AI-based tutoring or feedback with expert instruction or traditional teaching, using simulated technical performance scales as the main outcome (8, 20, 28, 31). Taken together, this subset shows that comparative evidence exists across multiple competency domains. However, the studies differ substantially in their comparators, outcomes, follow-up duration, and implementation contexts, so these findings should be interpreted as a descriptive mapping of reported comparative patterns, rather than as a basis for aggregate inferences about relative effectiveness (8, 9, 12, 13, 28).

3.6 Narrative synthesis of patterns and gaps

In general terms, recent evidence has focused on three main axes: (i) clinical interview and anamnesis training using simulated patients or chatbots, (ii) support for clinical reasoning in case-based scenarios, and (iii) training and evaluation of technical/procedural skills in simulation (8-10, 12, 14, 16, 20, 24, 28, 31).

Conversely, a lower relative density of studies was observed in large-scale automated assessment, sustained integration within real clinical contexts, and research evaluating the longitudinal retention of skills or their transfer to clinical performance with real patients (10, 18, 30). Likewise, although several studies used structured performance measures and some included direct comparators, the evidence remained heterogeneous in terms of outcomes, follow-up duration, comparators, and implementation scenarios (8, 9, 12, 13, 15, 17).

4. Discussion

This scoping review demonstrates that the integration of AI into the practical clinical training of undergraduate medical students is in a phase of rapid expansion, but is still undergoing conceptual, methodological, and pedagogical consolidation. Rather than conclusively demonstrating its superiority over traditional methods, the findings suggest that AI is redefining the conditions under which clinical competencies are acquired, practiced, and assessed, particularly in contexts where scalability and standardization are structural constraints.

One of the most relevant elements was the concentration of evidence in highly structuring and simulable domains, such as clinical interviewing, case-based diagnostic reasoning, and procedural skills in controlled environments, as shown in Figure 1 and Table 2. This pattern is consistent with previous, broader reviews, which show that AI applications in medical education tend to concentrate in areas such as interactive learning, surgical skills assessment, and diagnostic interpretation—that is, in contexts where tasks can be standardized, simulated, and repeated with relative ease (3, 5). In particular, recent literature on patients based on LLM models suggests that these systems show special promise for training in communication skills and conversational clinical simulations, precisely because they operate well on verbalizable tasks, with relatively limited scenarios and the possibility of immediate feedback (32). In this sense, rather than replacing traditional clinical teaching, AI appears to be optimizing those learning components that are most susceptible to standardization, which could favor more frequent, accessible, and potentially individualized deliberate practice for each student. However, this same affinity with highly simulable domains also helps explain why most of the evidence continues to focus on structured and simulated contexts, rather than real or longitudinal clinical scenarios (32). This distribution should also be interpreted considering that practical clinical training was conceptualized in this review as a continuum encompassing scenarios with varying degrees of pedagogical authenticity, environmental control, and proximity to clinical performance with real patients.

In addition to the above, this same concentration of evidence highlights a significant limitation: the scarcity of evidence in real-world clinical contexts and in more complex, dynamic, and interpersonal competencies that depend on uncertainty, patient variability, and contextualized

decision-making. This observation is consistent with previous reviews showing that most educational applications of AI have been developed in highly structured areas, such as interactive learning, procedural assessment, or diagnostic interpretation, with evidence predominantly generated in simulated or controlled environments (3, 5). Similarly, recent reviews on measurable educational outcomes indicate that the available evidence continues to be based primarily on single-center studies with small sample sizes and limited duration, making it difficult to determine whether the observed benefits translate into longitudinal retention or improved clinical performance in real-world studies (6). More specifically, the literature on virtual patients based on LLM suggests that, although these systems show potential for training and developing communication skills, they do not replace real-world interactions or fully reproduce the relational and contextual complexity of the clinical encounter (32). In this sense, the limited evaluation of transfer to the clinical care setting suggests that, to date, much of the evidence remains at an intermediate level between simulated learning and authentic clinical performance.

From an educational framework that articulates deliberate practice, self-regulated learning, and instructional scaffolding, this distribution suggests that AI seems best suited, for now, to competencies that can be broken down into explicit, observable, and repeatable tasks, where performance criteria are relatively stable and feedback can be delivered immediately and in a standardized manner. From the perspective of deliberate practice, these tools could expand opportunities for goal-oriented repetition and frequent feedback; from the perspective of self-regulated learning, they could support performance monitoring, gap identification, and progressive adjustment of learning; and from the perspective of instructional scaffolding, they could offer graduated support that facilitates progression toward greater autonomy (3, 32, 34). This helps to explain its increased presence in structured history taking, case-based reasoning, and procedural simulation training, as well as the recent interest in conversational virtual patients and more personalized learning experiences (3, 32, 34). In these domains, its main contribution would not seem to lie in replacing human teaching, but rather in expanding opportunities for deliberate practice, offering more consistent scaffolding, and supporting more accessible and adaptive learning pathways. However, this same logic raises a significant limitation: when clinical competence depends on ambiguity, contextual judgment, relational negotiation, or the dynamic integration of multiple environmental cues, excessive standardization risks unduly simplifying learning and diminishing the complexity of the clinical encounter. In this sense, rather than conceiving AI as a one-size-fits-all solution for clinical training, the findings suggest interpreting its pedagogical value in a situated manner, according to the type of competence involved and the training context in which it is implemented, while remaining cautious about broad extrapolations regarding its educational scope (3, 5, 6).

On the other hand, a subset of comparative studies reported comparable or, in some cases, favorable results for AI-based interventions, particularly in scenarios where AI was used to support feedback, tutoring, or clinical reasoning (8, 9, 17, 28, 31). However, these findings should be interpreted with caution, as the studies differed in comparators, outcomes, implementation conditions, and the intensity of educational support offered, which limits the possibility of drawing aggregate conclusions about relative effectiveness. In this comparative subset, several studies used comparators corresponding to teaching strategies with less feedback intensity, less standardization, or less availability of immediate support, making it difficult to attribute the observed benefits exclusively to the AI component of the intervention (8, 9, 17, 28, 31). This characteristic limits the ability to establish causal relationships between the use of AI and the reported results, since the differences could reflect, at least in part, inequalities in the quality, structure, or intensity of the educational intervention rather than a specific effect of the technological component. This is consistent with recent reviews that, while describing promising findings in certain domains, also underscore the heterogeneity of comparators, outcomes, and designs, as well as the need for methodologically more robust studies before drawing conclusions about the overall superiority or

comparative effectiveness of AI versus conventional teaching strategies (5-6). Consequently, the favorable results reported in this subset should be interpreted as preliminary comparative signals and not as direct causal evidence of AI superiority, especially when comparators do not control for training intensity in an equivalent manner.

Furthermore, a significant shift is emerging in the pedagogical role of AI: from a tool for accessing content to an active agent within the educational process. In particular, LLMs have fostered the development of conversational simulated patients and automated feedback systems that not only deliver information but also interact, respond, and adjust the learning experience based on student performance (9, 10, 17, 22, 25, 32). This shift is consistent with the literature, which describes LLMs as technologies with the potential to offer more personalized learning experiences, immediate feedback, and more continuous support for self-directed learning, especially in simulation and structured clinical training contexts (32, 34). In this sense, AI appears to be bringing medical education closer to models more focused on active and personalized learning, in line with the principles of self-regulated learning and deliberate practice. However, this change also introduces important challenges in terms of reliability, transparency, biases and pedagogical control, particularly because of the probabilistic and, at times, non-deterministic nature of these systems, which reinforces the need for supervised curriculum integration and rigorous pedagogical evaluation before assuming their sustained benefit (5, 6, 32, 33, 35).

Another critical aspect was the methodological heterogeneity observed in the included studies. The variability in designs, sample sizes, outcomes, and comparators limits the possibility of drawing robust conclusions about the effectiveness of these interventions. This finding is consistent with previous reviews that describe a broad but methodologically heterogeneous literature, with a predominance of studies in specific domains and comparators that are difficult to compare (3, 5). Along the same lines, a recent systematic review on measurable educational outcomes in health professional education concluded that the available evidence remains overall poor, based mainly on single-center studies with small samples and focused on outcomes corresponding to intermediate levels of Kirkpatrick's hierarchy, without evaluating more complex workplace behaviors or the impact on healthcare outcomes (6). In this context, the predominance of short-term studies makes it difficult to interpret the real impact of these interventions beyond immediate performance, which is particularly relevant in medical education, where outcomes of interest also include retention, transfer to clinical practice and impact on the quality of patient care (5-6).

This heterogeneity also constitutes a structural limitation of the present review that deserves to be explicitly highlighted. Since no formal assessment of methodological quality or risk of bias was conducted—a decision consistent with the exploratory purpose of a scoping review—it is not possible to rank the robustness of the included studies or compare the relative strength of their findings. Consequently, the reported comparative results should be read strictly as preliminary descriptive signals, especially within a corpus that is heterogeneous in terms of designs, comparators, outcomes, and educational contexts. The lack of evaluation also limits the possibility of assessing susceptibility to bias and drawing inferences about comparative effectiveness; therefore, any interpretation of the reported patterns should remain at the level of a descriptive mapping of the field.

In this context, the decision to exclude studies focused exclusively on satisfaction, perception, or attitudes toward AI stemmed from the desire to prioritize evidence with greater practical relevance to this review—that is, evidence linked to more substantive educational or implementation outcomes. Although perceptions and acceptability are relevant dimensions in emerging fields such as AI, broadly including only perceptual studies could have artificially inflated the volume of available evidence without providing sufficient information on learning, performance, or effective educational integration. However, this decision also means that aspects

such as acceptability, adoption, and perceived barriers to implementation were indirectly represented in the final corpus.

From a systemic perspective, the geographical distribution of studies suggests that the implementation of AI in medical education is strongly conditioned by the availability of digital infrastructure, technological capacity, and institutional resources. This pattern, also observed in other reviews in the field, raises the risk of widening gaps between educational contexts, disproportionately favoring institutions with greater access to platforms, connectivity, technical support, and curricular integration capacity. Consequently, the discussion on AI in medical education cannot be separated from considerations of equity, access, transparency, and governance, especially when the evidence comes predominantly from environments with high technological availability and when recent ethical frameworks insist on the need to implement these tools under principles of justice, safety, accountability, and pedagogical oversight (3, 32).

Table 1 shows a synthesis of the current state of knowledge: what we know and what we don't know about the use of AI in the practical clinical training of undergraduate medical students.

Table 1. Synthesis of the current state of knowledge: what we know and what we do not know about the use of AI in the practical clinical training of undergraduate medical students.

Dimension	What do we know?	What we don't know / Gaps identified
Predominant technologies	LLMs and generative tools are the most frequently studied technologies, especially in clinical interviewing, diagnostic reasoning, and conversational simulated patients.	There is little evidence on non-LLM AI applications in complex and integrated domains; most of the tools evaluated are prototypes or experimental versions.
Implementation contexts	The evidence focuses on structured simulation (OSCE, Mini-CEX, virtual patients, skills laboratories)	There is very limited evidence in real-world supervised clinical practice; it is unknown how these tools behave outside of controlled environments.
Immediate educational results	Some comparative studies report comparable or favorable results for AI in immediate performance measures in simulation	There is no robust evidence on longitudinal retention of skills or on transfer to the real clinical setting with patients
Comparative effectiveness	Comparative evidence exists in multiple domains (clinical interview, technical skills, clinical reasoning, integrated competence)	The comparators are heterogeneous and frequently less feedback-intensive, which makes it impossible to attribute the observed effects specifically to the technological component.
Pedagogical and ethical framework	The need for governance frameworks, curriculum oversight, and ethical principles for the integration of AI is recognized.	There are no validated formal pedagogical or ethical frameworks; the implications for curriculum design remain mostly implicit in the literature

Equity and context	The geographical distribution shows concentration in contexts with high technological capacity and digital infrastructure	Lack of education in resource-limited environments or low and middle-income contexts; risk of widening structural educational gaps

From a curriculum design perspective, the findings suggest that integrating AI into clinical training requires explicit pedagogical decisions about its role: whether as a substitute for scarce practice opportunities, as scaffolding for self-directed learning, or as a complementary formative assessment tool. In each case, the implications for instructional design, faculty supervision, and competency assessment differ substantially. Regardless of the curriculum model—whether competency-based, outcome-based, or based on progressive training stages—the responsible incorporation of AI will require clearly defining the pedagogical objectives to be achieved, the quality criteria for the tools adopted, and the mechanisms for monitoring and continuously evaluating its educational impact.

Finally, the findings reinforce the need to move towards a more mature research agenda that transcends feasibility or efficacy studies in controlled environments. In line with recent reviews, future research should prioritize longitudinal designs, evaluation in real clinical contexts, measurement of skills transfer, and comparison with optimized pedagogical interventions—and not solely with traditional methods of less intensive training. Likewise, greater standardization of outcomes, comparators, and reporting criteria is desirable so that future literature can more clearly distinguish between effects attributable to the technological component and those resulting from greater structuring, frequency of feedback, or intensity of educational support (5-6).

Taken together, the available evidence suggests that AI could make a significant contribution to practical clinical training, although some of these benefits may be mediated by greater structuring and intensity of educational support, rather than by the technological component in isolation. In this context, AI should not necessarily be understood as a replacement for existing teaching models, but rather as a resource capable of complementing and redesigning certain components of teaching. Its value seems to be concentrated, for now, in the possibility of scaling up practice opportunities, standardizing feedback, and fostering more accessible and, in some cases, more personalized learning experiences. However, its effective integration will require a critical, pedagogically guided, and evidence-based approach that considers not only its technological promise but also its methodological, ethical, and contextual limitations (3, 5-6, 35).

5. Conclusions

- This scoping review allowed us to map and characterize the current evidence on the use of AI in the practical clinical training of undergraduate medical students. The findings show that AI has been integrated primarily into simulated scenarios and structured tasks, such as clinical interviewing, diagnostic reasoning, and procedural training, with a growing predominance of tools based on large-scale language models.
- In summary, the evidence suggests that these technologies can expand opportunities for deliberate practice, standardize feedback, and support more accessible and, in some cases, more personalized learning processes. However, their current implementation is concentrated in intermediate stages of learning, with limited evidence regarding transfer to real-world clinical settings, longitudinal impact, and performance in complex, unstructured, or highly variable contexts.
- Furthermore, the heterogeneity of designs, outcomes, and comparators, along with the scarcity of evaluations in authentic contexts and consistent pedagogical and ethical frameworks, highlights the need to move toward more methodologically robust and

pedagogically informed research. In this context, future research should prioritize the evaluation of AI in real-world clinical settings, the measurement of skills transfer, and the development of ethical and educational frameworks to guide its responsible integration. In short, AI is emerging as a promising tool to complement and redesign certain components of clinical training, although its effective adoption will require critical, contextualized, and evidence-based implementation.

Funding: There has been no funding.

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

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of the following work, Artificial Intelligence was used in a limited way to improve the writing. Subsequently, the content was reviewed and edited, and the authors assume full responsibility for the content of the publication.

Authors' contributions: JB participated in the literature search, literature review, manuscript writing and revision, and the creation of figures and images. DP and CM participated in the literature search, literature review, manuscript writing and revision. TC participated in the literature search, literature review, manuscript writing and revision, and the creation of figures and images. AH participated in the review and editing of the manuscript as a medical education expert. All authors read and approved the final version of the manuscript.

6. References

1. Dewan P, Khalil S, Gupta P. Objective structured clinical examination for teaching and assessment: Evidence-based critique. *Clin Epidemiol Glob Health.* **2024,** 25, 101477. https://doi.org/10.1016/j.cegh.2023.101477
2. Tozsin A, Ucmak H, Soyturk S, Aydin A, Gozen AS, Al Fahim M, Güven S, Ahmed K. The role of artificial intelligence in medical education: A systematic review. *Surg Innov.* **2024,** 31(4), 415-423. https://doi.org/10.1177/15533506241248239
3. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, Hanson J, Haas M, Spadafore M, Grafton-Clarke C, Gasiea RY, Michie C, Corral J, Kwan B, Dolmans D, Thammasitboon S. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach.* **2024,** 46(4), 446-470. https://doi.org/10.1080/0142159x.2024.2314198
4. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - Scoping literature review. *Med Sci Educ.* **2024,** 35(1), 555-567. https://doi.org/10.1007/s40670-024-02206-6
5. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. *Med Sci Educ.* **2025,** 35(3), 1803-1816. https://doi.org/10.1007/s40670-025-02373-0
6. 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
7. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol.* **2005,** 8(1), 19-32. https://doi.org/10.1080/1364557032000119616
8. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: A randomized clinical trial. *JAMA Network Open.* **2022,** 5(2), e2149008. https://doi.org/10.1001/jamanetworkopen.2021.49008
9. Co M, Yuen THJ, Cheung HH. Using clinical history taking chatbot mobile app for clinical bedside teachings – A prospective case control study. *Heliyon.* **2022,** 8(6), e09751. https://doi.org/10.1016/j.heliyon.2022.e09751
10. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. 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

11. Lippitsch A, Steglich J, Ludwig C, Kellner J, Hempel L, Stoevesandt D, 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.* **2024,** 244, 107964. https://doi.org/10.1016/j.cmpb.2023.107964
12. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, et al. Enhancing medical interview skills through AI-simulated patient interactions: Nonrandomized controlled trial. *JMIR Med Educ.* **2024,** 10, e58753. https://doi.org/10.2196/58753
13. Ba H, Zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: A comparative study of ChatGPT-assisted and traditional teaching methods. *BMC Med Educ.* **2024,** 24(1), 558. https://doi.org/10.1186/s12909-024-05565-1
14. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Geist M, Jeselsohn M, 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.* **2024,** 24(1), 1391. https://doi.org/10.1186/s12909-024-06399-7
15. Zheng K, Shen Z, Chen Z, et al. Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education. *BMC Med Educ.* **2024,** 24(1), 1003. https://doi.org/10.1186/s12909-024-05977-z .
16. Luo MJ, Bi S, Pang J, Liu L, Tsui CK, Lai Y, 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 .
17. Hui Z, Zewu Z, Jiao H, et al. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. *BMC Med Educ.* **2025,** 25, 50. https://doi.org/10.1186/s12909-024-06321-1 .
18. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu 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.* **2025,** 11, e73419. https://doi.org/10.2196/73419 .
19. 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
20. Giglio B, Albeloushi A, Alhaj AK, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial intelligence-augmented human instruction and surgical simulation performance: A randomized clinical trial. *JAMA Surg.* **2025,** 160(9), 993-1003. https://doi.org/10.1001/jamasurg.2025.2564 .
21. Lau YH, Acharyya S, Wee CWL, Xu H, Pulido Saclolo R, Cao K, et al. Effectiveness of traditional, artificial intelligence-assisted, and virtual reality training modalities for focused cardiac ultrasound skill acquisition: A randomized controlled study. *Ultrasound J.* **2025,** 17(1), 61. https://doi.org/10.1186/s13089-025-00469-7 .
22. Sheth U, Lo M, McCarthy J, Baath N, Last N, Guo E, et al. Understanding the role of large language model virtual patients in developing communication and clinical skills in undergraduate medical education. *Int Med Educ.* **2025,** 4, 39. https://doi.org/10.3390/ime4040039
23. Öncü S, Torun F, Ülkü HH. AI-powered standardized patients: Evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med Educ.* **2025,** 25(1), 278. https://doi.org/10.1186/s12909-025-06877-6 .
24. Turner L, Kelleher M, Overla S, Zheng W, Gregath A, Gharib M, et al. Harnessing the generative power of AI to move closer to personalized medical education. *Acad Med.* **2025,** 100(12), 1447-1451. https://doi.org/10.1097/ACM.0000000000006185 .

25. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: A randomized clinical trial. *BMC Med Educ.* **2025,** 25(1), 1030. https://doi.org/10.1186/s12909-025-07614-9 .
26. Xie W, Yuan Z, Si Y, Huang Z, Li Y, Wu F, et al. Enhancing medical students' diagnostic accuracy of infectious keratitis with AI-generated images. *BMC Med Educ.* **2025,** 25(1), 1027. https://doi.org/10.1186/s12909-025-07592-y
27. Chen Y. Evaluation of the impact of AI-driven personalized learning platform on medical students' learning performance. *Front Med (Lausanne).* **2025,** 12, 1610012. https://doi.org/10.3389/fmed.2025.1610012
28. Sun Y, Liu F. Evaluating the impact of AI-tutoring versus expert human instruction on surgical skills in medical students. *Educ Inf Technol.* **2025,** 30(18), 26413-26431. https://doi.org/10.1007/s10639-025-13779-z .
29. Sun Y, Liu F. Real-world implementation of an AI learning tool-MetaGP-Edu in medical education: A multi-center cohort study. *Comput Educ.* **2025,** 237, 105388. https://doi.org/10.1016/j.compedu.2025.105388 .
30. Li J, Zhao H. Workflow-embedded AI as a cognitive scaffold: A randomized trial on knowledge retention and diagnostic competence in undergraduate radiology education. *Eur J Radiol Open.* **2026,** 16, 100724. https://doi.org/10.1016/j.ejro.2026.100724 .
31. Gomez C, Seenivasan L, Zou X, Yoon J, Chu S, Leong A, et al. Explainable AI for automated user-specific feedback in surgical skill acquisition. In: Guo X, Jin Y, Lamdouar H, Ouyang C, Men Q, Sahu M, Vedula SS, editors. Human-AI collaboration—First international workshop, HAIC 2025, held in conjunction with MICCAI 2025, proceedings. Lecture Notes in Computer Science, vol. 16214. Cham: Springer; **2026,** 25-34. https://doi.org/10.1007/978-3-032-08970-0_3 .
32. Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, et al. Embracing the future of medical education with large language model-based virtual patients: Scoping review. *J Med Internet Res.* **2025,** 27, e79091. https://doi.org/10.2196/79091 .
33. Mavrych V, Yousef EM, Yaqinuddin A, Bolgova O. Large language models in medical education: A comparative cross-platform evaluation in answering histological questions. *Med Educ Online.* **2025,** 30(1), 2534065. https://doi.org/10.1080/10872981.2025.2534065 .
34. Fortuna A, Prasetya F, Samala AD, Rawas S, Criollo-C S, Kaya D, et al. Artificial intelligence in personalized learning: A global systematic review of current advancements and shaping future opportunities. *Soc Sci Humanit Open.* **2025,** 12, 102114. https://doi.org/10.1016/j.ssaho.2025.102114 .
35. Tran M, Balasooriya C, Jonnagaddala J, Leung GKK, Mahboobani N, Ramani S, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *NPJ Digit Med.* **2025,** 8, 315. https://doi.org/10.1038/s41746-025-01721-z