



ARTIFICIAL INTELLIGENCE IN UNDERGRADUATE MEDICAL EDUCATION: A SCOPING REVIEW OF EDUCATIONAL APPLICATIONS

INTELIGÊNCIA ARTIFICIAL NA GRADUAÇÃO MÉDICA: UMA REVISÃO DE ESCOPO DAS APLICAÇÕES EDUCACIONAIS

Flávia Silva e Oliveira^{1*}, Beatriz Pires Carcute², Emily Caroline Alves Martins², Júlia Castro Andrade², Luana Travassos Batista³

¹PhD in Nursing, Federal University of Goiás (UFG), Goiânia (GO), Brazil; Faculty of Medicine, Federal University of Catalão (UFCAT), Catalão (GO), Brazil; ²Medical student, Faculty of Medicine, Federal University of Catalão (UFCAT), Catalão (GO), Brazil; ³PhD in Basic and Applied Immunology, Federal University of Amazonas (UFAM), Manaus (AM), Brazil; Substitute Professor, University Center of Mineiros (UNIFIMES), Trindade Campus, Trindade (GO), Brazil.

***Corresponding author:** Flávia Silva e Oliveira – **Email:** flavia04silva@gmail.com

Received: 11 aug. 2025

Accepted: 06 jan. 2026

Editors-in-Chief: Dr. Leonardo Pestillo de Oliveira and Dr. Mateus Dias Antunes

This is an open-access article distributed under the terms of the Creative Commons Attribution license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



ABSTRACT: The objective was to map the application of artificial intelligence (AI) in undergraduate medical education, focusing on the tools used, learning outcomes, performance, and gaps in the literature. Searches were conducted in PubMed, Scopus, Web of Science, EBSCOhost (CINAHL/MEDLINE), ScienceDirect, and Google Scholar, including empirical studies in English, Spanish, and Portuguese, without temporal restrictions. Study selection was performed in duplicate using Rayyan, and findings were described based on the TPACK framework. Of 605 records, 58 studies (2018–2024) were included, mostly from high-income countries. Quantitative designs predominated, with small to medium sample sizes. AI was primarily applied to theoretical instruction, educational assessment, and simulation-based training, often linked to improvements in academic performance and objective skills. It is concluded that AI demonstrates educational potential but remains limited by methodological heterogeneity, theoretical gaps, and contextual inequalities, highlighting the need for structured, ethical, and longitudinal curricular integration.

KEYWORDS: Artificial Intelligence. Learning. Medical Education. Student. Teaching.

RESUMO: O objetivo foi mapear a aplicação da inteligência artificial (IA) na graduação médica, com ênfase nas ferramentas utilizadas, desfechos de aprendizagem, desempenho e lacunas da literatura. As buscas foram realizadas no PubMed, Scopus, Web of Science, EBSCOhost (CINAHL/MEDLINE), ScienceDirect e Google Scholar, incluindo estudos empíricos em inglês, espanhol e português, sem restrição temporal. A seleção dos estudos foi realizada em duplicata utilizando o Rayyan, e os achados foram descritos a partir do referencial TPACK. Dos 605 registros, 58 estudos (2018–2024) foram incluídos, majoritariamente de países de alta renda. Predominaram delineamentos quantitativos, com amostras de pequeno a médio porte. A IA foi aplicada principalmente ao ensino teórico, avaliação educacional e treinamento em simulações, frequentemente associada a melhorias no desempenho acadêmico e habilidades objetivas. Conclui-se que a IA demonstra potencial educacional, mas permanece limitada por heterogeneidade metodológica, lacunas teóricas e desigualdades contextuais, ressaltando a necessidade de integração curricular estruturada, ética e longitudinal.

PALAVRAS-CHAVE: Aprendizagem. Educação Médica. Ensino. Estudante. Inteligência Artificial.

1. INTRODUCTION

Artificial intelligence (AI) is rapidly transforming health care and medical education by replicating human cognitive functions and improving efficiency.¹ In health care, AI contributes to accurate diagnoses and prognosis through large-scale data analysis² and has gained relevance in health promotion and disease prevention, key priorities of contemporary public health. AI-based initiatives have been widely used to encourage healthy behaviors, lifestyle changes, and preventive actions, especially in chronic noncommunicable diseases.³

In medical education, AI has enabled safe, scalable, and interactive learning environments, including simulations, automated feedback, and adaptive assessment tools.³ International studies report their use in undergraduate medical education across North America, Europe, Asia, and other regions, supporting exam preparation, clinical skills training, feedback, and curricular activities. These technologies may democratize access to educational resources, standardize training, and promote equity in learning opportunities.^{4,5} When integrated into undergraduate curricula, AI can help prepare future physicians for preventive care, population health strategies, and digitally mediated clinical practice.^{6,7}

Despite growing global interest, AI initiatives in medical education vary widely across countries and institutions. Most studies originate from high-income countries, with limited evidence from low- and middle-income settings, revealing geographic and contextual imbalances. Additionally, initiatives range from isolated pilots and elective courses to simulation-based training and AI-supported assessments.

Within this rapid expansion, AI adoption has often occurred in an uncoordinated and conceptually diffuse manner, driven more by technological availability than pedagogical need. This has resulted in fragmented implementations with heterogeneous instructional designs, learning objectives, and evaluation strategies.^{8–11} Such variability limits comparability between studies and hinders the translation of findings into coherent curricular guidance, leaving educators without clear direction on aligning AI with curricula, learning outcomes, and professional competencies.^{8,12,13}

Many studies have compared AI tools, particularly large language models such as ChatGPT, with medical students' performance, mainly in examinations or question-answering tasks.¹⁴ However, these comparisons are rarely framed within broader curricular or pedagogical contexts, limiting their relevance for educational decision-making. Although AI may achieve performance comparable to or exceeding that of students, results are heterogeneous and sensitive to factors such as prompts, language, and assessment formats, raising concerns about reliability, validity, and educational meaning.¹⁵

Furthermore, existing evidence often prioritizes isolated performance metrics, with limited attention to learning processes, pedagogical integration, or curricular coherence.^{5,14,16–18} The lack of explicit educational frameworks has led to fragmented findings, constraining evidence-informed decisions on curriculum design, faculty development, and responsible AI adoption.^{8,13,19,20} Faculty expertise, institutional readiness, and local contexts are also critical for effective AI integration but remain underexplored.^{8,12,21}

In this context, conceptual models grounded in learning theory are essential to clarify how AI interacts with pedagogy and medical content. The Technological Pedagogical Content Knowledge (TPACK) framework²² provides a useful lens by emphasizing the integrated relationship between technology, pedagogy, and disciplinary knowledge.

Accordingly, this scoping review aims to map the application of AI in undergraduate medical education, identifying tools used, effects on student learning and performance, and gaps in the literature. By organizing evidence from an educational perspective, the review seeks to inform curriculum design

and teaching practice, supporting educators and institutions in navigating the opportunities and challenges of AI in increasingly digital academic and clinical environments.

2. METHODOLOGY

The review protocol was developed according to PRISMA-P guidelines and refined by the research team, and was registered in the Open Science Framework (<https://doi.org/10.17605/OSF.IO/PZERT>). This scoping review was conducted and reported in accordance with the PRISMA-ScR guidelines. A comprehensive search strategy using relevant keywords and index terms was adapted for each database, and the study selection process is presented in a PRISMA-ScR flow diagram (Fig. 1). Studies published in English, Spanish, and Portuguese were included, with no date restrictions.

2.1 REVIEW QUESTION

The Population–Concept–Context (PCC) framework was applied to systematically map evidence on artificial intelligence (AI) in undergraduate medical education. The guiding question: “What is the current state of AI application in the teaching–learning process for medical educators and students, and how does it enhance educational outcomes?”, was operationalized as follows. The Population included medical educators and/or undergraduate medical students. The Concept comprised AI technologies (e.g., adaptive tutoring systems, automated assessment platforms, intelligent feedback mechanisms), their curricular integration, and pedagogical functions. The Context focused on AI’s effects on student performance, instructor workload, and curriculum design, as well as technical, pedagogical, and ethical implementation challenges and opportunities.

2.2 TYPE OF SOURCES

The review included peer-reviewed quantitative, qualitative, and mixed-methods studies reporting empirical data on educational aspects related to AI in undergraduate medical education, such as learning outcomes, skills development, curriculum integration, teaching methods, assessment, or engagement. Excluded were studies involving only postgraduate learners or professionals; non-empirical publications; research protocols; retracted articles; and studies mentioning AI without examining its educational application. Non-academic sources were also excluded. Eligible systematic reviews were included, and their references were screened. When multiple publications reported the same dataset, the most methodologically comprehensive study was selected.

2.3 SOURCES AND SEARCH STRATEGY

On 9 December 2024, comprehensive searches were conducted in PubMed, Scopus, Web of Science, EBSCOhost (CINAHL/MEDLINE), ScienceDirect, and Google Scholar. Searches used MeSH terms and keywords related to medical educators, medical students, teaching–learning, artificial intelligence, machine learning, and deep learning, customized for each database (Appendix I). Gray literature and conference proceedings were retrieved via Google Scholar, and reference lists of relevant reviews were hand-searched. All records were imported into Zotero, and duplicates were removed.

2.4 STUDY SELECTION

Before screening, the four-member team piloted 20 references to align inclusion criteria and data-extraction procedures, with calibration meetings after each phase. Two reviewers independently screened titles and abstracts in Rayyan, followed by duplicate full-text assessment of eligible studies. Disagreements were resolved by consensus or by a third reviewer.

2.5 DATA EXTRACTION AND SYNTHESIS

A structured, pilot-tested form was used to extract key study information, including authorship, year and country of publication, study design, methodological characteristics, AI tools employed, main findings, and educational relevance. Data extraction was independently performed by two reviewers, with discrepancies resolved through discussion with a third reviewer to ensure consistency and accuracy.

To support the educational interpretation of the evidence presented in Table 1, predefined analytical fields were established. Study designs followed the terminology reported by the original authors; when unclear, classifications were based on the predominant methodological approach. Main findings corresponded to the primary educational outcomes reported in the Results section of each study, identified through full-text analysis and reflecting the principal effect of the AI intervention as defined by the original authors.

AI tool(s) were identified from methodological descriptions and referred to the artificial intelligence technologies used in each intervention. These included large language models (e.g., GPT-3.5, GPT-4, ChatGPT, Bard, Claude, Bing Chat, Copilot, PaLM-2, and Google Gemini); domain-specific or customized LLMs (e.g., AnatomyGPT); classical machine-learning models (e.g., support vector machines, k-nearest neighbours, and case-based reasoning engines); deep-learning architectures (e.g., convolutional, recurrent, and ensemble neural networks); hybrid AI systems combining data-driven and mechanistic approaches (e.g., ANN-PBPK models); integrated AI platforms providing automated feedback or simulation-based tutoring (e.g., Virtual Operative Assistant, ICEMS, Hepius, and the GE EF-quality indicator); and AI-related curricular content addressing AI concepts or system design without direct operational use of a specific AI tool (e.g., course content only, courses about AI, or simulated AI-CDSS design).

Studies were thematically grouped into four categories: (1) large language models; (2) automated feedback systems; (3) simulation platforms and examination assessment tools; and (4) curricular integration modules. Results were presented in descriptive and tabular formats, and a qualitative analysis was conducted to identify key trends, patterns, and gaps in the literature. The TPACK (Technological Pedagogical Content Knowledge) framework²² was used as an analytical lens to implicitly categorize studies across three domains—technological (use of AI tools), pedagogical (teaching and assessment strategies), and content (clinical and medical knowledge)—as well as their intersections, including the pedagogical application of technology and its integration with curricular content.

3. RESULTS

3.1 STUDIES' CHARACTERISTICS

A total of 555 studies were identified across selected databases, supplemented by 50 studies from grey literature. The thoroughness of the literature search strategy was confirmed, as no additional

studies emerged from the reference lists of articles. Zotero was used for duplicate removal, followed by verification in Rayyan to ensure data integrity.

Out of the 536 titles and abstracts screened, 145 were selected for full-text reading, resulting in the inclusion of 58 studies in this scoping review (Fig 1).

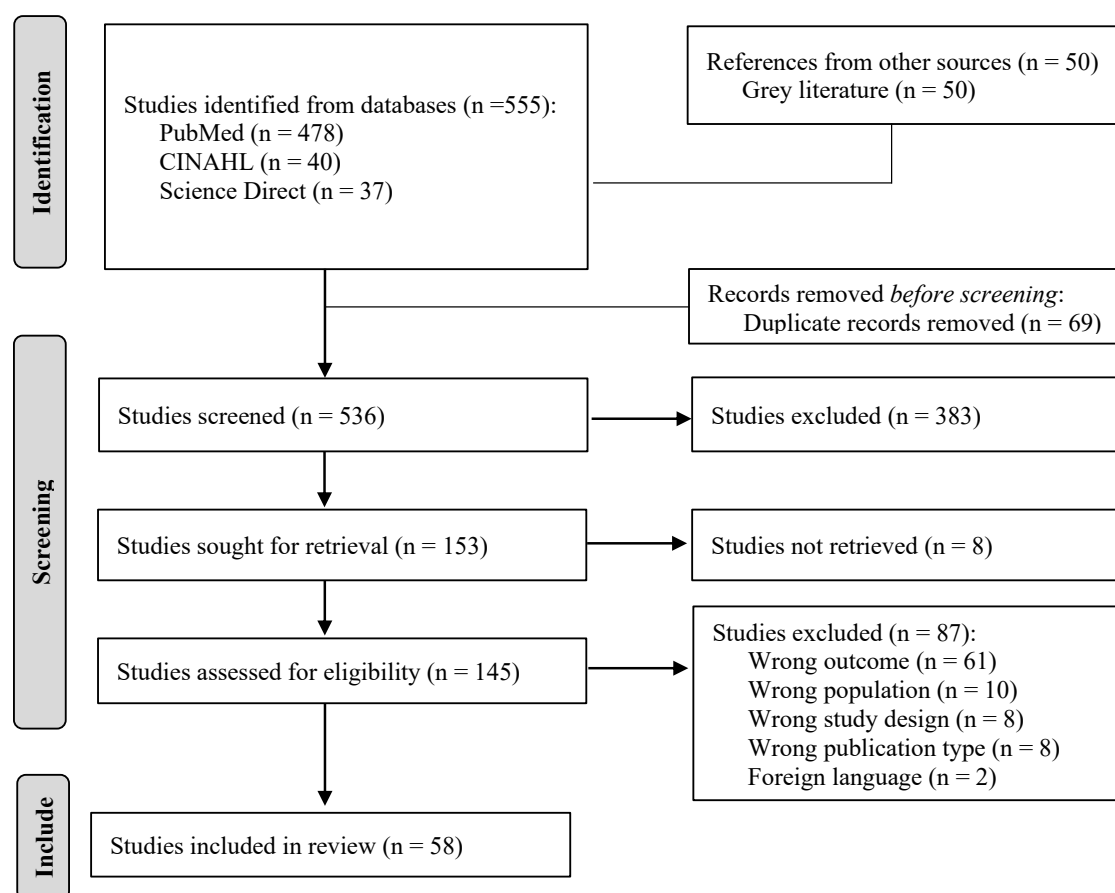


Figure 1. PRISMA-ScR flow diagram of study selection process.

Descriptive analysis of the 58 included studies (Table 1) revealed that most were conducted between 2018 and 2024, with a predominance in North America, accounting for approximately 33% (19/58) of the total. This region included studies from the United States and Canada.^{8, 9, 16, 19, 23-37}

Europe followed, with approximately 28% (16/58) of studies.^{5, 10-13, 17, 20, 38-44} Asia accounted for about 22% (13/58).^{18, 21, 45-52} Latin America contributed 5% (3/58), with studies conducted in Brazil, Chile, and Mexico.⁵³⁻⁵⁵ The Middle East represented approximately 5% (3/58)⁵⁶⁻⁵⁸ while Africa accounted for 3% (2/58).^{59, 60} Finally, Oceania was represented by 2% (1/58), with one study conducted in Australia.⁶¹

Regarding study design, cross-sectional studies were most common, representing 26% (15/58) of the total.^{14, 16, 18, 48, 53, 54, 57-59, 62} Next, quasi-experimental designs made up 17% (10/58).^{9, 25, 26, 32, 43, 45, 51, 52, 55, 63} Observational studies accounted for about 14% (8/58).^{5, 12, 17, 20, 27, 30, 31, 42} Randomized controlled trials represented 14% (8/58).^{34, 35, 39, 41, 46, 50, 56, 67} Experimental studies totaled 9% (5/58),^{19, 23, 28, 29, 33} and mixed-methods studies also accounted for 9% (5/58).^{10, 11, 17, 27, 63} Descriptive studies made up 7% (4/58),^{13, 20, 61, 64} while the only theoretical-conceptual study was by Houssaini et al., 2024 (2%; 1/58).⁶⁰

Regarding methodological approach, quantitative studies predominated, present in 76% (44/58) of cases.^{23, 36, 48, 49, 53, 59, 62, 65} Mixed-methods studies represented 17% (10/58),^{10, 11, 17, 27, 63} and qualitative studies accounted for 7% (4/58).^{21, 40, 60, 64}

Sample sizes ranged from 6 participants³² to 500,^{19, 44} with a mean of 132.4 participants and a median of 74.0. Half of the studies (50%) had small to moderate samples (fewer than 100 participants), while 8 studies (≈21%) included more than 200 participants. About 20% (12/58) of the studies did not report sample size.

A predominance of third-year medical students was observed.^{26, 41, 59} Some studies included first-year students,⁴⁷ fourth-year students,^{56, 64} and, less frequently, fifth-year students.⁶⁶

Beyond undergraduates, some studies involved standardized patients,⁴¹ AI models such as chatbots,⁶³ experienced healthcare professionals,³² surgical teams,²⁹ and medical residents.⁵¹

Chart 1. Synthesis of evidence on artificial intelligence in undergraduate medical education: characteristics and findings of studies.

Authors (Year/Country)	Study design	Methods	Main findings	AI tool(s)	Educational Relevance
Meyer; Riese; Streichert, 2024⁵ / Germany	Observational comparative study	GPT-3.5 and GPT-4 answered 937 text-only questions from three German licensing exams; scores were compared with those of the real examinee cohorts.	GPT-4's score fell within the top student percentiles on every exam; GPT-3.5 lagged behind but still mirrored student pass rates.	GPT-3.5, GPT-4	GPT-4 can be a high-level practice partner if students verify answers.
Zare et al., 2024^{14, 23} / Iran	Cross-sectional evaluation	Three sets of 200 MCQs (Persian + English) were put to GPT-3.5, GPT-4, Bard and compared with anonymised marks from 436 Iranian students.	GPT-4 clearly outperformed students in Persian and matched them in English; GPT-3.5 failed all Persian exams.	GPT-3.5, GPT-4, Bard	GPT-4 supports bilingual exam practice; language matters for performance.
Abbas; Rehman; Rehman, 2024⁸ / United States	Experimental computational study	163 publicly available NBME sample MCQs were entered one-by-one into four large language models.	GPT-4 answered 100 % correctly; the next best (GPT-3.5) reached 82 %.	GPT-4, GPT-3.5, Claude, Bard	GPT-4 is the current benchmark but still needs human checking for rare errors.
Abid et al., 2023²⁴ / United States	Quasi-experimental course evaluation	Nineteen final-year students completed a one-month, self-guided AI elective; pre/post confidence surveys and project deliverables were analysed.	Self-reported confidence in AI/ML concepts rose by 66 %; ¼ of students kept working on their AI projects afterwards.	(course content only)	Short, project-based electives can boost AI literacy at scale.
Burke et al., 2024²⁵ / United States	Retrospective observational study	ChatGPT-3.5 and standardised patients independently scored 168 first-year clinical notes against an 85-item rubric.	ChatGPT's scoring error rate was 1 % versus 7 % for human raters.	GPT-3.5	AI can provide fast, consistent feedback on free-text documentation.
Cianciolo; Lavoie; Parker, 2022²⁶ / United States	Retrospective observational study	700 diagnostic-justification essays were rescored by humans; an NLP model was trained on text metrics to predict those scores.	Machine scores correlated strongly with human ratings ($r \approx 0.7$) and with other academic indicators.	Custom NLP / ML	Automated marking of clinical reasoning essays is feasible and valid.
Collins; Black; Rarey, 2024¹⁹ / United States	Experimental computational study	A custom "AnatomyGPT" built from open-source texts was tested against ChatGPT on 60 NBME-style questions.	Accuracy was similar, but AnatomyGPT supplied more source citations in its explanations.	AnatomyGPT (GPT-4)	Domain-tailored GPTs can act as citation-rich tutors.
Fazlollahi et al., 2023³⁷ / Canada	Retrospective cohort study	Students performed five VR tumour resections; an AI tutor (VOA) selected	Targeted skills improved, but some non-targeted metrics	Virtual Operative Assistant	Expert oversight is needed when AI steers surgical

Authors (Year/Country)	Study design	Methods	Main findings	AI tool(s)	Educational Relevance
		four feedback targets and tracked 54 motion variables.	drifted, highlighting unintended learning effects.		practice.
Fazlollahi et al., 2022³⁶ / Canada	Randomised controlled trial	Seventy students were allocated to AI feedback (VOA), remote expert feedback, or no feedback during five VR neurosurgical tasks.	The VOA group achieved the highest objective performance without added cognitive load.	Virtual Operative Assistant	Metric-driven AI tutoring can outperform live experts in simulation labs.
Gin et al., 2022²⁷ / United States	Mixed-methods study	A bidirectional LSTM model was trained on 3 328 clerkship narrative feedback entries to predict entrustment levels; explainable AI highlighted language features.	The model linked specific feedback patterns to higher or lower entrustment ratings.	LSTM neural network	AI can mine large feedback sets to guide faculty on coaching language.
Krive et al., 2023⁸ / United States	Prospective cohort study	Twenty final-year students took a 4-week online elective integrating AI cases; quizzes, projects and reflections were analysed.	Average quiz score was 97 %; students reported moving from theory to practical AI application.	Course about AI	Well-structured electives can embed real-world AI thinking into curricula.
Ledwos et al., 2022²⁸ / Canada	Experimental study	Fifty surgeons-to-students repeated a VR subpial tumour resection five times; AI-selected metrics tracked learning curves.	Five of six metrics showed clear progress and distinguished expertise levels.	k-Nearest-Neighbours metric selection	Objective AI metrics support adaptive progression tracking.
Mirchi et al., 2020²⁹ / Canada	Experimental validation study	Fifty participants completed a VR task; an SVM-based VOA classified expert vs novice performance and generated feedback dashboards.	Classifier accuracy 92 %, sensitivity 100 %, specificity 82 %.	Support Vector Machine	Explainable AI feedback fosters trust and self-directed skill repair.
Morjaria et al., 2023³⁰ / Canada	Retrospective, non-controlled study	ChatGPT-3.5 answered 30 Bloom-stratified short-answer questions; six tutors graded its output against student minima and historical averages.	ChatGPT beat borderline students but scored below the historical class mean.	GPT-3.5	Un-proctored SAQs risk validity when LLMs are accessible.
Nagaraj et al., 2022³¹ / United States	Retrospective observational study	229 student knot-tying videos were coded pass/fail; 2D and 3D CNNs were trained to detect instrument-holding and knot errors.	Accuracy ~90 %, but lower sensitivity meant some poor performers were missed.	Convolutional neural networks	Automated video coaching works but must minimise false-negatives.
Park et al., 2024⁹ / United States	Quasi-experimental pilot	Ten first-year students took a parallel, non-credit AI elective integrated with four preclinical blocks; confidence surveys per block.	Confidence in AI topics rose significantly in all blocks; overall satisfaction 4.3/5.	Classic ML + CNN demos	Even tiny pilots show how AI themes can weave into basic-science teaching.
Ruberto et al., 2021³² / Canada	Proof-of-concept experimental study	Two physicians and two students ran a 10-min AR resus simulation; a multitask DNN used ECG + GSR to adjust patient severity in real time.	System successfully modulated scenario difficulty and was rated engaging by users.	Multitask deep neural network	Adaptive AR can tailor stress levels to optimise learning.
Scherr et al.,	Descriptive	Prompt-engineered	ChatGPT generated	GPT-3.5	Free, on-

Authors (Year/Country)	Study design	Methods	Main findings	AI tool(s)	Educational Relevance
2023 ³³ / United States	case study	ChatGPT-3.5 produced three interactive ACLS/ICU simulations; authors iteratively refined realism and accuracy.	dynamic, branching scenarios responsive to learner input.		demand clinical simulations are now within reach.
Sridharan; Sequeira, 2024 ¹⁶ / Kingdom of Bahrain	Cross-sectional tool audit	ChatGPT-3.5, Sage and Claude were asked to create pharmacology learning objectives and exam items on hypertension; outputs were rated for quality.	Learning objectives were consistently Bloom-aligned; item quality varied, requiring faculty editing.	GPT-3.5, Sage, Claude	LLMs can draft curriculum content but need expert review before use.
Yilmaz et al., 2023 ³⁴ / Canada	Randomised controlled trial	120 students in four groups practised VR bimanual resection with numerical, visual, visuospatial or no feedback.	Visual and visuospatial feedback groups achieved the most expert benchmarks by repetition five.	SVM-based metric engine	Visual AI feedback accelerates surgical skills compared with numbers alone.
Yilmaz et al., 2024 ³⁵ / Canada	Double-blind randomised controlled trial	97 students received real-time AI feedback (ICEMS), real-time expert feedback, or post-hoc only during five VR resections.	ICEMS group had the highest composite scores and lowest cognitive load.	Intelligent Continuous Expertise Monitoring System	Real-time AI feedback can beat traditional expert instruction.
Bajcetic et al., 2024 ³⁸ / Serbia	Methodological narrative review	Discussed ANN- and PBPK-driven simulations using real patient data to model drug ADME/PD.	AI-enhanced models produced high-fidelity PK/PD predictions, reducing animal tests.	ANN + PBPK hybrids	Virtual-patient drug sims enrich pharmacology teaching.
Birks et al., 2024 ¹⁰ / United Kingdom	Mixed-methods pre-post study	68 students learned prompt-engineering with Google Gemini to write SBA questions; confidence survey and focus group.	Confidence in using AI for question writing rose significantly; students valued the tool yet worried about accuracy.	Google Gemini	“Flipped assessment” workshops build AI literacy and question-design skills.
Borchert et al., 2023 ²⁰ / United Kingdom	Descriptive evaluation study	ChatGPT solved the entire 2023 UK SJT practice bank; answers were scored with the official rubric and mapped to GMC domains.	Overall 76 % score, evenly across professionalism domains; strongest in ranking tasks.	GPT-3.5	ChatGPT can support ethics and professionalism revision but is rarely perfect.
Coskun et al., 2024 ³⁹ / Turkey	Randomised, double-blind controlled trial	74 students compared ChatGPT-generated vignettes with human vignettes and sat 15 ChatGPT MCQs; psychometrics analysed.	Students rated vignettes equivalently; 11/15 AI MCQs met acceptable item metrics.	GPT-3.5	ChatGPT can draft usable clinical cases, saving educator time.
Driesnack et al., 2024 ¹² / Germany	Prospective observational study	46 volunteers joined a gamified, GPT-4-supported antibiotic-stewardship short course; pre-, post- and 12-mo tests.	Correct-answer rates jumped 74–83 %; gains held up to one year; 92 % found AI feedback helpful.	GPT-4	Gamification plus GPT brings durable knowledge and confidence.
Friederichs et al., 2023 ¹⁷ / Germany	Observational evaluation	ChatGPT answered 400 Progress-Test MCQs; results compared with German year-1–6 norms.	65.5 % correct—better than most years 1–3 and similar to late-stage students.	GPT-3.5	LLMs are suitable for factual self-quizzing but need critical framing.
Holderried et al., 2024 ⁴⁰ /	Prospective single-site	GPT-4 chatbot role-played a patient in 106 8-	99 % of answers medically plausible;	GPT-4	GPT-4 can deliver realistic,

Authors (Year/Country)	Study design	Methods	Main findings	AI tool(s)	Educational Relevance
Germany	study	min written interviews and generated structured feedback; compared with a human rater.	feedback agreement $\kappa = 0.83$ with the human evaluator.		structured bedside-history drills at scale.
Lippitsch, et al., 2024⁴¹ / Germany	Randomised, controlled crossover trial	168 students alternated between ViPATalk (rule-based avatar chatbot) and classic peer role-play across three cases.	ViPATalk was non-inferior to role-play on completeness and communication checklists.	AIML rule-based chatbot	Virtual-patient chatbots can safely substitute peer role-play sessions.
Moglia et al., 2022⁴² / Italy	Observational predictive-model study	176 novices did five VR robot-surgery tasks; ensemble DNNs predicted time/attempts to reach proficiency and were compared to RF/GBRT.	DNNs predicted struggling trainees earlier and more accurately than other models.	Ensemble deep neural networks	Early AI prediction lets educators tailor coaching resources.
Riedel et al., 2023⁴³ / Germany	Quasi-experimental computational study	ChatGPT answered two sets of OB/GYN MCQs; experts rated accuracy and depth ("density of insights").	Accuracy comparable to students; AI explanations contained many non-obvious insights.	GPT-3.5	ChatGPT adds explanatory richness that aids self-study.
Soulage et al., 2024¹³ / France	Quantitative comparison study	GPT-3.5 sat a master-level physiology exam alongside 24 students; blind grading.	Median AI score 75 % vs 59 % human; AI passed 100 % of the time.	GPT-3.5	Assessment formats must evolve to account for LLM capabilities.
Thomae et al., 2024¹¹ / Switzerland	Exploratory mixed-methods study	ChatGPT embedded in three placebo/nocebo teaching units; pre/post perception surveys.	Students found AI-generated texts clear and trustworthy; course deemed helpful.	GPT-3.5	Simple prompt-based integrations build AI-critical-thinking skills.
Tomova et al., 2024⁴⁴ / Germany	Computational experimental study	GPT-4 and Bing Chat produced topic-level feedback for 18 Progress-Test MCQs; six experts scored relevance (1-5).	Both models averaged > 4/5 relevance; no hallucinations detected.	GPT-4, Bing Chat	LLMs can mass-produce high-quality formative feedback.
Barui et al., 2022⁴⁵ / India	Observational comparison study	A VGG16/Inception CNN trained on 1 000 histology images, tested on 300 images; 25 students took same test.	CNN accuracy 89 % vs 78 % students; AI was quicker and more consistent.	CNN (VGG16, Inception)	Image-recognition AI strengthens histology drilling sessions.
Cheng et al., 2020⁴⁶ / Taiwan	Randomised controlled trial	36 fifth-years split to AI-assisted or conventional hip-fracture radiograph teaching; pre/post accuracy tests.	AI group improved from 76 % to 89 %; control showed no significant change.	DenseNet-121 "HipGuide"	Heat-map AI tutors boost image-interpretation accuracy, notably for lower achievers.
Ghosh et al., 2023⁴⁷ / India	Quasi-experimental classroom study	ChatGPT-3.5 answered 79 biochem exam items in new chat windows; responses graded with the cohort of 100 first-year students.	AI ranked 5th in class (140/200) but was slower on reasoning-type MCQs.	GPT-3.5	LLMs can act as an additional "student" for instant answer checking.
Huang et al., 2024⁴⁸ / Taiwan	Cross-sectional performance study	GPT-4 tackled 600 bilingual Stage-1 licensing MCQs (10 basic-science subjects).	Overall 87.5 % accuracy, clearing the 60 % pass mark in every subject.	GPT-4	Viable bilingual self-study tool for basic-science revision.
Huh, 2023¹⁸ / Korea	Cross-sectional comparison	GPT-3.5 and 77 first-years answered 79 parasitology MCQs; one expert graded	AI scored 61 % vs students 91 %; correct AI answers had acceptable	GPT-3.5	GPT-3.5 underperforms in niche domains but can

Authors (Year/Country)	Study design	Methods	Main findings	AI tool(s)	Educational Relevance
		explanations.	explanations.		still model explanation quality.
Kinoshita et al., 2024⁴⁹ / Japan	Retrospective observational study	U-Net trained on 740 labelled surgical images highlighted nerves; 1-min annotated clip shown to students vs lecture.	Students with AI clip doubled nerve-recognition scores compared with lecture alone.	U-Net	Short annotated videos markedly improve anatomical recognition.
Li et al., 2023⁵⁰ / China	Randomised controlled trial + interviews	30 fifth-years learned blood-cell morphology with or without a heat-map AI platform; accuracy tests and thematic interviews.	AI group accuracy rose from 76 % to 89 %; students valued visual cues.	DeepCyto CNN	AI-guided heat-maps shorten image-based learning curves.
Su et al., 2024⁵¹ / Taiwan	Quasi-experimental validation	54 ENT trainees used the MEKAS adaptive testing system; control group used standard learning.	Experimental group showed significant knowledge gain and rated usability high.	Knowledge-based CBR engine	Competency-adaptive AI tests personalise study paths.
Zheng et al., 2024⁵² / China	Quasi-experimental classroom study	66 students received either conventional or AI-enhanced scenario-based cardiology teaching (ChatGPT + Midjourney images); multiple assessments.	AI group scored higher in theory, Mini-CEX, and critical-thinking scales; satisfaction also higher.	GPT-3.5, Midjourney	AI-powered scenarios deepen engagement and skills in clinical topics.
Al-Worafi et al., 2023²¹ / Malaysia	Qualitative content-analysis study	Professors asked ChatGPT about designing an infectious-disease pharmacotherapy module; responses rated for appropriateness and accuracy.	Ratings ranged 50–92 %; helpful for structuring but sometimes shallow.	ChatGPT	Use ChatGPT as an idea generator, not a turnkey curriculum writer.
Alessi et al., 2024⁵³ / Brazil	Cross-sectional performance study	GPT-3.5 answered 333 Progress-Test MCQs from 2021–23; compared with national student averages.	AI averaged 68 % vs students 51 %; strongest area Public Health (78 %).	GPT-3.5	ChatGPT surpasses all years of Brazilian students on factual tests.
Alonso-Silveiro et al., 2018⁵⁴ / Mexico	Cross-sectional developmental study	A Raspberry-Pi laparoscopic box trainer tracked tool motion; an ANN classified user expertise over five days.	Classification accuracy 91 %; clear learning curves for novices.	Artificial neural network	Low-cost AI trainers objectively track psychomotor skill acquisition.
Castellano et al., 2024⁵⁵ / Chile	Quasi-experimental course study	131 first-years used a gamified mobile anatomy app with a neural-network recommender and NLP virtual assistant; pre/post tests.	Gamified AI tool improved scores and engagement; students valued tournaments and personalised tips.	Seq. NN recommender, Dialogflow NLP	AI-powered gamification strengthens self-regulated anatomy learning.
Aronovitz et al., 2024⁵⁶ / Israel	Randomised controlled parallel-group trial	60 fourth-years did a 6-min cardiac POCUS image-acquisition test with or without a real-time AI quality indicator.	AI group scored higher ($p = 0.002$) and was $2.3 \times$ likelier to capture correct apical views.	GE EF-quality indicator	Real-time AI feedback significantly upgrades ultrasound image quality.
Meo et al., 2024⁵⁷ / Saudi Arabia	Cross-sectional diagnostic study	Google Bard answered 60 OSPE image questions covering six organs; answers scored correct/incorrect.	Overall 57 % accuracy, ranging from 28 % (kidney) to 90 % (bone).	Google Bard	Bard shows moderate visual-ID skill; needs refinement before high-stakes use.
Mavrych;	Cross-sectional	Six LLMs answered 50	GPT-4 highest MCQ	GPT-4, GPT-3.5,	Advanced LLMs

Authors (Year/Country)	Study design	Methods	Main findings	AI tool(s)	Educational Relevance
Ganguly; Bolgova, 2024⁵⁸ / Saudi Arabia	comparison	upper-limb anatomy MCQs and generated related scenarios; outputs rated by two professors.	score (61 %); Bard/Gemini lowest (33 %).	Copilot, PaLM-2, Bard	can aid anatomy item writing but cannot yet replace faculty.
Cherif et al., 2024⁵⁹ / Tunisia	Cross-sectional performance study	ChatGPT-3.5 answered 45 pulmonology MCQs under two prompt formats (plain vs contextual); student scores N = 244.	Context-rich prompt passed the exam (51 %) and outscored 62 % of students; plain prompt failed.	GPT-3.5	Prompt engineering strongly influences ChatGPT's exam success.
Houssaini et al., 2024⁶⁰ / Morocco	Conceptual design study	Developed the DRIIPT framework (Design Thinking + AI) for a first-year innovation course; qualitative evaluation.	Reported boosts in creativity, critical thinking and GenAI fluency.	ChatGPT	Combining design thinking with GenAI tools nurtures innovation skills.
Xie et al., 2024⁶¹ / Australia	Descriptive expert-rating study	ChatGPT-4, Bard and BingAI answered tiered surgical scenarios; readability, DISCERN and expert Likert ratings recorded.	ChatGPT scored highest for readability, information quality and expert accuracy.	GPT-4, Bard, Bing AI	ChatGPT serves as a reliable adjunct for junior-doctor self-learning queries.
Gan et al., 2024⁶² / China	Parallel-design double-blind RCT	129 third-years studied orthopaedics with or without ChatGPT help for two weeks; post-test and semester exam scores tracked.	ChatGPT group scored higher on the immediate test and on later surgery and OB/GYN exams.	GPT-4	ChatGPT improves both short- and longer-term exam performance.
Baglivo et al., 2023⁶³ / Italy	Feasibility classroom study	36 fifth-years compared five chatbots' vaccination answers to their own during a 2-h session; discussion and satisfaction captured.	Chatbots out-scored students (12.2 vs 8.2/15) and sparked lively critical debates.	ChatGPT, Bing Chat, Chatsonic, Bard, YouChat	Guided chatbot use heightens engagement and critical appraisal skills.
Tsopra et al., 2022⁶⁴ / France	Descriptive curriculum-development study	Fifteen fourth-years designed AI clinical-decision prototypes in an elective; satisfaction survey and follow-up activities recorded.	Satisfaction 4.4/5; several students pursued further AI projects.	Simulated AI-CDSS design	Putting students in the designer role fosters deep digital-health literacy.
Bharatha et al., 2024⁶⁵ / Barbados	Cross-sectional performance study	GPT-4 and early- and mid-course students answered 304 MCQs tagged by Bloom level.	GPT-4 overall 74 % vs students 67 %; AI excelled in factual cells, less in higher-order reasoning.	GPT-4	LLMs are potent for knowledge recall but weaker in evaluation-level tasks.
Furlan et al., 2021⁶⁶ / Italy	Pre- and post-experimental study	Fifteen undergrads used the Hepius virtual-patient simulator with ITS; 22-item MCQ test before and after.	Total and core-question scores improved significantly post-simulation (P<.001).	SLSTM NLP + ITS	Mobile VPS can quickly boost diagnostic-reasoning knowledge.

ACLS – Advanced Cardiac Life Support; ADME – Absorption, Distribution, Metabolism and Excretion; AI – Artificial Intelligence; AIML – Artificial Intelligence Markup Language; ANN – Artificial Neural Network; AR – Augmented Reality; Bard – Google Bard/Gemini chatbot; CBR – Case-Based Reasoning; CDSS – Clinical Decision Support System; CNN – Convolutional Neural Network; DNN – Deep Neural Network; DRIIPT – Design-Thinking, Constructive Alignment and Generative-AI Framework; EF – Ejection Fraction; ENT – Ear, Nose and Throat; GBRT – Gradient Boosted Regression Trees; GMC – General Medical Council; GSR – Galvanic Skin Response; GPT-3.5 – Generative Pre-trained Transformer version 3.5; GPT-4 – Generative Pre-trained Transformer version 4; ICEMS – Intelligent Continuous Expertise Monitoring System; ICU – Intensive Care Unit; ITS – Intelligent Tutoring System; k-NN – k-Nearest Neighbours; LLM – Large Language Model; LSTM – Long Short-Term Memory; MCQ – Multiple-Choice Question; Mini-CEX – Mini Clinical Evaluation Exercise; ML – Machine Learning; NBME – National Board of Medical Examiners; NLP – Natural Language Processing; OB/GYN – Obstetrics and Gynecology; OSPE – Objective Structured Practical Examination; PBPK – Physiologically Based Pharmacokinetic; PD – Pharmacodynamics; POCUS – Point-of-Care Ultrasound; RF – Random Forest; RCT – Randomised Controlled Trial; SJT – Situational Judgement

Test; SAQ – Short-Answer Question; Seq. NN – Sequential Neural Network; SLSTM – Siamese Long Short-Term Memory; SVM – Support Vector Machine; U-Net – U-shaped Convolutional Neural Network; VPS – Virtual Patient Simulator; VOA – Virtual Operative Assistant; VR – Virtual Reality.

3.2 THEMATIC CATEGORIZATION OF THE STUDIES

3.2.1 LARGE LANGUAGE MODELS (LLMs)

Among the 58 included studies, 20 (34.5%) evaluated large language models (LLMs) in formal examination settings, comparing the scores and pass rates of ChatGPT-3.5, ChatGPT-4, Google Bard, and other LLMs with medical student performance.^{5, 14, 48, 59}

Across standardized medical examinations, GPT-4 demonstrated superior performance, with accuracy ranging from 70.6% to 100%, consistently outperforming ChatGPT-3.5 and other models such as Bard and Claude.^{23, 62} Student performance varied widely, with average scores between 50.8% and 90.8%, depending on exam type and language context.^{18, 53} In contrast, ChatGPT-3.5 achieved accuracy rates from 51.1%, below the 60% passing threshold in a pulmonology exam without contextual guidance, to 82.2% in NBME-style assessments, and performed below student averages in disciplines such as parasitology (60.8% vs. 90.8%)^{18, 23, 59}

3.2.2 AUTOMATED FEEDBACK SYSTEMS

Sixteen studies (27.6%) addressed automated feedback systems for clinical and surgical skills, assessing composite scores (e.g., ICEMS and OSATS), segmentation metrics (Dice, IoU), and image quality evaluations in POCUS.^{34, 40, 49, 56}

Clinical and surgical feedback systems, such as the Virtual Operative Assistant (VOA) and the Intelligent Continuous Expertise Monitoring System (ICEMS), led to error reduction and performance improvements between 10% and 20%, with statistical significance ($P < .01$).^{34, 67} Other studies also highlighted the use of AI for automated feedback, including GPT-4-based simulations,⁴⁰ correction of clinical notes,²⁵ formative feedback in surgical training,²⁹ and AI-based analysis of narrative feedback.²⁷

3.2.3 SIMULATION PLATFORMS AND ASSESSMENT TOOLS FOR EXAMINATIONS

Nine studies (15.5%) described simulation and virtual tutoring platforms, employing pre-post testing, cognitive load adaptation measures, and machine learning classifiers to assess pedagogical effectiveness.^{29, 32, 51, 54, 66}

AI-driven platforms such as Hepius (a virtual patient simulator with ITS, showing an average 15 percentage point gain in clinical reasoning scores);⁶⁶ HipGuide (heatmaps on hip radiographs, improving accuracy from 75.7% to 88.9%);⁴⁶ ICEMS (real-time surgical feedback in tumor resection simulation, with a 12-point improvement);³⁴ Virtual Operative Assistant (VOA) (an SVM-based tutor for neurosurgical simulation, achieving up to 15-point gains without increasing cognitive load);³⁶ EF Quality Indicator (a POCUS feedback tool making correct apical image acquisition 2.3× more likely);⁵⁶ DeepCyto morphology AI system (a CNN for blood cell classification, raising accuracy by 12 points);⁵⁰ and U-Net (a surgical video segmentation model doubling anatomical recognition scores),⁴⁹ all showed gains between 12 and 20 percentage points ($P < .01$).

These assessment-oriented applications demonstrated improved consistency, scalability, and efficiency in evaluating student performance, while also raising important considerations regarding validity, reliability, and the risk of overreliance on automated evaluation systems.³⁴

3.2.4 CURRICULAR INTEGRATION MODULES

Eleven studies (19.0%) examined the integration of AI tools into curricula and hybrid courses, using validated questionnaires to assess student satisfaction, perceived learning progress, and self-confidence in AI-related concepts.^{8,9,11,24} Two additional studies (3.4%) employed qualitative approaches, including content analysis, to explore perceptions of the benefits, risks, and factors influencing AI adoption in medical education.^{21,64}

The integration of AI modules into curricula—through elective courses,^{8,24} practical workshops focused on question generation using large language models,¹⁰ and journal club sessions combining article discussion with coding demonstrations⁹—was associated with a 50% to 66% increase in students' confidence in using AI tools, as well as high academic performance, with mean scores $\geq 90\%$ in quizzes and project evaluations.^{8,24}

Qualitative findings from instructors and faculty indicated that 80% to 90% perceived AI solutions as useful for curriculum planning and educational material development, while emphasizing the need for critical supervision to prevent inaccurate or superficial outputs. In one content analysis study, 85% of educators recommended ChatGPT as a support tool, provided that learning objectives and assessment items were carefully reviewed.²¹ Similarly, 80% of students and instructors participating in an AI-CDSS design course recommended the activity, underscoring the importance of continuous human oversight.⁶⁴

3.3 TPACK FRAMEWORK

The analysis of these studies reveals a practical convergence with the domains of the TPACK model (Fig. 2), although none of the studies explicitly referenced its application. The AI-driven interventions reflect technological knowledge (e.g., use of LLMs, simulators, feedback tools), pedagogical knowledge (e.g., real-time feedback, personalized learning, clinical simulations), and content knowledge (e.g., teaching anatomy, diagnostics, surgical skills). Studies that integrated all three domains—such as workshops using LLMs to address medical content, or intelligent tutors providing clinical feedback—illustrate the potential of TPACK as a guiding framework for future AI-enhanced curriculum design in medical education.

3.4 APPLICATIONS IN MEDICAL EDUCATION

Evidence suggests that AI tools can be employed to (1) support self-directed learning and exam preparation by reinforcing knowledge acquisition;^{5,53,59} (2) provide immediate and standardized feedback in clinical and surgical skills training;^{34,37,56} (3) enhance interactive simulations and the teaching of imaging techniques;^{33, 49, 57} (4) automate the grading of objective assessments and personalize learning trajectories;^{14, 19, 44} and (5) assist in the development of critical AI competencies and curricular planning.^{8, 11, 21, 39} However, continuous pedagogical supervision is required to ensure educational validity and ethical standards.^{13, 30}

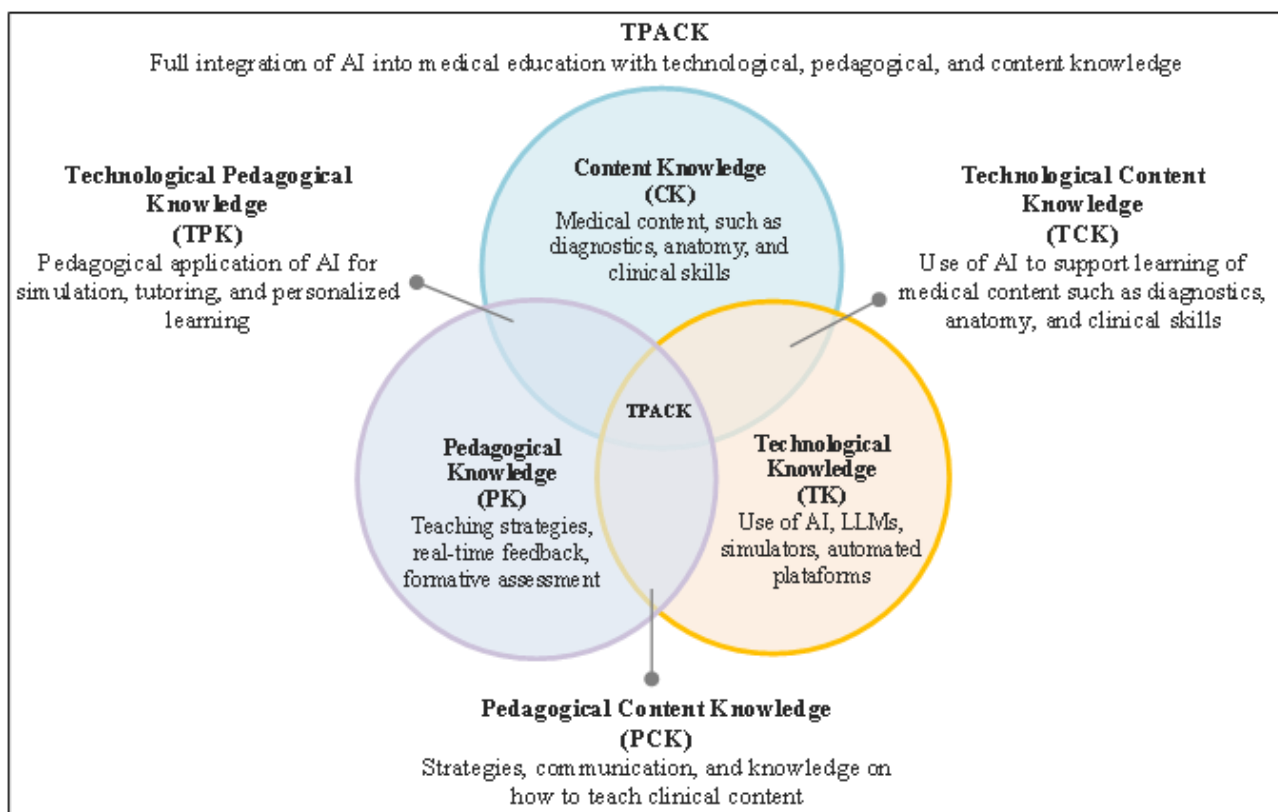


Figure 2. TPACK Conceptual Framework for AI in Medical Education.

4. DISCUSSION

Of the 58 studies analyzed (2018–2024), 83% originated from North America, Europe, and Asia, while only 15% came from Latin America, the Middle East, Africa, and Oceania, highlighting significant global research disparities likely related to differences in technological access, funding, and institutional priorities.

Most studies adopted cross-sectional or quasi-experimental designs and focused on AI tools such as large language models, automated feedback systems, and AI-oriented curricular modules. However, many investigations were limited by small sample sizes, single-center designs, and short study durations, restricting external validity and long-term outcome assessment.^{8, 66} There remains a notable lack of longitudinal studies conducted in real clinical settings, theoretically grounded curricula, and evaluations of knowledge retention, interprofessional competencies, and ethical dimensions of AI integration.

Students consistently reported high awareness of AI's relevance to medical practice (94.8–99.1%),^{68–72} yet formal AI instruction remains limited, resulting in insufficient familiarity with technical terminology (50–60%) and inadequate preparedness for clinical application.^{69,70,72} Despite this gap, student interest in AI education is strong, with support for mandatory curricula,⁷⁰ preferences for content addressing AI mechanisms, ethics, and risks,⁷¹ and advocacy for interdisciplinary, hands-on workshops.⁷⁰ Concerns persist regarding legal liability and AI limitations in unpredictable clinical scenarios, particularly when adverse outcomes may occur.⁶⁸

A critical limitation involves the reduced performance of LLMs in generating accurate, context-sensitive content in languages other than English, raising concerns about cross-cultural adaptation and technical terminology when training data are predominantly English-based.⁵⁹ Without targeted mitigation strategies, AI's educational benefits may remain concentrated in English-speaking regions, further marginalizing low- and middle-income and linguistically diverse settings. Conversely, AI has

demonstrated potential in automating patient education materials, enabling rapid production of personalized resources that may improve treatment adherence and health outcomes.¹¹

Beyond educational efficiency, AI integration in undergraduate medical education has important implications for health promotion and disease prevention. AI-supported tools can enhance competencies related to preventive counseling, risk stratification, patient education, and evidence-based communication, central to population health strategies.^{3,6,7} Through personalized learning pathways, early identification of knowledge gaps, and simulation-based training focused on preventive decision-making, AI may indirectly support improved health behaviors, adherence, and equity in care delivery.^{3,11,73} These benefits, however, depend on ethically grounded curricula and strong faculty oversight to prevent bias amplification, misinformation, and technological dependency, particularly among vulnerable populations.^{5,59}

In theoretical assessments, LLMs have shown satisfactory performance in multiple-choice examinations and progress tests, supporting their use as self-study and revision tools.^{16,41,48} Nonetheless, limitations persist, including inconsistent responses,²⁰ algorithmic bias,⁵ underperformance in complex clinical reasoning,⁴⁷ inability to express uncertainty,¹⁷ and insufficient justification of answers.¹⁸ Although recent algorithmic improvements¹⁴ and technological advances⁶⁵ have addressed some of these issues, uncritical AI use without parallel development of reflective and critical competencies risks undermining clinical reasoning.¹⁷

Beyond theoretical testing, AI has been effectively applied in clinical simulations and self-directed skills training by providing real-time, contextualized feedback, accelerating learning curves, particularly in early training stages.^{12,40,46,56} These tools also support early identification of struggling learners, improve academic performance, and reduce cognitive load.^{34,42}

Educators view AI as a valuable pedagogical resource that can enhance student engagement,¹¹ automate content generation,²¹ reduce assessment workload, and deliver frequent, targeted feedback.³¹ AI also enables realistic clinical vignettes and optimized instructional strategies.^{16,39,44} Successful implementation, however, requires faculty training, time for curricular revision, and institutional consensus on AI's role, ensuring it complements rather than replaces human interaction and the emotional dimensions of teaching.^{8,12}

There is growing consensus that AI technologies should support, not replace, medical practice and education by reinforcing clinical judgment and experiential learning. Achieving this balance requires collaboration among AI specialists, medical educators, and clinicians to address technical, ethical, and cultural challenges.^{13, 20, 63, 65}

Medical professionals anticipate that specialties such as radiology, dermatology, and pathology will lead AI adoption, followed by surgery and primary care within the next decade.⁶⁸ In clinical contexts, AI has demonstrated prognostic and therapeutic prediction accuracy comparable to or exceeding that of human experts, reducing subjectivity and interobserver variability.⁷⁴ Predictive models have been applied to in vitro fertilization, neonatal outcomes, preterm birth, fetal weight, and aneuploidy risk,⁷⁵ while AI-based decision support systems are established in imaging diagnostics, oncology, therapy response prediction, and clinical data management.⁷⁶ In population health, AI supports forecasting preventable hospitalizations, screening gaps, and automates triage, chart review, administrative workflows, and patient counseling.⁷³

Overall, these findings underscore the urgent need for structured, multicenter, interdisciplinary curricula addressing not only technical AI skills but also ethical, legal, and digital literacy competencies, alongside continuous physician training. Without such adaptations, academic institutions risk obsolescence amid rapid transformations in medical education and practice.^{13,76}

This review is limited by the absence of formal methodological quality appraisal, the inability to quantify effect sizes, and the rapidly evolving nature of the field. Additionally, the predominance of small, single-center studies limits external validity and generalizability across diverse educational and health system contexts.

4.1 PRACTICAL IMPLICATIONS

Undergraduate medical programs should establish clear institutional policies governing the educational use of large language models, particularly in assessment contexts, to protect academic integrity and ensure appropriate verification of AI-generated outputs.^{5, 14, 30} When adequately regulated, these tools may also support competencies relevant to health promotion, such as preventive counseling and evidence-based health communication, while reducing risks associated with misinformation and inequities.^{3, 77}

AI-based feedback systems and simulation tools can be integrated as adjuncts to clinical and surgical skills training, providing standardized, timely feedback and objective performance monitoring. Beyond technical skill acquisition, these applications may contribute to health promotion and disease prevention by enhancing early risk recognition and preventive decision-making, provided that faculty supervision ensures alignment with human-centered principles and mitigates unintended learning effects.^{6, 37, 56, 67}

Finally, effective AI integration requires longitudinal, framework-based curricular models rather than isolated electives. Such approaches should incorporate AI literacy, ethical training, and faculty development, with adaptation to local and linguistic contexts to promote equitable implementation and avoid reinforcing health disparities observed in current AI-driven health promotion initiatives.^{3, 7}

5. CONCLUSION

This scoping review mapped 58 studies (2018–2024) on artificial intelligence in undergraduate medical education and grouped the evidence into four thematic categories: large language models used primarily for exam preparation, question-solving, and content generation; automated feedback systems supporting clinical and surgical skills development; simulation platforms and assessment-oriented tools, including virtual tutoring and AI-assisted evaluation; and curricular integration modules, such as electives and workshops focused on AI literacy and application.

Across studies, AI-supported interventions were most consistently associated with improved performance in knowledge-based assessments and, in simulation-based contexts, gains in objective performance metrics supported by real-time feedback. Curricular integration modules frequently reported increases in student confidence and perceived preparedness regarding AI concepts and use. Several studies also emphasized that AI-generated outputs require expert review to ensure educational validity, particularly when used for feedback or assessment.

5.1 RECOMMENDATIONS FOR FUTURE STUDIES

Based on the gaps identified in this scoping review, future research should use multicenter, multilingual designs across various medical specialties and academic levels, and compare different LLM

versions and prompt strategies to evaluate their educational impact. Incorporating multimodal data and analyzing students' autonomous use of prompts can optimize human–AI interaction.^{5, 20, 48, 59}

To address the limited evidence on competency development, longitudinal studies in real clinical settings using validated metrics are needed to compare AI-driven and traditional feedback methods^{34, 56} and assess clinical reasoning, interprofessional skills, and long-term knowledge retention.^{34, 56}

Finally, the absence of theory-driven curricular models highlights the need for AI integration grounded in established learning theories, such as ACT-R and andragogy, to guide both instructional design and risk management.^{19, 40} Ethical integration also requires evaluating algorithmic biases, behavioral effects, and risks like cognitive dependency and diminished critical thinking.^{37, 52}

5.2 FUNDING

This study did not receive any funding from public, private, or institutional sources.

REFERENCES

1. Dave M, Patel N. Artificial intelligence in healthcare and education. *British Dental Journal*. 2023;234(10): 761–764. <https://doi.org/10.1038/s41415-023-5845-2>.
2. Jaworski A, Jasiński D, Jaworski W, Hop A, Janek A, Sławińska B, et al. Comparison of the Performance of Artificial Intelligence Versus Medical Professionals in the Polish Final Medical Examination. *Cureus*. 2024;16(8): e66011. <https://doi.org/10.7759/cureus.66011>.
3. Yousefi F, Naye F, Ouellet S, Yameogo AR, Sasseville M, Bergeron F, et al. Artificial Intelligence in Health Promotion and Disease Reduction: Rapid Review. *Journal of Medical Internet Research*. 2025;27: e70381–e70381. <https://doi.org/10.2196/70381>.
4. Alqahtani T, Badreldin HA, Alrashed M, Alshaya AI, Alghamdi SS, bin Saleh K, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy*. 2023;19(8): 1236–1242. <https://doi.org/10.1016/j.sapharm.2023.05.016>.
5. Meyer A, Riese J, Streichert T. Comparison of the Performance of GPT-3.5 and GPT-4 With That of Medical Students on the Written German Medical Licensing Examination: Observational Study. *JMIR Medical Education*. 2024;10(1). <https://doi.org/10.2196/50965>.
6. Smith A, Arena R, Bacon SL, Faghy MA, Grazi G, Raisi A, et al. Recommendations on the use of artificial intelligence in health promotion. *Progress in Cardiovascular Diseases*. 2024;87: 37–43. <https://doi.org/10.1016/j.pcad.2024.10.003>.
7. Qian Y, Siau KL. Advances in IoT, AI, and Sensor-Based Technologies for Disease Treatment, Health Promotion, Successful Ageing, and Ageing Well. *Sensors*. 2025;25(19): 6207. <https://doi.org/10.3390/s25196207>.
8. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: artificial intelligence in medical education. *JAMIA Open*. 2023;6(2). <https://doi.org/10.1093/jamiaopen/ooad037>.
9. Park SH, Pinto-Powell R, Thesen T, Lindqwister A, Levy J, Chacko R, et al. Preparing healthcare leaders of the digital age with an integrative artificial intelligence curriculum: a pilot study. *Medical Education Online*. 2024;29(1). <https://doi.org/10.1080/10872981.2024.2315684>.
10. Birks S, Gray J, Darling-Pomranz C. Using artificial intelligence to provide a 'flipped assessment' approach to medical education learning opportunities. *Medical Teacher*. 2024;47(8): 1377. <https://doi.org/10.1080/0142159X.2024.2434101>.
11. Thomae A V., Witt CM, Barth J. Integration of ChatGPT Into a Course for Medical Students: Explorative Study on Teaching Scenarios, Students' Perception, and Applications. *JMIR Medical Education*. 2024;10. <https://doi.org/10.2196/50545>.

12. Driesnack S, Rücker F, Dietze-Jergus N, Bondarenko A, Pletz MW, Viehweger A. A practice-based approach to teaching antimicrobial therapy using artificial intelligence and gamified learning. *JAC-Antimicrobial Resistance*. 2024;6(4). <https://doi.org/10.1093/jacamr/dlae099>.
13. Soulage CO, Van Coppenolle F, GUEBRE-EGZIABHER F. The conversational AI 'ChatGPT' outperforms medical students on a physiology university examination. *Advances in Physiology Education*. 2024;48(4): 677–684. <https://doi.org/10.1152/advan.00181.2023>.
14. Zare S, Vafaeian S, Amini M, Farhadi K, Vali M, Golestani A. Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams. *Scientific Reports*. 2024;14(1): 28456. <https://doi.org/10.1038/s41598-024-79335-w>.
15. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Academic Medicine*. 2024;99(1): 22–27. <https://doi.org/10.1097/ACM.0000000000005439>.
16. Sridharan K, Sequeira RP. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. *BMC Medical Education*. 2024;24(1):431. <https://doi.org/10.1186/s12909-024-05365-7>.
17. Friederichs H, Friederichs WJ, März M. ChatGPT in medical school: how successful is AI in progress testing? *Medical Education Online*. 2023;28(1). <https://doi.org/10.1080/10872981.2023.2220920>.
18. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *Journal of Educational Evaluation for Health Professions*. 2023;20: 1. <https://doi.org/10.3352/jeehp.2023.20.1>.
19. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: A customized artificial intelligence application for anatomical sciences education. *Clinical Anatomy*. 2024;37(6): 661–669. <https://doi.org/10.1002/ca.24178>.
20. Borchert RJ, Hickman CR, Pepys J, Sadler TJ. Performance of ChatGPT on the Situational Judgement Test—A Professional Dilemmas–Based Examination for Doctors in the United Kingdom. *JMIR Medical Education*. 2023;9. <https://doi.org/10.2196/48978>.
21. Al-Worafi YM, Goh KW, Hermansyah A, Tan CS, Ming LC. The Use of ChatGPT for Education Modules on Integrated Pharmacotherapy of Infectious Disease: Educators' Perspectives. *JMIR Medical Education*. 2024;10(1). <https://doi.org/10.2196/47339>.
22. Mishra P, Koehler MJ. Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. *Teachers College Record: The Voice of Scholarship in Education*. 2006;108(6): 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>.
23. Abbas A, Rehman MS, Rehman SS. Comparing the Performance of Popular Large Language Models on the National Board of Medical Examiners Sample Questions. *Cureus*. 2024;16(3). <https://doi.org/10.7759/cureus.55991>.
24. Abid A, Murugan A, Banerjee I, Purkayastha S, Trivedi H, Gichoya J. AI Education for Fourth-Year Medical Students: Two-Year Experience of a Web-Based, Self-Guided Curriculum and Mixed Methods Study. *JMIR Medical Education*. 2024;10(1). <https://doi.org/10.2196/46500>.
25. Burke HB, Hoang A, Lopreiato JO, King H, Hemmer P, Montgomery M, et al. Assessing the Ability of a Large Language Model to Score Free-Text Medical Student Clinical Notes: Quantitative Study. *JMIR Medical Education*. 2024;10. <https://doi.org/10.2196/56342>.
26. Cianciolo AT, Lavoie N, Parker J. Machine Scoring of Medical Students' Written Clinical Reasoning: Initial Validity Evidence. *Academic Medicine*. 2021;96(7): 1026–1035. <https://doi.org/10.1097/ACM.0000000000004010>.
27. Gin BC, ten Cate O, O'Sullivan PS, Hauer KE, Boscardin C. Exploring how feedback reflects entrustment decisions using artificial intelligence. *Medical Education*. 2022;56(3): 303–311. <https://doi.org/10.1111/medu.14696>.
28. Ledwos N, Mirchi N, Yilmaz R, Winkler-Schwartz A, Sawni A, Fazlollahi AM, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. *Journal of Neurosurgery*. 2022;137(4): 1160–1171. <https://doi.org/10.3171/2021.12.JNS211563>.
29. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS ONE*. 2020;15(2). <https://doi.org/10.1371/journal.pone.0229596>.

30. Morjaria L, Burns L, Bracken K, Ngo QN, Lee M, Levinson AJ, et al. Examining the Threat of ChatGPT to the Validity of Short Answer Assessments in an Undergraduate Medical Program. *Journal of Medical Education and Curricular Development*. 2023;10. <https://doi.org/10.1177/23821205231204178>.
31. Nagaraj MB, Namazi B, Sankaranarayanan G, Scott DJ. Developing artificial intelligence models for medical student suturing and knot-tying video-based assessment and coaching. *Surgical Endoscopy*. 2023;37(1): 402–411. <https://doi.org/10.1007/s00464-022-09509-y>.
32. Ruberto AJ, Rodenburg D, Ross K, Sarkar P, Hungler PC, Etemad A, et al. The future of simulation-based medical education: Adaptive simulation utilizing a deep multitask neural network. *AEM Education and Training*. 2021;5(3). <https://doi.org/10.1002/aet2.10605>.
33. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT Interactive Medical Simulations for Early Clinical Education: Case Study. *JMIR Medical Education*. 2023;9(1). <https://doi.org/10.2196/49877>.
34. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, Wang A, Makhani HH, Alsayegh A, et al. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems– A Four-Arm Randomized Control Trial. *Journal of Surgical Education*. 2023;81(2): 275–287. <https://doi.org/10.6084/m9.figshare.15132507.v1>.
35. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, Wang A, Makhani HH, Alsayegh A, et al. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems– A Four-Arm Randomized Control Trial. *Journal of Surgical Education*. 2024;81(2): 275–287. <https://doi.org/https://doi.org/10.1016/j.jsurg.2023.11.001>.
36. 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). <https://doi.org/10.1001/jamanetworkopen.2021.49008>.
37. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in Surgical Curriculum Design and Unintended Outcomes for Technical Competencies in Simulation Training. *JAMA network open*. 2023;6(9): e2334658. <https://doi.org/10.1001/jamanetworkopen.2023.34658>.
38. Bajčetić M, Mirčić A, Rakočević J, Đoković D, Milutinović K, Zaletel I. Comparing the performance of artificial intelligence learning models to medical students in solving histology and embryology multiple choice questions. *Annals of Anatomy - Anatomischer Anzeiger*. 2024;254: 152261. <https://doi.org/https://doi.org/10.1016/j.aanat.2024.152261>.
39. Coşkun Ö, Kiyak YS, Budakoğlu İİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: A randomized controlled experiment. *Medical Teacher*. 2024;(2): 268–274. <https://doi.org/10.1080/0142159X.2024.2327477>.
40. 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 Medical Education*. 2024;10. <https://doi.org/10.2196/59213>.
41. 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. *Computer Methods and Programs in Biomedicine*. 2024;244: 107964. <https://doi.org/https://doi.org/10.1016/j.cmpb.2023.107964>.
42. Moglia A, Morelli L, D'Ischia R, Fatucchi LM, Pucci V, Berchiolli R, et al. Ensemble deep learning for the prediction of proficiency at a virtual simulator for robot-assisted surgery. *Surgical Endoscopy*. 2022;36(9): 6473–6479. <https://doi.org/10.1007/s00464-021-08999-6>.
43. Riedel M, Kaefinger K, Stuehrenberg A, Ritter V, Amann N, Graf A, et al. ChatGPT's performance in German OB/GYN exams – paving the way for AI-enhanced medical education and clinical practice. *Frontiers in Medicine*. 2023;10. <https://doi.org/10.3389/fmed.2023.1296615>.
44. Tomova M, Roselló Atanet I, Sehy V, Sieg M, März M, Mäder P. Leveraging large language models to construct feedback from medical multiple-choice Questions. *Scientific Reports*. 2024;14(1). <https://doi.org/10.1038/s41598-024-79245-x>.
45. Barui S, Sanyal P, Rajmohan KS, Malik A, Dudani S. Perception without preconception: comparison between the human and machine learner in recognition of tissues from histological sections. *Scientific Reports*. 2022;12(1). <https://doi.org/10.1038/s41598-022-20012-1>.

46. Cheng CT, Chen CC, Fu CY, Chaou CH, Wu YT, Hsu CP, et al. Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights into Imaging*. 2020;11(1). <https://doi.org/10.1186/s13244-020-00932-0>.
47. Ghosh A, Maini Jindal N, Gupta VK, Bansal E, Kaur Bajwa N, Sett A. Is ChatGPT's Knowledge and Interpretative Ability Comparable to First Professional MBBS (Bachelor of Medicine, Bachelor of Surgery) Students of India in Taking a Medical Biochemistry Examination? *Cureus*. 2023;15(10). <https://doi.org/10.7759/cureus.47329>.
48. Huang CH, Hsiao HJ, Yeh PC, Wu KC, Kao CH. Performance of ChatGPT on Stage 1 of the Taiwanese medical licensing exam. *Digital Health*. 2024;10. <https://doi.org/10.1177/20552076241233144>.
49. Kinoshita K, Maruyama T, Kobayashi N, Imanishi S, Maruyama M, Ohira G, et al. An artificial intelligence-based nerve recognition model is useful as surgical support technology and as an educational tool in laparoscopic and robot-assisted rectal cancer surgery. *Surgical Endoscopy*. 2024;38(9): 5394–5404. <https://doi.org/10.1007/s00464-024-10939-z>.
50. Li J, Ouyang J, Liu J, Zhang F, Wang Z, Guo X, et al. Artificial Intelligence-based online platform assists blood cell morphology learning: A mixed-methods sequential explanatory designed research. *Medical Teacher*. 2023;45(6): 596–603. <https://doi.org/10.1080/0142159X.2023.2190483>.
51. Su JM, Hsu SY, Fang TY, Wang PC. Developing and validating a knowledge-based AI assessment system for learning clinical core medical knowledge in otolaryngology. *Computers in Biology and Medicine*. 2024;178. <https://doi.org/10.1016/j.compbiomed.2024.108765>.
52. Zheng K, Shen Z, Chen Z, Che C, Zhu H. Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education. *BMC Medical Education*. 2024;24(1). <https://doi.org/10.1186/s12909-024-05977-z>.
53. Alessi MR, Gomes HA, Lopes de Castro M, Terumy Okamoto C. Performance of ChatGPT in Solving Questions From the Progress Test (Brazilian National Medical Exam): A Potential Artificial Intelligence Tool in Medical Practice. *Cureus*. 2024;16(7). <https://doi.org/10.7759/cureus.64924>.
54. Alonso-Silverio GA, Pérez-Escamirosa F, Bruno-Sanchez R, Ortiz-Simon JL, Muñoz-Guerrero R, Minor-Martinez A, et al. Development of a Laparoscopic Box Trainer Based on Open Source Hardware and Artificial Intelligence for Objective Assessment of Surgical Psychomotor Skills. *Surgical Innovation*. 2018;25(4): 380–388. <https://doi.org/10.1177/1553350618777045>.
55. Castellano MS, Contreras-McKay I, Neyem A, Farfán E, Inzunza O, Ottone NE, et al. Empowering human anatomy education through gamification and artificial intelligence: An innovative approach to knowledge appropriation. *Clinical Anatomy*. 2024;37(1): 12–24. <https://doi.org/10.1002/ca.24074>.
56. Aronovitz N, Hazan I, Jedwab R, Shitrit I Ben, Quinn A, Wacht O, et al. The effect of real-time EF automatic tool on cardiac ultrasound performance among medical students. *PLoS ONE*. 2024;19(3). <https://doi.org/10.1371/journal.pone.0299461>.
57. Meo SA, AbuKhalaf AA, Meo MZS, Meo MOS, Ayub R, ElToukhy RA, et al. Role of artificial intelligence (Google bard) in morphological, histopathological, and radiological image identifications: Objective Structured Practical Examination (OSPE) type-based performance. *Saudi medical journal*. 2024;45(5): 531–536. <https://doi.org/10.15537/smj.2024.45.5.20240141>.
58. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy course: Comparative analysis. *Clinical Anatomy*. 2024;38(2): 200–210. <https://doi.org/10.1002/ca.24244>.
59. Cherif H, Moussa C, Missaoui AM, Salouage I, Mokaddem S, Dhahri B. Appraisal of ChatGPT's Aptitude for Medical Education: Comparative Analysis With Third-Year Medical Students in a Pulmonology Examination. *JMIR Medical Education*. 2024;10: e52818. <https://doi.org/10.2196/52818>.
60. Squalli Houssaini M, Aboutajeddine A, Toughrai I, Ibrahim A. Development of a design course for medical curriculum: Using design thinking as an instructional design method empowered by constructive alignment and generative AI. *Thinking Skills and Creativity*. 2024;52: 101491. <https://doi.org/https://doi.org/10.1016/j.tsc.2024.101491>.
61. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. *ANZ Journal of Surgery*. 2024;94(1–2): 68–77. <https://doi.org/10.1111/ans.18666>.

62. Gan W, Ouyang J, Li H, Xue Z, Zhang Y, Dong Q, et al. Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial. *Journal of Medical Internet Research*. 2024;26. <https://doi.org/10.2196/57037>.
63. Baglivo F, De Angelis L, Casigliani V, Arzilli G, Privitera GP, Rizzo C. Exploring the Possible Use of AI Chatbots in Public Health Education: Feasibility Study. *JMIR Medical Education*. 2023;9. <https://doi.org/10.2196/51421>.
64. Tsopra R, Peiffer-Smadja N, Charlier C, Campeotto F, Lemogne C, Ruzsniowski P, et al. Putting undergraduate medical students in AI-CDSS designers' shoes: An innovative teaching method to develop digital health critical thinking. *International Journal of Medical Informatics*. 2023;171: 104980. <https://doi.org/https://doi.org/10.1016/j.ijmedinf.2022.104980>.
65. Bharatha A, Ojeh N, Rabbi AMF, Campbell MH, Krishnamurthy K, Layne-Yarde RNA, et al. Comparing the Performance of ChatGPT-4 and Medical Students on MCQs at Varied Levels of Bloom's Taxonomy. *Advances in Medical Education and Practice*. 2024;15: 393–400. <https://doi.org/10.2147/AMEP.S457408>.
66. Furlan R, Gatti M, Menè R, Shiffer D, Marchiori C, Giaj Levra A, et al. A Natural Language Processing-Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study. *JMIR Medical Informatics*. 2021;9(4). <https://doi.org/https://doi.org/10.2196/24073>.
67. Fazlollahi AM, Bakhaider 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). <https://doi.org/10.1001/jamanetworkopen.2021.49008>.
68. AlZaabi A, AlMaskari S, AalAbdulsalam A. Are physicians and medical students ready for artificial intelligence applications in healthcare? *Digital Health*. 2023;9. <https://doi.org/10.1177/20552076231152167>.
69. Buabbas AJ, Miskin B, Alnaqi AA, Ayed AK, Shehab AA, Syed-Abdul S, et al. Investigating Students' Perceptions towards Artificial Intelligence in Medical Education. *Healthcare*. 2023;11(9): 1298. <https://doi.org/10.3390/healthcare11091298>.
70. Ejaz H, McGrath H, Wong BL, Guise A, Vercauteren T, Shapey J. Artificial intelligence and medical education: A global mixed-methods study of medical students' perspectives. *DIGITAL HEALTH*. 2022;8: 205520762210890. <https://doi.org/10.1177/20552076221089099>.
71. Kimmerle J, Timm J, Festl-Wietek T, Cress U, Herrmann-Werner A. Medical Students' Attitudes Toward AI in Medicine and their Expectations for Medical Education. *Journal of Medical Education and Curricular Development*. 2023;10. <https://doi.org/10.1177/23821205231219346>.
72. Al Saad MM, Shehadeh A, Alanazi S, Alenezi M, Abu alez A, Eid H, et al. Medical Students' Knowledge and Attitude Towards Artificial Intelligence: An Online Survey. *The Open Public Health Journal*. 2022;15. <https://doi.org/https://doi.org/10.2174/18749445-v15-e2203290>.
73. Lin S. A Clinician's Guide to Artificial Intelligence (AI): Why and How Primary Care Should Lead the Health Care AI Revolution. *The Journal of the American Board of Family Medicine*. 2022;35(1): 175–184. <https://doi.org/10.3122/jabfm.2022.01.210226>.
74. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*. 2019;16(11): 703–715. <https://doi.org/10.1038/s41571-019-0252-y>.
75. Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of Artificial Intelligence Reported in Obstetrics and Gynecology Journals: Systematic Review. *Journal of Medical Internet Research*. 2022;24(4): e35465. <https://doi.org/10.2196/35465>.
76. Rösler W, Altenbuchinger M, Baeßler B, Beissbarth T, Beutel G, Bock R, et al. An overview and a roadmap for artificial intelligence in hematology and oncology. *Journal of Cancer Research and Clinical Oncology*. 2023;149(10): 7997–8006. <https://doi.org/10.1007/s00432-023-04667-5>.
77. Uddin J, Feng C, Xu J. Health Communication on the Internet: Promoting Public Health and Exploring Disparities in the Generative AI Era. *Journal of Medical Internet Research*. 2025;27: e66032. <https://doi.org/10.2196/66032>.