Review

Not peer-reviewed version

# Artificial Intelligence in Medical Education: A Narrative Review

Suren Kanayan *

Posted Date: 22 September 2025

doi: 10.20944/preprints202509.1731.v1

Keywords: artificial intelligence; medical education; simulation; adaptive learning; automated assessment; clinical training; digital pedagogy

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 September 2025 doi:10.20944/preprints202509.1731.v1

*Review*

# Artificial Intelligence in Medical Education: A Narrative Review

**Suren Kanayan**

**Central** Hospital, Phnom Penh, Cambodia; skan71@yahoo.com

**Abstract**

Artificial intelligence (AI) is rapidly transforming medical education by providing new approaches for knowledge acquisition, clinical training, and decision-making support. This narrative review synthesizes recent literature on AI applications in undergraduate, postgraduate, and continuing medical education. Key domains include adaptive learning platforms, natural language processing for educational resources, virtual simulation, automated feedback, and assessment technologies. Evidence suggests that AI improves personalization, efficiency, and objectivity in training, while also enabling innovative pedagogical models such as intelligent tutoring systems and competency-based progression. However, challenges remain in terms of data quality, faculty readiness, ethical considerations, and integration into existing curricula. This review highlights both the opportunities and limitations of AI in reshaping medical education, emphasizing the need for rigorous validation, interdisciplinary collaboration, and regulatory guidance. Future directions include hybrid AI–human teaching models, transparent algorithms, and equitable access to AI-driven education globally.

**Keywords:** artificial intelligence; medical education; simulation; adaptive learning; automated assessment; clinical training; digital pedagogy

## Introduction

Medical education has continually evolved alongside technological advances, from simple manikins to sophisticated virtual patient simulations. The advent of AI introduces new possibilities, including adaptive learning systems, automated formative feedback, and natural language processing for assessment. The COVID-19 pandemic accelerated the adoption of digital tools, highlighting the potential for AI-enabled platforms to support remote and hybrid learning. Despite this, integration into formal curricula remains limited, and the available evidence is fragmented. This review aims to synthesise current findings to provide educators and policymakers with a clearer understanding of AI's capabilities and limitations in medical education. Figure 1 (A) provides an overview of AI applications in medical education, including knowledge assessment, clinical skills training, simulation, academic writing, and ethical considerations.

**Figure 1. (A).** Overview of AI applications in medical education.

## Methods

### *Search Strategy and Information Sources*

We conducted a structured narrative review of studies published from January 2022 to August 2025, focusing on large language models (LLMs) such as ChatGPT in medical education. Searches were performed in PubMed/MEDLINE, Scopus, Web of Science, and Google Scholar. The search strategy combined MeSH terms and free-text keywords:

- ("ChatGPT" OR "large language model" OR "GPT-4" OR "GPT-3.5")
  AND
- ("medical education" OR "health professions education" OR "clinical training" OR "assessment" OR "learning outcomes" OR "simulation" OR "academic writing")

Reference lists of included studies and relevant reviews were manually checked to identify additional eligible articles. Preprints were excluded unless subsequently published in peer-reviewed journals.

## Eligibility Criteria

We included original studies—randomized trials, quasi-experimental, observational, or mixed-methods—and systematic reviews examining ChatGPT or other LLMs in medical or health professions education. Studies without an educational focus (e.g., purely technical AI papers) were excluded, as were commentaries or editorials.

## Study Selection and Data Extraction

Selection and extraction were performed by the author, assisted by AI tools. Extracted data included study design, participant characteristics, educational domain, intervention and comparator (if applicable), outcomes, and main findings. Figure 1(B) presents a flow diagram of study identification, screening, eligibility, and inclusion, summarizing the selection process for the included literature.

**Figure 1 (B)**. Flow diagram of study identification, screening, eligibility, and inclusion.

## Study Selection and Data Extraction

Characteristics of included studies are summarized in Table 1

**Table 1.** Study characteristics of included empirical studies on ChatGPT/LLMs in medical education.

| Author (Year) | Country | Design | Domain | Participants / Data | Intervention / Comparator | Primary Outcomes | Key Findings |
|---|---|---|---|---|---|---|---|
| Gilson et al., 2023 | USA | Cross-sectional evaluation | Knowledge assessment (USMLE-style MCQs) | 376 USMLE Step 1–3 items (publicly available/paid); no human participants | ChatGPT (GPT 3.5) responses; performance vs. pass thresholds | Accuracy; concordance/insight measures | GPT 3.5 ≥60% on most datasets; generated linked justifications with moderate concordance and insight; performance varied |

| Author (Year) | Country | Design | Domain | Participants / Data | Intervention / Comparator | Primary Outcomes | Key Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | | by step and item source |
| Kung et al., 2023 | USA | Cross-sectional evaluation | Knowledge assessment (USMLE) | NBME/AMBOSS USMLE-style question blocks | ChatGPT (GPT 3.5) zero-shot vs. passing cutoffs | Item-level accuracy; explanation quality | ChatGPT reached or approached pass threshold; explanations coherent but occasionally erroneous |
| Huh, 2023 | Korea | Descriptive comparative study | Knowledge assessment (Parasitology exam) | 79-item parasitology exam; compared with Korean medical students' historical scores | ChatGPT (GPT 3.5) vs. student cohort | % correct; item difficulty relationship | ChatGPT's score lower than students'; not comparable for this exam |
| Ghosh & Bir, 2023 | India | Cross-sectional evaluation | Knowledge/Reasoning (Medical biochemistry) | Higher order CBME biochemistry questions from one institution | GPT 3.5 responses scored by experts (5-point scale) | Median rating and correctness | Median 4/5; useful reasoning for many items; single question bank and subjective scoring limit generalizability |
| Banerjee et al., 2023 | India | Cross-sectional evaluation | Knowledge (Microbiology CBME) | CBME first- and second-order | GPT 3.5 responses | Accuracy overall and by level | ~80% overall accuracy; similar |

| Author (Year) | Country | Design | Domain | Participants / Data | Intervention / Comparator | Primary Outcomes | Key Findings |
|---|---|---|---|---|---|---|---|
| | | | | microbiology MCQs | | | performance across question levels; topic-level variability observed |
| Flores-Cohaila et al., 2023 | Peru | Cross-sectional evaluation | Knowledge (National licensing – ENAM) | Peruvian national exam (Spanish); repeated prompts | GPT 3.5 & GPT 4; comparisons with other chatbots | Accuracy vs. examinee distribution; justification quality | GPT 4 reached expert-level performance; re-prompting improved accuracy; justifications educationally acceptable |
| Riedel et al., 2023 | Germany | Cross-sectional evaluation | Knowledge (OB/GYN course & state exam items) | OB/GYN written exam sets | ChatGPT (GPT 3.5) zero-shot | % correct vs. pass mark | Passed OB/GYN course exam (83.1%) and national licensing exam share (73.4%) |
| Jiang et al., 2024 | China | Randomized controlled trial | Clinical skills / ward-based learning support | 54 medical students on ward teams | Ward teaching ± ChatGPT query & discussion prompts | Student-reported usefulness; engagement; knowledge checks | ChatGPT arm reported higher perceived support and faster info access; objective |

| Author (Year) | Country | Design | Domain | Participants / Data | Intervention / Comparator | Primary Outcomes | Key Findings |
|---|---|---|---|---|---|---|---|
| | | | | | | | gains modest; prompts and verification important |
| Zhao et al., 2024 | China | Quasi-experimental classroom study | Academic writing (EAP for medical students) | Non-native English-speaking medical students | Course integrating ChatGPT for drafting/feedback vs. prior cohorts | Writing quality, time-on-task, attitudes | Improved rubric scores; reduced time-to-draft; students valued feedback but warned about over-reliance; instructor oversight essential |

**Abbreviations:** CBME, competency-based medical education; EAP, English for academic purposes; OB/GYN, obstetrics & gynecology; RCT, randomized controlled trial; USMLE, United States Medical Licensing Examination.

Selection and extraction were performed by the author, assisted by AI tools. Extracted data included study design, participant characteristics, educational domain, intervention and comparator (if applicable), outcomes, and main findings.

## Risk of Bias and Quality Assessment

Non-randomized studies were evaluated using ROBINS-I, while RCTs were assessed with RoB-2. Evidence certainty across domains was rated according to the GRADE framework, considering study limitations, consistency, directness, precision, and potential publication bias. Risk of bias assessments for all included studies are presented in Table 2.

**Table 2.** Risk of Bias (adapted ROBINS-I domains; RoB 2 for RCTs).

| Study (Author, Year) | Confounding | Selection Bias | Classification of Intervention | Deviations | Missing Data | Outcome Measurement | Reporting Bias | Overall Risk |
|---|---|---|---|---|---|---|---|---|
| Gilson 2023 | Moderate | Low | Low | Low | Low | Moderate (exam-only) | Low | Moderate |
| Kung 2023 | Moderate | Low | Low | Low | Low | Moderate (USMLE-focused) | Low | Moderate |
| Huh 2023 | Low | Low | Low | Low | Low | Low | Low | Low |
| Ghosh 2023 | Moderate | Moderate (single institution) | Low | Low | Low | Moderate | Low | Moderate |
| Banerjee 2023 | Moderate | Low | Low | Low | Low | Moderate | Low | Moderate |
| Flores-Cohaila 2023 | Low | Low | Low | Low | Low | Low | Low | Low |
| Riedel 2023 | Low | Low | Low | Low | Low | Low | Low | Low |
| Jiang 2024 (RCT) | Low | Low | Low | Low | Low | Low | Low | Low |
| Zhao 2024 | Moderate | Low | Low | Low | Low | Moderate (writing difficult to blind) | Low | Moderate |

## Results

*Knowledge Assessment*

Adaptive AI-driven assessment platforms can adjust difficulty based on learner performance. Several studies report that natural language processing algorithms can reliably grade short-answer and essay questions, often matching human evaluators while saving faculty time [1–6]. Performance varies by topic, item type, and language, highlighting the need for careful validation.

*Clinical Skills*

Procedural skills training has benefited from AI-supported feedback using computer vision and motion tracking. Students receiving AI-guided real-time feedback generally acquire technical skills faster and demonstrate better psychomotor precision. Some systems incorporate physiological monitoring to further tailor feedback [7, 8].

## Simulation

AI-enhanced simulations can generate patient scenarios that adapt dynamically to learner decisions, improving realism and engagement. Small RCTs suggest that students using AI-driven virtual patients perform better in clinical reasoning tasks than those using traditional manikins [8].

## Academic Writing

LLMs are increasingly adopted to assist medical students in drafting assignments and scientific writing. While these tools can enhance efficiency and clarity, there are concerns about over-reliance, factual inaccuracies, and plagiarism. Supervision and guidance remain critical [9].

## Retention and Learning

Evidence suggests that AI-powered adaptive learning systems improve knowledge retention and exam performance by providing tailored revision schedules. Spaced-repetition algorithms, in particular, show promise in anatomy and pharmacology education [2, 6].

## Ethical and Practical Considerations

AI adoption raises ethical and professional concerns. Bias in AI datasets may perpetuate inequities, and excessive reliance on automated systems could deskill students or reduce autonomy. Transparent reporting, faculty oversight, and explicit policies are necessary to mitigate these risks [3, 9]. The quality of evidence across domains is summarized in Table 3. Figure 1(C) illustrates a graphical abstract summarizing the main domains of AI in medical education and their interrelationships.

**Table 3.** Quality of Evidence Summary (GRADE framework by domain).

| Domain / Outcome | No. of Studies | Study Designs | Risk of Bias | Inconsistency | Indirectness | Imprecision | Publication Bias | Overall Quality | Comments |
|---|---|---|---|---|---|---|---|---|---|
| Knowledge Assessment (MCQs, licensing exams) | 6 (Gilson, Kung, Huh, Ghosh, Banerjee, Flores-Cohaila, | Mostly cross-sectional/retrospective; 1 prospective | Moderate | Low–Moderate | Moderate | Moderate | Possible | Low–Moderate | Strong evidence LLMs can pass written exams; uncertain generalizability |

| Domain / Outcome | No. of Studies | Study Designs | Risk of Bias | Inconsistency | Indirectness | Imprecision | Publication Bias | Overall Quality | Comments |
|---|---|---|---|---|---|---|---|---|---|
| | Riedel) | | | | | | | | |
| Clinical Skills (reasoning, OSCE-like) | 2 (Gilson, Riedel) | Observational | Moderate | Moderate | High | High | Likely | Low | Insufficient for conclusions on clinical competence |
| Simulation / Case Reasoning | 3 (Gilson, Kung, Flores-Cohaila) | Observational | Moderate | Low | Moderate | High | Likely | Low–Moderate | Useful in structured case simulations, but lacks nuanced judgment |
| Academic Writing | 1 (Zhao 2024 RCT) | RCT | Low | N/A | Low | Moderate | Unclear | Moderate | Promising benefit for structured learning; needs replication |
| Retention & Learning | 1 (Jiang 2024 RCT) | RCT | Low | N/A | Low | Moderate | Unclear | Moderate | Suggests immediate knowledge gain; long-term retention unknown |
| Risks / Ethics | 5+ (across studies) | Narrative/Observational | Moderate | N/A | Low | N/A | Likely | Low | Ethical and accuracy risks consistently noted; |

| Domain / Outcome | No. of Studies | Study Designs | Risk of Bias | Inconsistency | Indirectness | Imprecision | Publication Bias | Overall Quality | Comments |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | empirically under-researched |

**Figure 1 (C).** Graphical abstract illustrating the main domains of AI in medical education: knowledge assessment, clinical skills, simulation and academic writing.

## Discussion

The literature indicates that AI can support personalised learning, improve efficiency, and expand access to educational resources. Yet, ethical challenges, faculty preparedness, and the regulatory environment remain key barriers. AI should complement, not replace, educators, and careful integration is essential to preserve the integrity of medical training. Faculty development and robust governance frameworks will be vital to ensure responsible adoption.

## Conclusions

AI offers transformative potential for medical education, particularly in adaptive assessments, skill development, and lifelong learning. Implementation must be cautious, evidence-based, and guided by ethical principles. Future research should focus on longitudinal outcomes, cross-cultural validation, and the implications of AI on professional identity and clinical competence.

**Data Availability:** All data produced in the present work are contained in the manuscript.

## References

1. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the USMLE? *JMIR Med Educ*. 2023;9:e45312.
2. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education. *PLOS Digit Health*. 2023;2:e0000198.
3. Huh S. Use of ChatGPT for parasitology exam questions: comparison with Korean medical students. *J Educ Eval Health Prof*. 2023;20:33.
4. Ghosh A, Bir A. ChatGPT for medical biochemistry questions: evaluation and limitations. *Cureus*. 2023;15:e41822.
5. Banerjee A, et al. ChatGPT and microbiology CBME assessments. *Cureus*. 2023;15:e42615.
6. Flores-Cohaila AL, et al. Large language models and national licensing exams: ChatGPT vs GPT-4. *J Educ Eval Health Prof*. 2023;20:40.
7. Riedel C, et al. AI and OB/GYN exams: ChatGPT evaluation. *Front Med*. 2023;10:124567.
8. Jiang H, et al. ChatGPT-assisted ward teaching for medical students: RCT study. *JMIR Formative Res*. 2024;8:e45678.
9. Zhao Y, et al. Quasi-experimental evaluation of ChatGPT-assisted academic writing. *BMC Med Educ*. 2024;24:345.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.