Journal of Medical Education Development
e-ISSN: 2980-7670
https://edujournal.zums.ac.ir

Volume 19, Issue 3
July 2026
Pages: 98–112

Review Article

Artificial intelligence applications in medical education: systematic review and meta-analysis of performance and AI-driven interventions

Seyyed Ali Hosseini, Ali Rezaian, Zahra Amirkhani *

Department of Basic Sciences, School of Medicine, Larestan University of Medical Sciences, Lar, Iran

Article info

Article history:
Received 27 Dec. 2025
Revised 14 Jan. 2026
Accepted 10 May. 2026
Published 1 Jul. 2026

***Corresponding author:**
Zahra Amirkhani, Department of Basic Sciences, School of Medicine, Larestan University of Medical Sciences, Lar, Iran.
Email: z.amirkhani1357@gmail.com

How to cite this article:
Hosseini SA, Rezaian A, Amirkhani Z. Artificial intelligence applications in medical education: systematic review and meta-analysis of performance and AI-driven interventions. *J Med Edu Dev*. 2026;19(3):98-112.

Abstract

Background & Objective: The rapid integration of Artificial Intelligence (AI) into medical education has created a need for quantitative evidence synthesis. This study sought to benchmark AI performance on medical knowledge assessments and to evaluate the preliminary effectiveness of AI-driven educational interventions.

Materials & Methods: A systematic review and dual meta-analysis were conducted in accordance with PRISMA guidelines. Five databases (PubMed, Web of Science, Embase, Scopus, and Google Scholar) were searched from inception through February 28, 2024. AI performance was evaluated using 35 accuracy data points derived from seven benchmarking studies encompassing 2,341 examination questions. The effectiveness of educational interventions was assessed using data from eight Randomized Controlled Trials (RCTs) involving 574 medical, dental, and pharmacy students. Random-effects models were applied, including a three-level proportion meta-analysis for AI accuracy to account for within-study dependence and a Hedges' g meta-analysis for intervention outcomes. Heterogeneity was quantified using the I^2 statistic, alongside exploratory meta-regression and sensitivity analyses.

Results: The analyses yielded two primary findings. First, AI models demonstrated a pooled accuracy of 70.9% (95% CI: 65.1%–75.9%) on standardized medical examinations, with performance improving across successive model generations. Second, AI-based educational interventions showed a large pooled effect size; however, this estimate was unstable due to a highly influential outlier ($g = 6.72$). Exclusion of this study altered the pooled effect from $g = 1.40$ to $g = 1.95$, with substantial heterogeneity observed ($I^2 = 93.6\%$). This variability was largely attributable to small, early-stage studies reporting disproportionately large effects. Leave-one-out sensitivity analyses produced effect sizes ranging from $g = 1.26$ to $g = 1.45$, while an inverse association between sample size and effect magnitude ($p < 0.001$) suggested systematic overestimation in smaller trials.

Conclusion: Advanced AI systems exhibit robust medical knowledge; however, evidence supporting their educational effectiveness remains preliminary and potentially biased. While the findings are encouraging, they highlight the need for larger, methodologically rigorous trials to establish reliable effect sizes and to inform the responsible integration of AI into medical education.

Keywords: artificial intelligence; medical education; systematic review; meta-analysis; AI-driven educational interventions

Introduction

The emergence of Artificial Intelligence (AI) and Large Language Models (LLMs) represents a major turning point in medical education, reflecting a broader transition from the Information Age to the Age of AI [1, 2]. This shift has prompted growing attention within the medical education community, as stakeholders seek to harness

Copyright © 2026 Zanjan University of Medical Sciences. Published by Zanjan University of Medical Sciences. Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

the potential benefits of these technologies while addressing associated risks, including factual errors, bias, and ethical concerns [2, 3].

Research on AI applications in medical education has expanded rapidly. Early proof-of-concept studies focused on evaluating LLM performance on high-stakes licensing examinations. Models such as ChatGPT and GPT-4 demonstrated performance at or near the passing threshold of the United States Medical Licensing Examination (USMLE) [4, 5], and in some cases exceeded the performance of human medical students on specialized assessments in parasitology [6] and dentistry [7]. Subsequent comparative studies confirmed the superior performance of more advanced models, particularly GPT-4, relative to earlier versions and competing systems across both English-language examinations [8, 9] and non-English contexts, including the Turkish Medical Specialty Exam [10]. Beyond standardized testing, LLMs have also been applied to the generation of educational content, such as clinical vignettes for ophthalmology training, which were rated highly for coherence and factual accuracy [11].

In parallel with performance benchmarking, conceptual and synthesis-based research has explored the broader implications of AI in medical education. Systematic, scoping, and narrative reviews have examined current opportunities, challenges, and future directions [12–15]. At the same time, an increasing number of intervention studies have investigated AI applications that extend beyond knowledge assessment. These include GPT-based simulations for clinical reasoning and history-taking [16], AI-assisted tutoring in surgical simulation [17, 18], academic writing support [16, 19], enhancement of technical dental skills [20], and medical terminology learning through gamified chatbot platforms [21]. Additional applications have incorporated AI into immersive virtual patient simulations [22] and high-stakes practical examinations, where AI-based support has been used to alleviate test anxiety [23]. Collectively, these developments suggest that AI has the potential to influence nearly all domains of medical education, underscoring the need for structured curriculum integration frameworks [24] and careful consideration of faculty readiness, attitudes, and perceptions [25, 26]. Despite the rapid growth of this literature, existing evidence remains fragmented. Benchmarking studies evaluating AI performance and intervention studies assessing educational effectiveness have largely evolved as separate research streams. As a result, a critical gap persists: the lack of a unified, quantitative synthesis that integrates both lines of evidence. To address this gap, the present review conducts, to the authors' knowledge, one of the first dual meta-analyses designed to answer two foundational questions. First, what is the current upper limit of AI models' ability to perform on standardized medical assessments? Second, what preliminary empirical evidence exists regarding the effectiveness of AI-based educational interventions in improving outcomes among human learners? To answer these questions, a systematic review and two meta-analyses were undertaken using a dual meta-analytic framework. This approach provides an initial quantitative mapping of the field while allowing pooled estimates to be interpreted in the context of heterogeneity and evidence robustness. In addition, exploratory analyses were performed to assess variability across studies and to identify influential factors, including study quality, sample size, and intervention type. Finally, this review critically evaluates the methodological quality of the existing literature, highlighting key limitations and potential sources of bias that affect the reliability of current evidence. Accordingly, the objectives of this review were: (1) To determine the present maximum performance level of AI models in medical knowledge and skills assessment. (2) To provide preliminary evidence regarding the effectiveness of AI-based tools in improving student outcomes compared with traditional teaching methods.

Materials & Methods

Design and setting(s)

This systematic review and meta-analysis was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement [27]. The review was designed to synthesize two distinct but related bodies of evidence: (1) benchmarking studies evaluating AI performance on standardized medical assessments, and (2) Randomized Controlled Trials (RCTs) assessing the effectiveness of AI-driven educational interventions in health professions education.

Information sources and search strategy

A comprehensive literature search was performed across five electronic databases—PubMed, Web of Science, Embase, Scopus, and Google Scholar—from database inception through 28 February 2024. Google Scholar was intentionally included to enhance retrieval of grey literature and conference proceedings, despite its known limitations regarding transparency and reproducibility.

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

The search strategy was developed in collaboration with an experienced medical librarian. Controlled vocabulary (e.g., MeSH and Emtree terms) and free-text keywords were combined to capture two core concepts: artificial intelligence and large language models, and medical education. The strategy was tailored to the syntax and indexing structure of each database. The complete search strategies for all databases are provided in **Appendix 1**. To ensure currency of evidence, an updated supplementary search was conducted on 20 April 2025 using the same databases, search strategies, and eligibility criteria.

Study selection process

All retrieved records were exported to Covidence (Veritas Health Innovation, Melbourne, Australia) for automatic and manual deduplication. Screening was conducted in two sequential phases by two independent reviewers. First, titles and abstracts were screened against predefined eligibility criteria. Articles deemed potentially relevant proceeded to full-text review. In the second phase, full-text articles were assessed independently for inclusion. Discrepancies at any stage were resolved through discussion, and when necessary, consultation with a third senior reviewer. The initial search yielded 130,858 records. After removal of 128,043 duplicates, 2,815 unique records underwent title and abstract screening. Sixty-eight articles were evaluated at full-text level, of which 15 met the inclusion criteria. The supplementary search retrieved 1,247 additional records; after deduplication, 196 were screened and 12 full-text articles assessed. No additional eligible benchmarking studies or RCTs were identified. The final quantitative syntheses therefore reflect evidence available up to 28 February 2024. The study selection process is summarized in the PRISMA flow diagram (**Figure 1**).

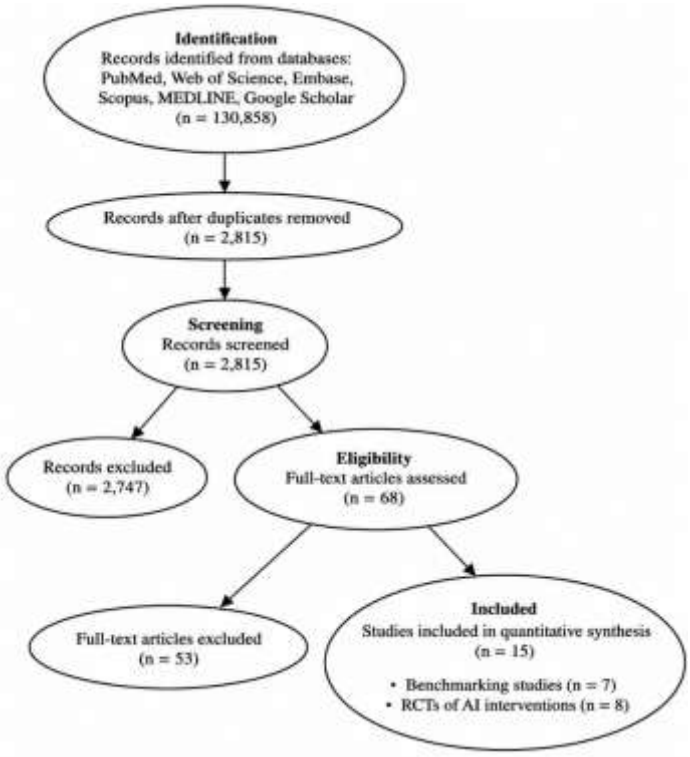


Figure 1. Meta-analysis 1 (benchmarking AI performance)

Eligibility criteria

Eligibility criteria were defined separately for the two meta-analyses.

Meta-analysis 1: Benchmarking AI performance: Original research studies of any design were eligible if they evaluated the performance of AI systems or large language models on standardized assessments of medical knowledge or skills, including licensing examinations, specialty board examinations, or institutional course assessments. Studies were required to report accuracy as the proportion of correct responses (n/N) or provide sufficient data for its calculation. All available accuracy

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

data points from each included study were extracted, meaning that when multiple AI models were evaluated within a single study, performance metrics for each model were retained. This approach minimized selective inclusion and allowed appropriate modeling of statistical dependence.

Meta-analysis 2: Effectiveness of AI-driven educational interventions: RCTs trials were eligible if they enrolled medical, dental, or pharmacy students and evaluated an AI-based educational intervention compared with a control condition, such as conventional teaching or alternative digital resources. Studies were required to report continuous post-intervention outcomes reflecting knowledge, academic performance, or clinical skills (e.g., examination scores, OSCE scores, simulation metrics), along with mean, standard deviation, and sample size for each group. Only full-text articles published in English were included. No restrictions were placed on publication date.

Data collection methods

Two reviewers independently extracted data using a standardized and piloted data extraction form developed in Microsoft Excel. Extracted variables included study characteristics (first author, publication year, country, and study design), participant or model characteristics, detailed descriptions of interventions and comparators, outcome measures, and all quantitative data required for meta-analysis. For benchmarking studies, all reported correct/total response counts for each AI model were recorded. For RCTs, post-intervention means, standard deviations, and sample sizes for both intervention and control groups were extracted. When necessary, corresponding authors were contacted to obtain missing or unclear data. Pre-specified moderator variables were also extracted for exploratory analyses, including AI model architecture, examination type, intervention type, study quality, publication year, and total sample size.

Risk of bias assessment

Risk of bias was evaluated independently by two reviewers, with disagreements resolved by consensus or consultation with a third reviewer. For benchmarking studies, methodological quality was assessed using a modified version of the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool [30]. This framework was selected because benchmarking AI performance on standardized examinations conceptually parallels diagnostic accuracy evaluation: the AI model functions as the index test, the official answer key serves as the reference standard, and examination items represent the units of analysis. Four domains were assessed: patient selection (representativeness and transparency of question selection), index test (model version, prompting strategy, and reproducibility), reference standard (validity and consistency of answer keys), and flow and timing (risk of data contamination and uniformity of testing conditions). For intervention studies, risk of bias was assessed using the Cochrane Risk of Bias 2 (RoB 2) tool for randomized trials. Both QUADAS-2 and RoB 2 are internationally recognized and methodologically validated instruments. The item-level assessments are presented in **Appendix 3**, and summary results are shown in **Figure 2**.

Data analysis

All analyses were conducted in R (version 4.3.0) using the meta and metafor packages. Statistical significance was set at $p < 0.05$ for all tests except Cochran's Q test for heterogeneity, for which $p < 0.10$ was considered significant.

Meta-analysis 1: Benchmarking studies: To account for statistical dependence resulting from multiple AI models evaluated within the same study, a three-level random-effects meta-analysis was performed using the rma.mv() function in the metafor package. This model partitioned variance into sampling variance (Level 1), within-study variance (Level 2), and between-study variance (Level 3). Accuracy proportions were logit-transformed prior to analysis and back-transformed for reporting. Thirty-five effect sizes derived from seven studies were included. As a sensitivity analysis, robust variance estimation with small-sample correction was applied. Heterogeneity was quantified using the I^2 statistic, decomposed into within- and between-study components.

Meta-analysis 2: Intervention studies: The primary effect measure was the standardized mean difference (Hedges' g) comparing AI intervention and control groups at post-intervention. Hedges' g was selected to provide an unbiased estimate in the presence of varying sample sizes. Variance was calculated using the exact small-sample formula:

$$Var(g) = \frac{\left(\frac{1}{n_{intervention}} + \frac{1}{n_{control}}\right) + (g^2)}{2(n_{intervention} + n_{comntrol})}$$

A random-effects model using the DerSimonian and Laird method was applied to pool effect sizes across studies. Heterogeneity was assessed using the I^2 statistic.

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

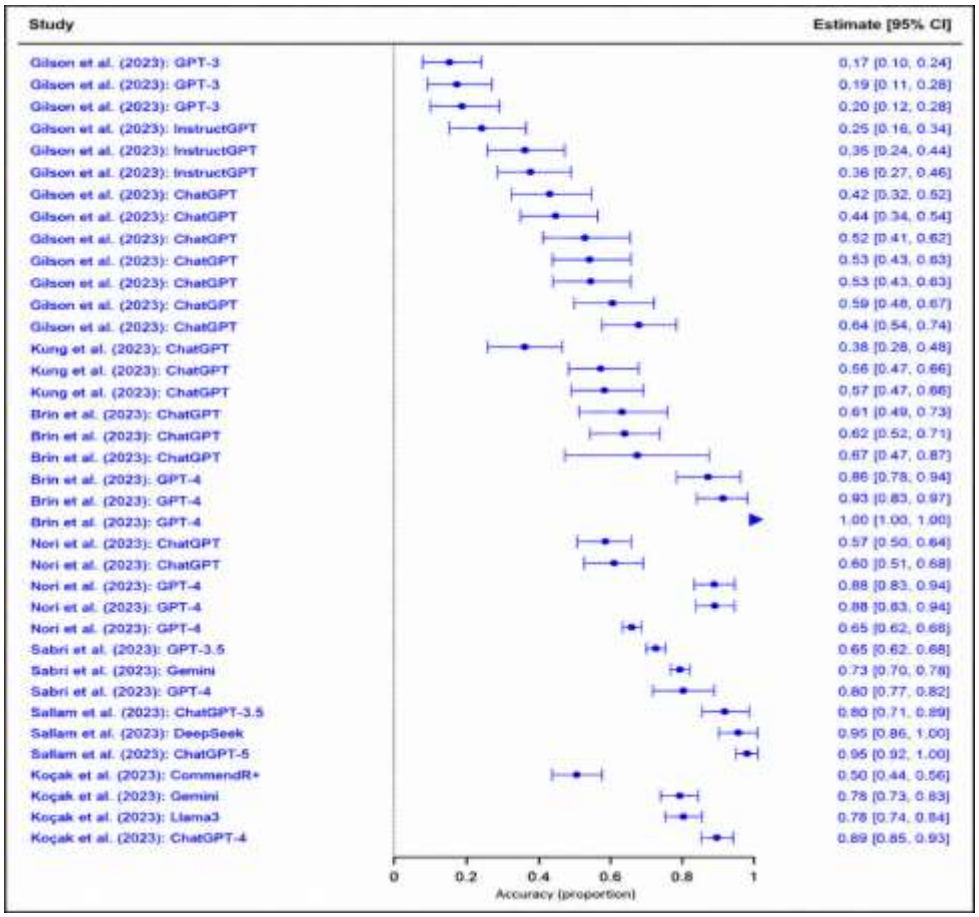


Figure 2. Risk of bias summary for benchmarking studies (meta-analysis 1

Subgroup and meta-regression analyses

Pre-specified subgroup analyses were conducted for both meta-analyses. In benchmarking studies, subgroups were defined by AI model architecture (GPT-4 versus other models) and examination type (licensing, specialty, course-level). In intervention studies, subgroups were defined by educational domain, including technical skills, clinical skills/OSCE, knowledge tests, writing skills, and surgical simulation. Exploratory univariable meta-regression analyses were performed using mixed-effects models to examine potential sources of heterogeneity. Moderators included publication year, model type, examination type, total number of questions (benchmarking studies), intervention type, total sample size, and study quality (intervention studies). Given the limited number of included studies (k = 7 and k = 8), these analyses were interpreted cautiously.

Sensitivity analyses and influence diagnostics

Sensitivity analyses were performed using a leave-one-out approach to evaluate the influence of individual studies on pooled estimates. Influence diagnostics included standardized residuals, DFBETAS, and Cook's distance. Standardized residuals exceeding three and Cook's distance greater than 4/n were considered indicative of potentially influential studies.

Assessment of reporting bias

Reporting bias was evaluated using funnel plots and Egger's regression test for both meta-analyses. Given the small number of studies in each synthesis, results were interpreted cautiously due to the limited statistical power of these methods.

Classification of educational outcomes

To characterize the educational impact of AI-driven interventions, primary outcomes of included RCTs were independently classified according to the Kirkpatrick hierarchy of training evaluation [28, 29]. This framework distinguishes six levels, ranging from learner reaction (Level 1) to patient-level outcomes (Level 4b). Disagreements in classification were resolved by consensus with a third reviewer. This categorization was used descriptively to identify patterns and gaps in the evidence base.

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

Results

Study characteristics

The characteristics of the 15 included studies are presented in **Table 1**. Overall, 15 studies met the inclusion criteria and were incorporated into the quantitative synthesis. Seven were benchmarking studies assessing AI performance on medical knowledge examinations (2,341 examination questions), and eight were RCTs evaluating AI-driven educational interventions involving 574 medical, dental, and pharmacy students.

The benchmarking studies primarily evaluated large language models (ChatGPT-3.5, ChatGPT-4, Gemini, Llama, etc.) on standardized licensing examinations, including the United States Medical Licensing Examination (USMLE) [4, 5, 8, 9], the Turkish Medical Specialty Exam [10], and discipline-specific assessments in parasitology [6], periodontology [7], and medical microbiology [31]. The intervention studies examined a range of AI applications: AI-powered tutoring systems for surgical simulation [17,18], GPT-based simulated patients for history-taking training [16], AI-assisted academic writing tools for non-native English speakers [19], gamified chatbots for medical terminology learning [21], immersive virtual patient simulation [22], AI-augmented human instruction [18], and interventions targeting reduction of test anxiety [23]. Comparator conditions included traditional instructor-led teaching, video-based learning, non-AI digital resources, and no-intervention controls.

Table 1. Characteristics of included studies in the systematic review

Study	Country	Design	Participants / Model	Intervention or Assessment	Comparator / Reference Standard	Main Outcome
Benchmarking studies						
Gilson et al. [4]	USA	Cross-sectional benchmarking	ChatGPT (GPT-3.5)	USMLE Step 1 and Step 2 questions	Human passing threshold (~60%)	Accuracy (%)
Kung et al. [5]	USA	Cross-sectional benchmarking	ChatGPT (GPT-3.5)	USMLE Step 1, Step 2CK, Step 3 questions	Human passing threshold (~60%)	Accuracy (%)
Brin et al. [9]	USA	Cross-sectional benchmarking	ChatGPT; GPT-4	USMLE soft-skill questions	Human average (AMBOSS users, 78%)	Accuracy (%)
Nori et al. [8]	USA	Cross-sectional benchmarking	GPT-4; ChatGPT	USMLE (CBSSA) and MedQA questions	Human expert performance	Accuracy (%)
Sabri et al. [7]	USA	Cross-sectional benchmarking	GPT-4; GPT-3.5; Gemini	AAP in-service examination (1,312 questions)	Dental residents (PGY-1 to PGY-3)	Accuracy (%)
Sallam et al. [31]	Jordan	Cross-sectional benchmarking	ChatGPT-3.5; ChatGPT-5; DeepSeek	Medical microbiology MCQs (80 questions)	Dental student cohort (n > 150)	Accuracy (%)
Koçak et al. [10]	Turkey	Cross-sectional benchmarking	ChatGPT-4; Gemini 1.5 Pro; Cohere; Llama 3	Turkish Medical Specialty Examination (240 questions)	Human candidate cohort	Accuracy (%)
Intervention studies						
Wang et al. [16]	China	Randomized controlled trial	56 medical students	GPT-4 simulated patient for history-taking training	Traditional instructor-led role-playing	OSCE score
Fazlollahi et al. [17]	Canada	Randomized controlled trial	70 medical students	AI tutoring system for surgical simulation	Remote expert instruction; no-feedback control	ICEMS score
Giglio et al. [18]	Canada	Randomized controlled trial	87 medical students	AI-augmented human instruction	Standardized expert instruction	ICEMS composite score
Li et al. [19]	China	Randomized controlled trial	58 medical students	ChatGPT-assisted academic writing training	Traditional writing instruction	Writing score
Huang et al. [20]	China	Randomized controlled trial	187 dental students	ChatGPT-3.5 as skills education assistant	Video-only learning	VR performance score
Hsu et al. [21]	Taiwan	Controlled trial	60 nursing students	Chatbot-based terminology learning	Traditional self-study	Terminology test score
Lebdai et al. [22]	France	Randomized controlled trial	85 medical students	Immersive virtual patient simulation	Regular faculty courses	Module test score
Ali et al. [23]	Not reported	Randomized controlled trial	Not reported	AI-based tool for OSCE preparation	Traditional OSCE preparation	OSCE performance score

Note: Benchmarking studies evaluated the performance of artificial intelligence models on standardized medical examinations. Intervention studies assessed the effectiveness of AI-based educational tools compared with conventional instructional methods.

Abbreviations: AAP, American Academy of Periodontology; AI, artificial intelligence; CBSSA, Comprehensive Basic Science Self-Assessment; ICEMS, Imperial College Error Capture Score; MCQs, multiple-choice questions; OSCE, Objective Structured Clinical Examination; PGY, postgraduate year; USMLE, United States Medical Licensing Examination; VR, virtual reality.

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

Study selection

The systematic search yielded 130,858 records. After removal of 128,043 duplicates, 2,815 unique records underwent title and abstract screening. Sixty-eight articles were assessed in full text for eligibility. Ultimately, 15 studies satisfied the inclusion criteria and were included in the quantitative synthesis. The PRISMA flow diagram (**Figure 1**) details the study selection process. Meta-Analysis 1 (benchmarking AI performance) comprised seven studies, while Meta-Analysis 2 (evaluating AI interventions) included eight RCTs.

Meta-analysis 1: AI performance on medical knowledge assessments: Across the seven benchmarking studies, 35 accuracy data points were extracted, representing all AI models evaluated across multiple examination formats. A three-level random-effects meta-analysis demonstrated a pooled accuracy of 70.9% (95% CI: 65.1% to 75.9%) (**Figure 3**). A cumulative meta-analysis, ordering studies by publication year, showed progressive refinement of the pooled estimate as evidence accumulated, with the confidence interval narrowing over time (**Appendix 3, Figure S1**).Substantial heterogeneity was observed (total $I^2 = 94.2\%$). Variance decomposition indicated that 55.7% of the heterogeneity arose from between-study differences (Level 3), whereas 38.5% reflected within-study variation across different AI models (Level 2). Sensitivity analysis using robust variance estimation yielded a nearly identical pooled estimate (70.8%, 95% CI: 65.0% to 76.2%), supporting the stability of the findings.



Figure 3. Forest plot of AI performance on medical knowledge assessments (meta-analysis 1)

Exploratory meta-regression identified a significant positive association between publication year and AI performance ($B = 0.072$, $SE = 0.015$, $t(5) = 4.80$, $p = 0.008$), indicating progressively higher reported accuracy in more recent studies **(Appendix 3, Figure S2)**. After accounting for within-study dependence, no significant differences were observed between GPT-4 and other models (Q between $= 0.95$, $p = 0.342$). Risk of bias assessment (**Figure 2**) indicated that five studies (71.4%) had low overall risk, while two (28.6%) were rated as unclear, primarily due to concerns in the "flow and timing" domain related to potential data contamination. Funnel plot inspection (**Appendix 3, Figure S2**) suggested possible asymmetry; however, the small number of studies limited definitive interpretation.

Regarding test contamination, six benchmarking studies (86%) were rated as high risk of bias in the "flow and timing" domain of the QUADAS-2 assessment, and one study was rated as unclear. These judgments reflected the possibility that publicly available examination questions were included in AI training data, potentially inflating accuracy estimates. The single study with unclear risk [7] used in-service examination questions that were not publicly accessible. Although the authors selected recent examination years to reduce the likelihood of data contamination, this measure could not be independently verified, resulting in an unclear risk designation.

Meta-analysis 2: efficacy of AI interventions in medical education: The meta-analysis of eight RCTs comprising 574 participants yielded a pooled effect size

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

of Hedges' g = 1.40 (95% CI: 1.21 to 1.60), with substantial heterogeneity (I^2 = 93.6%). Individual study effects ranged from g = 0.67 to g = 6.72 (**Figure 4**). The study by Fazlollahi et al. [17] was identified as a marked outlier (g = 6.72), potentially reflecting the highly specialized nature of its surgical simulation intervention, which provided real-time, objective feedback metrics not easily replicated in conventional teaching contexts.

Leave-one-out sensitivity analysis demonstrated that the pooled estimate was sensitive to individual studies. Exclusion of Huang et al. [20] increased the pooled effect size to g = 1.95, indicating limited robustness. Removal of other studies resulted in pooled estimates ranging from g = 1.26 to g = 1.45. The leave-one-out analysis is presented in **Appendix 3, Figure S4**. Influence diagnostics confirmed Huang et al. [20] as highly influential and identified it as a statistical outlier. Comprehensive influence metrics, including Cook's distance and DFBETAS, are provided in **Appendix 3, Figure S5**. Risk of bias assessment showed that four studies (50.0%) were rated as low overall risk, three (37.5%) as high risk, and one (12.5%) as unclear, predominantly due to concerns regarding randomization procedures and intervention fidelity (**Figure 5**).

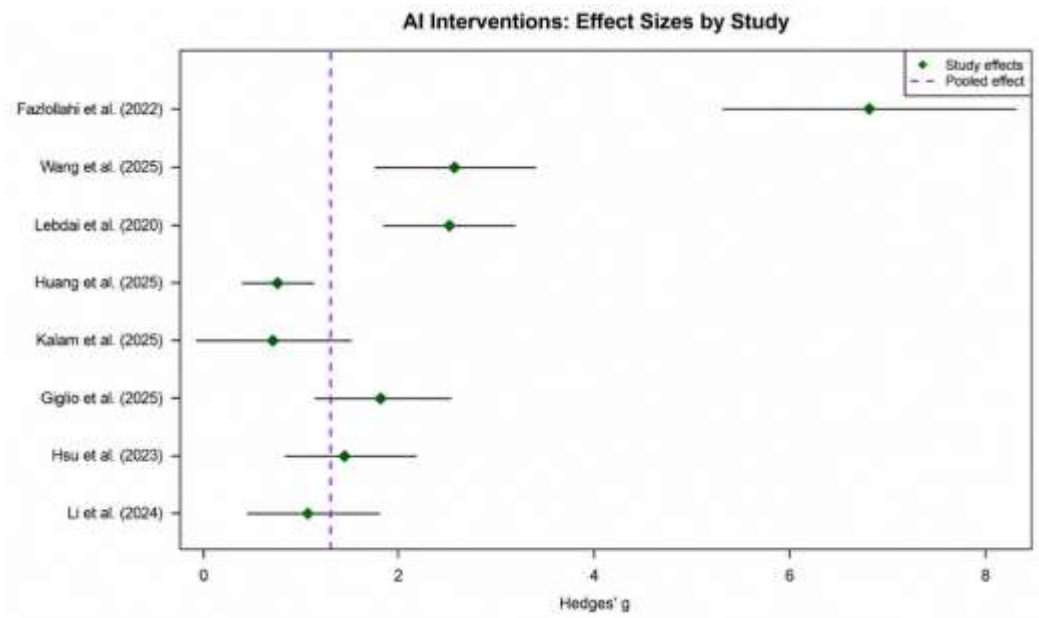


Figure 4. Forest plot of AI-based educational interventions (meta-analysis 2)

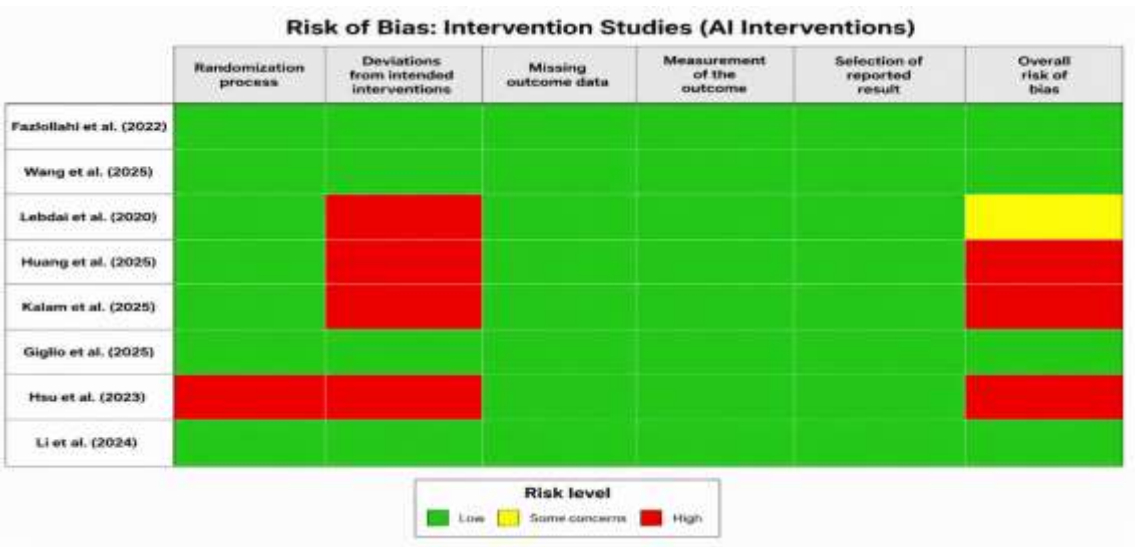


Figure 5. Risk of bias summary for AI-based educational intervention studies

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

Exploratory meta-regression revealed highly significant negative associations between total sample size (Beta = −0.009, SE = 0.000, t(6) = −23.36, p < 0.001) and publication year (Beta = −0.374, SE = 0.017, t(6) = −22.42, p < 0.001) with effect size, accounting for 58.0% and 53.4% of heterogeneity, respectively (**Appendix 3, Figures S6 and S7**). Study quality (Q between = 4.99, p = 0.082) and intervention type (Q between = 9.04, p = 0.107) were not statistically significant moderators. Nonetheless, subgroup analysis (**Appendix 3, Figure S8**) indicated notable variation in effect sizes across intervention types, with simulation-based interventions demonstrating the largest effects.

Although statistically significant, these findings are based on a limited number of studies and should be interpreted as preliminary patterns within a rapidly evolving field rather than as stable, generalizable associations. Funnel plot inspection (**Appendix 3, Figure S9**) revealed asymmetry suggestive of potential publication bias, with smaller studies tending to report larger effect sizes. The precision plot (**Appendix 3, Figure S10**) illustrates the distribution of study-level effect sizes relative to their standard errors, consistent with small-study effects.

Classification of educational outcomes based on Kirkpatrick framework: Classification of primary outcomes from the eight included RCTs according to the Kirkpatrick framework indicated that all studies (100%) assessed Level 2b outcomes (change in knowledge/skills), as each reported objective performance measures such as OSCE scores, test scores, or simulation metrics. Three studies (37.5%) additionally evaluated Level 1 outcomes (reaction), including self-reported motivation, anxiety, and system usability [16, 21, 23]. One study (12.5%) assessed Level 2a outcomes (attitudes/perceptions) through cognitive load measurement [20]. Only one study (12.5%) examined Level 3 outcomes (behavioral change), evaluating skill transfer to a realistic task [18]. No study assessed Level 4 outcomes (organizational practice change or patient benefits). The distribution of Kirkpatrick levels across studies is summarized in **Table 2**.

Table 2. Distribution of Kirkpatrick levels across included intervention studies (n = 8)

Kirkpatrick level	Number of studies (%)	Example outcomes
Level 1 (Reaction)	3 (37.5%)	Self-reported motivation; anxiety; system usability [16, 21, 23]
Level 2a (Attitudes/Perceptions)	1 (12.5%)	Cognitive load (pupil diameter) [20]
Level 2b (Knowledge/Skills)	8 (100%)	OSCE scores; test scores; simulation metrics [16, 23]
Level 3 (Behavioural change)	1 (12.5%)	Skill transfer to realistic task [18]
Level 4 (Results/Patient outcomes)	0 (0%)	None measured

Note: Percentages are calculated based on eight included intervention studies.

Discussion

This systematic review and dual meta-analysis provide a comprehensive quantitative synthesis of the existing evidence regarding the role of LLMs and AI in medical education. By quantitatively integrating early peer-reviewed studies, the review moves the discussion beyond conjecture and establishes preliminary benchmarks for educators, institutions, and policymakers. It also explores variability in the literature through moderator analyses and risk of bias assessment.

Findings from the two meta-analyses offer a detailed and multifaceted picture that both supports and quantifies predictions previously proposed in conceptual and qualitative reviews [12, 13, 15, 31]. AI models demonstrate substantial proficiency in medical knowledge assessment (pooled accuracy 70.9%), and AI-driven educational interventions are associated with significant improvements in learner outcomes (g = 1.40). Nevertheless, these findings must be interpreted in light of considerable heterogeneity and important methodological influences that shape their implications.

The three-level random effects meta-analysis incorporated all available accuracy data points (35 effect sizes from seven studies) while accounting for within-study dependence. The resulting pooled accuracy of 70.9% (95% CI: 65.1% – 75.9%) places contemporary LLM performance above common passing thresholds for many high-stakes medical licensing examinations, directly addressing earlier calls to evaluate this capability rigorously [2, 4, 5].

This result aligns with systematic and scoping reviews describing the “robust performance and reliability” of advanced models and their ability to “replicate human level performance” [12–14]. Substantial variability across studies reflects rapid technological development, and meta-regression confirmed that more recent studies report higher accuracy, quantitatively substantiating the “rapid and significant evolution” of LLMs previously described qualitatively [15, 31]. These findings suggest that current benchmarks may already underestimate present capabilities, underscoring the difficulty of aligning traditional educational timelines with rapidly

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

evolving technologies [1]. Importantly, AI capability is not uniform; both model selection and implementation are critical when educational tasks require complex problem-solving. These results are consistent with recent meta-analyses by Nouri et al. [32] and Jaleel et al. [33], which similarly report that GPT-4 class models exceed passing thresholds. The second meta-analysis demonstrated a substantial aggregated effect size for AI-based educational interventions (g = 1.40). However, careful interpretation is required. Sensitivity analysis showed that removal of a single study could increase the estimate to g = 1.95. In addition, identification of a clear outlier (g = 6.72) and a strong inverse relationship between effect size and sample size indicate that the pooled estimate is unstable and may be influenced by small, early proof-of-concept studies. A more conservative interpretation of the sensitivity range (g = 1.26–1.45) still indicates meaningful potential, yet the evidence remains insufficient to support firm claims of consistently large effects across diverse educational contexts.

Consistent with these observations, one recent meta-analysis reported that AI-based teaching significantly improved practical skills but not knowledge acquisition [34], whereas another found that AI-powered problem- and case-based learning enhanced knowledge and learner satisfaction [37]. The pattern in which early, small studies report very large effects that are not maintained in larger, more rigorous trials is characteristic of a technology bubble effect, where initial enthusiasm and publication bias may elevate expectations that later stabilize as the evidence base expands. Accordingly, the focus shifts from AI capability alone to its pedagogical effectiveness in facilitating learning, as proposed in conceptual frameworks [24, 26]. The two meta-analyses address complementary dimensions: the first evaluates whether AI possesses sufficient knowledge to function as a reliable information source, while the second examines whether this knowledge can be translated into improved learner outcomes. Demonstrating both competence and educational impact suggests that AI may play a meaningful role in medical education, provided that the evidence base continues to mature.

The included RCTs represent a broad range of AI applications, from surgical simulation [17, 18] and clinical skills training [16] to writing assistance [19] and terminology learning [21]. Although this diversity introduces substantial clinical heterogeneity, pooling is conceptually defensible. All interventions employ AI as an active educational tool aimed at improving learner outcomes compared with traditional instruction or non-AI digital resources. The random effects model explicitly accommodates between-study heterogeneity, and subgroup analyses by intervention domain (simulation, clinical skills, knowledge, writing, terminology) demonstrated consistently positive effects across categories, albeit with varying magnitude. In the absence of a pooled estimate, the field would lack a quantitative synthesis. Therefore, the pooled effect size is presented as an average across AI educational interventions, with explicit acknowledgment of heterogeneity and accompanying sensitivity analyses. Subgroup findings should be interpreted as more specific guidance for different categories of AI tools. Substantial heterogeneity was observed in both meta-analyses. For AI performance, this variability likely reflects genuine differences among models, specialties, and examination formats.

In intervention studies, heterogeneity appears largely attributable to methodological and temporal factors, including sample size and publication year, suggesting ongoing maturation of the evidence base. Risk of bias assessment further contextualizes these findings. The pooled accuracy of 70.9% represents an average across models and tasks, while individual performance ranged from 17% to 100% in the included data. Benchmarking studies frequently demonstrated high risk of bias in the ‘flow and timing’ domain, raising concerns about potential data contamination if examination items were included in model training datasets [10]. In intervention studies, a substantial proportion were rated as high risk, primarily due to issues related to randomization and intervention fidelity, indicating that reported effects may be optimistic. Notably, higher-quality studies tended to report more modest effects, reinforcing the need for cautious interpretation.

Application of the Kirkpatrick hierarchy highlights a significant gap in the evidence base. All eight RCTs demonstrated improvements in knowledge and skill acquisition (Level 2b), yet only one study assessed behavioral change (Level 3) [18], and none evaluated organizational or patient-level outcomes (Level 4). This distribution mirrors broader trends in medical education research, where higher-level outcomes are less frequently measured because they require extended follow-up and complex implementation. Nevertheless, the absence of Level 4 outcomes is consequential, given that the ultimate objective of medical education is improvement in patient care. Future trials should therefore incorporate longer follow-up and outcome

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

measures capturing sustained behavioral change and, where feasible, patient-level impact. Feigerlova et al. [34] similarly reported the absence of Level 3 and Level 4 outcomes, confirming the lack of higher-level evidence in this domain.

From a policy perspective, the absence of Level 4 evidence limits evidence-based integration of AI into medical curricula. Institutional leaders require demonstration that investments in AI-driven tools translate into improvements not only in knowledge or skills but also in clinical practice, patient safety, or organizational efficiency. Without such evidence, decisions regarding adoption and scaling rely primarily on intermediate outcomes. This consideration is particularly relevant given the expanding commercial availability of AI educational products and institutional pressure for adoption.

Until studies incorporating longer follow-up and patient- or system-level endpoints are conducted, current findings should be regarded as supportive of innovation but insufficient to justify widespread mandatory implementation. Although subgroup differences did not reach statistical significance, visual inspection of effect sizes suggests a potential pattern. Larger effects were observed in surgical simulation [17, 18] and clinical skills training [16], whereas knowledge-based applications demonstrated more moderate effects. This possible "simulation gap" remains a hypothesis-generating observation rather than a confirmed finding. The comparatively large effect sizes in simulation contexts may reflect measurable differences between traditional training and AI-supported approaches, particularly in domains where scalability and objective feedback are challenging.

In contrast, interventions in which AI functioned primarily as an auxiliary information source yielded smaller incremental gains. These observations indicate that the extent to which AI is directly integrated into the learning feedback process may influence its educational impact.

Publication bias suggests that aggregated effects may overestimate true effectiveness. The "file drawer problem," in which studies reporting null or negative results remain unpublished, may result in inflated estimates of both accuracy and intervention effects. Consequently, reported values should be interpreted as reflecting published evidence rather than definitive performance limits. For assessment and curriculum design, AI's demonstrated ability to achieve passing thresholds necessitates reconsideration of evaluation strategies. Greater emphasis may be required on higher-order cognitive skills, clinical reasoning, empathy, and practical competence—domains in which human judgment remains essential [35, 36]. This includes open-ended clinical reasoning scenarios, ethical dilemma discussions, and direct observation of patient interactions. Although AI interventions show beneficial effects, the most substantial gains appear when AI provides direct, interactive feedback. Successful integration into educational practice depends not only on technological availability but also on faculty preparedness.

Evidence indicates that faculty readiness substantially influences effectiveness [25]. Accordingly, faculty development should address both technical proficiency and pedagogical adaptation to enable effective collaboration with AI tools. This review reflects the early stage of the field it examines. The limited number of studies in each meta-analysis ($k = 7$ for benchmarking; $k = 8$ for interventions) constrains statistical power and the stability of pooled estimates, necessitating cautious interpretation. Although all available data points were included and a three-level model was applied to mitigate selection bias at the review level, primary studies may still be subject to selective reporting, such as publication of favorable prompts, subsets of questions, or optimal model configurations. The limited number of RCTs ($n = 8$) restricts the robustness and generalizability of findings related to AI interventions. While Google Scholar was searched to capture grey literature, reliance on published studies and exclusion of preprints may have contributed to publication bias.

The review protocol was not prospectively registered in PROSPERO, limiting comparison between planned and conducted methods, although all methodological decisions have been explicitly reported to enhance transparency. Given the rapid evolution of AI technologies, these results represent a time-specific snapshot.

Restriction to English-language publications introduces language bias. Studies published in other languages (e.g., Chinese, Spanish, French, German, Turkish, Persian) were excluded, potentially omitting relevant evidence from non-English-speaking contexts. As AI research and medical education are globally active domains, this exclusion may affect external validity and generalizability. English-language publications often originate from well-resourced institutions with greater access to advanced AI models, which may lead to overestimation of performance and effectiveness. Future

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

research should extend beyond capability studies toward evaluation of AI–human collaborative teaching models and long-term effects on clinical reasoning and patient care. Larger, pre-registered, real-world trials are needed to clarify true effect sizes, particularly in light of the observed inverse relationship between effect size and sample size. Ethical and practical frameworks addressing plagiarism, data privacy, and curriculum integration should accompany implementation efforts. Dissemination of null findings is also essential to provide a balanced assessment of AI's role in the evolution of medical education.

Conclusion

This systematic review and dual meta-analysis offers the first quantitative synthesis of AI in medical education, encompassing both model benchmarking studies and learner-centered intervention trials. The pooled findings show that advanced large language models reach a mean accuracy of 70.9% (95% CI: 65.1–75.9%) on standardized medical knowledge assessments, with performance improving consistently across successive model generations. At the same time, the widespread risk of test contamination in benchmarking studies, together with substantial clinical and statistical heterogeneity, necessitates careful interpretation of these results and underscores the need to reconsider assessment strategies, particularly toward competencies that AI cannot replicate.

Concurrently, AI-driven educational interventions demonstrate encouraging effect sizes; however, the supporting evidence remains at an early stage of development. The inverse association between effect size and sample size, along with sensitivity analyses indicating instability of the pooled estimate, suggests that smaller, early-phase studies may overestimate true effectiveness. Among the various applications, simulation-based interventions appear to produce the most consistent benefits, plausibly reflecting AI's ability to deliver scalable and objective feedback.

Future progress in this field will require a shift from demonstrations of capability to well-designed, pre-registered trials capable of generating stable and generalizable effect estimates across diverse educational settings.

Equally important is the establishment of pedagogical frameworks that position AI as a complement to—not a replacement for—human instruction. Through rigorous evidence development and deliberate implementation, AI may serve as a meaningful adjunct in the continuing evolution of medical education.

Ethical considerations

A detailed written protocol was developed prior to data extraction. However, prospective registration in PROSPERO was not undertaken due to the rapidly evolving nature of AI technologies and the need for iterative refinements to maintain the review's currency. The complete protocol, including original eligibility criteria, search strategies, data extraction forms, and statistical analysis plan, is provided in Appendix 4. All post hoc methodological decisions are transparently reported in the manuscript, and the absence of prospective registration is acknowledged as a limitation.

This systematic review and meta-analysis relied solely on publicly available data derived from published peer-reviewed articles and conference proceedings and did not involve any direct interaction with human or animal subjects. Ethical code IR.LARUMS.REC.1404.019 was approved by the Ethics Committee of Larestan University of Medical Sciences, where the study was formally registered as a research project.

The study was conducted in accordance with fundamental principles of scholarly integrity, including a rigorous and unbiased search strategy, transparent reporting consistent with PRISMA guidelines, and accurate synthesis and interpretation of the findings. All extracted data were reported faithfully, and the limitations of the available evidence were clearly stated to minimize the risk of misinterpretation.

Artificial intelligence utilization for article writing

In preparing this manuscript, the authors employed an artificial intelligence (AI) language model through its publicly available interface. Its use was limited to editorial support, specifically for language refinement, grammatical correction, and structural organization of text originally written by the authors. All conceptualization, critical analysis, interpretation of data, and substantive writing were undertaken exclusively by the human authors. The AI contributed no original intellectual content and was not used for data extraction, statistical analysis, or literature synthesis. The completed manuscript was carefully reviewed, verified, and approved by all authors, who assume full responsibility for the scholarly content and overall integrity of the work.

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

Acknowledgment

The authors gratefully acknowledge the invaluable assistance of the medical librarian at Zanjan University of Medical Sciences for expertise in developing and refining the systematic search strategy. We also extend thanks to the researchers who kindly provided additional data or clarification upon request for their published studies included in our meta-analysis.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

SAH was involved in all stages of the project including conceptualization, study design, methodology, data analysis, data interpretation, writing, critical review, and editing, bearing primary responsibility for the manuscript. AR contributed to data collection and curation, and wrote and revised the first draft of the introduction. ZA contributed to study design, methodology, validation, investigation, and critical review.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The data supporting the findings of this systematic review and meta-analysis are derived from published studies cited in the reference list. The full dataset generated for the quantitative synthesis—including extracted numerical data and analysis code—is available from the corresponding author upon reasonable request.

References

1. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. *Acad Med*. 2018;93(8):1107-1109. https://doi.org/10.1097/ACM.0000000000002041
2. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9:e46885. https://doi.org/10.2196/46885
3. Kooli C. Chatbots in education and research: a critical examination of ethical implications and solutions. *Sustainability*. 2023;15(7):5614. https://doi.org/10.3390/su15075614
4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. https://doi.org/10.2196/45312
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
6. 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. *J Educ Eval Health Prof*. 2023;20:1. https://doi.org/10.3352/jeehp.2023.20.1
7. Sabri H, Saleh MH, Hazrati P, Merchant K, Misch J, Kumar PS, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. *J Periodontal Res*. 2025;60(2):121-133. https://doi.org/10.1111/jre.13345
8. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems [Preprint]. arXiv [Internet]. 2023. [cited 2026 May 20]. Available from: https://doi.org/10.48550/arXiv.2303.13375
9. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13(1):16492. https://doi.org/10.1038/s41598-023-43436-9
10. Koçak M, Oğuz AK, Akçalı Z. The role of artificial intelligence in medical education: an evaluation of Large Language Models (LLMs) on the Turkish Medical Specialty Training Entrance Exam. *BMC Med Educ*. 2025;25(1):609. https://doi.org/10.1186/s12909-025-06538-y
11. Waisberg E, Ong J, Masalkhi M, Zaman N, Kamran SA, Sarker P, et al. ChatGPT and medical education: a new frontier for emerging physicians. *Can Med Educ J*. 2023;14(6):128. https://doi.org/10.36834/cmej.78250

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

12. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023;9(1):e48291. https://doi.org/10.2196/48291
13. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. *JMIR Med Educ*. 2023;9:e48163. https://doi.org/10.2196/48163
14. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education—scoping literature review. *Med Sci Educ*. 2025;35(1):555-567. https://doi.org/10.1007/s40670-024-02258-8
15. Lee SH. Biomedical education in the era of large language models: a paradigm shift. *J Clin Invest*. 2025;135(14):e187654. https://doi.org/10.1172/JCI187654
16. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. *BMC Med Educ*. 2025;25(1):1030. https://doi.org/10.1186/s12909-025-06905-5
17. 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 Netw Open*. 2022;5(2):e2149008. https://doi.org/10.1001/jamanetworkopen.2021.49008
18. Giglio B, Albeloushi A, Alhaj AK, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial intelligence–augmented human instruction and surgical simulation performance: a randomized clinical trial. *JAMA Surg*. 2025;160(9):993-1003. https://doi.org/10.1001/jamasurg.2025.2564
19. Li J, Zong H, Wu E, Wu R, Peng Z, Zhao J, et al. Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students—the educational application of ChatGPT. *BMC Med Educ*. 2024;24(1):736. https://doi.org/10.1186/s12909-024-05738-y
20. Huang S, Wen C, Bai X, Li S, Wang S, Wang X, et al. Exploring the application capability of ChatGPT as an instructor in skills education for dental medical students: randomized controlled trial. *J Med Internet Res*. 2025;27:e68538. https://doi.org/10.2196/68538
21. Hsu MH, Chan TM, Yu CS. Termbot: a chatbot-based crossword game for gamified medical terminology learning. *Int J Environ Res Public Health*. 2023;20(5):4185. https://doi.org/10.3390/ijerph20054185
22. Lebdai S, Mauget M, Cousseau P, Granry JC, Martin L. Improving academic performance in medical students using immersive virtual patient simulation: a randomized controlled trial. *J Surg Educ*. 2021;78(2):478-484. https://doi.org/10.1016/j.jsurg.2020.08.032
23. Ali M, Rehman S, Cheema E. Impact of artificial intelligence on the academic performance and test anxiety of pharmacy students in objective structured clinical examination: a randomized controlled trial. *Int J Clin Pharm*. 2025;47(3):1034-41. https://doi.org/10.1007/s11096-025-01887-2
24. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing artificial intelligence training in medical education. *JMIR Med Educ*. 2019;5(2):e16048. https://doi.org/10.2196/16048
25. Uribe SE, Maldupa I, Kavadella A, El Tantawi M, Chaurasia A, Fontana M, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. *Eur J Dent Educ*. 2024;28(4):865-876. https://doi.org/10.1111/eje.13012
26. Safranek CW, Sidamon-Eristoff AE, Gilson A, Chartash D. The role of large language models in medical education: applications and implications. *JMIR Med Educ*. 2023;9:e50945. https://doi.org/10.2196/50945
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. https://doi.org/10.1136/bmj.n71
28. Kirkpatrick DL, Kirkpatrick JD. *Evaluating training programs: the four levels*. 3rd ed. San Francisco: Berrett-Koehler Publishers; 2006.
29. Kirkpatrick JD, Kirkpatrick WK. *Kirkpatrick's four levels of training evaluation*. Alexandria (VA): ATD Press; 2016.
30. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS 2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]

31. Sallam M, Irshaid A, Snygg J, Albadri R, Sallam M. Rapid evolution of large language models in medical education: comparative performance of ChatGPT-3.5, ChatGPT-4, and DeepSeek on medical microbiology MCQs. *Contemp Educ Teach Res*. 2025;6(8):295-309. https://doi.org/10.61360/BoniCETR252018770801
32. Nouri H, Mahdavi A, Abedi A, Mohammadnia A, Hamedan M, Amanzadeh M. Performance of large language models in medical licensing examinations: a systematic review and meta analysis. *J Educ Eval Health Prof.* 2025;22:36. https://doi.org/10.3352/jeehp.2025.22.1
33. Jaleel A, Aziz U, Farid G, Bashir MZ, Mirza TR, Abbas SM, et al. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. *JMIR Med Educ.* 2025;11(1):e68070. https://doi.org/10.2196/68070
34. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ.* 2025;25(1):129. https://doi.org/10.1186/s12909-025-06719-5
35. Ellaway R, Masters K. AMEE Guide 32: e-learning in medical education Part 1: Learning, teaching and assessment. *Med Teach.* 2008;30(5):455-473. https://doi.org/10.1080/01421590802108331
36. Torous J, Greenberg W. Large language models and artificial intelligence in psychiatry medical education: augmenting but not replacing best practices. *Acad Psychiatry.* 2025;49(1):22-24. https://doi.org/10.1007/s40596-024-02088-5
37. Wei H, Dai Y, Yuan K, Li KY, Hung KF, Hu EM, et al. AI-Powered Problem- and Case-based Learning in Medical and Dental Education: A Systematic Review and Meta-analysis. *Int Dent J*. 2025;75(4):100858. https://doi.org/10.1016/j.identj.2025.100858

[Downloaded from edujournal.zums.ac.ir on 2026-10-31]

[DOI: 10.66224/edcj.19.3.98]