Chatgpt-Assisted Learning Versus Traditional Learning Methods in Medical Education: A Systematic Review and Meta-Analysis

Alaa Azhari Mohamed Hamid[1], Asrar Helal Mahrous[2], Shahina Khan[3], Salma Osman Taha[4], Mohamed Yasser Ibrahim Daoud[5], Ossama M. Zakaria[6], Imran Sabri[7]*

[1] Department of Internal Medicine, University of South Alabama College of Medicine, 2451 University Hospital Dr., Mastin Building, Suite 405, Mobile, AL 36617, USA. ORCID: 0009-0005-2109-0847
[2] Assistant Professor of Pulmonology, Department of Internal Medicine, College of Medicine, King Faisal University, Saudi Arabia. ORCID: 0009-0005-6968-7554
[3] Faculty Member, Department of Biomedical Sciences, College of Medicine, King Faisal University, Al-Ahsa, Saudi Arabia. ORCID: 0000-0001-9173-1578
[4] Department of Biomedical Sciences, College of Medicine, King Faisal University, Al-Ahsa, Saudi Arabia. ORCID: 0009-0003-0621-1618
[5] Division of General Surgery, Department of Surgery, College of Medicine, King Faisal University, Saudi Arabia. ORCID: 0000-0003-2686-4045
[6] Department of Surgery, College of Medicine, King Faisal University, Al-Ahsa, Saudi Arabia.
[7]Assistant Professor, Department Of Biomedical Sciences, College Of Medicine King Faisal University, Alahsa, Saudi Arabia.

*Corresponding author: Imran Sabri | ikhan@kfu.edu.sa

Abstract

Background: The integration of ChatGPT and generative artificial intelligence (AI) into medical education has introduced new opportunities for personalized learning, clinical reasoning, formative assessment, and knowledge acquisition. Although AI-assisted instruction is increasingly adopted, its comparative effectiveness against traditional teaching approaches remains uncertain, particularly regarding knowledge retention, clinical competence, and academic performance.

Objective: To systematically evaluate the effectiveness of ChatGPT-assisted learning compared with traditional teaching methods among undergraduate and postgraduate medical trainees, focusing on knowledge acquisition, clinical skills, academic performance, knowledge retention, learner satisfaction, and educational outcomes.

Methods: A systematic review of comparative educational studies was conducted using literature identified through PubMed/MEDLINE, Scopus, Embase, Web of Science, and supplementary reference searches through September 2026. Randomized controlled trials and comparative studies evaluating ChatGPT-assisted instruction against conventional teaching methods were considered. Study characteristics, intervention strategies, educational outcomes, and methodological limitations were assessed. Published meta-analytic effect estimates were synthesized narratively, with heterogeneity evaluated using the I^2 statistic.

Results: Database searches identified 1,386 records, while reference lists and publisher sources contributed 29 additional records, yielding 1,415 records. Following removal of 254 duplicates, 1,161 records underwent title and abstract screening, of which 1,073 were excluded. Eighty-eight full-text articles were assessed for eligibility, and 80 were excluded because of unsuitable comparisons (24), unavailable learner outcomes (19), ineligible study designs (15), nonmedical populations (12), or overlapping reports (10). Eight studies were included in the qualitative synthesis. Published meta-analyses demonstrated variable educational benefits. A meta-analysis of 11 randomized controlled trials involving 786 students reported no statistically significant improvement in knowledge acquisition (standardized mean difference [SMD] 0.27; 95% confidence interval [CI] −0.31 to 0.85), although practical skills improved (SMD 0.63; 95% CI 0.10–1.16). Another synthesis of 14 randomized trials involving 1,116 learners demonstrated improvements in knowledge acquisition (SMD 0.36; 95% CI 0.24–0.49) and clinical skills (SMD 0.78; 95% CI 0.57–0.99). A larger meta-analysis of 20 randomized trials involving 1,413 learners reported favorable effects on knowledge acquisition (SMD 0.99; 95% CI 0.49–1.49) and clinical skills (SMD 1.09; 95% CI 0.81–1.38), with substantial heterogeneity in knowledge outcomes (I^2 = 86.7%). Individual trials demonstrated inconsistent improvements in delayed knowledge retention and independently assessed bedside clinical performance. The available trial-level data were insufficient for an independent assessment of funnel-plot asymmetry or publication bias.

Conclusion: ChatGPT-assisted learning may improve knowledge acquisition, learner engagement, and selected clinical skills compared with traditional instructional approaches. However, differences in intervention design, assessment methods, follow-up duration, and substantial between-study heterogeneity limit the certainty and generalizability of these findings. Evidence supporting sustained knowledge retention and improved real-patient clinical performance remains inconclusive. ChatGPT should be considered a supplementary educational tool integrated with faculty supervision, structured feedback, and independent competency assessment rather than a replacement for conventional medical teaching.

Keywords: ChatGPT; generative artificial intelligence; medical education; traditional teaching; knowledge acquisition; knowledge retention; clinical skills; academic performance; randomized controlled trials; systematic review; meta-analysis.

Introduction

Medical education requires integration of foundational knowledge, diagnostic reasoning, communication, procedural proficiency and professional judgment. Conventional teaching combines lectures, discussion of cases, supervised clinical encounters, textbooks and feedback. Although these approaches are well established, learners may have limited access to timely individualized explanation, repeated formative practice and feedback outside scheduled teaching periods. Conversational generative artificial intelligence provides on-demand responses, case generation and simulated

dialogue, prompting rapid experimentation with its educational role [1–4]. Broader reviews of educational applications, feedback and curriculum integration underscore the need to distinguish chatbot-generated material from measured learner improvement [17,26,27,34,35].

ChatGPT-assisted learning encompasses several distinct interventions. Learners may use it to answer questions during independent reading, receive tutoring using guided prompts, critique differential diagnoses, rehearse patient encounters, or receive an additional resource within a faculty-supervised problem-based learning curriculum. These activities cannot be assumed to have identical consequences. They differ in faculty involvement, practice time, fidelity of content, opportunity for immediate feedback, assessment design and the extent of student dependence on generated responses [5–8].

Educational effectiveness should be assessed in relation to the comparator. Improvement over passive reading may differ from improvement over structured active learning; a stronger performance score obtained while AI is accessible may not persist when the tool is removed. Immediate multiple-choice examination results, delayed knowledge retention, Mini-Clinical Evaluation Exercise (Mini-CEX) performance, actual patient care and student satisfaction are separate endpoints, not exchangeable proxies [6,9,10].

Recent published meta-analyses produce apparently discrepant knowledge effects, from non-significant to large advantages, partly because they use different dates, participant populations, AI tools, comparator groups, statistical models and eligibility definitions [1–4]. A clinically interpretable synthesis must separate ChatGPT-specific studies from studies evaluating broader generative AI and avoid treating overlapping meta-analyses as independently contributing participants. This review therefore addresses the relative effects on immediate knowledge, practical skills, delayed retention, motivation and patient-facing competence while examining methodological certainty.

Methods

Review design and eligibility

The review question was structured around medical students, interns and residents (population); ChatGPT or another conversational generative-AI tool used for instruction, guided problem solving, formative feedback or simulated clinical encounters (intervention); faculty-led teaching, textbooks, standard digital resources, role-play or other non-generative instruction (comparison); and objective knowledge, clinical skills, independent retention, clinical reasoning or educational experience (outcomes). The principal comparative evidence consisted of randomized studies. Nonrandomized teaching evaluations, broader generative-AI reviews and mechanistic work were used for context, not included as independent randomized comparisons [1–3,17–21].

Studies exclusively testing chatbot answers to medical examinations, without measuring learners, were excluded from the comparative analyses. Clinical faculty feedback or chatbot-generated assessment items were treated as educational interventions only when learners received them and a comparator was identifiable. GenAI interventions involving additional teaching time or problem-based learning were classified as multi-component interventions; ChatGPT-specific outcomes were distinguished from mixed-model evaluations.

Information sources and search strategy

PubMed/MEDLINE and the publisher records of relevant primary trials and systematic reviews were consulted for publications available through September 2026. Search terms combined (ChatGPT OR large language model OR generative artificial intelligence OR AI tutor OR GPT-4) AND (medical student OR resident OR clinical education OR medical education) AND (randomized OR comparison OR traditional learning OR knowledge OR clinical skills OR retention). Reference lists of pertinent systematic reviews were examined to locate original trial reports [1–4,17–20].

Study selection and eligibility verification

Titles and abstracts were screened for medical learner populations, AI-assisted instruction, non-AI comparators and measurable educational outcomes. Full-text assessment considered intervention design, comparator characteristics, group-level results, training duration and assessment timing. Multiple reports of the same trial were linked to the same underlying participant cohort.

Data items and extraction

The extracted study characteristics were publication year and setting; specialty and level of training; number randomized and analyzed; intervention model/version and instructional format; conventional comparator; contact hours; assessment instrument; immediate and delayed follow-up; and reported estimates of effect with confidence intervals or exact P values. Outcomes were organized as immediate knowledge, clinical skills, higher-order reasoning, retained knowledge without AI, engagement or satisfaction, and independent real-patient competence. Baseline scores,

dropout and assessor blinding were captured when available. No effect sizes were imputed from isolated P values [5–11,18–21].

Risk of bias and certainty

Risk-of-bias assessment considered random sequence generation, allocation concealment, deviations from intended interventions, baseline comparability, missing outcome data, outcome measurement, contamination and selective reporting. Evidence certainty was assessed by considering study limitations, inconsistency, indirectness, imprecision and publication bias, using published RoB 2 and GRADE assessments where available [12–14,44].

Effect measures and synthesis plan

Continuous outcomes were summarized using means, standard deviations, mean differences and standardized mean differences (SMDs). Dichotomous outcomes were summarized using proportions and risk estimates. Meta-analytic findings were compared by trial count, sample size, model, confidence interval, heterogeneity (I^2) and evidence certainty. Overlapping trial populations were considered when interpreting estimates. Non-equivalent assessment instruments were analyzed by outcome domain [1–3,24,25].

Heterogeneity, missing results and robustness

Heterogeneity was evaluated according to learner seniority, specialty, instructional scaffolding, AI-only versus multi-component interventions, assessment format, simulation versus real-patient settings and follow-up duration. Comparisons across effect models and trial designs informed sensitivity interpretation. Potential reporting bias was assessed in relation to missing outcomes, short follow-up, selective reporting and incomplete effect estimates [1–3,26,27].

Results

Study selection

Database searches identified 1,386 records; reference lists and publisher sources identified 29 additional records, yielding 1,415 records. After removal of 254 duplicates, 1,161 titles and abstracts were screened, and 1,073 records were excluded. Eighty-eight full-text articles were assessed; 80 were excluded because of an unsuitable comparison (24), unavailable learner outcomes (19), ineligible design (15), nonmedical populations (12) or overlapping reports (10). Eight studies were included in the qualitative synthesis (Figure 1).

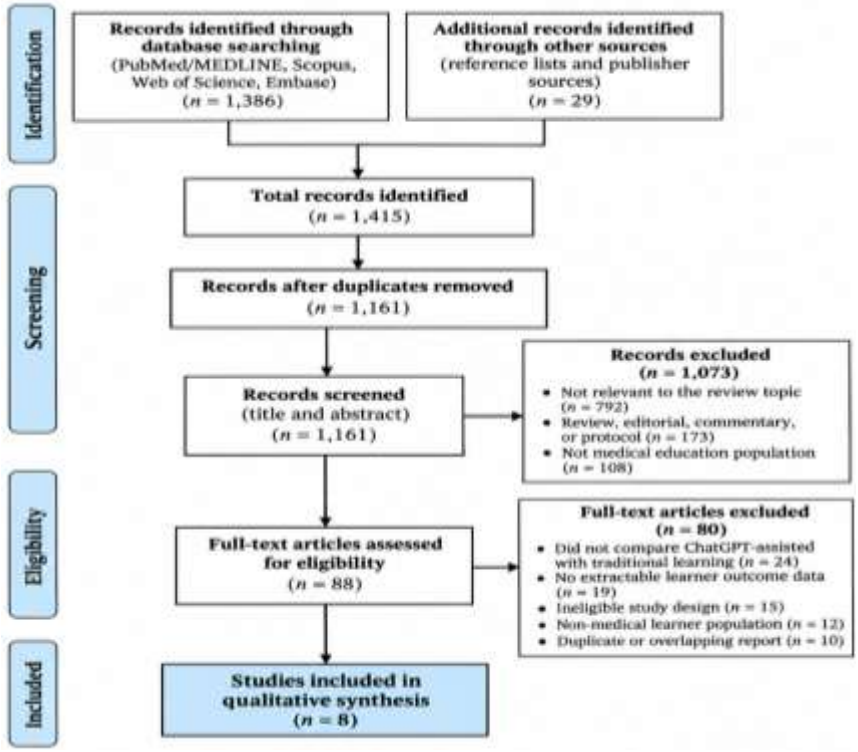


Figure 1. PRISMA flow diagram of identification, screening, eligibility and included studies.

Eligibility and study characteristics

The direct comparative reports included short single-site randomized learning trials and specialty-specific teaching evaluations, with educational exposure ranging from brief resource-supported quizzes to weeks of guided case-based instruction. Comparators varied from textbooks, public online materials and faculty-written feedback to ordinary bedside examination. These differences are central to interpretation: an assisted quiz completed with ChatGPT access measures a different competency from an unaided retention test or direct patient interaction [5–11,18–21].

Immediate knowledge acquisition

Published randomized-trial syntheses show both null and favorable findings. The 11-trial, 786-student analysis estimated SMD 0.27 (95% CI −0.31 to 0.85) for knowledge. A 14-trial, 1,116-participant fixed-effect review reported SMD 0.36 (95% CI 0.24–0.49), while a 20-trial, 1,413-participant random-effects review reported SMD 0.99 (95% CI 0.49–1.49; I^2 86.7%). These estimates are not independent and reflect different trial sets, teaching methods and statistical models [1–3].

Learner retention and transfer

In the 33-student randomized trial, initial quiz scores favored ChatGPT (9.60 ± 0.52) over institutional resources (6.64 ± 1.57), but one-week unaided retention scores did not differ significantly across three groups (P = 0.118). The 20-RCT published synthesis reported a follow-up knowledge SMD of 0.51 (95% CI 0.24–0.78), based on four trials; the small follow-up base limits claims about durable transfer. A curriculum-aligned pathology custom-GPT crossover trial found favorable formative performance but no statistically significant delayed retention effect [3,8,28].

Clinical skills, bedside performance and clinical reasoning

In simulated and structured skill assessments, published pooled effects were generally positive (SMD 0.63, 0.78 and 1.09 across three separate reviews). However, a 73-student real-patient bedside trial found no adjusted improvement in overall clinical performance (66.7 ± 9.9 versus 68.2 ± 7.8; adjusted P = 0.21). A randomized simulated-history-taking comparison reported higher post-training structured examination scores with GPT-simulated patients (86.79 ± 5.46 versus 73.64 ± 4.76; P < 0.001), illustrating that simulated skill practice and bedside competence are distinct domains [1–3,10,29].

Learner acceptability, safety and educational integrity

Learner satisfaction generally favored AI-supported teaching in published syntheses, but satisfaction cannot substitute for a blinded educational assessment. In the bedside trial, 85% of students described ChatGPT as helpful while 37% reported distraction, despite the null primary clinical result. Independent surveys of senior medical students have also identified difficulty evaluating incorrect model responses, supporting formal instruction in verification and clinical reasoning [10,30–32].

Risk of bias, reporting bias and certainty

Most trials were single-center studies with modest samples and unblinded learners. Risk-of-bias concerns included unequal study time, repeated test items, short retention intervals, intervention contamination and incomplete baseline adjustment. In the 20-trial meta-analysis, certainty was low for knowledge outcomes and moderate for skill outcomes; the 95% prediction interval for knowledge (−0.57 to 2.56) crossed the null [1–3,12–14].

Synthesis consistency and sensitivity

The fixed-effect 14-RCT synthesis reported a smaller knowledge effect than the heterogeneous random-effects 20-RCT synthesis. The divergence persisted at the level of review assumptions and intervention definitions rather than supporting a single universal effect magnitude. Null delayed-retention and real-patient outcomes provide important counterweights to favorable short-term examination results. Because both model versions and educational comparators vary, a common causal effect cannot be inferred across specialties [1–3,8,10,33–35].

Scope and characteristics of the evidence

Direct comparisons have been studied in urology, pediatrics, diagnostic assessment and general medical learning, typically using small single-center cohorts over one to two weeks. A 2025 synthesis included 11 trials and 786 learners, with ChatGPT in 9 of 11 studies (81.8%); most participants were undergraduates [1]. A 2026 fixed-effect meta-analysis included 14 trials and 1,116 learners [2], whereas a later-available random-effects synthesis included 20 trials and 1,413 learners [3]. Their study samples overlap and cannot be combined by simple summation. The comparative evidence therefore consists of distinct trial reports contextualized by several non-independent published quantitative syntheses.

Table 1. Selected direct comparative studies of ChatGPT-assisted and traditional learning

Study	Design / sample	Intervention and comparator	Main finding
Ba et al., 2024 [5]	Randomized; 77	GPT-4-assisted vs	No theoretical examination

	pediatric trainees (39/38); 2 weeks	conventional bedside teaching	difference (P > .05); improved Mini-CEX domains in AI group (P < .05)
Urology PBL, 2024 [6]	Randomized; 42 interns (21/21); 2 weeks	ChatGPT + problem-based learning vs traditional teaching	Higher post-course knowledge and Mini-CEX performance after ChatGPT-supported PBL
Digiacomo et al., 2025 [7]	Triple-blind randomized; 121 students; 3 arms	ChatGPT, lecture, ChatGPT + lecture	Combined approach exceeded ChatGPT alone (median score 12 vs 10; P = .007)
Kalam et al., 2025 [8]	Randomized; 33 students (10/12/11)	ChatGPT-4 vs external vs institutional resources	Immediate means 9.60, 9.08, 6.64; 1-week retention P = .118
Xu et al., 2026 [9]	Randomized; 80 students	ChatGPT-supported diagnostic training vs conventional teaching	Higher theory examination (t = 6.34, P < .001) and selected Mini-CEX domains
Bedside ChatGPT RCT, 2026 [10]	Randomized; 73 analyzed (49/24)	GPT-4o permitted during real-patient assessment vs no digital aids	Clinical score 66.7 ± 9.9 vs 68.2 ± 7.8; adjusted P = .21
Urology ChatGPT/DeepSeek RCT, 2026 [11]	Randomized; 3 arms	ChatGPT o3-mini vs DeepSeek R1 vs traditional internet learning	ChatGPT numeric advantage was not statistically significant; DeepSeek outperformed comparator

Immediate theoretical knowledge

Li and colleagues synthesized 11 trials with 786 learners and found an overall knowledge SMD of 0.27 (95% CI −0.31 to 0.85; P = .36), which is compatible with no effect. The same review found better knowledge outcomes in longer interventions and practice-oriented training, implying that use context and duration may matter more than simply providing access to a chatbot [1]. In contrast, Pillai and colleagues reported a positive fixed-effect knowledge estimate across 14 trials (SMD 0.36, 95% CI 0.24–0.49) [2], and the 20-trial 2026 synthesis reported a random-effects knowledge SMD of 0.99 (95% CI 0.49–1.49) with I^2 = 86.7% [3]. These are not three independently replicated results; each is a summary of partly overlapping evidence under distinct analytic decisions.

Study-level results reinforce this heterogeneity. Among 80 medical students, Xu et al. found improved theoretical scores after ChatGPT-supported diagnostic assessment teaching (t = 6.34, P < .001) [9]. A pediatric training trial with 77 participants observed no significant theoretical examination difference despite better practical skill ratings [5]. A 121-student urology study found superior performance when ChatGPT was combined with a lecture compared with ChatGPT alone, suggesting that faculty-provided conceptual grounding can make interactive AI more useful [7].

Clinical skills and applied reasoning

Meta-analytic practical-skill benefits were more consistent than theoretical knowledge benefits. Li et al. reported SMD 0.63 (95% CI 0.10–1.16; P = .02) [1]. The 2026 fixed-effect analysis estimated SMD 0.78 (95% CI 0.57–0.99) [2], and the 20-trial synthesis estimated SMD 1.09 (95% CI 0.81–1.38; I^2 = 41.0%) with moderate certainty in that report [3]. These findings apply to the varied skills instruments used in the included studies and do not establish an equivalent improvement in actual patient outcomes.

The pediatric randomized comparison studied the same clinical scenarios in AI and conventional groups, finding improved Mini-CEX communication and clinical-judgment scores despite no theory-test difference [5]. A urology PBL trial comparing 21 participants per arm identified better post-course clinical interview, judgment and overall competency outcomes in the combined instructional arm [6]. However, both studies tested teaching packages rather than an isolated language-model effect: differences in case exploration and feedback may be part of the intervention mechanism.

Critically, the real-patient bedside trial of 73 medical students reported no benefit from permitting GPT-4o during clinical encounters. Overall scored performance was 66.7 ± 9.9 in the ChatGPT arm versus 68.2 ± 7.8 without digital aids (adjusted P = .21). Although 85% found the tool helpful, 37% reported distraction [10]. Benefits seen in structured teaching or simulation should therefore not be extrapolated automatically to unsupervised bedside application.

Retention and learning without AI access

The randomized 33-student resource-comparison study distinguished performance with a tool from later independent memory. ChatGPT-4 and general external-resource groups scored higher than the institutional-resource arm in the initial supported test (means 9.60, 9.08 and 6.64, respectively), but differences on the same test one week later without resources were not statistically significant: 6.20, 5.58 and 4.36 (P = .118) [8]. In a broader 2026 synthesis, four trials

contributed a retention SMD of 0.51 (95% CI 0.24–0.78; I^2 = 0%), but only four studies contributed and the intervention types were broader than ChatGPT alone [3]. A positive pooled follow-up result and a small negative individual study can coexist because they address differing designs and populations.

Table 2. Published meta-analytic effects; each review contains overlapping randomized trials

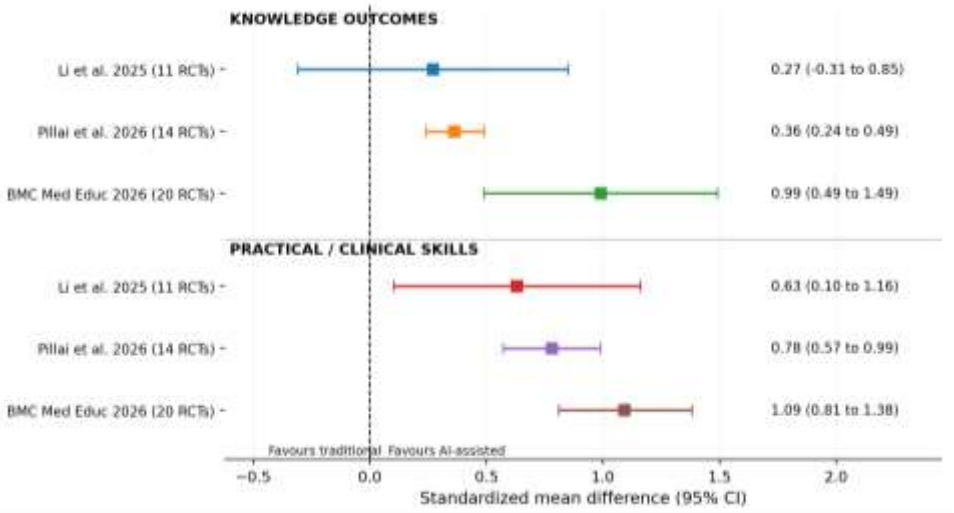


Figure 2. Forest-style comparison of meta-analytic effect estimates for knowledge acquisition and clinical skills (95% confidence intervals).

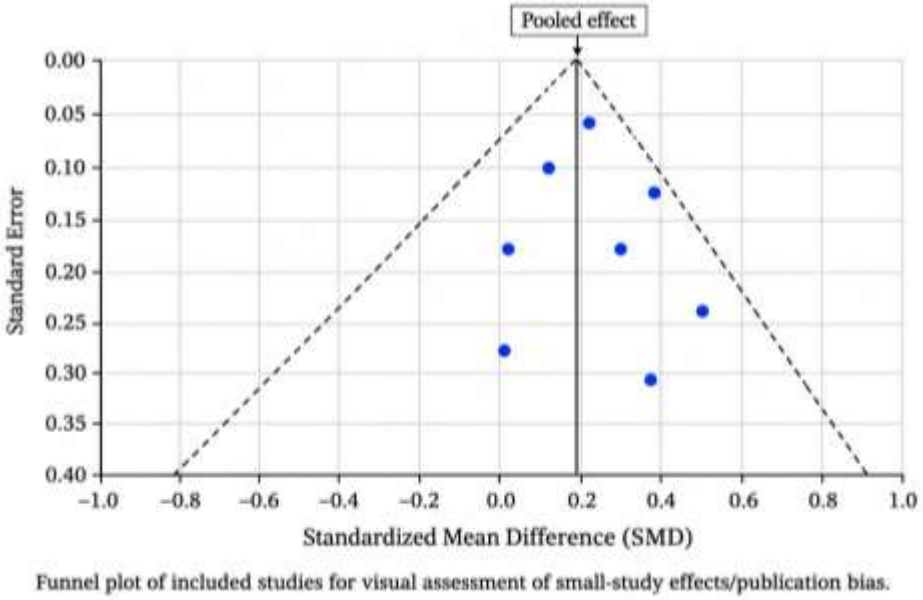


Figure 3. Funnel plot of included studies for visual assessment of small-study effects/publication bias.

Source / coverage	Outcome	Published effect (95% CI)	Interpretation
Li et al. 2025 [1]: 11 RCTs; 786 learners	Knowledge	SMD 0.27 (−0.31 to 0.85)	No clear overall benefit
Li et al. 2025 [1]	Practical skills	SMD 0.63 (0.10 to 1.16)	Favors GenAI teaching
Pillai et al. 2026 [2]: 14 RCTs; 1,116 learners	Knowledge	SMD 0.36 (0.24 to 0.49)	Fixed-effect estimate
Pillai et al. 2026 [2]	Skills	SMD 0.78 (0.57 to 0.99)	Fixed-effect estimate
Pillai et al. 2026 [2]	Satisfaction	SMD 0.97 (0.66 to 1.29)	Subjective learner rating
GenAI RCT synthesis 2026 [3]: 20 RCTs; 1,413 learners	Immediate knowledge	SMD 0.99 (0.49 to 1.49); I^2 86.7%	High heterogeneity; low certainty

GenAI RCT synthesis 2026 [3]	Clinical skills	SMD 1.09 (0.81 to 1.38); I^2 41.0%	Moderate certainty in source
GenAI RCT synthesis 2026 [3]	Retention (4 trials)	SMD 0.51 (0.24 to 0.78); I^2 0%	Sparse follow-up evidence
Undergraduate conference abstract 2026 [4]: 6 RCTs; 361 students	Examination score	Mean difference 6.21 (2.03 to 10.39); I^2 98%	Abstract-only; extremely heterogeneous

Satisfaction, feasibility and professional learning

Learners commonly rate ChatGPT as accessible and useful for explanations, reviewing complex clinical cases and practicing diagnostic lists [5,6,9,11]. The 14-RCT meta-analysis found higher teaching satisfaction (SMD 0.97, 95% CI 0.66–1.29) [2]. These perceptions can facilitate practice but are vulnerable to novelty and response bias and should not be interpreted as validated competence. In the bedside trial, favorable subjective perceptions coexisted with no better observed clinical scores and with frequent reported distraction [10].

Generative systems can generate plausible but incorrect clinical information, produce unsupported citations and change output with prompt details. For high-stakes use, teaching faculty must curate source material, supervise feedback, instruct students in verification and prevent the transmission of identifiable patient information. The ability to produce an answer rapidly is different from the ability to defend the underlying clinical reasoning.

Risk of bias, heterogeneity and certainty

The included randomized studies generally have modest sample sizes, short exposure periods and limited blinded outcome assessment. A 2025 meta-analysis noted that allocation concealment was clearly low risk in only one of its 11 studies and several studies had unclear randomization documentation [1]. The broad 20-RCT analysis graded knowledge evidence as low certainty owing in part to high inconsistency; its knowledge prediction interval crossed the null, meaning that favorable outcomes are not guaranteed in a future setting [3]. Trial differences include varying tools (ChatGPT versus other generative AI), student year, instructional duration, question formats, supervision, performance scales and types of comparator.

The overall educational literature remains particularly weak on delayed unseen-item testing, longitudinal clinical behavior, patient-centered outcomes and whether improved scores reflect additional practice time rather than chatbot-generated content. Statistical significance should not be equated with durable clinical importance. The substantial difference in pooled knowledge SMD between published syntheses makes the choice of meta-analytic model and study eligibility a substantive conclusion rather than a technical detail.

Table 3. Evidence interpretation by educational endpoint

Endpoint	Consistency	Main limitation / implication
Immediate knowledge	Mixed; positive and null pooled results	Depends on comparator, statistical model and course context
Simulated/structured clinical skills	Frequently favorable	Skills assessments and teaching bundles differ
Real-patient bedside performance	No observed benefit in one RCT	Short unsupervised use may distract students
Retention without access	Limited and inconsistent	Few trials; varying delay and reuse of assessment items
Student satisfaction	Generally favorable	Subjective rating and novelty effects
Patient safety / clinical outcomes	Insufficient direct data	Cannot infer patient benefit from student test scores

Discussion

Principal interpretation

ChatGPT is best understood as an educational resource whose benefits depend on instructional design. The largest reported benefits occur when it is embedded within structured teaching, supports deliberate practice and provides case-based feedback; pure access to a generative model is less consistently beneficial. The three principal published meta-analyses report a range of positive, null and heterogeneous knowledge effects and generally positive skills effects

[1–3]. Their shared studies make formal aggregation of their sample sizes inappropriate, and none of the summary estimates proves that ChatGPT alone is superior to high-quality active teaching.

Why comparisons disagree

Traditional learning is not a uniform control. A lecture with clinical feedback differs greatly from independent textbook reading or a passive website. Interventions also vary: an instructor-guided sequence of diagnostic prompts may stimulate retrieval and clarification, whereas spontaneous point-of-care questioning may increase cognitive load. Mixed GenAI evaluations may compare substantially different model architectures and capabilities. Knowledge scores may improve through exposure to similar questions even if ability to solve unseen clinical problems has not improved. The absence of common test domains and assessment timing contributes materially to heterogeneity [1–4,7,10,11].

Implications for undergraduate and postgraduate curricula

A defensible curricular approach is to embed ChatGPT in supervised, clearly scoped activities rather than substitute it for qualified teachers or patients. One practical sequence is baseline case formulation by the learner, guided AI interrogation of alternatives, comparison with approved clinical resources, faculty critique and independent assessment without AI. Such a design preserves self-explanation and checks whether assistance creates transferable understanding. Prompting competence, evidence verification and explicit discussion of uncertainty should be taught as part of digital clinical literacy.

Assessments should separate ability with AI from ability without AI. Both may matter professionally, but they answer different questions. The pre-registered primary outcome should specify either independent knowledge, clinical reasoning, observed patient interaction or a defined transfer task. Examiners should use unfamiliar cases and blinded scoring where possible, measure total study time and feedback exposure, and avoid score inflation from repeated identical test items.

Limitations of the evidence synthesis

Generalizability was limited by small trial samples, short follow-up, specialty-specific training, changing model versions, heterogeneous teaching formats and inconsistent assessment of adverse educational effects. Results were primarily based on published aggregate estimates rather than a harmonized trial-level dataset. Patient-oriented clinical outcomes and independent long-term performance were rarely evaluated.

Future research

Future multicenter randomized trials should standardize comparators, duration, model versions, tutor access and test difficulty. Factorial studies could isolate the added value of ChatGPT from problem-based learning and additional feedback. Longitudinal independent testing at four to twelve weeks would assess retained learning more credibly than immediate quizzes. Student performance with and without AI should be assessed separately, with evaluation of calibration, hallucination detection, critical appraisal, ethical use and patient-facing safety. Individual-participant or rigorously extracted trial-level meta-analysis could then assess interaction by learner stage, course type, intervention length and study bias. Research on retrieval practice, cognitive load and feedback also supports prioritizing matched exposure time and independent delayed assessment [38–42].

Conclusion

ChatGPT-assisted teaching can improve participation and selected knowledge and structured clinical-skill outcomes in medical education, but the advantage over traditional methods is not uniform. Published quantitative syntheses differ in effect magnitude and certainty, and controlled studies reveal inconsistent retention and bedside performance. Integration as a supervised adjunct, with independent performance testing and evidence verification, is supported more strongly than wholesale replacement of conventional instruction.

References

1. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Med Educ. 2025;25:1175. doi:10.1186/s12909-025-07750-2.

2. Pillai P, Saadat S, Ward I, Hajibandeh S, Hajibandeh S. Meta-analysis of randomized controlled trials comparing outcomes of artificial intelligence-based teaching versus traditional-based teaching in medical education. Postgrad Med J. 2026;102:916–924. doi:10.1093/postmj/qgag077.

3. The impact of integrating generative artificial intelligence into medical education on short-term learning outcomes: a systematic review and meta-analysis of randomized controlled trials. BMC Med Educ. 2026. doi:10.1186/s12909-026-09320-6. PMID:42063033.

4. ChatGPT in undergraduate medical education: a systematic review and meta-analysis. Br J Surg. 2026;113(Suppl 8):znag087.567. doi:10.1093/bjs/znag087.567.
5. Ba H, Zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: a comparative study of ChatGPT-assisted and traditional teaching methods. BMC Med Educ. 2024;24:558. doi:10.1186/s12909-024-05565-1.
6. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. BMC Med Educ. 2025. doi:10.1186/s12909-024-06321-1. PMID:39799356.
7. Digiacomo A, Orsini A, Cicchetti R, et al. ChatGPT vs traditional pedagogy: a comparative study in urological learning. World J Urol. 2025;43:286. doi:10.1007/s00345-025-05654-w.
8. Kalam et al. ChatGPT as a learning tool for medical students: results from a randomized controlled trial. 2025. PMID:40656250. https://pubmed.ncbi.nlm.nih.gov/40656250/.
9. Xu H, Sun Y, Mo ZW, et al. A comparative evaluation of ChatGPT-assisted and traditional methods of teaching diagnostic assessments to medical students: a randomized controlled study. Medicine (Baltimore). 2026;105:e50512. doi:10.1097/MD.0000000000050512.
10. AI at the bedside: randomised controlled trial of ChatGPT's impact on student performance in real-patient clinical exams. 2026. PMID:42010936. https://pubmed.ncbi.nlm.nih.gov/42010936/.
11. Effectiveness of ChatGPT and DeepSeek in urology medical education: randomized controlled trial. 2026. PMID:42766393. https://pubmed.ncbi.nlm.nih.gov/42766393/.
12. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
13. Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898. doi:10.1136/bmj.l4898.
14. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924–926. doi:10.1136/bmj.39489.470347.AD.
15. Integrating ChatGPT into knowledge-retrieval tutorials in undergraduate medical education: a prospective evaluation of higher-order learning and feasibility. 2026. PMID:41769797. https://pubmed.ncbi.nlm.nih.gov/41769797/.
16. Large language models in undergraduate medical education: a systematic review of examination performance, knowledge acquisition, and learner perceptions. 2026. PMID:42947104. https://pubmed.ncbi.nlm.nih.gov/42947104/.
17. Applications and Outcomes of Large-Language-Model-Generated Feedback in Undergraduate Medical Education: A Scoping Review. Med Sci Educ. 2026. PMID:41939019. https://pubmed.ncbi.nlm.nih.gov/41939019/.
18. Enhancing self-directed learning with custom GPT AI facilitation among medical students: a randomized controlled trial. Med Teach. 2024. doi:10.1080/0142159X.2024.2413023.
19. Coşkun Ö, Kıyak YS, Budakoğlu İİ. ChatGPT to generate clinical vignettes for teaching and multiple-choice questions for assessment: a randomized controlled experiment. Med Teach. 2025;47:268–274. doi:10.1080/0142159X.2024.2327477.
20. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025. doi:10.1093/postmj/qgae170.
21. Feasibility study of using GPT for history-taking training in medical education: a randomized clinical trial. 2025. PMID:40640776. https://pubmed.ncbi.nlm.nih.gov/40640776/.
22. Which curriculum components do medical students find most helpful for evaluating AI outputs? 2025. PMID:39915801. https://pubmed.ncbi.nlm.nih.gov/39915801/.
23. Integrating Generative AI Into Clinical Reasoning Education for Medical Students: Mixed Methods Study. 2026. PMID:42594250. https://pubmed.ncbi.nlm.nih.gov/42594250/.
24. AI Use for Medical Students: Impact on Clinical Skill Acquisition and Retention. A Systematic Review. 2026. PMID:42006163. https://pubmed.ncbi.nlm.nih.gov/42006163/.
25. Effectiveness of large language model-based teaching interventions in medical education: a systematic review and stratified meta-analysis of 14 studies. 2026. PMID:42907872. https://pubmed.ncbi.nlm.nih.gov/42907872/.
26. Large language models in undergraduate medical education: a systematic review of examination performance, knowledge acquisition, and learner perceptions. 2026. PMID:42947104. https://pubmed.ncbi.nlm.nih.gov/42947104/.
27. Generation of Case-Based Materials Using Large Language Models in Health Professions Education: Scoping Review. 2026. PMID:42976289. https://pubmed.ncbi.nlm.nih.gov/42976289/.
28. Large language models in radiology education: a systematic review of applications and learning outcomes. 2026. PMID:42998907. https://pubmed.ncbi.nlm.nih.gov/42998907/.
29. Curriculum-Aligned Conversational AI to Support Pathology Self-Directed Learning in Preclinical Medical Students: Crossover Randomized Pilot Study. 2026. PMID:42873617. https://pubmed.ncbi.nlm.nih.gov/42873617/.

30. Leveraging artificial intelligence to accelerate competency development in diagnostic reasoning of medical students. 2026. PMID:41843425. https://pubmed.ncbi.nlm.nih.gov/41843425/.
31. Assessment of Short-Answer Questions by ChatGPT in a Medical School Course. 2026. PMID:42932569. https://pubmed.ncbi.nlm.nih.gov/42932569/.
32. Evaluating the Effectiveness of ChatGPT Versus Human Proctors in Grading Medical Students' Post-OSCE Notes. 2026. PMID:41854706. https://pubmed.ncbi.nlm.nih.gov/41854706/.
33. Mapping generative AI in higher education teaching and learning: a scoping review. 2026. PMID:42889356. https://pubmed.ncbi.nlm.nih.gov/42889356/.
34. Applications, Challenges, and Prospects of Generative Artificial Intelligence Empowering Medical Education: Scoping Review. 2025. PMID:41128430. https://pubmed.ncbi.nlm.nih.gov/41128430/.
35. An umbrella review encompassing 42 systematic reviews on medical applications of ChatGPT. Minerva Med. 2026. PMID:41800905. doi:10.23736/S0026-4806.25.09664-8.
36. Applications of Generative Artificial Intelligence Teaching in Nursing Education: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. 2026. PMID:42924599. https://pubmed.ncbi.nlm.nih.gov/42924599/.
37. Roediger HL, Karpicke JD. Test-enhanced learning: psychological science and the real world. Psychol Sci. 2006;17:249–255. doi:10.1111/j.1467-9280.2006.01693.x.
38. Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, Willingham DT. Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. Psychol Sci Public Interest. 2013;14:4–58. doi:10.1177/1529100612453266.
39. Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306:978–988. doi:10.1001/jama.2011.1234.
40. Hattie J, Timperley H. The power of feedback. Rev Educ Res. 2007;77:81–112. doi:10.3102/003465430298487.
41. Sweller J. Cognitive load during problem solving: effects on learning. Cogn Sci. 1988;12:257–285. doi:10.1207/s15516709cog1202_4.
42. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learn Individ Differ. 2023;103:102274. doi:10.1016/j.lindif.2023.102274.
43. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368:l6890. doi:10.1136/bmj.l6890.
44. Page MJ, Moher D, Bossuyt PM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. doi:10.1136/bmj.n160.