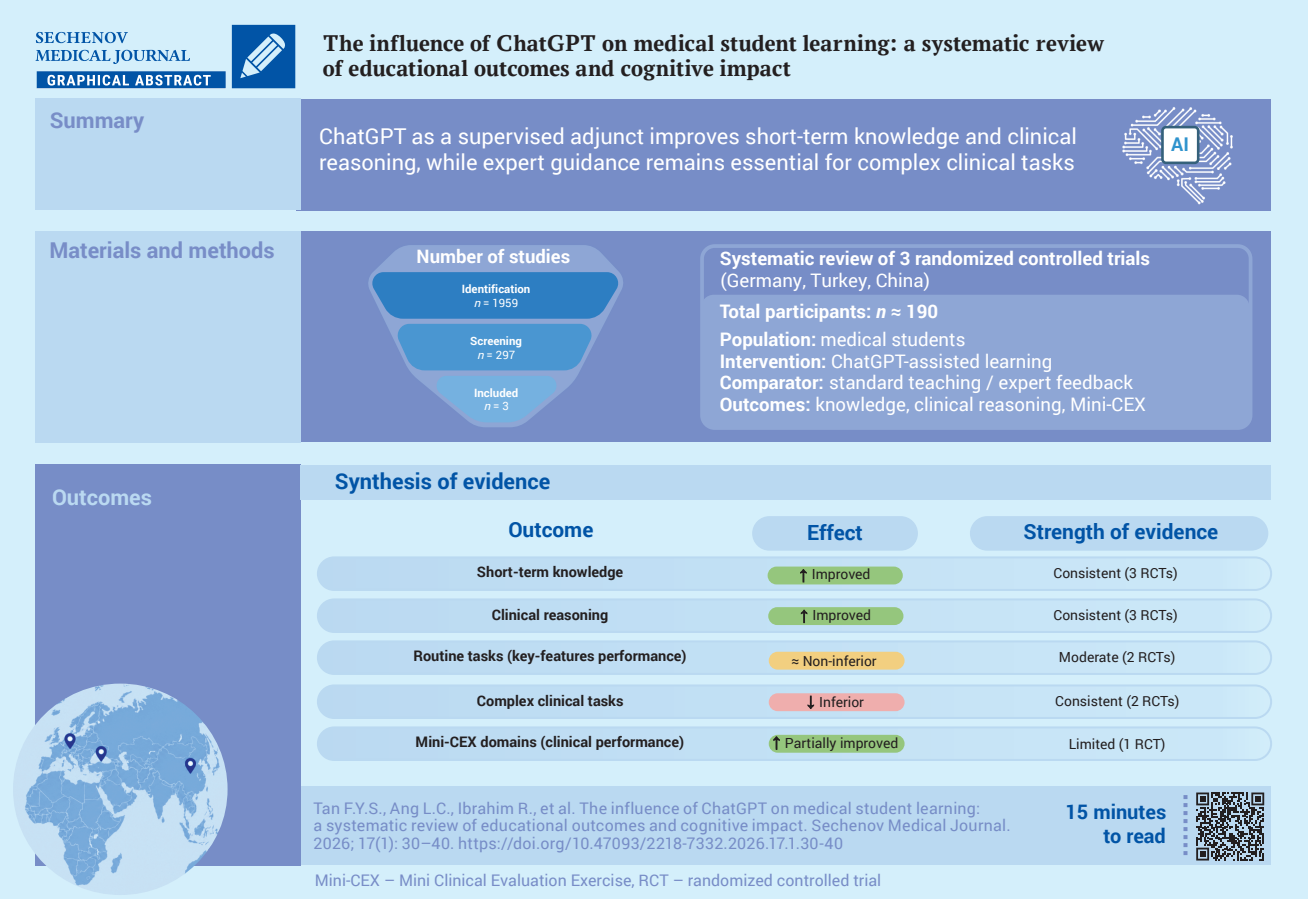
Review / Обзор
https://doi.org/10.47093/2218-7332.2026.17.1.30-40

The influence of ChatGPT on medical student learning: a systematic review of educational outcomes and cognitive impact

Felix Y.S. Tan, Lydia C. Ang, Rahima Ibrahim, Munawwarah Abdul Majeed, Mohd Aliff H. Mohd Aris, Sherly D. George✉
Manipal University College Malaysia (MUCM)
2, Jalan Batu Hampar, Bukit Baru, Melaka, 75150, Malaysia



Abstract

Aim. To evaluate whether the use of ChatGPT as a supplement to usual teaching improves medical students' short-term knowledge, clinical reasoning, and near-term performance.

Materials and methods. We systematically searched PubMed, Scopus, ScienceDirect, SpringerLink, and Web of Science on 25 June 2025, for studies involving medical students that evaluated ChatGPT as a supplement to teaching and reported objective educational outcomes. Two independent reviewers screened records, extracted data, and assessed the risk of bias. A narrative synthesis was then conducted due to the level of heterogeneity in interventions and outcome measures across the studies.

Results. Three randomized trials conducted in Germany, Turkey, and China met the inclusion criteria. ChatGPT-supported interventions improved or at least maintained short-term educational outcomes over the control groups for knowledge tests, clinical reasoning, and some of the Mini-Clinical Evaluation Exercise (Mini-CEX) domains.

Structured and immediate ChatGPT feedback improved Clinical Reasoning Indicator-History Taking Inventory scores after a simulated patient encounter, and ChatGPT-generated explanations were not inferior to expert feedback in overall key-features question performance but were less effective for more complex items, where expert feedback remained superior. Overall, the risk of bias was judged to be low to some concerns, with likely unblinded Mini-CEX assessment noted as a significant limitation.
Conclusion. ChatGPT used as a supervised adjunct to teaching showed value for short-term knowledge acquisition and clinical reasoning development.

Keywords: artificial intelligence; undergraduate medical training; medical education; clinical reasoning; problem-based learning; educational innovation
MeSH terms:
EDUCATION, MEDICAL – METHODS
EDUCATIONAL TECHNOLOGY – METHODS
MACHINE LEARNING
For citation: Tan F.Y.S., Ang L.C., Ibrahim R., Abdul Majeed M., Mohd Aris M.A.H., George S.D. The influence of ChatGPT on medical student learning: a systematic review of educational outcomes and cognitive impact. Sechenov Medical Journal. 2026; 17(1): 30–40. https://doi.org/10.47093/2218-7332.2026.17.1.30-40

CONTACT INFORMATION:
Sherly D. George, Cand. of Sci. (Medicine), Department of Physiology, Faculty of Medicine, Manipal University College Malaysia (MUCM).
Address: 2, Jalan Batu Hampar, Bukit Baru, Melaka, 75150, Malaysia
E-mail: sherly.george@manipal.edu.my

Conflict of interest. The authors declare that there is no conflict of interests.
Financing. The study was unfunded (own resources).
Use of artificial intelligence. No artificial intelligence tools were used in the preparation of this manuscript.

Received: 29.12.2025
Accepted: 28.03.2026
Date of publication: 29.05.2026

УДК [616:378]:004.8

Влияние ChatGPT на обучение студентов медицинских вузов: систематический обзор образовательных результатов и когнитивного воздействия

Ф.И.Ш. Тан, Л.Ч. Анг, Р. Ибрахим, М. Абдул Маджид, М.А.Х. Мохд Арис, Ш.Д. Джордж✉
Малайзийский университетский колледж Манипал
ул. Джалан Бату Хампар, д. 2, район Букит Бару, Малакка, 75150, Малайзия

Аннотация

Цель. Оценить, способствует ли использование ChatGPT в качестве дополнения к традиционному обучению улучшению краткосрочного уровня знаний студентов-медиков, клинического мышления и ближайших учебных результатов.
Материалы и методы. 25 июня 2025 г. был проведен систематический поиск исследований в базах данных PubMed, Scopus, ScienceDirect, SpringerLink и Web of Science. В обзор включались работы с участием студентов-медиков, в которых ChatGPT использовался как дополнение к обучению и оценивались объективные образовательные результаты. Два независимых рецензента осуществляли отбор публикаций, извлечение данных и оценку риска систематической ошибки. В связи с неоднородностью вмешательств и показателей исходов был выполнен нарративный синтез данных.

Результаты. Критериям включения соответствовали три рандомизированных исследования, проведенных в Германии, Турции и Китае. Использование ChatGPT в образовательных вмешательствах способствовало улучшению либо по меньшей мере сохранению краткосрочных учебных результатов по сравнению с контрольными группами в отношении тестов знаний, клинического мышления и ряда доменов Mini-Clinical Evaluation Exercise (Mini-CEX). Структурированная и немедленная обратная связь с использованием ChatGPT улучшала показатели Clinical Reasoning Indicator–History Taking Inventory после взаимодействия с симулированным пациентом. Объяснения ответов, сгенерированные ChatGPT, не уступали экспертной обратной связи по общему результату выполнения заданий формата key-features, однако были менее эффективны при решении более сложных вопросов, где экспертная обратная связь имела преимущество. Общий риск систематической ошибки был оценен как низкий или вызывающий некоторые опасения; существенным ограничением являлась вероятная неослепленная оценка Mini-CEX.
Заключение. Использование ChatGPT в качестве контролируемого дополнительного инструмента обучения продемонстрировало эффективность в отношении краткосрочного усвоения знаний и формирования клинического мышления.

Ключевые слова: искусственный интеллект; додипломная подготовка врачей; медицинское образование; клиническое мышление; проблемно-ориентированное обучение; образовательные инновации
Рубрики MeSH:
ОБРАЗОВАНИЕ МЕДИЦИНСКОЕ – МЕТОДЫ
ОБРАЗОВАНИЯ ТЕХНОЛОГИЯ – МЕТОДЫ
МАШИННОЕ ОБУЧЕНИЕ
Для цитирования: Тан Ф.И.Ш., Анг Л.Ч., Ибрахим Р., Абдул Маджид М., Мохд Арис М.А.Х., Джордж Ш.Д. Влияние ChatGPT на обучение студентов медицинских вузов: систематический обзор образовательных результатов и когнитивного воздействия. Сеченовский вестник. 2026; 17(1): 30–40. https://doi.org/10.47093/2218-7332.2026.17.1.30-40

КОНТАКТНАЯ ИНФОРМАЦИЯ:
Джордж Шерли Дебора, канд. мед. наук, кафедра физиологии медицинского факультета Малайзийского университетского колледжа Манипал.
Адрес: ул. Джалан Бату Хампар, д. 2, район Букит Бару, Малакка, 75150, Малайзия
E-mail: sherly.george@manipal.edu.my

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Финансирование. Исследование выполнено без спонсорской поддержки (за счет собственных средств).
Использование искусственного интеллекта. Инструменты искусственного интеллекта не использовались при подготовке данной рукописи.

Поступила: 29.12.2025
Принята: 28.03.2026
Дата печати: 29.05.2026

Abbreviations:
AI – artificial intelligence
CRI-HTI – Clinical Reasoning Indicator-History Taking Inventory
KFQ – key-features questions
LLM – large language model
Mini-CEX – Mini-Clinical Evaluation Exercise
PBL – problem-based learning
RCT – randomized controlled trial

HIGHLIGHTS	КЛЮЧЕВЫЕ ПОЛОЖЕНИЯ
Controlled trial evidence suggests that ChatGPT can improve or at least maintain short-term learning outcomes in medical students when used as a supervised adjunct.	Данные контролируемых исследований свидетельствуют о том, что ChatGPT при использовании в качестве контролируемого вспомогательного инструмента обучения способен улучшать или как минимум поддерживать краткосрочные результаты обучения студентов-медиков.
Structured, immediate ChatGPT feedback after simulated histories improves observable clinical reasoning behaviours on CRI-HTI scoring.	Структурированная и немедленная обратная связь от ChatGPT после моделируемого сбора анамнеза улучшает наблюдаемые показатели клинического мышления при оценке по шкале CRI-HTI.
ChatGPT-generated explanations are non-inferior to expert feedback for overall key-features performance on routine clinical reasoning tasks.	Объяснения, сгенерированные ChatGPT, не уступают экспертной обратной связи по общим результатам выполнения заданий формата key-feature questions при решении стандартных задач клинического мышления.
Expert feedback remains superior for more complicated items, supporting human oversight for high-complexity reasoning.	При выполнении более сложных заданий экспертная обратная связь остается более эффективной, что подтверждает необходимость участия преподавателя при решении клинических задач высокой сложности.
ChatGPT-assisted problem-based learning improves short-term knowledge and selected Mini-CEX domains during a clinical rotation.	Проблемно-ориентированное обучение с использованием ChatGPT улучшает краткосрочное усвоение знаний и отдельные оцениваемые домены Mini-CEX во время клинической ротации.

The advent of artificial intelligence (AI) applications, such as ChatGPT, has had a major impact on many industries, including medical education. As medical schools advance to include innovative technologies, the literature on how AI can be employed to enhance the learning outcomes of medical trainees continues to grow. OpenAI's sophisticated large language model (LLM), ChatGPT, has demonstrated competence in various domains of medical education. It has been reported to be useful in creating experiences for individuals who learn, developing critical thinking skills, and supporting problem-based learning (PBL) interventions among medical students [1, 2]. In addition to providing tailored learning experiences, ChatGPT enables the consistent delivery of quality learning across contexts.

Despite these benefits, empirical evidence of the effectiveness of ChatGPT in medical education remains limited. Surapaneni [3] reported a performance gap between medical students and responses produced by ChatGPT for certain tests, which raises questions regarding accuracy and applicability. Moreover, the use of AI-based clinical reasoning and skills assessment raises important ethical and pedagogical issues, a theme that is further illustrated by the duality in perceptions of AI technology in clinical learning settings [4].

This review aims to synthesize controlled evidence on whether ChatGPT, used as a supplement to usual instruction, improves medical students' knowledge, clinical reasoning, and near-term performance, with the goal of identifying educational effects, boundary conditions, and implications for safe curricular use.

MATERIALS AND METHODS

Protocol and reporting framework

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines. The review question was framed using the PICO structure: What is the impact of ChatGPT on the educational outcomes and cognitive development of medical students? The protocol specified the inclusion/exclusion criteria, information sources, screening methods, data extraction fields, risk-of-bias assessment by study design, and a data synthesis plan. This review was not prospectively registered with PROSPERO. Registration was not completed at inception due to resource and time constraints during the initial protocol development phase. We will prospectively register any future update or extension of this review to strengthen transparency. Nevertheless, the eligibility criteria, outcomes of interest, and synthesis approach were finalised before screening commenced and were applied consistently throughout the review process.

Eligibility criteria

We included original empirical studies published in English from 2022 onwards that enrolled medical students and evaluated ChatGPT as part of an educational intervention with measurable results. Eligible designs included randomized controlled trials (RCT) or nonrandomized controlled trials, controlled cohort studies, and mixed-methods evaluations with quantitative endpoints. We excluded editorials, letters, opinion pieces, reviews, purely technical descriptions

without analysable educational data, studies not involving medical students (unless a distinct medical-student subgroup could be extracted), non-learning uses of ChatGPT (e.g., administrative tasks), and non-English publications.

Information sources and search strategy

A comprehensive search was performed in five bibliographic databases with pre-specified Boolean logic adapted to each platform: PubMed, Scopus, ScienceDirect, SpringerLink, and Web of Science. Search strings combined terms for the technology (e.g., "ChatGPT," "large language model," "artificial intelligence") with medical education terms ("medical students," "medical education") and outcome terms ("learning outcomes," "performance," "clinical reasoning," "critical thinking", Table 1). The literature search was conducted on 25 June 2025. Searches were limited to English-language records and publication years from 2022 onwards, where database filters permitted. No study design filter was applied during the search stage. The results were exported on the same search wave, merged in a reference manager, and de-duplicated prior to screening.

Study selection

The screening was conducted in two stages by two independent reviewers (L.A.C. and R.I.). Titles/abstracts were screened against the eligibility criteria, followed by full-text assessment of potentially relevant reports. Disagreements were resolved through discussion or, if required, by a third reviewer (M.A.M.). The screening workflow was supported by Rayyan[1], a web-based platform used to facilitate blinded title/abstract screening [5]. The citation-checking workflow was supported by scite.ai[2], a web-based citation analysis platform used as a supplementary tool for verification, including checking citation contexts. The PRISMA flow diagram documents the process and counts: 1481 unique records were screened after the removal of 478 duplicates; 297 full texts were retrieved and assessed; 294 were excluded based on eligibility (review articles, n = 74; no analyzable data, n = 132; wrong study design, n = 87; non-English, n = 1). Three studies met all the criteria and were included in the narrative synthesis (Fig. 1).

Data extraction and management

A structured data-extraction form, piloted on a subset of studies, captured the following: study setting and design; participant characteristics (stage of training, sample size); description of the ChatGPT/LLM intervention (role in the learning activity, prompt structure, timing relative to instruction, presence/absence of human oversight); comparator condition(s); outcome measures and instruments (e.g., written knowledge tests, Key-Features Questions, Clinical Reasoning Indicator–History Taking Inventory (CRI-HTI), Mini-Clinical Evaluation Exercise (Mini-CEX)); follow-up interval(s); statistical results (effect estimates, measures of variability, p-values); and author-reported implementation details (e.g., disclosure of AI use, debriefing, verification steps). Two reviewers independently extracted the data; inconsistencies were reconciled by consensus with reference to the source report.

Outcomes

The primary outcomes were objective educational measures aligned with the intervention's learning aims, including near-term written knowledge performance and validated indices of clinical reasoning and skills (e.g., CRI-HTI subdomains and Mini-CEX ratings). Secondary outcomes included learner-reported indicators relevant to cognitive impact (e.g., perceived learning effectiveness, breadth of clinical exposure, and critical stance toward AI following disclosure and debriefing). Where available, immediate and short-delay (⩽10 days) post-intervention results were collected to gauge the early retention.

Risk of bias and study quality assessment

Risk-of-bias judgements were performed at the study level by design category using validated tools. RCTs were appraised using Version 2 of the Cochrane tool for assessing the risk of bias in randomized trials (RoB 2) (domains: randomization process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported

Table 1. **Search strategy across databases**

Database	Search string
ScienceDirect	"ChatGPT" AND "medical students" AND ("learning" OR "education" OR "academic performance" OR "critical thinking")
PubMed	((ChatGPT) OR (artificial intelligence)) AND ((medical students[Title/Abstract]) OR (medical education[Title/Abstract])) AND ((performance) OR (academic) OR (clinical reasoning)) AND (2022/1/1:2025/6/25[pdat])
SpringerLink	"ChatGPT" AND "medical education" AND ("student performance" OR "reasoning skills" OR "AI-assisted learning")
Web of Science	TS=("ChatGPT" OR "language models") AND TS=("medical students" OR "medical education") AND TS=("learning outcomes" OR "academic performance" OR "cognitive skills")
Scopus	TITLE-ABS-KEY("ChatGPT" OR "large language model") AND TITLE-ABS-KEY("medical education" OR "medical students") AND TITLE-ABS-KEY("performance" OR "critical thinking" OR "learning outcomes")

[1] Rayyan – Intelligent Systematic Review. https://www.rayyan.ai (access date: 25.06.2025).
[2] scite.ai – Smart Citations. https://scite.ai (access date: 25.06.2025).

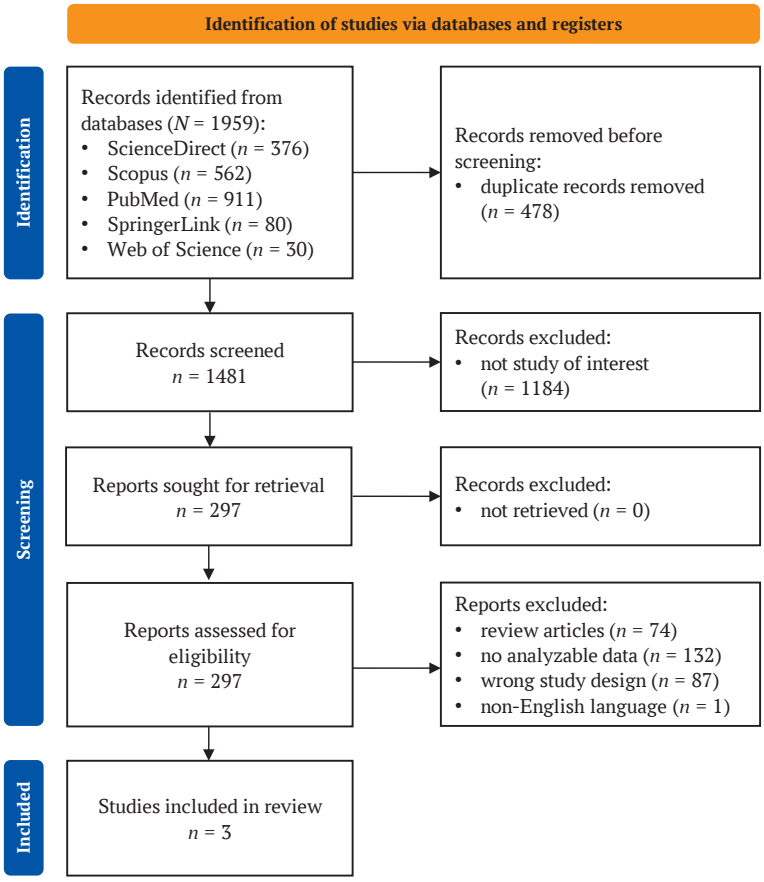


FIG. 1. PRISMA flow diagram of study selection.

result). ROBINS-I (Risk Of Bias In Non-randomized Studies-of Interventions) and MMAT (Mixed Methods Appraisal Tool) were pre-specified for non-randomized and mixed-methods studies but were not applied because all included studies were randomized. Disagreements were resolved by discussion; where reporting limited firm judgements, domains were rated "some concerns" with justification.

Data synthesis and analysis

Given the small number of eligible studies and heterogeneity in interventions (LLM role, presence of structured feedback, task complexity), comparators, and outcome instruments, a narrative synthesis was specified as the primary approach. Planned quantitative synthesis required ⩾ 5 trials with commensurate outcome measures; this threshold was not met because of non-overlapping instruments (CRI-HTI, Key-Features Questions (KFQ), Mini-CEX) across studies. If conditions permitted, continuous outcomes would have been summarized as mean differences or standardized mean differences with 95% confidence intervals using random-effects models, with heterogeneity quantified by I^2 and explored via subgroup analyses (e.g., pre-clinical vs. clinical stage, routine vs. complex case tasks, human oversight vs. AI-only feedback). Instead, we synthesized findings by mapping interventions to learning mechanisms (e.g., immediate task-specific feedback, rehearsal of history-taking, PBL scaffolding), outcome family (knowledge, clinical reasoning, clinical performance), and task complexity, noting consistencies and divergences across settings. Where immediate and short-delay outcomes were both reported, we qualitatively described patterns to indicate early retention.

RESULTS

Study selection and characteristics

Of the 1481 unique records screened, three RCTs met the inclusion criteria and were included in the review of educational outcomes in medical students. The studies were conducted in Germany, Turkey, and China, and each evaluated ChatGPT as a supplement to established teaching formats rather than as a replacement. The validated indices included CRI-HTI, KFQ, and Mini-CEX. The interventions fell into three pedagogic roles: (i) simulated patient encounters with LLM-delivered formative feedback, (ii) AI-generated written feedback on text-based clinical reasoning problems, compared directly with expert feedback, and (iii) ChatGPT-assisted PBL integrated into a brief clinical rotation. Outcomes comprised validated indices of clinical reasoning (CRI-HTI; KFQ), theory examinations, and Mini-CEX

domains, with short-delay assessments reported. The sample sizes ranged from 21 to 129 randomized participants, with follow-up completion reported in all trials. Reporting of model version, prompting and output verification was limited across trials.

Interventions and comparators

In the German trial, pre-clinical students completed four six-minute LLM-simulated history-taking encounters; only the intervention arm received structured ChatGPT feedback after each case, generated against CRI-HTI rubric [6]. Two blinded human raters scored the transcripts (intraclass correlation coefficient was 0.924). The comparator was an identical simulation without feedback.

In Turkey, first-year students undertook five days of spaced ContExtended Question practice for urinary tract infections; arms differed only in the source of explanations provided after each step (expert-written vs. ChatGPT-generated) [7]. Clinical reasoning was measured using the KFQs immediately and 10 days later.

In China, fifth-year interns on a two-week urology rotation were randomized to PBL+ChatGPT (pre-class exploration and in-class discussion with AI support, with instructor oversight) versus traditional teaching [8]. The primary outcomes were a 100-point theory exam (pre-and three days post-lecture) and Mini-CEX ratings across seven domains; student satisfaction was secondary (Table 2).

Primary outcomes

Across the two reasoning-focused trials, ChatGPT improved or matched outcomes when deployed as a feedback engine, with caveats for complex cases.

In the German simulation RCT (n = 21 analyzed), after four sessions, LLM feedback improved final CRI-HTI scores versus simulation without feedback (3.60 ± 0.13 vs. 3.02 ± 0.12; F(1,18) = 4.44, p = 0.049; partial η^2 = 0.198), with gains concentrated in 'creating context' (p = 0.046) and 'securing information' (p = 0.018), but not focusing questions (p = 0.265). The rater agreement for transcript scoring was excellent (intraclass correlation coefficient was 0.924). These findings indicate that immediate task-specific LLM feedback can shift observable reasoning behaviors over a short training period.

In the Turkish RCT (n = 129 randomized; ⩾115 tested), ChatGPT feedback vs. expert feedback produced no difference in overall KFQ scores at immediate testing (ChatGPT 74.7 ± 15.1 vs. Expert 78.5 ± 20.6; p = 0.26) or at 10 days (ChatGPT 76.0 ± 14.5 vs. Expert 78.0 ± 21.2; p = 0.57). For complicated urinary tract infection items in delayed testing, expert feedback outperformed ChatGPT (p < 0.001). Notably, the disclosure of AI use increased students' critical approach to AI, with medium to significant effects. Together, these data suggest that ChatGPT can provide non-inferior formative explanations for routine problems, while expert oversight remains advisable for nuanced scenarios.

In the Chinese rotation RCT (n = 42), both groups improved from baseline, but PBL+ChatGPT achieved a higher post-course theory score at three days (93.90 ± 3.65) than traditional teaching (90.33 ± 4.08; p < 0.01). This short-term knowledge advantage was accompanied by higher Mini-CEX ratings in medical interviewing, clinical judgement, and overall competence, with no differences in other domains. Mini-CEX gains favouring PBL+ChatGPT were concentrated in interviewing, judgement, and overall competence. These domain-specific improvements may reflect the intervention's emphasis on structured question framing and synthesis, however, this interpretation remains hypothetical. All assessments were completed within standardized time windows and by a single assessor, which supports procedural consistency but may limit blinding. Because a single, likely unblinded assessor completed the Mini-CEX, these skill domain gains warrant cautious interpretation.

Secondary outcomes (acceptability and perceived learning)

Medical students rated the LLM-supported activities favorably. In the German study, participants described the simulated encounters as realistic and feasible, mirroring the observed behavioral gains when feedback was present [6]. In Turkey, students in the ChatGPT feedback group developed a notably more critical outlook [7]. As a result, they became more skeptical about the accuracy and reliability of AI-generated information and more cautious in accepting AI content without question. In the Chinese trial, satisfaction was uniformly high after PBL+ChatGPT, with no reports of dissatisfaction [8].

Synthesis of effects and certainty

Because interventions and instruments differed across trials, meta-analysis was not appropriate; instead,

Table 2. **Characteristics of the included randomized studies**

Study	Country	Intervention/comparator	Outcome(s)
Brügge et al., 2024 [6]	Germany	LLM-simulated patient interviews ± ChatGPT feedback	CRI-HTI
Çiçek et al., 2025 [7]	Turkey	ChatGPT-generated vs. expert-written feedback on KFQs	Immediate and 10-day scores
Hui et al., 2025 [8]	China	ChatGPT-assisted PBL vs. traditional teaching	Theory test and Mini-CEX

Note: CRI-HTI – Clinical Reasoning Indicator-History Taking Inventory; KFQ – key-features questions; LLM – large language model; Mini-CEX – Mini-Clinical Evaluation Exercise; PBL – problem-based learning.

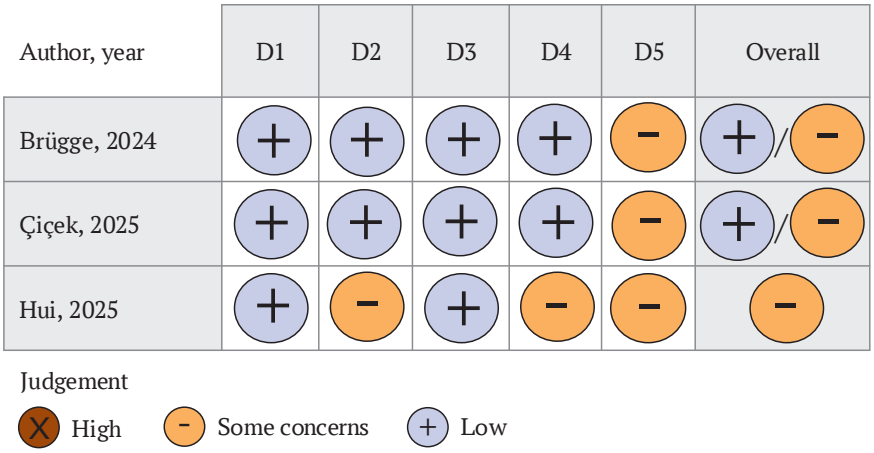


FIG. 2. Risk-of-bias assessment of randomized controlled trials.

Note: D1 – randomization process; D2 – deviations from intended interventions; D3 – missing outcome data; D4 – measurement of the outcome; D5 – selection of the reported result.

narrative synthesis showed consistent benefits of immediate, task-specific feedback, parity with expert feedback for routine problems, and short-term knowledge and Mini-CEX gains with ChatGPT-assisted PBL under supervision.

Risk of bias

We appraised the risk of bias using RoB 2 across five domains (Fig. 2). Overall, the included trials showed a generally acceptable methodological profile for short-term educational outcomes. The randomization process was judged to be at low risk of bias in all three studies: Brügge et al. used a two-group randomized design with balanced baseline characteristics, Çiçek et al. applied computer-generated randomization with comparable groups, and Hui et al. used randomized allocation with equal group sizes.

Deviations from intended interventions were unlikely to materially affect the results in the German and Turkish trials, where training procedures, exposure, and assessment conditions were standardized across arms. In the Turkish trial, participants were blinded to the feedback source, and the only intended difference between groups was whether the explanations were expert-written or ChatGPT-generated. By contrast, the Chinese trial raised some concerns because intervention delivery was unblinded, which may have influenced performance-related outcomes.

Missing outcome data were judged to be at low risk of bias across the three studies, with minimal exclusions or complete post-course assessment reported and no evidence of differential attrition. Measurement of outcomes was also robust in two trials: the German study used two independent blinded raters and the validated CRI-HTI instrument, with excellent inter-rater agreement, whereas the Turkish study relied on objective standardized key-features scoring. In the Chinese trial, the theory test was objective; however, Mini-CEX was assessed by a single likely unblinded evaluator, raising concerns about detection bias. Therefore, Mini-CEX findings, although directionally favorable, should be interpreted with caution.

The domain “selection of the reported result” was conservatively rated as “some concerns” in all three studies because none reported prospective registration or a pre-specified analysis plan. Taken together, the overall risk of bias was judged as low to some concerns for the German and Turkish trials and as some concerns for the Chinese trial. The evidence base is therefore methodologically acceptable for evaluating short-term educational outcomes, but conclusions regarding clinical skills should remain cautious when based on unblinded performance ratings.

Summary of main findings

Taken together, the best available randomized evidence supports an optimistic but circumspect conclusion: when used to augment existing teaching, it provides immediate, structured feedback; by scaffolding PBL or streamlining explanations, ChatGPT improves or maintains near-term educational outcomes for medical students, with clear added value in efficiency and reach. The signal is strongest for short-cycle learning targets (history-taking behaviors, script refinement, and theory recall), while complex clinical problems still benefit from human calibration. Across trials, disclosure and supervision varied. In the Turkish trial, disclosure of AI use was reported to increase students’ critical stance toward AI. Contrastingly, other trials provided supervision and structured feedback but did not quantify disclosure-related effects.

DISCUSSION

Principal findings and interpretation

Across three RCTs, ChatGPT used alongside usual teaching equaled or improved the learning outcomes.

The small number of eligible trials reflects the early stage of controlled research on ChatGPT in medical education rather than selective inclusion. Effects were observed in written knowledge, clinical reasoning during history taking, and selected Mini-CEX domains. Two themes recur. These findings are consistent with established learning theory. Feedback is one of the strongest influences on learning, but its effects depend on timing, specificity, and whether it directs attention to the task and process rather than to the learner alone [9]. These findings are educationally plausible because the included interventions emphasized repeated practice, timely feedback, and structured support. In the reviewed trials, ChatGPT appeared most useful when it was embedded within clearly defined learning activities rather than used as an unrestricted answer-generating tool.

First, feedback is the engine of the learning process. For example, ChatGPT provided prompt, task-based feedback after simulated patient encounters, resulting in improved clinical reasoning (measured on a validated instrument) compared to a no-feedback control group of students. When used to replace experts as the source of feedback on clinical key-features problems, ChatGPT was found to be as effective as expert-written feedback overall at immediate and 10-day testing, demonstrating scalable parity for routine problems. This pattern suggests that ChatGPT may be better suited to routine, structured tasks than to complex, ambiguity-rich clinical reasoning. In such settings, expert feedback still appears to provide a meaningful advantage. This pattern is also compatible with retrieval practice theory, which predicts that repeated testing and corrective explanation can strengthen learning more effectively than passive review alone [10].

Second, there is the question of context and complexity. In the same RCT, expert feedback outperformed ChatGPT for more complex items at late testing, indicating that cognitive scaffolding for complex multi-step reasoning is still better provided by experienced clinicians or by blended human-AI feedback calibrated to the task [7]. Beyond test scores, the Chinese controlled cohort suggests the spillover into the performance domain: ChatGPT-assisted PBL not only improved near-term theory scores but also enhanced Mini-CEX ratings for interviewing and clinical judgment, outcomes that may be relevant to supervised clinical performance [8]. In addition, ChatGPT increases the pace of formative feedback and expands case exposure. When integrated into formalized activities, it enhances knowledge and chosen performance indicators. Where issues require subtlety, the insight of an expert provides an edge [7]. The discovered limits are informative and represent design choices rather than intrinsic constraints.

Strengths, limitations, and implications

This synthesis is based on the use of validated tools (e.g., CRI-HTI, Mini-CEX) and authentic assessment tasks (key features, observed encounters). Positive results were observed across three settings (Germany, Turkey, and China), supporting further study across diverse curricula, however, they do not establish generalizability.

However, there are conditions under which highlighting is necessary for the reader. First, most outcomes were short-term, and we have limited information on longer-term retention, transfer to authentic clerkship performance, or downstream patient-centered outcomes. One plausible risk of AI-supported learning is over-reliance on the tool, whereby improved immediate task performance may not translate into durable competence if students defer reasoning rather than internalize it. This concern is particularly relevant when outcomes are measured over short intervals. Second, the learning tasks studied on history taking, key features reasoning in one topic, and short-horizon PBL represent important but narrow sections of the early curriculum [6]. Third, the primacy of expert feedback on challenging cases at late testing is a salutary reminder that blind use of LLM results to "replace" faculty is unwise; complexity calibration and human oversight seem to be essential for safety and quality [7]. LLM outputs can be fluent yet incorrect, and learners may inadvertently internalise inaccuracies or biased reasoning if responses are accepted uncritically [11]. Accordingly, educational use should be paired with explicit verification practices, particularly for safety-critical or high-complexity content. Lastly, our approach to mixed-population studies may introduce selection bias if medical-student subgroup data were selectively reported or not extractable.

Methodologically, future studies could pre-register protocols, determine power for non-inferiority or superiority prespecified outcomes, report adherence to intervention protocols (e.g., prompt templates, versioning), and disclose guardrails (instructional hints, verification steps), which may modulate both efficacy and safety. In parallel, curricular integration should acknowledge the “black box” features of LLMs and the risk of plausible-sounding inaccuracies or biased outputs.

Implementation also occurs within a pre-existing landscape of academic integrity challenges in medical training. Generative AI may amplify familiar risks (e.g., unacknowledged assistance, plagiarism, and inappropriate collaboration), particularly where institutional policies are unclear or inconsistently applied [12, 13]. In controlled, supervised interventions, the educational signal is clearer. However, in unsupervised settings, it becomes harder to distinguish AI-supported learning from AI-facilitated misconduct. Therefore, for this reason, clear permissible-use guidance, assessment design that samples unaided reasoning, and teaching students how to document and justify AI use are therefore essential complements to any educational deployment of ChatGPT [12, 13].

Three design choices appear to be important. First, feedback should be structured, timely, and task-focused; in the simulation trial, ChatGPT’s outputs were mapped

to CRI-HTI behaviors and delivered immediately after the task was completed [6]. Second, align support to task complexity: routine problems tolerate automated feedback, whereas complex or safety-critical scenarios merit clinician oversight or blended experts, which is the AI commentary [7]. Third, LLMs should be used to widen exposure and rehearsal opportunities, but they should be treated as adjuncts to live bedside teaching, not substitutes for patient contact and mentorship [7, 8].

Generative AI also raises academic and assessment integrity concerns, as unsupervised use may blur the boundary between supported learning and unpermitted assistance. Assessment policy should evolve with time. Where ChatGPT is available, assessment for learning can plausibly include AI-mediated practice (e.g., formative key features with ChatGPT feedback), while assessment of learning must ensure that summative tasks adequately sample unaided reasoning and performance. Explicit instruction on when to use AI (e.g., when setting up for PBL but not while responding to objective structured clinical examination) will allow students to benefit from AI while maintaining academic integrity and skill building.

Key questions now are the durability of gains at 3–6 months, transfer to objective structured clinical examinations and workplace-based assessments, the complexity thresholds at which AI-generated feedback underperforms expert guidance, and whether blended expert, the AI feedback narrows that gap. Adequately powered multicenter randomized trials should compare ChatGPT-only, expert-only, and blended models stratified by case complexity and accompanied by cost-effectiveness analyses. Reproducibility requires clear reporting of the model version, prompt templates, guardrails, and protocol adherence.

CONCLUSION

Across three randomized studies on medical students, ChatGPT used as a supplement to usual instruction matched or improved short-term outcomes in knowledge, clinical reasoning, and selected workplace-based domains. Benefits were clearest when the model delivered immediate, structured feedback or scaffolded PBL; by contrast, complex, nuanced problems continued to favor expert guidance. On balance, the evidence favors an adjunctive role for ChatGPT, which is expanding practice and feedback at a low marginal cost while preserving human oversight for complex reasoning and high-stakes competencies.

AUTHOR CONTRIBUTIONS

Felix Y.S. Tan conceived and designed the study, developed the methodology, performed the formal analysis, and prepared the original draft of the manuscript. Lydia C. Ang and Rahima Ibrahim participated in the investigation and screening process and contributed to the review and editing of the manuscript. Munawwarah Abdul Majeed participated in the investigation, curated the data, performed adjudication, and contributed to the review and editing of the manuscript. Mohd Aliff H. Mohd Aris performed validation procedures and contributed to the review and editing of the manuscript. Sherly D. George conceived and designed the study, developed the methodology, supervised the project, performed validation procedures, and contributed to the review and editing of the manuscript. All authors approved the final version of the article.

ВКЛАД АВТОРОВ

Ф.И.Ш. Тан разработал концепцию и дизайн исследования, сформировал методологию, выполнил формальный анализ и подготовил первоначальный вариант рукописи. Л.Ч. Анг и Р. Ибрахим участвовали в проведении исследования и скрининге публикаций, а также внесли вклад в редактирование и доработку рукописи. М. Абдул Маджид участвовала в проведении исследования, осуществляла курирование данных, выполняла экспертную оценку материалов и внесла вклад в редактирование и доработку рукописи. М.А.Х. Мохд Арис выполнял процедуры валидации и участвовал в редактировании и доработке рукописи. Ш.Д. Джордж разработала концепцию и дизайн исследования, сформировала методологию, осуществляла общее руководство проектом, выполняла процедуры валидации и внесла вклад в редактирование и доработку рукописи. Все авторы одобрили окончательную версию статьи.

REFERENCES / ЛИТЕРАТУРА

1. *Abouammoh N., Alhasan K., Aljamaan F., et al.* Perceptions and earliest experiences of medical students and faculty with chatgpt in medical education: qualitative study. JMIR Med Educ. 2025; 11: e63400. https://doi.org/10.2196/63400. PMID: 39977012
2. *Sallam M.* ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023; 11(6): 887. https://doi.org/10.3390/healthcare11060887. PMID: 36981544
3. *Surapaneni K.M.* Assessing the Performance of ChatGPT in medical biochemistry using clinical case vignettes: observational study. JMIR Med Educ. 2023; 9: e47191. https://doi.org/10.2196/47191. PMID: 37934568
4. *Hisan U.K., Amri M.M.* ChatGPT and medical education: a double-edged sword. Journal of Pedagogy and Education Science. 2023; 2(01): 71–89. https://doi.org/10.56741/jpes.v2i01.302
5. *Ouzzani M., Hammady H., Fedorowicz Z., Elmagarmid A.* Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016 Dec; 5(1): 210. https://doi.org/10.1186/s13643-016-0384-4. PMID: 27919275
6. *Brügge E., Ricchizzi S., Arenbeck M., et al.* Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. BMC Med Educ. 2024; 24(1): 1391. https://doi.org/10.1186/s12909-024-06399-7. PMID: 39609823
7. *Çiçek F., Ülker M., Özer M., Kıyak Y.S.* ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025; 101(1195): 458–463. https://doi.org/10.1093/postmj/qgae170. PMID: 39656920
8. *Hui Z., Zewu Z., Jiao H., Yu C.* Application of ChatGPT-assisted problem-based learning teaching method in clinical

medical education. BMC Med Educ. 2025; 25(1): 50. https://doi.org/10.1186/s12909-024-06321-1. PMID: 39799356
9. *Hattie J., Timperley H.* The Power of feedback. Review of Educational Research. 2007; 77(1): 81–112. https://doi.org/10.3102/003465430298487
10. *Roediger H.L., Karpicke J.D.* Test-enhanced learning: taking memory tests improves long-term retention. Psychol Sci. 2006; 17(3): 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x. PMID: 16507066
11. *Farquhar S., Kossen J., Kuhn L., et al.* Detecting hallucinations in large language models using semantic entropy. Nature. 2024; 630(8017): 625–630. https://doi.org/10.1038/s41586-024-07421-0
12. *Kasneci E., Sessler K., Küchemann S., et al.* ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 2023; 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274
13. *Bittle K., El-Gayar O.* Generative AI and academic integrity in higher education: a systematic review and research agenda. Information. 2025; 16(4): 296. https://doi.org/10.3390/info16040296

INFORMATION ABOUT THE AUTHORS / ИНФОРМАЦИЯ ОБ АВТОРАХ

Felix Y.S. Tan, physician, Faculty of Medicine, Manipal University College Malaysia (MUCM).
ORCID: https://orcid.org/0009-0006-1086-6790

Lydia C. Ang, physician, Faculty of Medicine, Manipal University College Malaysia (MUCM).
ORCID: https://orcid.org/0009-0006-9369-8656

Rahima Ibrahim, physician, Faculty of Medicine, Manipal University College Malaysia (MUCM).
ORCID: https://orcid.org/0009-0004-6937-1608

Munawwarah Abdul Majeed, physician, Faculty of Medicine, Manipal University College Malaysia (MUCM).
ORCID: https://orcid.org/0009-0009-2363-7358

Mohd Aliff H. Mohd Aris, physician, Faculty of Medicine, Manipal University College Malaysia (MUCM).
ORCID: https://orcid.org/0009-0003-7178-1668

Sherly D. George✉**,** Cand. of Sci. (Medicine), Department of Physiology, Faculty of Medicine, Manipal University College Malaysia (MUCM).
ORCID: https://orcid.org/0000-0002-8836-5746

Тан Феликс И Шуэн, врач, медицинский факультет Малайзийского университетского колледжа Манипал.
ORCID: https://orcid.org/0009-0006-1086-6790

Анг Лидия Ченг, врач, медицинский факультет Малайзийского университетского колледжа Манипал.
ORCID: https://orcid.org/0009-0006-9369-8656

Ибрахим Рахима, врач, медицинский факультет Малайзийского университетского колледжа Манипал.
ORCID: https://orcid.org/0009-0004-6937-1608

Абдул Маджид Мунавварах, врач, медицинский факультет Малайзийского университетского колледжа Манипал.
ORCID: https://orcid.org/0009-0009-2363-7358

Мохд Арис Мохд Алифф Хайкал, врач, медицинский факультет Малайзийского университетского колледжа Манипал.
ORCID: https://orcid.org/0009-0003-7178-1668

Джордж Шерли Дебора✉, канд. мед. наук, кафедра физиологии медицинского факультета Малайзийского университетского колледжа Манипал.
ORCID: https://orcid.org/0000-0002-8836-5746

✉ Corresponding author / Автор, ответственный за переписку