



BASIC RESEARCH:

Effectiveness of Chatbot-Assisted Learning in Real Dental Student Cohorts: A Systematic Review, Exploratory Meta-analysis, and Functional Narrative Synthesis Efectividad del aprendizaje asistido por chatbots en cohortes reales de estudiantes de odontología: Revisión sistemática, metaanálisis exploratorio y síntesis funcional narrativa

Eliana Pineda-Vélez^{1,2,3} <https://orcid.org/0000-0002-2431-7489>

Diego Villada-Osorio¹ <https://orcid.org/0000-0002-3365-4022>

Carlos M. Ardila^{2,4} <https://orcid.org/0000-0002-3663-1416>

¹Doctorate in Diversity Education Program, Universidad de Manizales, Manizales 170003, Colombia.

²Basic Sciences Department, Biomedical Stomatology Research Group, Faculty of Dentistry, Universidad de Antioquia U de A, Medellín 050010, Colombia.

³Faculty of Dentistry, Institución Universitaria Visión de las Américas, Medellín 050040, Colombia.

⁴Department of Periodontics, Saveetha Dental College, and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai 600077, India.

Correspondence to: Carlos M. Ardila - martin.ardila@udea.edu.co

Received: 17-VI-2026

Accepted: 2-VII-2026

ABSTRACT: Artificial intelligence chatbots are increasingly being incorporated into dental education, but their effectiveness in real student cohorts remains uncertain. This systematic review, meta-analysis, and functional narrative synthesis evaluated the educational impact of chatbot-assisted learning interventions in dental students, focusing on student performance, self-efficacy, and domain-specific applications. A systematic search of PubMed/MEDLINE, Scopus, and Embase was conducted for studies published up to May 2026. Eligible studies included real dental students exposed to chatbot-assisted educational interventions and assessed using explicit comparators or meaningful pre-post designs. Risk of bias was assessed with RoB 2.0 and the JBI checklist, and certainty of evidence was evaluated using GRADE. Eight studies involving 716 undergraduate dental students were included in the qualitative synthesis, and two studies were included in an exploratory meta-analysis. The pooled estimate favored chatbot-assisted learning over comparator conditions for overall student educational performance (SMD=0.62, 95% CI: 0.37 to 0.87; $I^2=0\%$), although this finding should be interpreted cautiously because only two studies were pooled and they assessed different manifestations of educational performance. Certainty of evidence was low. Functional narrative synthesis identified four educational domains: knowledge acquisition and academic performance; clinical reasoning and diagnostic communication; operative and procedural skill support; and academic writing, evidence handling, and reflective tasks. Chatbot-assisted learning appears promising, but current evidence remains limited, heterogeneous, and context-dependent.



KEYWORDS: Chatbot-assisted learning; Dental education; Artificial intelligence in dentistry; Systematic review; Meta-analysis.

RESUMEN: Los chatbots de inteligencia artificial se están incorporando cada vez más en la educación odontológica, aunque su efectividad en cohortes reales de estudiantes sigue siendo incierta. Esta revisión sistemática, metaanálisis exploratorio y síntesis funcional narrativa evaluó el impacto educativo de las intervenciones de aprendizaje asistido por chatbots en estudiantes de odontología, con énfasis en el desempeño estudiantil, la autoeficacia y las aplicaciones específicas por dominio. Se realizó una búsqueda sistemática en PubMed/MEDLINE, Scopus y Embase de estudios publicados hasta mayo de 2026. Se incluyeron estudios con estudiantes reales de odontología expuestos a intervenciones educativas asistidas por chatbots y evaluados mediante comparadores explícitos o diseños pre-post relevantes. El riesgo de sesgo se evaluó con RoB 2.0 y la lista de verificación del JBI, y la certeza de la evidencia mediante GRADE. Ocho estudios, con 716 estudiantes de pregrado en odontología, fueron incluidos en la síntesis cualitativa, y dos estudios en un metaanálisis exploratorio. El efecto combinado favoreció el aprendizaje asistido por chatbots frente a las condiciones comparadoras para el desempeño educativo global (SMD=0,62; IC 95%: 0,37 a 0,87; $I^2=0\%$), aunque este hallazgo debe interpretarse con cautela porque solo se combinaron dos estudios y estos evaluaron diferentes manifestaciones del desempeño educativo. La certeza de la evidencia fue baja. La síntesis narrativa funcional identificó cuatro dominios educativos: adquisición de conocimientos y desempeño académico; razonamiento clínico y comunicación diagnóstica; apoyo a habilidades operativas y procedimentales; y escritura académica, manejo de evidencia y tareas reflexivas. El aprendizaje asistido por chatbots parece prometedor, pero la evidencia actual sigue siendo limitada, heterogénea y dependiente del contexto.

PALABRAS CLAVE: Aprendizaje asistido por chatbots; Inteligencia artificial; Educación odontológica; Revisión sistemática; Metaanálisis.

INTRODUCTION

The integration of artificial intelligence (AI) in dental education has witnessed exponential growth, particularly with the emergence of generative AI tools such as ChatGPT, Gemini, Copilot, and Claude. These generative AI chatbots, underpinned by large language models (LLMs), offer unprecedented opportunities for asynchronous, adaptive, and personalized learning. Early work in this area also showed that these systems could achieve strong performance on dental knowledge tasks and examination-style questions, which helped stimulate interest in their educational use and prompted broader discussion about their possible role in student learning (1,2).

As this field has evolved, the key question has shifted from whether chatbots can answer dental questions accurately to whether their use as learning supports produces meaningful educational benefits in real student populations. This distinction is important. High model performance on examinations or question banks may indicate technical capability, but it does not by itself demonstrate that chatbot-assisted learning improves student outcomes in authentic educational settings (1,2).

Despite the promising performance of generative AI models, their integration into dental education is not without challenges. Comparative educational research has already suggested that chatbot-

supported learning does not uniformly outperform conventional methods. In one study, conventional instruction produced significantly higher post-test scores than ChatGPT-assisted learning, underscoring the possibility that rapid access to generated content may not always translate into deeper understanding (3). In parallel, other investigations have shown that chatbot output may vary across dental disciplines and levels of cognitive demand, reinforcing the need for careful educational evaluation rather than uncritical adoption (4).

Beyond theoretical instruction, chatbots have shown value in enhancing students' clinical reasoning and communication skills. Student-centered applications have included virtual patient interviewing, where chatbot-based interactions were associated with improvements in perceived competence, participation, and practice opportunities in preclinical learning environments (5). Similarly, controlled educational research has shown that chatbot-assisted instruction may improve performance, learning motivation, and self-efficacy while also reducing cognitive load in operative skills training (6). These findings suggest that chatbots may function not merely as answer engines, but as educational supports capable of shaping how students engage with knowledge and skills development.

At the same time, the existing literature remains conceptually heterogeneous (7-9). Some studies have focused on chatbot-assisted learning as an educational intervention involving real dental students (3,5,6), whereas others have examined model performance on licensing or specialty examinations (10,11), or compared chatbot outputs against human respondents in more technical benchmarking contexts (12,13). Although all of these studies contribute to the wider conversation on AI in dental education, they do not answer the same pedagogical question. For a review specifically concerned with educational effectiveness, the most informative evidence

comes from studies in which real dental students are exposed to chatbot-assisted learning and their outcomes are measured directly against an explicit comparator or across a meaningful educational pre-post design.

This narrower focus is methodologically important. It reduces the risk of conflating educational effectiveness with model accuracy, improves conceptual coherence across included studies, and allows the review to concentrate on outcomes that matter most for curriculum development, including knowledge acquisition, skill development, self-efficacy, and educational experience. It also makes it possible to interpret the evidence in a way that is directly relevant to faculty decision-making and implementation in undergraduate dental training (14).

To date, most available studies are limited to single-center observational designs, often lacking comparators and focusing on narrow curricular areas. Moreover, even among student-centered studies, chatbot-assisted learning has been applied to diverse educational tasks, including didactic reinforcement, history-taking practice, scientific writing, and operative or diagnostic training. This diversity means that quantitative synthesis alone may not be sufficient to explain how these tools are being used pedagogically or which competency domains they appear to support most consistently. For this reason, a functional narrative synthesis is needed alongside meta-analysis to clarify the instructional roles, targeted competencies, and recurrent patterns of educational use across the included studies.

Accordingly, the objective of this systematic review and meta-analysis is to assess the effectiveness of chatbot-assisted learning interventions in improving academic and clinical outcomes among dental students. More specifically, this systematic review, meta-analysis, and functional narrative synthesis aims to evaluate the educa-

tional effectiveness of chatbot-assisted learning interventions in real dental student cohorts. The primary outcomes include academic performance, clinical reasoning and diagnostic learning, self-efficacy, and the development of clinical and communication competencies. Secondary outcomes include motivation, satisfaction, engagement, cognitive load, and perceived usefulness. By focusing on student-centered comparative educational evidence, this review seeks to provide a more methodologically coherent and pedagogically relevant basis for the integration of chatbot-based tools into dental curricula.

MATERIALS AND METHODS

PROTOCOL AND REGISTRATION

This systematic review, meta-analysis, and functional narrative synthesis was designed and reported in accordance with the PRISMA 2020 (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to ensure methodological rigor and transparency (15). The protocol was prospectively registered in the International Prospective Register of Systematic Reviews (PROSPERO) Database (CRD420251111207).

All methods, including eligibility criteria, search strategy, risk of bias assessment, and data synthesis plans, were defined a priori and followed as outlined in the protocol.

ELIGIBILITY CRITERIA

Studies were eligible for inclusion only if they involved real undergraduate or postgraduate dental students exposed to a chatbot-assisted educational

intervention. Eligible interventions included the use of chatbot-based tools (e.g., ChatGPT, Copilot, Bard, Gemini, Claude, Llama) for academic, diagnostic, clinical, or communication-related learning activities within dental education. Studies were required to report at least one student-level educational outcome, such as academic performance, diagnostic learning, self-efficacy, communication skills, cognitive load, motivation, user satisfaction, or perceived usefulness.

To ensure that the review addressed educational effectiveness rather than technical model benchmarking, eligible studies were required to include an explicit educational comparator. Acceptable comparators included conventional teaching, written materials, literature-based learning, non-AI instructional approaches, parallel control groups, or pre-post educational designs with student-level outcome assessment. Primary study types considered were randomized controlled trials (RCTs), quasi-experimental studies, and observational educational studies with a comparative or pre-post design.

Studies were excluded if they did not include real dental students; assessed chatbot performance only against examinations, experts, or human respondents without a student-centered educational intervention; focused exclusively on patient-facing applications; lacked a meaningful educational comparator or pre-intervention baseline; or failed to report extractable student-level outcomes. Editorials, narrative reviews, commentaries, letters, protocols, and purely technical validation studies were also excluded. There were no restrictions regarding language or publication date, provided that full-text access was available.

PICO FRAMEWORK

This review was guided by the following PICO elements:

Population (P): Real undergraduate or postgraduate dental students.

Intervention (I): Chatbot-assisted educational interventions used for academic, diagnostic, clinical, or communication training in dental education.

Comparator (C): Conventional teaching methods, written materials, literature-based learning, parallel control conditions, or pre-post educational designs.

Outcomes (O): Academic performance, diagnostic learning, self-efficacy, clinical or communication-related competencies, motivation, user satisfaction, engagement, cognitive load, and perceived usefulness.

This focused framework was adopted to ensure conceptual alignment between the review question and the included evidence base, specifically targeting the educational effectiveness of chatbot-assisted learning rather than the standalone technical performance of chatbot systems.

INFORMATION SOURCES AND SEARCH STRATEGY

A comprehensive and systematic literature search was performed across PubMed/MEDLINE, Scopus, and Embase. The final search in all databases was conducted on May 30, 2026. Embase was searched through Elsevier Embase.com. The search strategy incorporated both controlled vocabulary (e.g., MeSH and Emtree terms, when applicable) and free-text terms to ensure broad coverage of relevant literature. Keywords included: chatbot, ChatGPT, generative artificial intelligence, large language model, LLM, artificial intelligence in education, AI-assisted learning, conversational

agent, educational chatbot, dental chatbot, dental students, dental education, oral health education, intelligent tutoring systems, virtual teaching assistant, machine learning in dentistry, adaptive learning, asynchronous learning, digital learning tools, diagnostic AI in dentistry, and simulation-based learning. Boolean operators (AND, OR), truncation (e.g., educat), and proximity operators were applied to refine the search strategy. Search strings were adapted to the specific syntax of each database. The full search strategies for all databases are provided in Supplementary Table S1.

All retrieved records were exported to EndNote 21, where duplicates were identified using automated duplicate detection and then manually verified based on title, authors, year, journal, and DOI. Reference lists of all included studies were manually screened to identify additional eligible articles not captured through electronic searches. Grey literature sources, trial registries, preprint servers, and dissertation databases were not systematically searched.

STUDY SELECTION

Two reviewers independently screened the titles and abstracts retrieved from the search. Potentially eligible articles were retrieved in full and assessed against the inclusion and exclusion criteria. Particular attention was paid during full-text screening to confirm the presence of real dental student participants, a chatbot-assisted educational intervention, and a comparator or pre-post educational design consistent with the review objective. Disagreements were resolved through discussion or consultation with a third reviewer. The selection process was documented using a PRISMA flow diagram.

DATA EXTRACTION

Data were extracted independently by two reviewers using a standardized and pre-tested

form. Extracted data included study characteristics, participant demographics, chatbot model and instructional purpose, comparator, measured outcomes, assessment tools, and statistical results. Additional data relevant to the functional narrative synthesis were also extracted, including the primary instructional use of the chatbot, the targeted competency domain, the educational setting, and the role of the chatbot within the learning process (e.g., explanatory support, virtual patient interaction, writing assistance, or operative guidance). Discrepancies were resolved through consensus or arbitration.

OUTCOME MEASURES

Primary outcomes included academic performance (e.g., test scores, structured knowledge assessments), diagnostic learning, self-efficacy, and the development of clinical and communication competencies. Secondary outcomes included user satisfaction, engagement, cognitive load, motivation, and perceived usefulness. Outcomes related solely to chatbot accuracy against expert or faculty benchmarks were not treated as primary evidence of educational effectiveness and were considered only when directly linked to student-centered educational interventions.

RISK OF BIAS AND EVIDENCE CERTAINTY

Risk of bias was assessed according to study design. Randomized controlled trials and randomized comparative studies were evaluated using the Cochrane RoB 2 tool (16), with judgments made at the outcome level for the main student educational outcome reported by each study. Domain-level judgments, signaling-question responses, and rationales were recorded.

Non-randomized educational studies were assessed using the appropriate Joanna Briggs Institute (JBI) critical appraisal checklist according to study design (17). Quasi-experimental or

pre-post intervention studies were assessed using the JBI checklist for quasi-experimental studies, whereas analytical cross-sectional or comparative observational studies were assessed using the JBI checklist for analytical cross-sectional studies when applicable.

Certainty of evidence was evaluated using the GRADE approach, considering risk of bias, inconsistency, indirectness, imprecision, and publication bias (18). Full item-level RoB 2 and JBI assessments are provided in Supplementary Tables S2 and S3.

DATA SYNTHESIS AND EXPLORATORY META-ANALYSIS

Quantitative synthesis was planned using random-effects meta-analysis when at least two studies reported student-level continuous outcomes considered sufficiently comparable for standardized effect-size estimation. Standardized mean differences (SMDs) with 95% confidence intervals (95% CI) were pre-specified for continuous outcomes to accommodate differences in measurement scales. Heterogeneity was assessed using the I^2 and Tau^2 statistics.

Because the included studies were expected to vary in instructional format, comparator type, study design, and outcome domain, quantitative pooling was restricted to outcomes judged to be both statistically extractable and conceptually appropriate for synthesis. In particular, objective performance-based outcomes were considered for pooling when they reflected student educational performance after chatbot-assisted learning. However, differences in the educational construct assessed by each study were explicitly considered when interpreting the pooled estimate.

The primary quantitative synthesis was therefore treated as exploratory. It was intended to provide a cautious estimate of overall student

educational performance rather than evidence for a single homogeneous educational outcome. Operative/procedural skill performance and knowledge examination performance were considered different manifestations of student performance and not identical constructs. Accordingly, the pooled estimate was interpreted alongside the study-level findings, design characteristics, and certainty-of-evidence assessment.

Studies with substantial methodological, comparator, or outcome heterogeneity were not combined quantitatively. Subgroup analyses by educational domain and type of comparator were planned to explore potential sources of heterogeneity, but they were not performed because too few studies were available within each subgroup. When studies lacked sufficient or comparable quantitative data, meta-analysis was avoided. In such cases, findings were summarized narratively and stratified according to educational domain and study design, following the SWiM (Synthesis Without Meta-analysis) reporting framework.

In addition to the quantitative synthesis, a functional narrative synthesis was conducted to describe how chatbot-assisted learning was implemented across the included studies. For each study, data were extracted on the primary instructional purpose of the chatbot, the educational task in which it was used, the targeted student competency, the comparator or pre-post framework, and the main student-level outcome.

Studies were grouped according to their dominant educational function. The coding process was guided by the review objective and by the extracted intervention characteristics. When a study addressed more than one educational dimension, classification was based on the principal intended learning function of the chatbot intervention rather than on secondary or exploratory outco-

mes alone. This approach was used to identify recurrent patterns of chatbot-assisted learning across heterogeneous educational interventions and to support interpretation of findings that were not suitable for quantitative pooling.

Quantitative syntheses were conducted using Review Manager (RevMan) version 5.4.1 using a random-effects model and standardized mean differences (Hedges *g*) for continuous outcomes. Sensitivity checks, cross-validation of pooled estimates and heterogeneity metrics, and figure rendering were performed in R version 4.3.2 using standard meta-analytic plotting functions. The summary statistics used for effect-size estimation and quantitative pooling are provided in Supplementary Table S4.

RESULTS

STUDY SELECTION

A total of 2800 records were identified through systematic searches of PubMed/MEDLINE, Scopus, and Embase. After duplicate removal, records were screened by title and abstract, and studies were excluded when they were unrelated to chatbot-assisted learning in dental education, did not involve real dental students, lacked student-centered educational outcomes, focused exclusively on patient-facing applications, or corresponded to non-eligible publication types such as editorials, narrative reviews, and commentaries. Of the 39 full-text articles assessed for eligibility, 8 studies met the final inclusion criteria and were included in the qualitative synthesis (3-6,19-23). The 31 full-text studies excluded after application of the revised eligibility criteria, together with the primary reasons for exclusion, are listed in Supplementary Table S5. Among the included studies, 2 provided sufficiently comparable continuous student-level outcomes with extractable

summary statistics and were therefore included in the exploratory meta-analysis (6,22). The study selection process is summarized in Figure 1.

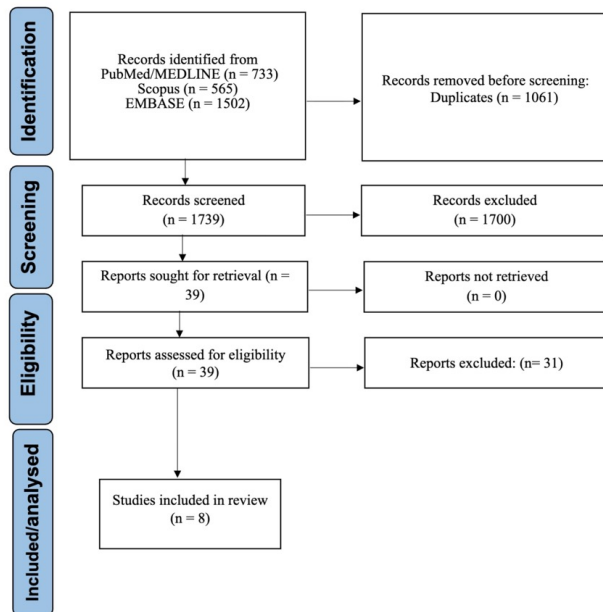


Figure 1. Study selection process.

CHARACTERISTICS OF INCLUDED STUDIES

The final review set comprised 8 studies published between 2024 and 2026, representing a total of 716 real dental students exposed to chatbot-assisted educational interventions (3-6,19-23). All included cohorts were undergraduate dental student cohorts and used comparative educational designs, including randomized controlled trials, mixed-methods interventions, observational pre-post studies, and comparative cross-sectional designs. A detailed overview of study characteristics is presented in Table 1.

The included studies varied in methodological structure but all evaluated chatbot-assisted learning within a clearly educational context. Three studies used randomized or randomized comparative designs (3,6,23), two used mixed-methods comparative/intervention designs (5,22), one was an observational pilot study with a pre-post struc-

ture (19), and the remaining studies used comparative educational designs (20,21). Comparators included conventional teaching, written materials, literature-based learning, video-based learning, traditional in-person workshops, and pre-post educational assessment. These design differences contributed to substantial heterogeneity in both intervention format and outcome reporting, as shown in Table 1.

For interpretation, findings were considered according to study design and strength of causal inference. Randomized or parallel comparative studies were interpreted as providing stronger evidence regarding educational effectiveness, whereas uncontrolled pre-post or primarily self-report studies were considered more informative for feasibility, learner experience, perceived competence, and hypothesis generation. Therefore, pre-post improvements were not treated as equivalent to controlled evidence of comparative effectiveness.

The chatbot interventions were dominated by the ChatGPT family of models, including ChatGPT, ChatGPT-3.5, ChatGPT-4o, and GPT-4-based architectures, while one study evaluated Cogniti, an institutional virtual patient system built on GPT-4 Turbo (5). Across the included studies, chatbot-assisted learning was applied to a range of educational functions, including knowledge acquisition, history-taking practice, operative skills training, scientific writing, and AI-driven virtual patient simulation. This diversity in instructional purpose later informed the functional narrative synthesis. Full details on intervention characteristics, comparators, and outcome domains are provided in Table 1.

Across the included studies, the most frequently reported outcomes were academic performance, self-efficacy, clinical or communication-related competence, motivation, cognitive load, and perceived usefulness or satisfaction. However, outcome structures were not uniform. Some

studies emphasized objective performance-based measures, such as post-test scores, knowledge examination grades, or operative skill performance (3,6,22,23), whereas others focused primarily on student-reported educational benefits, including perceived competence, confidence, usability, and motivational value (5,19-21). This heterogeneity limited the number of studies suitable for quanti-

tative pooling and supported the use of both an exploratory meta-analysis and a complementary functional narrative synthesis. A study-by-study breakdown of study characteristics is shown in Table 1, whereas detailed outcome-level results, assessment methods, sample sizes, and outcome classifications are provided in Supplementary Table S6.

Table 1. Characteristics of the included studies.

Study	Country	Student population	Design	Chatbot / platform	Comparator	Main outcome domain(s)
Bhatia <i>et al.</i> (2024) (3)	India	100 undergraduate dental students	Randomized comparative educational study	ChatGPT	Conventional teaching with textbooks after a common lecture	Academic performance
Or <i>et al.</i> (2025) (5)	Australia	31 dental students	Mixed-methods intervention study	Cogniti (GPT-4 Turbo-based virtual patient chatbot)	Traditional in-person history-taking workshops	Clinical reasoning and diagnostic communication training; perceived competence; engagement
Huang <i>et al.</i> (2025) (6)	China	187 undergraduate dental students	Randomized controlled trial	ChatGPT-3.5	Video-based learning only (blank control group)	Operative/procedural skill performance; cognitive load; self-efficacy; motivation
Prinz <i>et al.</i> (2025) (19)	Germany	92 undergraduate dental students	Observational pilot study with pre-post design	GPT-4-turbo-based virtual patient e-learning tool	Preintervention vs postintervention self-assessment	Clinical reasoning and diagnostic communication training; self-efficacy
Roganović (2024) (20)	Serbia	104 dental students	Comparative educational study	ChatGPT	Use of ChatGPT vs description-only exposure vs no exposure	Knowledge acquisition and academic performance; expectations toward AI
Saravia-Rojas <i>et al.</i> (2024) (21)	Perú	55 fourth-year undergraduate dental students	Comparative sequential educational study	ChatGPT	Conventional literature search in PubMed/scientific articles	Academic writing, evidence handling, and reflective task support
Kavadella <i>et al.</i> (2024) (22)	Cyprus	77 second-year dental students	Mixed-methods comparative study	ChatGPT-3.5	Literature research group	Knowledge acquisition and academic performance; knowledge examination
Keles (2026) (23)	Turkey	70 fourth-year dental students	Randomized pre-test/post-test educational study	ChatGPT-4o	Written materials	Knowledge acquisition and academic performance; motivation/perceptions

PRIMARY QUANTITATIVE SYNTHESIS

Only two studies provided objective continuous student-level outcomes with extractable summary statistics and were therefore included in the exploratory meta-analysis (6,22). These studies evaluated chatbot-assisted learning in real dental student cohorts using explicit educational comparators, but they assessed different manifestations of educational performance. Huang *et al.* (6) compared ChatGPT-assisted operative skills learning plus video instruction with video-based learning alone, whereas Kavadella *et al.* (22) compared ChatGPT-supported assignment preparation with traditional literature-based learning. Accordingly, the pooled construct was interpreted as an exploratory estimate of overall student educational performance rather than as evidence for a single homogeneous educational outcome.

In Huang *et al.* (6), the chatbot-assisted group demonstrated significantly higher performance than the control group, with mean scores of 73.12 (SD 10.06) versus 65.54 (SD 12.48), respectively. In Kavadella *et al.* (22), students assigned to the ChatGPT group also achieved higher knowledge examination grades than those in the literature research group, with mean scores of 7.54 (SD 1.18) and 6.94 (SD 1.12), respectively. These raw group means were transformed into standardized mean differences for quantitative pooling. Additional outcome-level details for

all included studies are provided in Supplementary Table S6.

A random-effects meta-analysis using standardized mean differences (Hedges *g*) showed an overall effect favoring chatbot-assisted learning over comparator conditions (SMD=0.62, 95% CI: 0.37 to 0.87). No statistical heterogeneity was detected ($\tau^2=0.00$; $I^2=0\%$; $Q=0.29$; $df=1$; $p=0.593$). However, this finding should be interpreted cautiously because only two studies were available for pooling, limiting the ability of heterogeneity statistics to detect between-study differences. Clinical, methodological, and educational heterogeneity therefore remained relevant despite the absence of detected statistical heterogeneity. The forest plot of the exploratory meta-analysis is presented in Figure 2.

Although both studies favored chatbot-assisted learning, the pooled estimate should be interpreted with caution. One study assessed operative/procedural skill performance and the other assessed knowledge examination performance. For this reason, the quantitative synthesis reflects a broad and exploratory construct of overall student educational performance rather than a single narrowly defined learning outcome. The quantitative findings were therefore interpreted alongside the narrative synthesis, the functional narrative synthesis, and the certainty-of-evidence assessment.

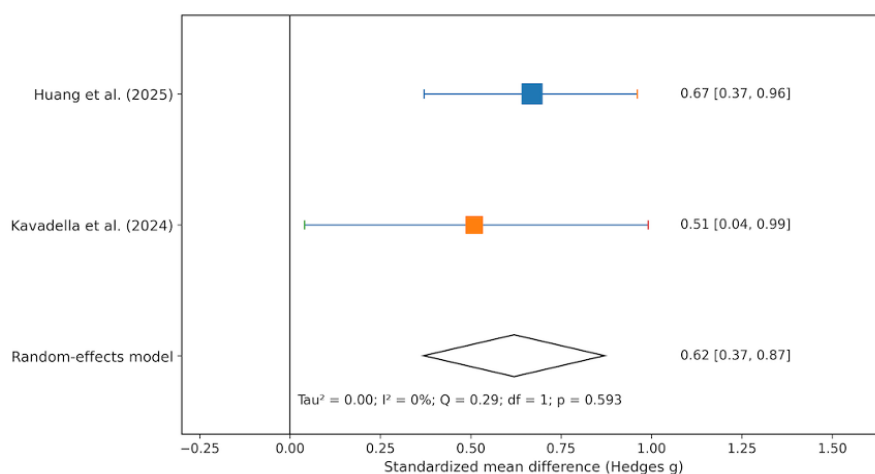


Figure 2. Forest plot of the exploratory meta-analysis comparing chatbot-assisted learning versus comparator conditions in real dental student cohorts. Effect sizes are presented as standardized mean differences (Hedges g) with 95% confidence intervals using a random-effects model. The included studies were Huang *et al.* (6) and Kavadella *et al.* (22). The pooled estimate should be interpreted cautiously because only two studies were included and they assessed different manifestations of student educational performance.

FUNCTIONAL NARRATIVE SYNTHESIS

The functional narrative synthesis grouped the 8 included studies according to the dominant educational purpose of the chatbot intervention rather than by chatbot platform or study design. Four functional domains were identified: knowledge acquisition and academic performance; clinical reasoning and diagnostic communication training; operative and procedural skill support; and academic writing, evidence handling, and reflective task support. This classification is summarized in Figure 3.

The largest domain was knowledge acquisition and academic performance, which included Bhatia *et al.* (3), Roganović [20], Kavadella *et al.* (22), and Keles (23). These studies used chatbots mainly as learning aids for didactic reinforcement, content review, assignment preparation, or structured educational support. However, the direction of findings was not uniform. Some studies reported better performance or favorable educational responses with chatbot support (20,22,23), whereas Bhatia *et al.* (3) found higher performance with conventional instruction. This indica-

tes that chatbot-assisted learning in knowledge-oriented tasks may be beneficial when embedded in a structured learning activity, but it does not consistently outperform conventional teaching.

The second domain comprised clinical reasoning and diagnostic communication training, represented by Or *et al.* (5) and Prinz *et al.* (19). In these studies, chatbots functioned primarily as virtual patients or simulated conversational partners, providing opportunities for repeated practice in history-taking, diagnostic questioning, and clinical communication. The findings were generally favorable, particularly for perceived competence, confidence, engagement, and self-assessed improvement. Nevertheless, the evidence in this domain was mainly based on subjective or self-reported outcomes and lacked strong objective comparative assessment, limiting the strength of inference regarding measurable competency gains.

Operative and procedural skill support was represented by Huang *et al.* (6). This study provided the clearest evidence of a direct performance-related educational benefit, showing that chatbot-assisted instruction combined with video-based

learning improved operative skills performance compared with video-based learning alone. The intervention also showed favorable effects on cognitive load, self-efficacy, and motivation. Although this domain was supported by only one study, its randomized design and objective skills-oriented assessment make it one of the more informative contributions to the current evidence base.

The final domain, academic writing, evidence handling, and reflective task support, was represented by Saravia-Rojas *et al.* (21). In this study, ChatGPT was used as a support tool for academic and evidence-related tasks, but conventional literature-based learning performed better across several rubric-based criteria. This finding suggests that chatbot use may not be sufficient, and may even be less effective, when the educational task requires critical evidence retrieval, appraisal, argument construction, and reflective integra-

tion. Therefore, this domain provides a cautionary signal that chatbot-assisted learning should not be assumed to improve all forms of academic work.

Overall, the functional narrative synthesis indicates that chatbot-assisted learning in dental education is best understood as a set of distinct educational applications rather than as a single homogeneous intervention. The most consistent benefits appeared when chatbots were used as adjunctive scaffolds for structured learning, simulation, or procedural support. In contrast, evidence was weaker or more mixed when chatbots were used for independent knowledge acquisition or evidence-based academic tasks. This functional heterogeneity helps explain why only a limited number of studies were suitable for quantitative pooling and supports interpreting the meta-analysis as a broad estimate of educational performance rather than as evidence of uniform effectiveness across all chatbot-assisted learning domains.

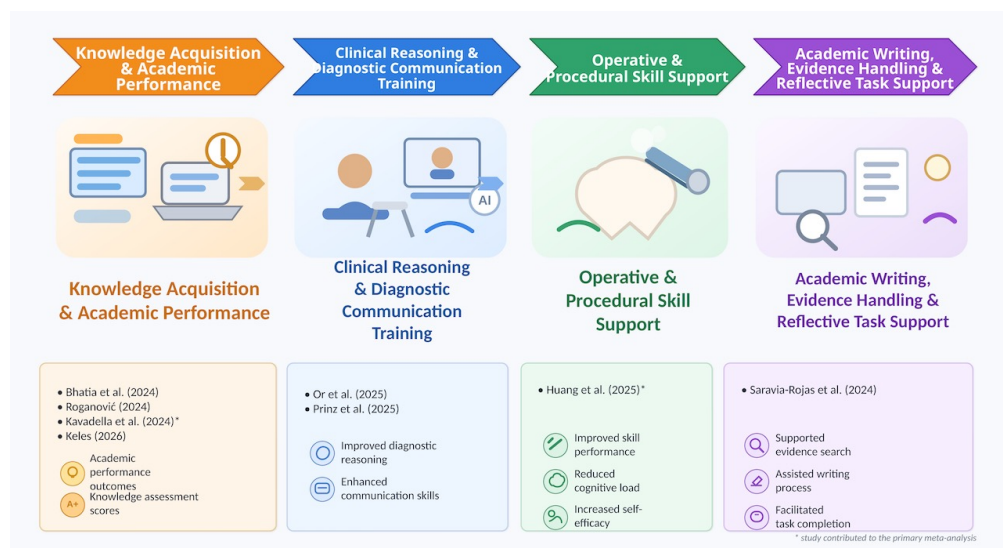


Figure 3. Functional narrative synthesis of chatbot-assisted learning in dental education. Included studies were classified according to their dominant educational purpose into four domains: knowledge acquisition and academic performance (3,20,22,23); clinical reasoning and diagnostic communication training (5,19); operative and procedural skill support [6]; and academic writing, evidence handling, and reflective task support (21). Studies marked with an asterisk contributed to the exploratory meta-analysis.

RISK OF BIAS AND EVIDENCE CERTAINTY

Risk of bias was assessed according to study design, using RoB 2.0 [16] for the randomized studies and the JBI checklist (17) for the non-randomized studies.

Overall, the methodological profile of the included evidence was mixed. The randomized studies generally showed more robust design features, whereas the non-randomized studies were more frequently limited by self-reported outcomes, non-equivalent comparators, sequential designs, or expectation-related confounding. The study-level assessments are summarized separately for randomized and non-randomized studies in Table 2 and Table 3, respectively.

Among the randomized studies, Huang *et al.* (6) and Keles (23) were judged to have some concerns overall. Both studies reported appropriate randomization procedures, clearly defined intervention and comparator conditions, and objective or structured educational outcome assessment. However, some concerns remained in the selective reporting domain because the availability of a prespecified analysis plan for all reported educational outcomes was unclear. By contrast, Bhatia *et al.* (3) was judged to have high risk of bias overall, because internal inconsisten-

cies were noted between the text, participant flow, and numerical reporting, and because allocation concealment and blinding were not described. These randomized-study judgments are summarized in Table 2, with item-level details provided in Supplementary Table S2.

Among the non-randomized studies, methodological quality varied from moderate to low. Kavadella *et al.* (22), Or *et al.* (5), and Prinz *et al.* (19) were judged to have moderate methodological quality overall. Kavadella *et al.* used parallel comparison groups and an objective examination outcome, although some participants did not complete the final assessment. Or *et al.* clearly described the intervention and student cohort, but the main outcomes were largely self-reported and not supported by a strong concurrent control structure. Prinz *et al.* reported a clearly defined educational intervention, but the pre-post design lacked a control group and relied mainly on self-perceived improvement. By contrast, Saravia-Rojas *et al.* (21) and Roganović (20) were judged to have lower methodological quality, mainly because of sequential within-cohort comparison in the former and self-selection plus expectation-related confounding in the latter. A summary methodological appraisal of the non-randomized studies, based on the most relevant JBI domains, is presented in Table 3.

Table 2. Risk of bias assessment of randomized studies using RoB 2.0.

Study	Randomization process	Deviations from intended interventions	Missing outcome data	Measurement of the outcome	Selection of the reported result	Overall judgment
Bhatia <i>et al.</i> (2024) (3)	Some concerns	High risk	Some concerns	Some concerns	High risk	High risk
Huang <i>et al.</i> (2025) (6)	Low risk	Low risk	Low risk	Low risk	Some concerns	Some concerns
Keles (2026) (23)	Low risk	Low risk	Low risk	Low risk	Some concerns	Some concerns

Table 3. Summary methodological appraisal of non-randomized studies based on JBI criteria. This table presents a condensed summary of the main methodological strengths and limitations of the included non-randomized studies, derived from the JBI checklist domains most relevant to comparative educational research. It should be interpreted as an overall methodological profile rather than as a full item-by-item JBI scoring matrix.

Study	Main methodological strengths	Main methodological concerns	Overall methodological quality
Or <i>et al.</i> (2025) (5)	Clearly described intervention, student cohort, and educational setting; appropriate mixed-methods design for exploratory educational evaluation	Main outcomes relied largely on self-reported perceived competence and experience; no strong concurrent comparator for the principal educational outcome	Moderate
Prinz <i>et al.</i> (2025) (19)	Clearly defined intervention and pre-post framework; focused educational objective; transparent reporting of student perceptions	No control group; reliance on self-perceived improvement; dropout reduced confidence in stability of findings	Moderate
Roganović (2024) (20)	Clear description of participant groups and quiz-based outcome; relevant educational question	Self-selection into groups, possible expectation-related confounding, and limited ability to isolate the specific effect of chatbot use	Low
Saravia-Rojas <i>et al.</i> (2024) (21)	Clearly structured educational task and rubric-based outcome assessment; direct comparison with a conventional strategy	Sequential within-cohort design, substantial risk of order and carryover effects, and unequal completion across phases	Low
Kavadella <i>et al.</i> (2024) (22)	Parallel comparison groups; objective knowledge examination outcome; clearly described intervention and comparator	Incomplete final examination data and limited control of potential confounding related to group differences	Moderate

Taken together, these assessments indicate that the evidence base includes a small number of comparatively stronger controlled studies, alongside several informative but more bias-prone educational studies. This pattern should be considered when interpreting both the pooled estimate and the domain-specific findings from the functional narrative synthesis.

CERTAINTY OF EVIDENCE (GRADE)

Certainty of evidence was assessed using GRADE (18), but only for outcomes that remained central to the revised review question and were supported by the final included studies. Because the evidence base was small, methodologically mixed, and outcome-heterogeneous, certainty was generally limited by risk of bias, indirectness, and

imprecision. A summary of certainty judgments is provided in Table 4.

For the primary exploratory meta-analytic outcome, overall student educational performance, the certainty of evidence was judged as low. The evidence was based on only two studies, one randomized study (6) and one non-randomized comparative study (22). Although the pooled estimate favored chatbot-assisted learning, confidence in this estimate was limited by study limitations, indirectness, and imprecision. Indirectness was considered serious because the two pooled studies assessed different manifestations of educational performance, namely operative/procedural skill performance and knowledge examination performance. Imprecision was also considered because only two studies contributed to the pooled

estimate. Publication bias could not be formally assessed because of the very small number of studies included in the quantitative synthesis.

For clinical reasoning and diagnostic communication training, the certainty of evidence was judged as very low, because the evidence was based mainly on Or *et al.* (5) and Prinz *et al.* (19), both of which suggested educational benefit but relied largely on self-perceived improvement and lacked robust parallel controlled comparisons. For self-efficacy, the certainty was also judged as very low, because the evidence came from a small number of studies with mixed designs and incomplete pooled-ready quantitative reporting.

For academic writing, evidence handling, and reflective task support, the certainty of evidence was judged as very low, because this domain was informed essentially by a single high-risk study, Saravia-Rojas *et al.* (21), in which the conventional

strategy outperformed ChatGPT on several rubric-based criteria, but the sequential design prevented confident causal interpretation.

Secondary outcomes such as motivation, cognitive load, and perceived usefulness were not supported by a sufficiently large and methodologically consistent body of evidence to justify ratings above low or very low, depending on the domain. In particular, cognitive load and motivation were informed primarily by Huang *et al.* (6) and, in a more limited perceptual form, by Keles (23), whereas perceived usefulness was derived largely from self-report data across non-randomized educational studies.

Overall, the certainty profile indicates that chatbot-assisted learning in dental education is promising, but the current evidence remains too limited and heterogeneous to support strong conclusions across all educational domains.

Table 4. GRADE summary of findings for the main educational outcome domains.

Outcome domain	Contributing studies / sample size	Main finding	Starting certainty	Domains downgraded	Final certainty (GRADE)	Justification
Overall student educational performance	Huang <i>et al.</i> (6) and Kavarella <i>et al.</i> (22); 264 students	Exploratory meta-analysis favored chatbot-assisted learning (SMD=0.62, 95% CI: 0.37 to 0.87)	High for randomized evidence; lower confidence due to mixed randomized and non-randomized evidence	Risk of bias: -1; indirectness: -1; imprecision: -1; publication bias: not formally assessable	Low	Downgraded because the evidence was based on only two studies, including one randomized and one non-randomized study. Indirectness was serious because the pooled studies assessed different manifestations of educational performance: operative/procedural skill performance and knowledge examination performance. Imprecision was considered because only two studies contributed to the estimate. Publication bias could not be formally assessed.
Clinical reasoning and diagnostic communication training	Or <i>et al.</i> (5) and Prinz <i>et al.</i> (19); 123 students	Findings were directionally favorable, mainly for perceived competence, confidence, and self-reported improvement	Low for non-randomized evidence	Risk of bias: -1; indirectness: -1; imprecision: -1; publication bias: not formally assessable	Very low	Downgraded because the evidence came from non-randomized or uncontrolled educational designs, relied mainly on subjective outcomes, lacked robust concurrent controls, and included few students. Publication bias could not be formally assessed.
Self-efficacy	Huang <i>et al.</i> (6) and Prinz <i>et al.</i> (19); 279 students	Both studies suggested improvement in self-efficacy or confidence	High for randomized evidence; lower confidence due to mixed randomized and non-randomized evidence	Risk of bias: -1; indirectness: -1; imprecision: -1; publication bias: not formally assessable	Very low	Downgraded because the evidence came from mixed study designs, outcome definitions were not fully consistent, and quantitative reporting was insufficient for pooling. The domain also relied partly on self-reported measures, and publication bias could not be formally assessed.
Academic writing, evidence handling, and reflective task support	Saravia-Rojas <i>et al.</i> (21); 55 students	Conventional literature based learning outperformed ChatGPT on several rubric-based criteria	Low for non-randomized evidence	Risk of bias: -1; imprecision: -1; publication bias: not formally assessable	Very low	Downgraded because the evidence was based on a single sequential study with low methodological quality, risk of order and carryover effects, and a small sample size. Publication bias could not be formally assessed.
Motivation, cognitive load, and perceived usefulness	Huang <i>et al.</i> (6), Keles (23), and supportive self-report evidence from non-randomized studies; at least 257 students for motivation/cognitive load outcomes	Evidence suggested favorable trends, particularly for reduced cognitive load and improved motivation in specific contexts	High for randomized evidence; lower confidence for self-reported evidence from non-randomized studies	Risk of bias: -1; indirectness: -1; imprecision: -1; publication bias: not formally assessable	Low to very low	Downgraded because outcome definitions and measurement instruments varied across studies, some outcomes were self-reported, and pooled estimates were not available for most secondary outcomes. Publication bias could not be formally assessed.

DISCUSSION

This systematic review, meta-analysis, and functional narrative synthesis examined the educational effectiveness of chatbot-assisted learning in real dental student cohorts, focusing only on studies with explicit educational comparators or meaningful pre-post designs. Compared with the broader literature on chatbot benchmark performance, this narrower synthesis offers a more educationally relevant view of how these tools function in authentic learning settings. Across 8 studies involving 716 students, the evidence suggested that chatbot-assisted learning may improve overall student educational performance, although the effect should be interpreted cautiously because the underlying outcomes were heterogeneous and the certainty of evidence was low. Consistent with the functional narrative synthesis presented in the Results, chatbot-assisted learning has not been used in a single uniform way; rather, it has been implemented across four educational domains: knowledge acquisition and academic performance; clinical reasoning and diagnostic communication training; operative and procedural skill support; and academic writing, evidence handling, and reflective task support.

The primary quantitative finding favored chatbot-assisted learning over comparator conditions for overall student educational performance. However, this result should be interpreted as exploratory and context-dependent. The pooled estimate was derived from only two studies and combined different manifestations of educational performance: operative/procedural skill performance in one study and knowledge examination performance in the other (6,22). Therefore, the meta-analysis should not be interpreted as demonstrating general superiority of chatbot-assisted learning across dental education. Rather, it suggests a preliminary positive signal in selected objective performance-based outcomes, which

requires confirmation in larger, methodologically stronger, and more outcome-consistent studies.

Within the domain of knowledge acquisition and academic performance, the findings were clearly mixed. Conventional instruction outperformed ChatGPT-assisted teaching in one randomized study, suggesting that rapid access to generated explanations may not necessarily translate into deeper conceptual understanding in all contexts (3). By contrast, another study found higher knowledge examination scores in the ChatGPT group than in the literature-search group, indicating that chatbot support may be helpful when the educational task requires quick synthesis and structured retrieval of information (22). This mixed picture is compatible with the broader literature showing that LLM-based tools can perform strongly in some dental educational settings while remaining highly dependent on task design, prompting strategy, and curricular context (24). Related comparative work has also shown that chatbot performance can vary by subject complexity and by whether the task privileges recall, interpretation, or clinically contextualized reasoning (25,26). Taken together, these findings suggest that chatbot-assisted learning is unlikely to replace conventional teaching, but it may enhance academic performance when used as a scaffold for guided study, clarification, and structured review. The newer randomized study by Keles reinforces the relevance of this domain by examining chatbot-supported learning in periodontal-systemic teaching, although the broader lesson remains that chatbot utility depends on how the tool is positioned pedagogically rather than on the technology alone (23).

The review also identified a distinct domain of clinical reasoning and diagnostic communication training, in which the main value of chatbot-assisted learning appeared to lie in practice opportunities, structured dialogue, and confidence-building rather than in objective performance superiority.

Virtual patient interaction and AI-driven simulated clinical encounters were associated with improved perceived competence, better organization of questioning, and more opportunities for repeated low-risk practice (5,19). These findings are educationally meaningful because they address common challenges in dental training, including limited time for individualized supervision and restricted opportunities for repeated rehearsal of communication skills. At the same time, broader studies outside the final review set have shown that chatbots can perform strongly in structured diagnostic tasks, particularly when prompts are well specified and the clinical problem is text-based (27). Those data are useful as contextual evidence, but they should not be conflated with educational effectiveness in student cohorts, because strong AI performance does not automatically mean that student learning improves to the same extent.

A third and particularly important domain was operative and procedural skill support. In this review, the strongest evidence for a direct educational effect came from the randomized controlled trial by Huang *et al.* (6), in which chatbot-assisted instruction was associated with higher performance in operative skills training, lower cognitive load, and better self-efficacy and motivation than video-based learning alone. This study is notable because it moves beyond knowledge recall and into technically oriented preclinical education. Its findings suggest that chatbots may support procedural learning by structuring information, reducing extraneous processing, and helping learners focus on relevant task elements. This is one of the clearest indications from the current evidence base that chatbot-assisted learning may have value not only in didactic reinforcement, but also in skills-oriented training. However, evidence from adjacent educational tasks remains mixed. Automated essay scoring and related evaluative applications, for example, suggest that LLMs may assist with feedback processes while still showing limitations in nuanced educational judgment (28).

Similarly, work examining the reliability of ChatGPT in orthodontic information delivery has raised concerns about consistency and content quality, reinforcing the need for human supervision when these tools are used in clinically sensitive educational contexts (29).

The final functional domain, academic writing, evidence handling, and reflective task support, produced a more cautionary pattern. In the study by Saravia-Rojas *et al.* (21), the conventional literature-search strategy outperformed ChatGPT across several rubric-based criteria, including evidence use, argument evaluation, and generation of alternatives. This finding is highly informative because it shows that chatbot-assisted learning is not automatically superior when students are expected to search, evaluate, and integrate evidence critically. That interpretation is strengthened by external findings showing persistent limitations of current AI systems in image-dependent and multimodal dental tasks, where performance may deteriorate substantially when reasoning must be linked to radiographic or visual information (30). At the same time, broader analyses of LLMs in standardized dental testing suggest that these tools can still serve as useful adjuncts for review and preparation, particularly in text-dominant formats (31).

Evidence from reflective assignments and qualitative educational tasks also suggests that chatbots may help structure initial ideas or organize material, but do not replace deeper interpretive work by students and educators (32). Thus, in this domain, chatbots are better understood as supportive tools rather than as replacements for rigorous evidence-based academic work.

An important implication of this review is that the educational value of chatbots is likely to be context-dependent. The strongest benefits were observed when the chatbot served as a scaffold, a practice partner, or a supportive adjunct, rather

than as an autonomous substitute for faculty guidance or evidence-based study. This interpretation is also consistent with the broader literature in oral and maxillofacial radiology and other image-intensive fields, where LLM-based systems may perform reasonably in text-based questioning but remain less dependable when interpretation requires integration of visual findings, contextual nuance, or specialty-specific judgment (33). In endodontic diagnosis, for example, strong AI performance against students has been reported in structured settings (34), but such findings again reflect benchmarking rather than direct evidence of learning gain. In this sense, the current review supports a model of assisted integration, not unrestricted substitution.

The interpretation of these findings should also account for study design. Randomized and parallel comparative studies provide stronger evidence regarding the comparative effectiveness of chatbot-assisted learning, whereas uncontrolled pre-post studies are more vulnerable to testing effects, maturation, regression to the mean, expectation-related bias, and other contextual influences. Accordingly, improvements observed in pre-post designs were interpreted as supportive or hypothesis-generating rather than as definitive evidence that chatbot-assisted learning is superior to conventional educational approaches.

This review has several strengths. First, it deliberately restricted inclusion to studies involving real dental student cohorts, thereby avoiding the conceptual conflation of educational effectiveness with technical benchmark performance. Second, it required explicit educational comparators or meaningful pre-post designs, which improved the pedagogical coherence of the evidence base. Third, it combined exploratory meta-analysis with a functional narrative synthesis, allowing both quantitative estimation and interpretive understanding of how chatbots are being used across different educational purposes. Finally, the methodolo-

gical appraisal was tailored to study design, using RoB 2.0 (16), JBI (17), and GRADE (18), which strengthened the rigor of interpretation.

Several limitations must also be acknowledged. The evidence base was small, and all included cohorts were undergraduate dental students, limiting generalizability to postgraduate training. Most studies were single-center and short-term, and the methodological quality was variable. Only two studies contributed to the exploratory meta-analysis. Although no statistical heterogeneity was detected, this finding should be interpreted cautiously because heterogeneity statistics are unreliable and underpowered when based on only two studies. Moreover, the pooled construct combined different manifestations of student performance, including operative/procedural skill performance and knowledge examination performance. Several outcomes, particularly in the communication and self-efficacy domains, were based mainly on self-report rather than standardized objective assessment. In addition, there was substantial heterogeneity in intervention format, comparator type, educational domain, study design, and outcome measurement, which limited the scope of quantitative pooling and reduced certainty of evidence. Because grey literature sources, trial registries, preprint servers, and dissertation databases were not systematically searched, unpublished, recently completed, or non-indexed educational studies may have been missed. Publication bias could not be formally or meaningfully assessed because too few studies contributed to the pooled analysis.

From an educational and curricular perspective, the findings support the selective and supervised incorporation of chatbots into dental training. These tools appear most useful when they are embedded in clearly defined instructional designs, such as preparatory review, guided discussion, virtual patient interaction, or structured practice support. They may be particularly

valuable in settings where faculty time is limited, where repeated low-risk rehearsal is needed, or where learners benefit from immediate clarification and feedback. However, chatbot use should remain aligned with curriculum objectives, supported by faculty oversight, and framed explicitly as an adjunct rather than a replacement for conventional teaching, critical appraisal, or clinical judgment. In clinically oriented teaching, their greatest immediate value may lie in simulation and structured reasoning support rather than in independent decision-making.

Future research should prioritize well-designed comparative trials with clearly defined chatbot interventions, robust educational comparators, and validated outcome measures. Greater emphasis is needed on objective student performance outcomes, not only on perceptions or satisfaction. Longitudinal studies should examine whether chatbot-assisted learning improves retention, transfer to clinical settings, and sustained competency development. It will also be important to study how these tools function across different curricular levels, including postgraduate dental training, and whether their effects vary by prior AI familiarity, learning style, or clinical experience. Finally, future studies should report outcomes with sufficient statistical detail to support secondary synthesis and should distinguish clearly between improvements in educational experience and improvements in measurable competence.

CONCLUSIONS

This systematic review, exploratory meta-analysis, and functional narrative synthesis suggests that chatbot-assisted learning may provide educa-

tional support in selected areas of dental education. However, the available evidence remains limited, heterogeneous, and generally low to very low in certainty. The exploratory quantitative synthesis indicated a favorable effect on overall student educational performance, but this estimate was based on only two studies assessing different educational outcomes and should not be interpreted as evidence of general superiority over conventional instruction.

At present, chatbots appear most appropriate as adjunctive and supervised educational tools that may support knowledge reinforcement, structured practice, communication rehearsal, and selected academic tasks. Their integration into dental curricula should therefore be cautious and pedagogically structured, with continued emphasis on faculty guidance, critical reasoning, evidence appraisal, and clinically grounded learning. Further well-designed comparative studies with objective outcomes and longer follow-up are needed before stronger conclusions can be drawn about their long-term educational value and curricular role.

SUPPLEMENTARY MATERIALS: Table S1. Complete Search Strategies for Each Database. Table S2. RoB 2 item-level assessment of randomized studies. Table S3. JBI item-level assessment of non-randomized studies. Table S4. Summary statistics used in the exploratory meta-analysis. Table S5. Excluded Studies and Reasons for Exclusion. Table S6. Detailed outcome-level results of included studies.

CONFLICT OF INTEREST: The authors declare that there is no conflict of interest.

FUNDING: This research received no external funding.

DATA AVAILABILITY STATEMENT: Datasets are available on request: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT: Not required.

AUTHOR CONTRIBUTION STATEMENT: Conceptualization, E.P.V., D.V.O. and C.M.A.; Methodology, E.P.V., D.V.O. and C.M.A.; Formal analysis, E.P.V. and D.V.O.; Investigation, E.P.V., D.V.O. and C.M.A.; Data curation, E.P.V., D.V.O. and C.M.A.; Writing—original draft preparation, E.P.V., D.V.O. and C.M.A.; Writing—review and editing, E.P.V., D.V.O. and C.M.A.; Project administration, E.P.V. All authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Wu Y.H., Tso K.Y., Chiang C.P. Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations. *J Dent Sci.* 2025; 20 (3): 1709-15. doi:10.1016/j.jds.2025.03.030
2. Nguyen H.C., Dang H.P., Nguyen T.L., Hoang V., Nguyen V.A. Accuracy of latest large language models in answering multiple choice questions in dentistry: A comparative study. *PLoS One.* 2025; 20 (1): e0317423. doi:10.1371/journal.pone.0317423
3. Bhatia A.P., Lambat A., Jain T. A comparative analysis of conventional and ChatGPT-assisted teaching methods in undergraduate dental education. *Cureus.* 2024 May 9; 16 (5): e60006. doi:10.7759/cureus.60006
4. Khareedi R., Fernandez D. The role of chatbots in enquiry-based learning for oral health students—An exploratory study. *Eur J Dent Educ.* 2025 May 15. doi:10.1111/eje.13115.
5. Or A., Sukumar S., Ma A., Ang D., Liu M., Ritchie H.E., Sarrafpour B. Enhancing dental students' history-taking skills with a generative artificial intelligence chatbot. *J Dent Educ.* 2025 Jun 11. doi:10.1002/jdd.13952
6. Huang S., Wen C., Bai X., Li S., Wang S., Wang X., Yang D. 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. doi:10.2196/68538
7. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. *Cureus.* 2025; 17 (1): e77292. doi:10.7759/cureus.77292
8. Kim W., Kim B.C., Yeom H.G. Performance of Large Language Models on the Korean Dental Licensing Examination: A Comparative Study. *Int Dent J.* 2025 Feb; 75 (1): 176-184. doi: 10.1016/j.identj.2024.09.002
9. Ullah R., Shaikh M.S., Shahani N., Lone M.A., Fareed M.A., Zafar M.S. Comparing ChatGPT and dental students' performance in an introduction to dental anatomy examination: A cross-sectional study. *Eur J Dent.* 2025 May 13. doi:10.1055/s-0045-1808254
10. Metin U., Goymen M. Information from digital and human sources: A comparison of chatbot and clinician responses to orthodontic questions. *Am J Orthod Dentofacial Orthop.* 2025 May 6. doi:10.1016/j.ajodo.2025.04.008
11. Öztürk Z., Bal C., Çelikkaya B.N. Evaluation of information provided by ChatGPT versions on traumatic dental injuries for dental students and professionals. *Dent Traumatol.* 2025; 41 (4): 427-36. doi:10.1111/edt.13042
12. Fukuda H., Morishita M., Muraoka K., Yamaguchi S., Nakamura T., Yoshioka I.,

- Awano S., Ono K. Evaluating the image recognition capabilities of GPT-4V and Gemini Pro in the Japanese national dental examination. *J Dent Sci.* 2025; 20 (1): 368-72. doi:10.1016/j.jds.2024.06.015
13. Durmazpinar P.M., Ekmekci E. Comparing diagnostic skills in endodontic cases: dental students versus ChatGPT-4o. *BMC Oral Health.* 2025; 25 (1):457. doi:10.1186/s12903-025-05857-y
 14. Ali K., Barhom N., Tamimi F., Duggal M. ChatGPT—a double-edged sword for healthcare education? Implications for assessments of dental students. *Eur J Dent Educ.* 2024 Feb; 28 (1): 206-11. doi:10.1111/eje.12937
 15. Page M.J., McKenzie J.E., Bossuyt P.M., Boutron I., Hoffmann T.C., Mulrow C.D., et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021; 372: n71. doi:10.1136/bmj.n71
 16. Sterne J.A.C., Savović J., Page M.J., Elbers R.G., Blencowe N.S., Boutron I., et al. RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ.* 2019; 366: 14898. doi:10.1136/bmj.l4898
 17. Moola S., Munn Z., Tufanaru C., Aromataris E., Sears K., Sfetcu R., et al. Chapter 7: Systematic reviews of etiology and risk. In: Aromataris E., Munn Z., editors. *JBIM Manual for Evidence Synthesis.* Adelaide: JBI; 2020. doi:10.46658/JBIMES-20-08
 18. Guyatt G.H., Oxman A.D., Vist G.E., Kunz R., Falck-Ytter Y., Alonso-Coello P., et al. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ.* 2008 Apr 26; 336 (7650): 924-6. doi:10.1136/bmj.39489.470347.AD
 19. Prinz M., Schäfer E., Bürklein S., Donnermeyer D. Endodontic diagnostics training in undergraduate dental education: An observational pilot study on AI-driven virtual patient e-learning. *Int Endod J.* 2025 Jul 22. doi:10.1111/iej.14277
 20. Roganović J. Familiarity with ChatGPT features modifies expectations and learning outcomes of dental students. *Int Dent J.* 2024 Dec; 74 (6): 1456-62. doi:10.1016/j.identj.2024.04.012
 21. Saravia-Rojas M.Á., Camarena-Fonseca A.R., León-Manco R., Geng-Vivanco R. Artificial intelligence: ChatGPT as a disruptive didactic strategy in dental education. *J Dent Educ.* 2024 Jun; 88 (6): 872-6. doi:10.1002/jdd.13485
 22. Kavarella A., Dias da Silva M.A., Kaklamanos E.G., Stamatopoulos V., Giannakopoulos K. Evaluation of ChatGPT's real-life implementation in undergraduate dental education: mixed methods study. *JMIR Med Educ.* 2024 Jan 31; 10: e51344. doi:10.2196/51344
 23. Keles M. A Randomised Study of ChatGPT-4o vs. Written Materials for Teaching Periodontal-Systemic Relationships in Dental Education. *Eur J Dent Educ.* 2026 Mar 28. doi: 10.1111/eje.70149
 24. Aldukhail S. Mapping the landscape of generative language models in dental education: A comparison between ChatGPT and Google Bard. *Eur J Dent Educ.* 2025; 29 (1): 136-48. doi:10.1111/eje.13056
 25. Li C., Zhang J., Abdul-Masih J., Zhang S., Yang J. Performance of ChatGPT and dental students on concepts of periodontal surgery. *Eur J Dent Educ.* 2025; 29 (1): 36-43. doi:10.1111/eje.13047
 26. Künzle P., Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin Oral Investig.* 2024; 28 (11): 575. doi:10.1007/s00784-024-05968-w
 27. Tomo S., Lechien J.R., Bueno H.S., Cantieri-Debortoli D.F., Simonato L.E. Accuracy and consistency of ChatGPT-3.5 and -4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnosis

- tic performance analysis. *Clin Oral Investig.* 2024 Sep 24; 28 (10): 544. doi: 10.1007/s00784-024-05939-1
28. Quah B., Zheng L., Sng T.J.H., Yong C.W., Islam I. Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Med Educ.* 2024 Sep 3; 24 (1): 962. doi: 10.1186/s12909-024-05881-6.
29. Demirsoy K.K., Buyuk S.K., Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? *Angle Orthod.* 2024 Nov 1; 94 (6): 602-7. doi:10.2319/031224-207.1
30. Morishita M., Fukuda H., Muraoka K., Nakamura T., Hayashi M., Yoshioka I., et al. Evaluating GPT-4V's performance in the Japanese national dental examination: A challenge explored. *J Dent Sci.* 2024 Jul; 19 (3): 1595-600. doi:10.1016/j.jds.2023.12.007
31. Sabri H., Saleh M.H.A., Hazrati P., Merchant K., Misch J., Kumar P.S., 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 Feb; 60 (2): 121-33. doi:10.1111/jre.13323
32. Brondani M., Alves C., Ribeiro C., Braga M.M., Garcia R.C.M., Ardenghi T., et al. Artificial intelligence, ChatGPT, and dental education: Implications for reflective assignments and qualitative research. *J Dent Educ.* 2024 Dec; 88 (12): 1671-80. doi:10.1002/jdd.13663
33. Jeong H., Han S.S., Yu Y., Kim S., Jeon K.J. How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofac Radiol.* 2024 Sep 1; 53 (6): 390-5. doi:10.1093/dmfr/twae021
34. Qutieshat A., Al Rusheidi A., Al Ghammari S., Alarabi A., Salem A., Zelihic M. Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence. *Diagnosis (Berl).* 2024 May 3; 11 (3): 259-65. doi:10.1515/dx-2024-0034

SUPPLEMENTARY

Table S1. Search strategies used across databases.

Database	Search strategy	Database	Search strategy
PubMed/MEDLINE	("Artificial Intelligence"[Mesh] OR "Natural Language Processing"[Mesh] OR chatbot*[tiab] OR "chat bot"*[tiab] OR ChatGPT[tiab] OR Gemini[tiab] OR Copilot[tiab] OR Claude[tiab] OR Bard[tiab] OR "large language model"*[tiab] OR LLM[tiab] OR LLMs[tiab] OR "generative artificial intelligence"[tiab] OR "generative AI"[tiab] OR "conversational agent"*[tiab] OR "educational chatbot"*[tiab] OR "virtual teaching assistant"*[tiab] OR "intelligent tutoring system"*[tiab]) AND ("Education, Dental"[Mesh] OR "Students, Dental"[Mesh] OR "Schools, Dental"[Mesh] OR "Education"[Mesh] OR "dental education"[tiab] OR "oral health education"[tiab] OR "dental student"*[tiab] OR "undergraduate dental student"*[tiab] OR "postgraduate dental student"*[tiab] OR "dental curriculum"[tiab] OR "dental training"[tiab] OR "dental school"*[tiab] OR "dentistry education"[tiab]) AND (dentist*[tiab] OR dental[tiab]	OR dentistry[tiab] OR oral[tiab] OR orthodontic*[tiab] OR endodontic*[tiab] OR periodont*[tiab] OR prosthodontic*[tiab] OR radiolog*[tiab] OR "oral medicine"[tiab] OR "oral pathology"[tiab])	Scopus
		TITLE-ABS-KEY("artificial intelligence" OR "natural language processing" OR chatbot OR chatbots OR "chat bot" OR "chat bots" OR chatgpt OR gemini OR copilot OR claude OR bard OR "large language model" OR "large language models" OR llm OR llms OR "generative artificial intelligence" OR "generative ai" OR "conversational agent" OR "conversational agents" OR "educational chatbot" OR "educational chatbots" OR "virtual teaching assistant" OR "virtual teaching assistants" OR "intelligent tutoring system" OR "intelligent tutoring systems") AND TITLE-ABS-KEY("dental education"	

Database	Search strategy	Database	Search strategy
	OR "oral health education" OR "dental student" OR "dental students" OR "undergraduate dental student" OR "undergraduate dental students" OR "postgraduate dental student" OR "postgraduate dental students" OR "dental curriculum" OR "dental training" OR "dental school" OR "dental schools" OR "dentistry education") AND TITLE-ABS-KEY(dentist OR dentists OR dental OR dentistry OR oral OR orthodontic OR orthodontics OR endodontic OR endodontics OR periodontics OR periodontal OR prosthodontic OR prosthodontics OR radiology OR radiologic OR "oral medicine" OR "oral pathology") (		OR bard:ti,ab,kw OR 'large language model*':ti,ab,kw OR llm:ti,ab,kw OR llms:ti,ab,kw OR 'generative artificial intelligence':ti,ab,kw OR 'generative AI':ti,ab,kw OR 'conversational agent*':ti,ab,kw OR 'educational chatbot*':ti,ab,kw OR 'virtual teaching assistant*':ti,ab,kw OR 'intelligent tutoring system*':ti,ab,kw) AND (
			'dental education'/exp OR 'dental student'/exp OR 'education'/exp OR 'dental school'/exp OR 'dental education':ti,ab,kw OR 'oral health education':ti,ab,kw OR 'dental student*':ti,ab,kw OR 'undergraduate dental student*':ti,ab,kw OR 'postgraduate dental student*':ti,ab,kw OR 'dental curriculum':ti,ab,kw OR 'dental training':ti,ab,kw OR 'dental school*':ti,ab,kw OR 'dentistry education':ti,ab,kw) AND (
Embase	'artificial intelligence'/exp OR 'natural language processing'/exp OR chatbot*:ti,ab,kw OR 'chat bot*':ti,ab,kw OR chatgpt:ti,ab,kw OR gemini:ti,ab,kw OR copilot:ti,ab,kw OR claude:ti,ab,kw		'dentistry'/exp OR dentist*:ti,ab,kw OR dental:ti,ab,kw OR dentistry:ti,ab,kw OR oral:ti,ab,kw OR orthodontic*:ti,ab,kw OR endodontic*:ti,ab,kw OR periodont*:ti,ab,kw OR prosthodontic*:ti,ab,kw OR radiolog*:ti,ab,kw OR 'oral medicine':ti,ab,kw OR 'oral pathology':ti,ab,kw)

Table S2. RoB 2 item-level assessment of randomized studies.

Study	Outcome assessed	D1: Randomization process	D2: Deviations from intended interventions	D3: Missing outcome data	D4: Measurement of the outcome	D5: Selection of the reported result	Overall judgment	Main rationale
Bhatia <i>et al.</i> (3)	Academic performance/post-test score	Some concerns	High risk	Some concerns	Some concerns	High risk	High risk	Concerns were related to incomplete reporting of allocation concealment and blinding, inconsistencies in participant flow and numerical reporting, and insufficient information to exclude selective reporting.
Huang <i>et al.</i> (6)	Operative/procedural skill performance	Low risk	Low risk	Low risk	Low risk	Some concerns	Some concerns	Random allocation and clearly defined intervention/comparator conditions were reported. The primary outcome was objectively generated by the virtual-reality system. However, the absence of a clearly accessible prespecified statistical analysis plan for the specific educational outcome led to some concerns in the selective reporting domain.
Kales (23)	Knowledge acquisition/academic performance	Low risk	Low risk	Low risk	Low risk	Some concerns	Some concerns	Randomization, allocation concealment, and blinded/anonymized scoring were described. Some concerns remained in the selective reporting domain because the availability of a prespecified analysis plan for all reported educational outcomes was unclear.

D1, bias arising from the randomization process; D2, bias due to deviations from intended interventions; D3, bias due to missing outcome data; D4, bias in measurement of the outcome; D5, bias in selection of the reported result.

Table S3. JBI item-level assessment of non-randomized studies.

Study	Design/checklist applied	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Overall methodological quality	Main concerns
Or <i>et al.</i> (5)	JBI checklist for quasi-experimental studies	Yes	Yes	Unclear	No	No	Yes	Yes	Yes	Yes	Moderate	Clearly described intervention and educational setting, but main outcomes were largely self-reported and no strong concurrent control was available for the principal educational outcome.
Prinz <i>et al.</i> (19)	JBI checklist for quasi-experimental studies	Yes	Yes	Unclear	No	No	Yes	Yes	Yes	Yes	Moderate	Clearly defined pre-post educational intervention, but absence of a control group and reliance on self-perceived improvement limited causal inference.
Saravia-Rojas <i>et al.</i> (21)	JBI checklist for quasi-experimental studies	Yes	Yes	Unclear	No	No	Yes	Yes	Unclear	Yes	Low	Sequential within-cohort design created risk of order and carryover effects; unequal completion across phases further reduced confidence.
Kavadella <i>et al.</i> (22)	JBI checklist for quasi-experimental studies	Yes	Yes	Unclear	Yes	Unclear	Yes	Yes	Yes	Yes	Moderate	Parallel comparison groups and objective examination outcome were strengths, but incomplete final assessment data and limited control of baseline group differences remained concerns.
Roganović (20)	JBI checklist for analytical cross-sectional studies	Yes	Yes	Unclear	Unclear	No	Yes	Yes	Yes	N/A	Low	Self-selection into exposure groups and expectation-related confounding limited attribution of observed differences to chatbot use.

For quasi-experimental studies, Q1-Q9 correspond to the JBI checklist items for quasi-experimental studies. For analytical cross-sectional studies, Q1-Q8 correspond to the JBI checklist items for analytical cross-sectional studies; Q9 was not applicable.

Table S4. Summary statistics used in the exploratory meta-analysis.

Study	Outcome included in meta-analysis	Chatbot-assisted group (n)	Mean	SD	Comparator group (n)	Mean	SD
Huang <i>et al.</i> (2025) (6)	Operative/procedural skill performance	94	73.12	10.06	93	65.54	12.48
Kavadella <i>et al.</i> (2024) (22)	Knowledge examination performance	37	7.54	1.18	40	6.94	1.12

Outcome measures were transformed into standardized mean differences (Hedges g) and pooled using a random-effects model. In Huang *et al.* (6), the final analyzed sample comprised 187 students after exclusion of 5 participants who did not meet eye-tracking calibration requirements.

Table S5. Full-text studies excluded after application of the revised eligibility criteria and primary reasons for exclusion.

Study	Primary reason for exclusion from the revised review
Wu <i>et al.</i> (2025)	No real dental student cohort; benchmark study based on multiple-choice questions only.
Nguyen <i>et al.</i> (2025)	No real dental student intervention; head-to-head LLM benchmark on MCQs.
Khareedi <i>et al.</i> (2025)	No real dental student cohort; exploratory comparison of chatbot outputs evaluated by academics.
Kinikoglu (2025)	No real dental student educational intervention; AI benchmark on specialty examination questions.
Kim <i>et al.</i> (2025)	No real dental student intervention; licensing examination benchmark of LLMs.
Ullah <i>et al.</i> (2025)	Included real students, but compared ChatGPT performance against students rather than evaluating a chatbot-assisted educational intervention.
Metin & Goymen (2025)	Comparative diagnostic benchmark across chatbots, students, dentists, and specialists; no student-centered educational intervention.
Öztürk <i>et al.</i> (2025)	No real dental student intervention; evaluation of chatbot output quality, reliability, and readability.
Fukuda <i>et al.</i> (2025)	No real dental student cohort; benchmark on national examination questions.
Durmazpinar & Ekmekci (2025)	Included students, but assessed diagnostic performance of ChatGPT versus students rather than educational effectiveness of chatbot-assisted learning.
Ali <i>et al.</i> (2024)	No real dental student cohort; exploratory assessment of ChatGPT on examination items.
Davies (2025)	Comparative diagnostic benchmark of AI versus students; no chatbot-assisted educational intervention.
Katebzadeh <i>et al.</i> (2025)	Focused on GenAI performance versus students rather than student outcomes following chatbot-assisted learning.
Kim & Kim (2025)	Focused on item generation and question quality, not on chatbot-assisted learning outcomes in students.
Stephan <i>et al.</i> (2024)	Compared AI-generated and student-generated radiology reports; no chatbot-assisted educational intervention.
Aldukhail (2025)	No real dental student cohort; AI systems were evaluated as test subjects.
Li <i>et al.</i> (2025)	Compared chatbot versions and student performance, but did not evaluate a defined chatbot-assisted educational intervention against an educational comparator.
Künzle & Paris (2024)	No real dental student cohort; benchmark study based on test questions only.
Tomo <i>et al.</i> (2024)	Diagnostic performance comparison involving specialists, dentists, and students; no student-centered educational intervention.
Quah <i>et al.</i> (2024)	Focused on automated essay scoring by ChatGPT rather than chatbot-assisted learning as an educational intervention.
Demirsoy <i>et al.</i> (2024)	Evaluated perceived quality of ChatGPT responses among students, specialists, and patients; no educational intervention or learning comparator.
Morishita <i>et al.</i> (2024)	No real dental student cohort; image-based national examination benchmark.
Sabri <i>et al.</i> (2025)	No real dental student educational intervention; standardized examination benchmark of LLMs.
Brondani <i>et al.</i> (2024)	Did not evaluate chatbot-assisted learning effectiveness in students; focused on thematic analysis support.
Jeong <i>et al.</i> (2024)	Compared chatbot and student performance on examination questions; no chatbot-assisted educational intervention.
Qutieshat <i>et al.</i> (2024)	Compared diagnostic accuracy of ChatGPT and students; no student-centered educational intervention.
Rokhshad <i>et al.</i> (2024)	Comparative benchmark involving clinicians and students; no chatbot-assisted educational intervention.
Rahad <i>et al.</i> (2024)	Exploratory mixed-participant study without a clear educational comparator and without a defined student-centered intervention design.
Giannakopoulos <i>et al.</i> (2023)	No real dental student cohort; faculty-based evaluation of chatbot responses to evidence-based questions.
Hultgren <i>et al.</i> (2023)	Mixed students/teachers study focused on comparing ChatGPT with teacher responses and intended learning outcomes, rather than student-level educational effectiveness.
Danesh <i>et al.</i> (2023)	No real dental student cohort; benchmark study on board-style multiple-choice questions.

Table S6. Detailed outcome-level results of included studies.

Study	Educational domain	Outcome	Instrument / assessment method	Time point	Analyzed sample	Main result	Effect estimate / CI	Losses or missing data	Outcome type
Bhatia <i>et al.</i> (3)	Knowledge acquisition and academic performance	Post-test academic performance	Pre-validated online questionnaire / knowledge test	Post-intervention	100 allocated; 82 reported in analysis table	Conventional teaching showed higher post-test scores than ChatGPT-assisted learning (3.57 ± 0.498 vs. 1.70 ± 1.33 ; $p < 0.001$)	$t = 12.263$; 95% CI not reported	Not clearly reported	Objective test score
Or <i>et al.</i> (5)	Clinical reasoning and diagnostic communication training	Perceived competence, participation, practice opportunities, enjoyment	Post-chatbot student survey comparing chatbot experience with in-person learning	Post-intervention	31 students	Perceived competence improved after chatbot use; students reported more participation and practice opportunities, but most found the chatbot less interesting than in-person teaching	Exact effect estimate / CI not reported	Not reported	Self-reported
Huang <i>et al.</i> (6)	Operative and procedural skill support	Operative/procedural skill performance	Desktop virtual-reality operative test	Post-intervention after 1-week learning period	187 students; ChatGPT $n = 94$, control $n = 93$	ChatGPT group performed better than control group (73.12 ± 10.06 vs. 65.54 ± 12.48 ; $p < 0.001$)	$t = 4.569$; 95% CI not reported	5 excluded before analysis because of eye-tracking calibration failure	Objective performance score
Huang <i>et al.</i> (6)	Operative and procedural skill support	Cognitive load	Eye-tracking pupil-diameter change	During/post-learning task	187 students; ChatGPT $n = 94$, control $n = 93$	Cognitive load was lower in ChatGPT group (0.137 ± 0.036 vs. 0.312 ± 0.032 ; $p < 0.001$)	Exact CI not reported	5 excluded before analysis because of eye-tracking calibration failure	Objective physiological measure
Huang <i>et al.</i> (6)	Operative and procedural skill support	Self-efficacy and learning motivation	Questionnaire	Post-intervention	187 students; ChatGPT $n = 94$, control $n = 93$	ChatGPT group showed higher self-efficacy ($p = 0.04$) and motivation ($p = 0.02$)	Exact effect estimate / CI not reported	5 excluded before analysis because of eye-tracking calibration failure	Self-reported
Prinz <i>et al.</i> (19)	Clinical reasoning and diagnostic communication training	Self-perceived endodontic diagnostic confidence and diagnostic skills	Pre-post Likert-scale questionnaires	Baseline and after 3-month access to virtual patient e-learning tool	T1: 92 students; T2: 72 students	Self-perceived confidence in endodontic diagnostics improved after use of the AI-driven virtual patient tool	Exact effect estimate / CI not reported	20 students did not complete T2; dropout 21.7%	Self-reported
Roganović (20)	Knowledge acquisition and academic performance	Pharmacology learning performance	Pharmacology quiz after learning about side effects of drugs used in dental practice	Post-learning	104 students; VG $n = 13$, NG $n = 47$, NN $n = 44$ at initial grouping	ChatGPT users performed better than students who neither used ChatGPT nor read its description. Students who only read the ChatGPT description also performed better than the NN group	Exact effect estimate / CI not reported	Not reported	Objective quiz score

Study	Educational domain	Outcome	Instrument / assessment method	Time point	Analyzed sample	Main result	Effect estimate / CI	Losses or missing data	Outcome type
Roganović (20)	Knowledge acquisition and academic performance	Expectations and attitudes toward ChatGPT	Survey before and after reading ChatGPT feature description	Pre/post description exposure	104 students	Reading ChatGPT feature descriptions modified expectations toward AI and was associated with quiz performance	Exact effect estimate / CI not reported	Not reported	Self-reported
Saravia-Rojas et al. (21)	Academic writing, evidence handling, and reflective task support	Scientific writing task performance	Professor-rated rubric with five criteria	After ChatGPT task and after conventional literature-search task	ChatGPT task n=55; conventional task n=39	Conventional method produced higher total scores than ChatGPT (16.33±2.46 vs. 15.04±2.78; p=0.019)	95% CI not reported	Unequal completion between phases	Objective rubric-based score
Saravia-Rojas et al. (21)	Academic writing, evidence handling, and reflective task support	Rubric domains	Professor-rated rubric	After each task	ChatGPT task n=55; conventional task n=39	Conventional method scored higher for evidence use (p=0.042), argument evaluation (p=0.017), and generation of alternatives (p<0.001)	95% CI not reported	Unequal completion between phases	Objective rubric-based score
Saravia-Rojas et al. (21)	Academic writing, evidence handling, and reflective task support	Perceived usefulness of ChatGPT	Anonymous questionnaire	Post-task	42 questionnaire respondents	Most students considered ChatGPT useful or very useful for academic activities	Exact effect estimate / CI not applicable	42/55 completed questionnaire	Self-reported
Kavadella et al. (22)	Knowledge acquisition and academic performance	Knowledge examination grade	10-item multiple-choice knowledge examination, 0-10 scale	Post-assignment	70 students completed examination; ChatGPT n=37, literature research n=33/40 as reported across analysis sources	ChatGPT group achieved higher examination grades than literature research group; values used in meta-analysis were 7.54±1.18 vs. 6.94±1.12	p=0.045; 95% CI not reported	70/77 completed examination	Objective examination score
Kavadella et al. (22)	Knowledge acquisition and academic performance	Learning experience and perceived benefits/limitations	Open-ended evaluation questionnaire	Post-assignment	ChatGPT group respondents	Students reported benefits including immediate responses and broad knowledge base, and limitations including false citations and need for prompt rephrasing	Not applicable	Not clearly reported	Self-reported / qualitative
Keles (23)	Knowledge acquisition and academic performance	Post-test knowledge score	Test B, 20-item knowledge test; ANCOVA adjusted for Test A	Post-intervention after 30-min learning session	70 students; written materials n=35, ChatGPT-4o n=35	No significant between-group difference in adjusted post-test total score; both groups improved pre-post	ANCOVA; exact CI not reported	No losses reported	Objective knowledge test
Keles (23)	Knowledge acquisition and academic performance	Perceptions of learning method	7-item, 5-point Likert questionnaire	Post-intervention	70 students	ChatGPT group scored higher for accessibility/ease of information and motivation; written materials group reported better concentration/focus; no significant differences for other items	p<0.05 for selected items; CI not reported	No losses reported	Self-reported