

Concise Review

Accuracy of Large Language Models in Answering Dental Examination Questions: A Systematic Review and Meta-Analysis



Mahmood Dashti^{a,b*}, Farshad Khosraviani^c, Atieh Meyari^d,
 Mohammad Hosein Amirzade-Iranaq^e, Akhilanand Chaurasia^f,
 Delband Hefzi^g, Niloofar Ghadimi^h, Antonin Tichy^{i,j}, Zohaib Khurshid^{k,l},
 Falk Schwendicke^{i*}

^a Dentofacial Deformities Research Center, Research Institute of Dental Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

^b Department of Artificial Intelligence Engineering, Graduate School of Natural and Applied Sciences, Istinye University, Istanbul, Türkiye

^c UCLA School of Dentistry, CA, USA

^d Department of Esthetic & Restorative Dentistry, School of Dentistry, Tehran Azad University of Medical Sciences, Tehran, Iran

^e Universal Scientific Education and Research Network (USERN), Tehran University of Medical Sciences, Tehran, Iran

^f Department of Oral Medicine and Radiology, King George's Medical University, Lucknow, India

^g School of Dentistry, Tehran University of Medical Science, Tehran, Iran

^h Dental Materials Research Center, TeMS.C., School of Dentistry, Azad University of Medical Sciences, Tehran, Iran

ⁱ Clinic for Conservative Dentistry, Periodontology and Digital Dentistry, LMU University Hospital, LMU Munich, Munich, Germany

^j Institute of Dental Medicine, First Faculty of Medicine, Charles University, Prague, Czech Republic

^k Department of Prosthodontics and Dental Implantology, College of Dentistry, King Faisal University, Al-Ahsa, Saudi Arabia

^l Center for Artificial Intelligence and Innovation (CAII), Faculty of Dentistry, Chulalongkorn University, Bangkok, Thailand

ARTICLE INFO

Article history:

Received 18 December 2025

Received in revised form

11 March 2026

Accepted 30 March 2026

Available online 18 May 2026

Key words:

Artificial intelligence

Dental examination

Large language models

meta-analysis

ABSTRACT

Introduction: Large language models (LLMs), including OpenAI's GPT family accessed via interfaces such as ChatGPT and Microsoft Copilot, as well as non-GPT systems such as Google Gemini, are increasingly applied in healthcare and dental education. However, the accuracy of these systems in specialized tasks such as answering dental examination questions remains unclear.

Methods: This systematic review and meta-analysis evaluated LLM performance in answering dental questions. Databases searched were PubMed, Embase, Scopus, and Web of Science. Data on question type and number, LLM versions, and accuracy rates were extracted. Pooled accuracy was estimated using a random-effects model; heterogeneity and publication bias were assessed.

Results: A total of 39 studies were included, with ChatGPT-4 being the most frequently evaluated model. The pooled accuracy for LLMs was 63.7% (95% CI: 60.3%-67.1%), with high heterogeneity ($I^2 = 91.5\%$). Subgroup analysis revealed ChatGPT-4 and Copilot (a GPT-based interface) achieved the highest pooled accuracies (~73% and ~75%, respectively). Direct comparisons confirmed ChatGPT-4 significantly outperformed earlier versions and some competitor models. Sensitivity analyses supported the robustness of findings.

Conclusion: LLMs demonstrate moderate accuracy in answering dental examination questions and are currently insufficient for autonomous clinical decision-making. When their

* Corresponding author. Chair and Director Clinic, Conservative Dentistry, Periodontology and Digital Dentistry, University Hospital of the Ludwig-Maximilians-University, Munich, Germany; and Department of Artificial Intelligence Engineering, Graduate School of Natural and Applied Sciences, Istinye University, Istanbul, Türkiye.

E-mail addresses: 2533275099@stu.istinye.edu.tr (M. Dashti), falk.schwendicke@med.uni-muenchen.de (F. Schwendicke).

<https://doi.org/10.1016/j.identj.2026.109609>

0020-6539/© 2026 The Authors. Published by Elsevier Inc. on behalf of FDI World Dental Federation. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

limitations are explicitly recognized, however, these systems may serve as valuable adjuncts in dental education and examination preparation. Methodological strategies such as structured prompting and retrieval-augmented approaches warrant further investigation but were not the primary focus of the present analysis.

© 2026 The Authors. Published by Elsevier Inc. on behalf of FDI World Dental Federation.

This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>)

Introduction

Artificial intelligence (AI) refers to the simulation of human intelligence by machines, enabling them to perform tasks typically requiring human cognitive abilities, such as reasoning, problem-solving, and learning.^{1,2} Within AI, machine learning (ML) and deep learning represent critical subsets, employing data-driven algorithms to extract patterns, predict outcomes, and make decisions.

Large language models (LLMs) are generative AI systems trained on large text corpora that can interpret prompts and produce human-like text.^{3,4} In dentistry, LLMs are already being explored for education and assessment support (eg, tutoring, formative feedback, and examination preparation), patient-facing communication (eg, postoperative instructions and informed-consent explanations), evidence-based dentistry tasks (eg, summarizing guidelines and literature), and workflow support (eg, drafting clinical notes and referral letters).^{5,6} Because dental care is safety-critical and often guideline-driven, their use depends on demonstrable accuracy and reproducibility in dentistry-relevant tasks.

Various LLMs and their associated interfaces are presently available. To avoid conflating models with the user-facing platforms that use them, we list the foundation LLMs first, followed by the major conversational interfaces built on top of them.

Foundation LLMs:

- GPT model family (OpenAI): This includes the models commonly accessed through ChatGPT (eg, GPT-3.5, GPT-4, GPT-4o), which are transformer-based LLMs trained on large-scale corpora and widely applied in education, research, and healthcare.⁶⁻⁸
- Gemini (Google DeepMind): A multimodal LLM designed for text and image understanding, with strong analytical and scientific reasoning capabilities.⁹
- Claude (Anthropic): A safety-aligned LLM built to minimize hallucinations and improve reliability in sensitive fields such as healthcare.¹⁰
- LLaMA (Meta): An open-foundation family of LLMs trained on publicly available datasets, supporting transparent and reproducible research.¹¹
- DeepSeek LLM: An open-source suite trained on models ranging from 7 billion to 67 billion parameters, using large-scale token datasets and shown to outperform several existing open models on reasoning and coding benchmarks.¹²

Interfaces built on top of foundation LLMs include ChatGPT and Microsoft Copilot (formerly Bing Chat) which are conversational interfaces that provide access to OpenAI's GPT family models, with Copilot incorporating additional

search grounding and safety layers.^{3,6-8} Google's Gemini interface was formerly released under the name Bard and is powered by the Gemini family of models.⁹ In the present review, model performance is reported according to the platform and model designation used in each original study.

While LLMs can tackle complex queries and offer expert-level responses, their application in dentistry remains an emerging area warranting in-depth evaluation.^{6,10} LLMs exhibit significant promise in answering dental examination questions, a critical area in both academic and clinical practice. By leveraging extensive training datasets, these models can accurately interpret and respond to multiple-choice questions, case studies, and scenario-based queries. LLMs can support dental education, aid clinical decision-making, perform on standardized exams, and bridge language barriers, enhancing access to dental knowledge.^{2,5,13}

Despite their potential, questions about accuracy, consistency, and reliability in specialized fields like dentistry persist. Evaluating these models' performance on dental examinations is thus crucial for validating their applicability and establishing best practices. This systematic review and meta-analysis seeks to evaluate the accuracy and reliability of LLMs in answering dental examination questions. The null hypothesis was that there is no significant variation in accuracy among different LLMs when answering dental examination questions, and that their accuracy does not reach the expected benchmark levels typically required for passing standard dental examination.

Methods and material

This systematic review and meta-analysis adhered to the guidelines established by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) criteria.¹⁴ The process involved the systematic extraction, selection, and screening of relevant research articles.

This review was not registered in PROSPERO, and no protocol was published in advance. We therefore report the full search strategies, prespecified eligibility criteria, screening procedures, and analytical decisions in the Methods and Supplementary materials to support transparency.

We have implemented the PICO framework in accordance with the research question, which was: 'How accurately can LLMs answer the dental examinations and dental questions?', to enhance our search strategy:

- Population (P): Dental examination and question sets, including standardized licensing exams, specialty exams, and validated educational question banks in dentistry.

- Intervention (I): Use of LLMs (eg, ChatGPT, Bing AI/Copilot, Google Bard/Gemini, Claude, LLaMA) to answer dental questions.
- Comparison (C): Accuracy against a reference standard (official answer keys, validated question banks, or expert consensus). No consistent direct human comparator across studies.
- Outcome (O): Primary: Accuracy rate (percentage of correct answers). Secondary: (i) between-model/version differences, defined as the absolute difference in accuracy (percentage points) within studies that evaluated two or more LLM models/versions under the same question set; and (ii) subgroup differences in pooled accuracy by LLM model/version group, question type/format, and dental specialty.

Eligibility criteria

Studies were included if they¹ evaluated LLMs in answering dentistry-related examination or knowledge-based questions,² provided the model with a defined set of dental questions, and³ reported accuracy or performance outcomes.

Studies were excluded if they¹ were reviews (scoping, narrative, systematic, or meta-analyses),² generated questions using LLMs instead of assessing their ability to answer existing dental questions,³ lacked sufficient methodological detail (eg, unspecified model version or missing accuracy data), or⁴ were not focused on dental topics.

We included peer-reviewed full-text articles. Preprints were eligible only if they were indexed in Scopus (identified via the Scopus 'Secondary documents' option), had an accessible full text, and reported extractable accuracy outcomes. We did not search non-indexed grey literature sources (eg, websites, blogs, theses) and excluded conference abstracts without sufficient extractable data.

No language restrictions were applied during database searching. However, all studies that met the eligibility criteria and provided extractable accuracy outcomes were available as full texts in English; therefore, all included studies were English-language full texts. A comprehensive overview of the specific search terms employed, the databases accessed, and the number of results that each search yielded is provided in supplementary file 1.

After screening the retrieved records using their titles and abstracts in the first step, we evaluated the full texts of the included records to determine their eligibility. We compiled comprehensive summaries for each study included, detailing their characteristics, methodologies, and relevant outcomes (Table 1).

Research strategy and screening

Four databases were searched: MEDLINE via PubMed, Embase, Scopus, and Web of Science. Searches were conducted within the native interfaces/platforms of PubMed, Embase, Scopus, and Web of Science. Within Scopus, we additionally applied the 'Secondary documents' option to capture additional indexed record types, which may include items originating from preprint servers (eg, medRxiv) when indexed in Scopus. We did not conduct a separate search

directly within preprint repositories. All retrieved records were exported to a reference manager for deduplication prior to screening. Deduplication was then performed using automated matching (DOI, title, and author) followed by manual verification of uncertain matches, and only the deduplicated set proceeded to title/abstract screening, the search conducted in January 2025 using the queries listed in Supplementary Table 1. Titles and abstracts were independently screened by D.H. and M.H.A., with disagreements resolved by M.D. The same two reviewers (D.H. and M.H.A.) also performed the full-text assessment and data extraction independently, and any discrepancies were again adjudicated by M.D. Only studies meeting all eligibility criteria and providing complete data were included in the final analysis. Deduplication was performed using automated matching (DOI, title, and author) followed by manual verification of uncertain matches. After deduplication, each unique record was screened once to prevent double counting across databases.

The data extracted from the included articles are summarized in Table 1. These include author details, publication year, origin country, question type, number of questions, LLM model, and accuracy of correctly answered questions. A PRISMA flow chart, created using R software,¹⁵ was produced to illustrate the selection process (Figure 1).

Quality evaluation

Risk of bias and applicability concerns were assessed using a modified QUADAS-2 framework appropriate for studies evaluating large language models on question-answering tasks. Two reviewers (D.H. and M.H.A.) independently evaluated all included studies using predefined signalling questions, with disagreements resolved by a third reviewer (M.D.). Each domain was judged as low, high, or unclear risk of bias, with unclear ratings assigned when reporting was insufficient. Full domain-level judgments and the complete signalling-question checklist are provided in Supplementary Table 2.

Domain mapping and signalling questions

- Domain 1: Item selection and sampling of questions
Signalling questions: Were questions sourced from official examinations or validated question banks? Were inclusion and exclusion criteria for items prespecified? Was any selective or post-hoc inclusion likely to inflate accuracy?
Applicability focus: Does the question pool represent the intended dental examination domain?
- Domain 2: Index test procedures for LLM response generation and scoring
Signalling questions: Were prompts, temperature, and model versions prespecified and applied consistently? Were raters blinded to the reference standard during scoring? Were tie-break and partial-credit rules defined in advance?
Applicability focus: Do the prompting and evaluation conditions reflect realistic examination use?
- Domain 3: Reference standard
Signalling questions: Was the reference based on an official answer key or expert consensus? Was the reference determined independently of the LLM output? Were

Table 1 – Summary of data extracted from the included studies.

N0	Study	Country	Question field	Number of questions	LLM model	Correct answer (percentage) with SD and CI	Source of questions	Question format	Image or text questions
01	Acar ¹⁸	Turkey	Oral surgery	20	ChatGPT-3.5	88.40 ± 0.06	Author-constructed	Open-ended	Text-only
					Microsoft Bing	75.10 ± 0.05			
					Google Bard	70.50 ± 0.02			
02	Ali et al ¹⁹	Qatar	healthcare education	50	ChatGPT-3.5	88%	Author-constructed	Mixed	Text-only
03	Alsayed et al ²⁰	Saudi Arabia	Oral surgery	15	ChatGPT-4	78%	Author-constructed	Open-ended	Text-only
			Preventive dentistry	15	ChatGPT-4	86%			
			Oral cancer	20	ChatGPT-4	72%			
04	Azadi et al ²¹	Iran	Oral surgery	50	Bard	34%	Author-constructed	Mixed	Text-only
					GPT-3.5	36%			
					GPT-4	38%			
					Claude-Instant	38%			
					Microsoft Bing	26%			
05	Batool et al ²²	Pakistan	Operative/ Endodontics	40	ChatGPT3.5	60%	Author-constructed	Open-ended	Text-only
			Periodontics	40	ChatGPT3.5	50%			
			Oral Surgery	40	ChatGPT3.5	40%			
			Prosthodontics	40	ChatGPT3.5	60%			
06	Brozovic et al ²³	Croatia	General dentistry	532	Microsoft Bing	71.99%	Mixed/multiple sources	Mix	Text-only
07	Cai et al ²⁴	China	Oral surgery	30	ChatGPT	100%	Author-constructed	Open-ended	Text-only
08	Chau et al ²⁵	China	UK and US Dental Licensing exam/ Official exam materialinations	1461	ChatGPT 3.5	43.3%	Licensing exam/Official exam material	MCQ	Text-only
					ChatGPT 4	62.7			
09	Danesh et al ²⁶	Canada	Periodontic	311	ChatGPT3.5	57.9%	Licensing exam/Official exam material	MCQ	Text-only
					ChatGPT4	73.6%			
10	Danesh et al ²⁷	Canada	General Dentistry	143	ChatGPT3.5	61.3%	Mixed/multiple sources	MCQ	Text-only
					ChatGPT4	76.9%			
11	Daraqel et al ²⁸	China	Orthodontics	100	ChatGPT3.5	90% (80%-90%, 95%CI)	Author-constructed	Open-ended	Text-only
					Google Bard	80% (80%-90%, 95%CI)			
12	Dashti et al ²⁹	Iran	INBDE, ADAT, and DAT	122 (knowledge-based) 60 (Case history)	ChatGPT 3.5	78%	Licensing exam/Official exam material	MCQ	Text-only
					ChatGPT4	88%			
					ChatGPT 3.5	70%			
					ChatGPT4	71%			
13	Díaz-Flores García et al ³⁰	Spain	Endodontics	30	Google Gemini	37.11% ± 2% (34.02%-40.32%, 95%CI)	Author-constructed	Open-ended	Text-only
14	Dursun et al ³¹	Turkey	Orthodontics	20	ChatGPT-3.5	83% ± 15%	Author-constructed	Open-ended	Text-only
					ChatGPT-4	90% ± 12.2%			
					Gemini	82% ± 14.4%			

(continued on next page)

Table 1. (Continued)

N0	Study	Country	Question field	Number of questions	LLM model	Correct answer (percentage) with SD and CI	Source of questions	Question format	Image or text questions
15	Farajollahi et al ³²	Iran	Endodontics	100	Copilot ChatGPT-3.5	87% ± 16.2% 40%	Licensing exam/Official exam material	MCQ	Text-only
16	Freire et al ³³	Spain	Prosthodontics	30	ChatGPT-4	25.6% (22.9%-28.6%, 95%CI)	Author-constructed	Open-ended	Text-only
17	Fuchs et al ³⁴	Switzerland	Swiss Federal Licensing exam/Official exam material Dental Medicine (SFLEDM)	32	ChatGPT-3 ChatGPT 4	62.6% ± 3.3% 66.7% ± 3.2%	Licensing exam/Official exam material	MCQ	Text-only
18	Giannakopoulos et al ³⁵	Cyprus	General dentistry	20	ChatGPT 4 ChatGPT 3.5 Bard Bing Chat	72% ± 18% 59% ± 24% 57% ± 31% 54% ± 34%	Author-constructed	Open-ended	Text-only
19	Hatia et al ³⁶	Italy	Orthodontics	21	ChatGPT 4	81.6%	Author-constructed	Open-ended	Text-only
20	Jacobs et al ³⁷	USA	Oral surgery	25	ChatGPT 4 ChatGPT 3.5	81.6% 87.2%	Author-constructed	Open-ended	Text-only
21	Jaworski et al ³⁸	Poland	Polish Final Dentistry Examination (LDEK)	11 (Clinical case) 188 (Knowledge)	ChatGPT-4o	36.36% 72.87%	Licensing exam/Official exam material	MCQ	Text-only
22	Jeong et al ³⁹	Republic of Korea	General dentistry	16	ChatGPT 3.5 ChatGPT 4 Bard Bing	31.3% 93.8% 50.0% 68.8%	Licensing exam/Official exam material	Mixed	Text-only
23	Johnson et al ⁴⁰	India	Dental trauma	20	Gemini Claude AI ChatGPT 3.5 Bing	80% 80% 75% 40%	Author-constructed	Open-ended	Text-only
24	Künzle et al ⁴¹	Germany	Restorative Endodontics	151	ChatGPT-4.o Chat-GPT-4 ChatGPT-3.5 Gemini 1.0	72% 62% 25% 44%	Question bank	MCQ	Text-only
25	Makrygiannakis et al ⁴²	Greece	Orthodontics	10	Google's Bard ChatGPT-3.5 Chat-GPT-4 Microsoft's Bing	4.6/10 46% 3.8/10 38% 4.7/10 47% 7.1/10 71%	Author-constructed	Open-ended	Text-only
26	Mohammad-Rahimi et al ⁴³	USA	Oral pathology, oral medicine, oral radiology	66	Microsoft's Bing ChatGPT3.5 ChatGPT4 Google Bard Claude	61.82% ± 17.64% 77.48% ± 17.58% 81.32% ± 16.5% 68.48% ± 17.4% 77.08% ± 20.76%	Author-constructed	Open-ended	Text-only

(continued on next page)

[illegible]

(continued on next page)

Table 1. (Continued)

N0	Study	Country	Question field	Number of questions	LLM model	Correct answer (percentage) with SD and CI	Source of questions	Question format	Image or text questions
34	Suárez et al ⁵¹	Spain	Examination (2019-2023) Endodontic	60	Gemini ChatGPT-4	61.7% 57.33% ± 0.82% (55.72% -58.95%, 95%CI)	Author-constructed	MCQ (dichotomous)	Text-only
35	Suárez et al ⁵²	Spain	Oral surgery	30	ChatGPT-4	71.7% (68.9 –74.6%, 95%CI)	Mixed/multiple sources	Open-ended	Not reported
36	Turunç Oğuzman et al ⁵³	Turkey	Dental specialty Entrance exam in Turkey	Prosthodontics 126 Radiology 123	ChatGPT 3.5 Bard ChatGPT 3.5 Bard	35.7% 38.9% 52.8% 52.8%	Licensing exam/Official exam material	MCQ	Text-only
37	Vaira et al ⁵⁴	Italy	Oral Surgery	72 Open-ended 72 True/False	ChatGPT 4 ChatGPT 4	81.7% 84.7%	Author-constructed	Mixed	Text-only
38	Yamaguchi et al ⁵⁵	Japan	Japanese National Dental Hygienist Examination	73	ChatGPT 3.5 ChatGPT 4 Bard	63.0% 75.3% 66.7%	Licensing exam/Official exam material	MCQ	Text-only
39	Yunsun et al ⁵⁶	South Korea	Pediatric	32	Bing Chat ChatGPT 3.5 Gemini	68.5% 35.3% ± 5.6% 33.0% ± 4.0%	Licensing exam/Official exam material	MCQ	Text-only

CI, confidence interval; SD, standard deviation; Not reported, indicates the primary study did not explicitly describe the item source or format.

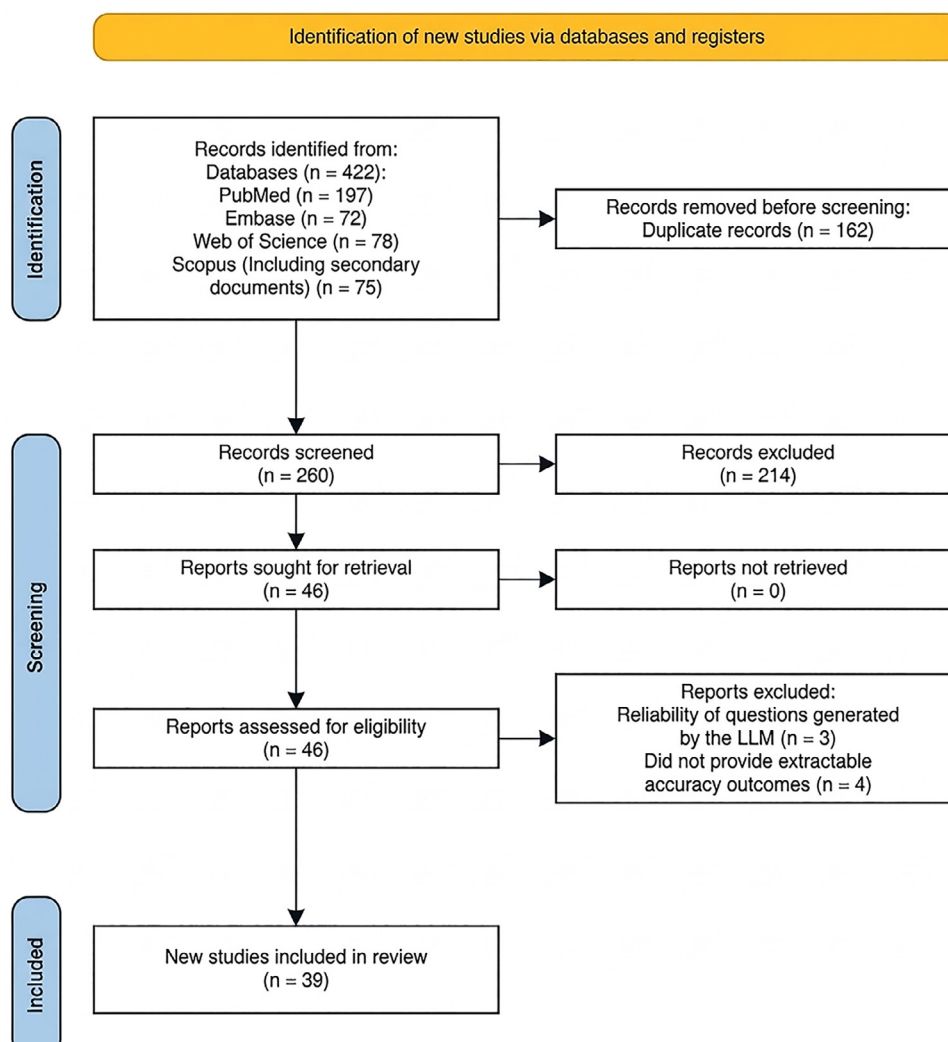


Fig. 1 – PRISMA flow diagram of study selection.

discrepancies resolved through a documented adjudication process?

Applicability focus: Does the reference standard accurately represent correct dental knowledge?

- Domain 4: Flow and timing

Signalling questions: Were all items included in the analysis? Was the test order consistent across models? Were repeated queries, model updates, or use of external tools controlled or reported?

Applicability focus: Does the test flow reflect a stable and unbiased evaluation process?

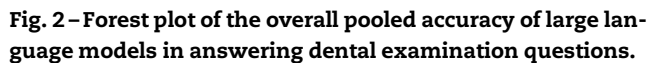
Summary measures and data synthesis

A meta-analysis was conducted using data from the included studies to determine the proportion of correct answers for each LLM. In cases where publications evaluated multiple AI models, each model was analysed individually and designated by the model version and author name.

To avoid conflating brand names with underlying models, we harmonized labels across studies. Microsoft Copilot (formerly called 'Bing Chat/AI' during our time window of

evaluation) is an interface powered by OpenAI GPT with Microsoft safety/policy layers; we analyse it as Copilot (GPT family interface). Google Bard was rebranded to Gemini; when studies reported Bard during the Bard to Gemini transition, we classified them under Gemini (Bard-era). Where a paper explicitly evaluated ChatGPT-4 within OpenAI's ChatGPT product, we label it ChatGPT-4 (OpenAI interface). These decisions preserve UI/prompting differences without mislabelling Copilot as a distinct base model.

Statistical analyses were performed using R 4.4.1 and RStudio, with the 'meta, metafor, metasens, dplyr, and ggplot2' packages. A random-effects model using the Inverse variance method with Freeman-Tukey double arcsine transformation was employed for the primary analysis. Heterogeneity was assessed using I^2 statistics, with values of 25%, 50%, and 75% indicating low, moderate, and high heterogeneity, respectively. Publication bias was evaluated using funnel plots and Egger's test.¹⁶ Subgroup analyses on different LLM models were performed to explore potential sources of heterogeneity. To ensure uniformity, proportion values expressed as percentages in the studies were transformed into decimal format.



Robustness was assessed through several methods, including leave-one-out influence analysis, alternative estimators such as DerSimonian-Laird and maximum likelihood, comparison of fixed-effect models, and Baujat diagnostic plots to identify influential studies.¹⁷ The fixed-effect

The 39 included studies used heterogeneous question sources and formats. Question provenance most commonly consisted of author-constructed question sets ($n = 21/39$, 53.8%), followed by licensing or official examination materials ($n = 13/39$, 33.3%). A smaller proportion used question banks ($n = 2/39$, 5.1%) or mixed/multiple sources ($n = 3/39$, 7.7%). Regarding question format, studies used multiple-choice questions (MCQ, including dichotomous items) in 17/39 studies (43.6%), open-ended formats in 17/39 (43.6%), and mixed formats in 5/39 (12.8%). Almost all studies evaluated text-only questions (38/39, 97.4%), while image involvement was not reported in one study (1/39, 2.6%). These variations represent plausible contributors to between-study heterogeneity and were therefore captured in [Table 1](#) for transparent interpretation of pooled estimates. These variables were reported descriptively to characterize heterogeneity; they were not used as prespecified moderators in meta-regression due to inconsistent reporting across studies.

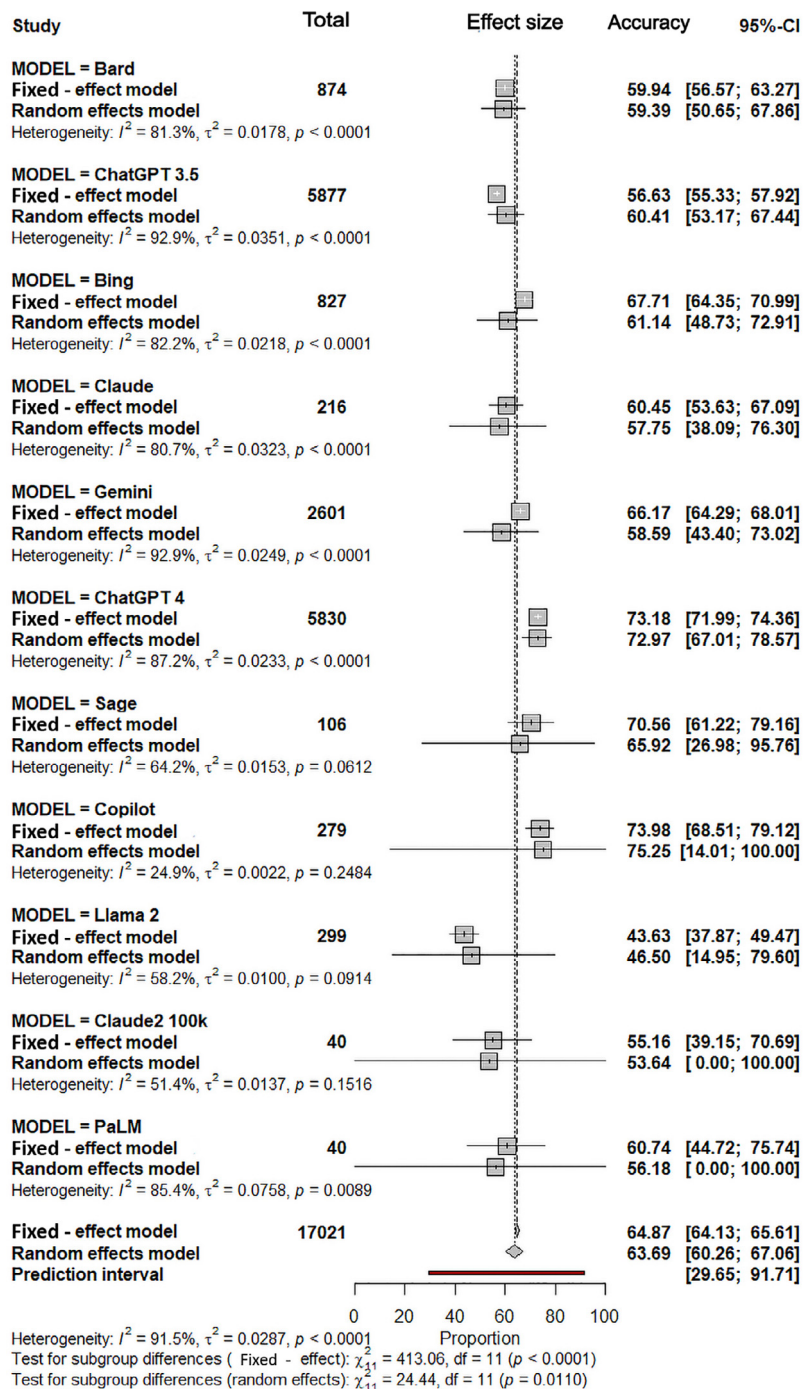


Fig. 3 – Forest plot of subgroup analysis according to large language model type.

Risk of bias assessment

Under the item selection and sampling domain, 22 of the 39 included studies (56.4%) were judged at high risk of bias, primarily due to the use of author-generated or non-standardized question sets rather than official or validated examination items. While such question sources may introduce subjectivity and reduce comparability across studies, they consistently aimed to assess the same construct, namely the accuracy of LLMs in answering dental knowledge

questions, and were therefore considered unlikely to alter the overall direction of the pooled results.

All studies (100%) were rated as low risk in the index test domain, reflecting clear documentation of prompting procedures, independent scoring, and reproducible conditions. The reference standard domain was low risk in 92.3% (36/39) of studies that used official answer keys or expert consensus; three studies (7.7%) had unclear risk because of insufficient reporting. The flow and timing domain was low risk in all studies, with complete inclusion of evaluated items and no evidence of selective omission.

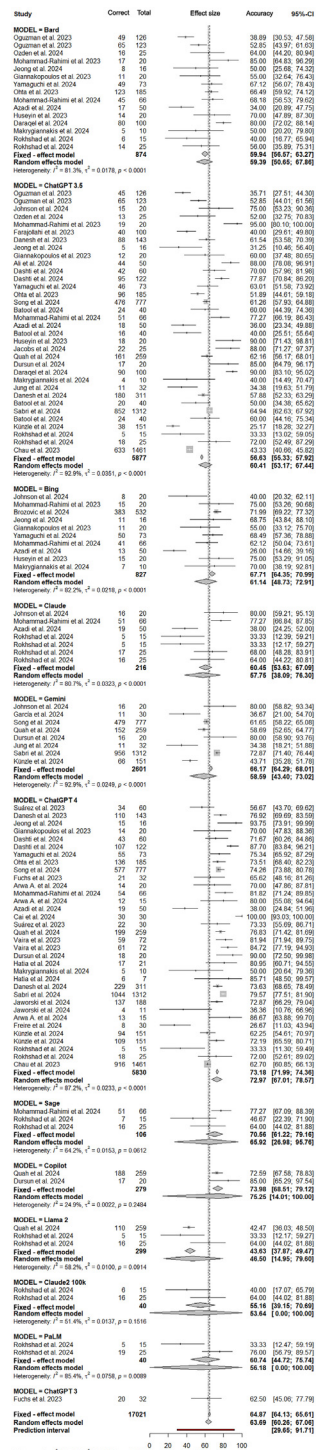


Fig. 4 – Forest plot of individual study estimates grouped by large language model subgroup.

Regarding applicability concerns, most studies were rated low concern across domains. The adapted 'item selection' domain showed moderate concern in studies using locally constructed question banks, as these may not fully represent standardized dental curricula. The index test and reference standard domains showed low concern, indicating that the evaluated models and scoring criteria were appropriate to the

intended research question. Overall, despite variability in question sources, methodological consistency and transparent reporting across studies support the robustness and applicability of the meta-analytic findings (Supplementary Table 2).

Meta-analysis

Pooled effect size estimates

Under the random-effects model, the meta-analysis of 39 studies yielded an overall pooled answer rate of 63.69% (95% CI: 60.26%-67.06%), with substantial between-study heterogeneity ($I^2 = 91.5\%$, $\tau^2 = 0.0287$; $P < .0001$). Prediction intervals spanning 29.65% to 91.71% highlight the variability in expected correct answer rate across future implementations. Sensitivity analyses comparing fixed- and random-effects models revealed similar pooled estimates, with fixed-effects results at 64.87% (95% CI: 64.13%-65.61%) (Figure 2).

Subgroup analysis by LLM model

Subgroup analyses stratified by architecture demonstrated statistically significant differences in performance (Q-between = 24.44, $P = .011$). Copilot (GPT family interface) and ChatGPT-4 (OpenAI interface) achieved the highest pooled estimates (Copilot: 75.25%; 95% CI: 14.01%-100%; ChatGPT4: 72.97%; 95% CI: 67.01%-78.57%), with limited heterogeneity ($I^2 = 24.9\%$, $P = .25$) for Copilot and higher heterogeneity ($I^2 = 87.2\%$, $P < .0001$) for ChatGPT. Overall, 34 studies employed these two LLM-based chatbots/interfaces. For the majority of comparisons of accuracy, significant differences between different LLMs were identified (Figures 3 and 4).

Publication bias

Quantitative assessments of publication bias yielded mixed evidence. While Egger's regression test indicated no small-study effects ($P = .56$), Begg's test results ($P = .31$) suggested consistency with funnel plot asymmetry. Trim-and-fill analysis imputed six hypothetical missing studies, adjusting the pooled accuracy to 65.43% (95% CI: 61.88%-68.89%), ie, very similar to the base-case analysis.

Sensitivity analysis

Methodological robustness was confirmed through multiple sensitivity checks. Leave-one-out analysis demonstrated stable pooled accuracy estimates ranging from 64.13% to 65.66%, with the most influential study (Chau et al²⁵) altering results by 1.96% upon exclusion (Figure 5).

Discussion

Since its launch in November 2022, ChatGPT has gained widespread popularity given its ease of use, accessibility, and ability to automate various cognitive tasks in a time-efficient manner. Ever since, ChatGPT and several other LLM-based

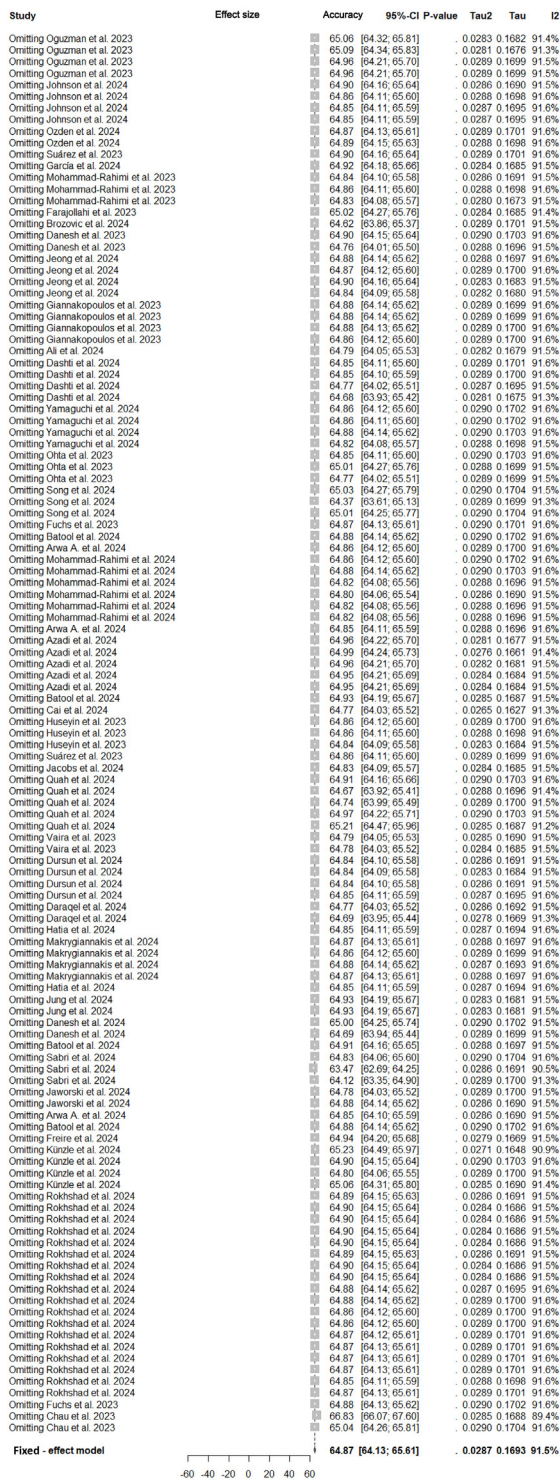


Fig. 5 – Leave-one-out sensitivity analysis of pooled accuracy.

tools have become increasingly adopted across many fields, including medical and dental sciences.^{6,57,58} Twelve of the included studies were published in 2023, compared with 27 in 2024. This reflects the great interest these LLMs have garnered in the dental community since their launch, a trend expected to persist as LLMs advance and new developments are tested for dental use. Notably, two studies also

investigated LLMs' accuracy in answering dental hygienist examination questions, highlighting that these models can benefit other professionals in oral healthcare.^{50,55}

Implementations vary from diagnosis and decision-making to education and research.^{59,60} The accuracy of the models' output has been a topic of debate, with numerous researchers designing studies to evaluate LLMs' performance against standardized examination questions.^{31,39,44,61} In detail, the accuracy of LLMs in tackling medical examination questions has been previously systematically reviewed.⁶¹ However, that is not the case for the dental field. The present study bridges this knowledge gap by synthesizing a meta-analytic estimate of the accuracy of contemporary LLMs in answering dental questions. As of January 2025, the results showed that OpenAI's ChatGPT was the most widely examined model, with ChatGPT-4 being the most accurate among those evaluated. The pooled accuracy of the LLMs amounted to 63.69% (95% CI: 60.26%-67.06%). Because Copilot is a GPT-powered interface rather than a distinct base LLM, its slightly higher accuracy likely reflects interface-level differences (guardrails, context windows, search/grounding, default prompting).

The results of the present study are in line with a previous systematic review investigating LLMs' accuracy in responding to medical questions,⁶² which concluded that ChatGPT was the most widely used and most accurate model, particularly in diagnosis and classification tasks.⁶² Furthermore, ChatGPT-4 outperformed version 3.5, achieving accuracies above 90% in some instances.^{31,39,44} This is comparable to recent studies that reported accuracies of up to 100% in diagnosing clinical cases.^{63,64} However, in some of these evaluations, the questions and their correct answers may have been publicly available, which introduces the possibility that such content was included in the training data of the evaluated LLMs, potentially inflating performance estimates. Notably, the LLMs demonstrated superior performance when answering dental questions compared to medical questions, with a recent systematic review reporting an accuracy of 56%.⁶⁵ Nonetheless, the models showed inconsistencies. Results varied widely, with some models scoring as low as 25%,⁴¹ which is in agreement with previous lower-range accuracy estimates of 18.3%.⁶⁴

The pooled accuracy of 63.69% falls within the statistically acceptable range but may be unacceptable for reliable use in clinical practice, where patient safety could be compromised by inaccurate LLM recommendations. Furthermore, because these models are trained on data available only up to a certain point, they may not always reflect the most recent clinical practice guidelines, nor are they necessarily able to comprehensively assess a clinical case scenario. Some newer systems offer optional online search or retrieval mechanisms that can supplement their training data, but this only partially mitigates these limitations. Therefore, such models should not be relied upon for real-world clinical decision-making, but may offer value as supporting tools in dental education or as an aid in preparing for dental examinations, provided their limitations are clearly recognized. As noted by Hager et al,⁶⁶ current LLMs fail to consistently follow diagnostic or treatment guidelines across pathologies and cannot reliably interpret laboratory results, which underscores their

unsuitability for autonomous clinical decision-making. In dentistry specifically, Giannakopoulos et al³⁵ warn that imprudent use of LLMs may adversely impact patient care. Additionally, given the potential of LLMs to produce misinformation,⁶⁷ they currently cannot replace human knowledge and skill.⁶⁸

Our review also differs from prior dentistry-focused syntheses that concentrated specifically on dental licensing examinations. For example, Liu et al⁶⁹ evaluated LLM performance in dental licensing examinations worldwide. In contrast, we used broader eligibility criteria that included licensing examinations, specialty board-style assessments, validated educational question banks, and dentistry-relevant knowledge assessments across multiple subdisciplines. We also separated model/version and interface labels (eg, Copilot as a GPT-based interface), incorporated a QUADAS-2–based risk-of-bias assessment tailored to LLM question-answering studies, and performed model-stratified pooling with multiple sensitivity diagnostics to test robustness.

Our pooled estimates suggest that LLM performance in dentistry has been evaluated mainly in controlled, text-based knowledge settings, most commonly using multiple-choice questions with a defined answer key. This format resembles exam preparation and foundational knowledge checks rather than real-world clinical decision-making. Consequently, the observed accuracy should not be interpreted as evidence that LLMs can safely perform independent clinical reasoning, treatment planning, or documentation without expert oversight.

A key finding is the substantial heterogeneity across studies, which is plausibly driven by differences in question format (multiple-choice versus open-ended), question provenance (official licensing examinations and validated banks versus author-constructed sets), and whether items were scenario-based. In particular, author-constructed question sets can systematically influence measured accuracy depending on topic selection, wording, and difficulty distribution. This aligns with our risk-of-bias assessment, where item selection and reporting practices were common concerns. These issues likely explain why pooled accuracy varies across studies even for the same named model family.

From an educational perspective, the results support cautious use of LLMs as supplementary learning tools (eg, generating explanations, highlighting key concepts, and supporting formative self-testing) while emphasizing that correctness must be verified, especially in high-stakes domains such as oral pathology/oncology, pharmacology, and medical emergencies. The strongest near-term value is therefore likely in education and structured knowledge support, rather than autonomous clinical recommendation. Our findings complement prior work focused on dental licensing examinations⁶⁹ by including a broader range of dentistry-relevant assessments and by explicitly examining heterogeneity sources across study designs. Broader evidence from licensing-exam evaluations in health professions similarly emphasizes variability by question type and the need for careful governance and oversight.⁷⁰

Jin et al⁷¹ also pooled ChatGPT-3.5 and GPT-4 performance across national health licensing examinations (medicine, pharmacy, dentistry, nursing) and reported an overall

accuracy around 70%, with GPT-4 consistently outperforming ChatGPT-3.5. Dentistry constituted only a small portion of their included evidence and the review was restricted to multiple-choice licensing items. In contrast, our review is dentistry-focused and includes broader question sources and formats, reinforcing that reported accuracy is highly context-dependent and should not be generalized across tasks or model releases.

Training data transparency and other ethical issues remain concerns with these LLMs.⁷² Recent work has highlighted that dental applications raise additional governance challenges, including data provenance, privacy, fairness, explainability, and responsibility during clinical implementation, underscoring the need for structured ethical frameworks when deploying LLMs in practice.⁷²

While it appears that some dental-related content was used in their training, it is not clear on which data they were trained. This makes it problematic to countercheck the reliability of LLMs' answers, as they may be based on inaccurate or outdated information and perpetuate biases in the training data.⁷³⁻⁷⁵ After all, these models function as sophisticated pattern recognition systems and do not generate new knowledge nor critically assess data before synthesizing an output. Nevertheless, recent evidence suggests that fine-tuning LLMs with carefully curated resources and using contextual prompting can significantly improve their accuracy in answering specialized dental examination questions.⁷⁶ The inherent variability in dental literature and clinical practice guidelines means that model outputs reflect the biases and inconsistencies present in their training data. The limited explainability of models' output is a relatively common issue with many AI applications,⁷⁷ making it difficult to assess the reasoning behind their specific answers or recommendations.

This review has several limitations that should be acknowledged. Substantial heterogeneity was observed across the included studies, reflecting variability in question sources, formats, dental specialties, model versions, and evaluation procedures. High heterogeneity is common in accuracy and prevalence meta-analyses; notably, a recent methodological review reported a median I^2 value of 96.9% across 134 meta-analyses,⁷⁸ underscoring that elevated heterogeneity alone does not invalidate pooled estimates when robustness is supported by sensitivity analyses, as observed in the present study. Limitations of the evidence base include the exclusion of non-English-language full texts, which may reflect language or indexing bias, and that most included studies evaluated textbook-style or examination-based questions rather than complex real-world clinical scenarios, which limits generalizability to routine clinical practice. In addition, incomplete reporting of model parameters, platform versions, and query conditions in several studies constrained reproducibility and risk-of-bias assessment. Finally, a further limitation is the absence of prospective protocol registration, which may reduce transparency and increase the risk of selective reporting despite our detailed reporting of methods and analyses.

It is also noteworthy that the rapid pace of LLM updates is an important interpretive constraint. Most primary studies assessed specific, time-stamped versions (commonly GPT-3.5

and GPT-4) because those were the publicly accessible models during their study periods. Our meta-analysis therefore estimates performance of the versions evaluated in the published literature up to 31 January 2025, rather than claiming that the same accuracy applies to newer model releases. Ongoing model updates may change performance; periodic (or 'living') updates of this review will be necessary as new dentistry-specific evaluations of newer models become available.

Taken together, these findings indicate that while current LLMs demonstrate moderate accuracy in dental knowledge assessments, they are not suitable for autonomous clinical decision-making. When their limitations are explicitly recognized, however, LLMs may serve as useful adjuncts in dental education and examination preparation, provided that outputs are interpreted critically and under professional supervision.

Conclusion

This systematic review and meta-analysis demonstrates that large language models currently achieve moderate accuracy when answering dental examination questions. Although newer GPT-4-class systems generally outperform earlier versions, overall performance remains insufficient for unsupervised clinical application. Given the precision required for dental diagnosis and treatment planning, LLMs should not be relied upon as autonomous decision-making tools. When their methodological and conceptual limitations are appropriately acknowledged, however, these systems may offer meaningful value as supportive resources in dental education and examination preparation. Future research should prioritize standardized evaluation protocols, transparent reporting of model configurations, and systematic investigation of performance-enhancing strategies such as structured contextual prompting and retrieval-augmented approaches to better define the appropriate role of LLMs in dentistry.

Author contributions

M.D.: Conception and design of study, Acquisition of data, Drafting of article and/or critical revision, Final approval of manuscript. F.Kh.: Conception and designed of study, Final approval of manuscript, Analysis of data. A.M.: Acquisition of data, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript. M.A.: Conception and design of study, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript. A.Ch.: Conception and design of study, Acquisition of data, Drafting of article and/or critical revision, Final approval of manuscript. D.H.: Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript. N.Gh.: Acquisition of data, Analysis of data, Drafting of article and/or critical revision, Final approval of manuscript. A.T.: Final approval of manuscript, Drafting of article and/or critical revision. Z.Kh.: Conception and design of study, Drafting of article and/or critical revision, Final approval of manuscript. F.Sch.: Final approval of manuscript, Drafting of article and/or critical revision. All

authors have critically reviewed and approved the final draft and are responsible for the content and similarity index of the manuscript.

Funding

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

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

The author is an Editorial Board Member for this journal and was not involved in the editorial review or the decision to publish this article. Given his role as editorial board member, had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

During the preparation of this work the authors used ChatGPT-5.2 (OpenAI, San Francisco, CA, USA) program software in order to improve the flow and grammar of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.identj.2026.109609](https://doi.org/10.1016/j.identj.2026.109609).

REFERENCES

1. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in health-care: past, present and future. *Stroke Vasc Neurol* 2017;2 (4):230–43. doi: [10.1136/svn-2017-000101](https://doi.org/10.1136/svn-2017-000101).
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44–56. doi: [10.1038/s41591-018-0300-9](https://doi.org/10.1038/s41591-018-0300-9).
3. Hariri W. Unlocking the potential of ChatGPT: a comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv*. 2023. doi: [10.48550/arXiv.2304.02017](https://doi.org/10.48550/arXiv.2304.02017).
4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023;2: e0000198. doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198).

5. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res* 2020;99(7):769–74. doi: [10.1177/0022034520915714](https://doi.org/10.1177/0022034520915714).
6. Alhaidry HM, Fatani B, Alrayes JO, Alman A, Alhaed NK. ChatGPT in dentistry: a comprehensive review. *Cureus* 2023;15:e38317. doi: [10.7759/cureus.38317](https://doi.org/10.7759/cureus.38317).
7. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv* 2023;55:248. doi: [10.1145/3571730](https://doi.org/10.1145/3571730).
8. Eraslan R, Ayata M, Yagci F, et al. Exploring the potential of artificial intelligence chatbots in prosthodontics education. *BMC Med Educ* 2025;25:321. doi: [10.1186/s12909-025-06849-w](https://doi.org/10.1186/s12909-025-06849-w).
9. Suh PS, Shim WH, Suh CH, et al. Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro vision using image inputs from diagnosis please cases. *Radiology* 2024;312(1):e240273. doi: [10.1148/radiol.240273](https://doi.org/10.1148/radiol.240273).
10. Ahmad B, Saleh K, Alharbi S, Alqaderi H, Jeong YN. Artificial intelligence in periodontology: performance evaluation of ChatGPT, claude, and gemini on the in-service examination. *medRxiv*. 2024. doi:[10.1101/2024.05.16.24307451](https://doi.org/10.1101/2024.05.16.24307451)
11. Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models (Version 1). *arXiv*. 2023. doi:[10.48550/arXiv.2302.13971](https://doi.org/10.48550/arXiv.2302.13971)
12. DeepSeek-AI, Bi X, Chen D, et al. DeepSeek LLM: scaling open-source language models with longtermism. *arXiv*. 2024. doi:[10.48550/arXiv.2401.02954](https://doi.org/10.48550/arXiv.2401.02954)
13. Dashti M, Londono J, Ghasemi S, et al. Attitudes, knowledge, and perceptions of dentists and dental students toward artificial intelligence: a systematic review. *J Taibah Univ Med Sci* 2024;19(2):327–37. doi: [10.1016/j.jtumed.2023.12.010](https://doi.org/10.1016/j.jtumed.2023.12.010).
14. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009;151:264–9. doi: [10.7326/0003-4819-151-4-200908180-00135](https://doi.org/10.7326/0003-4819-151-4-200908180-00135).
15. Haddaway NR, Page MJ, Pritchard CC, McGuinness LA. PRISMA2020: an R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst Rev* 2022;18:e1230. doi: [10.1002/cl2.1230](https://doi.org/10.1002/cl2.1230).
16. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315(7109):629–34. doi: [10.1136/bmj.315.7109.629](https://doi.org/10.1136/bmj.315.7109.629).
17. Baujat B, Mahé C, Pignon JP, Hill C. A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials. *Stat Med* 2002;21(18):2641–52. doi: [10.1002/sim.1221](https://doi.org/10.1002/sim.1221).
18. Acar AH. Can natural language processing serve as a consultant in oral surgery? *J Stomatol Oral Maxillofac Surg* 2024;125(3):101724. doi: [10.1016/j.jormas.2024.101724](https://doi.org/10.1016/j.jormas.2024.101724).
19. 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;28(1):206–11. doi: [10.1111/eje.12938](https://doi.org/10.1111/eje.12938).
20. Alsayed AA, Aldajani MB, Aljohani MH, et al. Assessing the quality of AI information from ChatGPT regarding oral surgery, preventive dentistry, and oral cancer: an exploration study. *Saudi Dent J* 2024;36:1305–14. doi: [10.1016/j.sdentj.2024.07.012](https://doi.org/10.1016/j.sdentj.2024.07.012).
21. Azadi A, Gorjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2024;137(6):587–93. doi: [10.1016/j.oooo.2024.03.013](https://doi.org/10.1016/j.oooo.2024.03.013).
22. Batool I, Naved N, Kazmi SMR, Umer F. Leveraging large language models in the delivery of post-operative dental care: a comparison between an embedded GPT model and ChatGPT. *BDJ Open* 2024;10(1):48. doi: [10.1038/s41405-024-00256-9](https://doi.org/10.1038/s41405-024-00256-9).
23. Brozovic J, Mikulic B, Tomas M, Juzbas M, Blaskovic M. Assessing the performance of Bing Chat artificial intelligence: dental exams, clinical guidelines, and patients' frequent questions. *J Dent* 2024;144:104981. doi: [10.1016/j.jdent.2024.104981](https://doi.org/10.1016/j.jdent.2024.104981).
24. Cai Y, Zhao R, Zhao H, Li Y, Gou L. Exploring the use of ChatGPT/GPT-4 for patient follow-up after oral surgeries. *Int J Oral Maxillofac Surg* 2024;53(10):867–72. doi: [10.1016/j.ijom.2024.07.003](https://doi.org/10.1016/j.ijom.2024.07.003).
25. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. *Int Dent J* 2024;74(3):616–21. doi: [10.1016/j.identj.2024.02.004](https://doi.org/10.1016/j.identj.2024.02.004).
26. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. *J Periodontol* 2024;95(7):682–7. doi: [10.1002/JPER.24-0015](https://doi.org/10.1002/JPER.24-0015).
27. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: a preliminary study on ChatGPT. *J Am Dent Assoc* 2023;154(11):970–4. doi: [10.1016/j.adaj.2023.07.018](https://doi.org/10.1016/j.adaj.2023.07.018).
28. Daraql B, Wafaie K, Mohammed H, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. *Am J Orthod Dentofacial Orthop* 2024;165(6):652–62. doi: [10.1016/j.ajodo.2023.11.008](https://doi.org/10.1016/j.ajodo.2023.11.008).
29. Dashti M, Ghasemi S, Ghadimi N, et al. Performance of ChatGPT 3.5 and 4 on U.S. dental examinations: the INBDE, ADAT, and DAT. *Imaging Sci Dent* 2024;54(3):271–5. doi: [10.5624/isd.20240047](https://doi.org/10.5624/isd.20240047).
30. Díaz-Flores García V, Freire Y, Tortosa M, Tejedor B, Estevez R, Suárez A. Google Gemini's performance in endodontics: a study on answer precision and reliability. *Appl Sci* 2024;14(15):6627. doi: [10.3390/app14156627](https://doi.org/10.3390/app14156627).
31. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak* 2024;24(1):211. doi: [10.1186/s12911-024-02598-x](https://doi.org/10.1186/s12911-024-02598-x).
32. Farajollahi M, Modaberi A. Can ChatGPT pass the "Iranian Endodontics Specialist Board" exam? *Iran Endod J* 2023;18(3):192. doi: [10.22037/iej.v18i3.38988](https://doi.org/10.22037/iej.v18i3.38988).
33. Freire Y, Laorden AS, Perez JO, Sanchez MG, García VDF, Suarez A. ChatGPT performance in prosthodontics: assessment of accuracy and repeatability in answer generation. *J Prosthet Dent* 2024;131(4):e59–67. doi: [10.1016/j.pros-dent.2023.09.021](https://doi.org/10.1016/j.pros-dent.2023.09.021).
34. Fuchs A, Trachsel T, Weiger R, Eggmann F. ChatGPT's performance in dentistry and allergy-immunology assessments: a comparative study. *Swiss Dent J* 2023;134(5):384–93. doi: [10.36477/sdj-2023-138](https://doi.org/10.36477/sdj-2023-138).
35. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. *J Med Internet Res* 2023;25:e51580. doi: [10.2196/51580](https://doi.org/10.2196/51580).
36. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. *J Clin Med* 2024;13(3):743. doi: [10.3390/jcm13030743](https://doi.org/10.3390/jcm13030743).
37. Jacobs T, Shaari A, Gazonas CB, Ziccardi VB. Is ChatGPT an accurate and readable patient aid for third molar extractions? *J Oral Maxillofac Surg* 2024;82(10):1239–45. doi: [10.1016/j.joms.2024.04.003](https://doi.org/10.1016/j.joms.2024.04.003).
38. Jaworski K, Jasiński D, Sławińska B, et al. GPT-4o vs. human candidates: performance analysis in the polish final dentistry examination. *Cureus* 2024;16(9):e68813. doi: [10.7759/cureus.68813](https://doi.org/10.7759/cureus.68813).

39. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofac Radiol* 2024;53(6):390–5. doi: [10.1259/dmfr.20240141](https://doi.org/10.1259/dmfr.20240141).
40. Johnson AJ, Singh TK, Gupta A, et al. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. *Dent Traumatol* 2025;41(2):187–93. doi: [10.1111/edt.12937](https://doi.org/10.1111/edt.12937).
41. Kuenzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin Oral Investig* 2024;28(11):604. doi: [10.1007/s00784-024-05996-6](https://doi.org/10.1007/s00784-024-05996-6).
42. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. *Eur J Orthod* 2024;46(2) cjae017. doi: [10.1093/ejo/cjae017](https://doi.org/10.1093/ejo/cjae017).
43. Mohammad-Rahimi H, Khoury ZH, Alamdari MI, et al. Performance of AI chatbots on controversial topics in oral medicine, pathology, and radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2024;137(5):508–14. doi: [10.1016/j.oooo.2023.12.017](https://doi.org/10.1016/j.oooo.2023.12.017).
44. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int Endod J* 2024;57(3):305–14. doi: [10.1111/iej.13894](https://doi.org/10.1111/iej.13894).
45. Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and bard on the Japanese National Dentist Examination: a comparison study. *Cureus* 2023;15(12):e50369. doi: [10.7759/cureus.50369](https://doi.org/10.7759/cureus.50369).
46. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. *Dent Traumatol* 2024;40(6):722–9. doi: [10.1111/edt.12920](https://doi.org/10.1111/edt.12920).
47. Quah B, Yong CW, Lai CWM, Islam I. Performance of large language models in oral and maxillofacial surgery examinations. *Int J Oral Maxillofac Surg* 2024;53(10):881–6. doi: [10.1016/j.ijom.2024.07.005](https://doi.org/10.1016/j.ijom.2024.07.005).
48. Rokhshad R, Fadul M, Zhai G, Carr K, Jackson JG, Zhang P. A comparative analysis of responses of artificial intelligence Chatbots in special needs dentistry. *Pediatr Dent* 2024;46(5):337–44.
49. Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. *J Periodontal Res* 2025;60(2):121–33. doi: [10.1111/jre.13205](https://doi.org/10.1111/jre.13205).
50. Song ES, Lee SP. Comparative analysis of the response accuracies of large language models in the Korean National Dental Hygienist Examination Across Korean and English Questions. *Int J Dent Hyg* 2024;22(4):542–53. doi: [10.1111/idh.12834](https://doi.org/10.1111/idh.12834).
51. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. *Int Endod J* 2024;57(1):108–13. doi: [10.1111/iej.13895](https://doi.org/10.1111/iej.13895).
52. Suárez A, Jiménez J, Llorente de Pedro M, et al. Beyond the scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. *Comput Struct Biotechnol J* 2024;24:46–52. doi: [10.1016/j.csbj.2024.01.005](https://doi.org/10.1016/j.csbj.2024.01.005).
53. Turunç Oğuzman R, Yurdabakan ZZ. Performance of chat generative pretrained transformer and bard on the questions asked in the dental specialty entrance examination in Turkey Regarding Bloom's Revised Taxonomy. *Curr Res Dent Sci* 2024;34(1):25–34. doi: [10.1016/j.crds.2023.11.002](https://doi.org/10.1016/j.crds.2023.11.002).
54. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. *Otolaryngol Head Neck Surg* 2024;170(6):1492–503. doi: [10.1177/01945998231217451](https://doi.org/10.1177/01945998231217451).
55. Yamaguchi S, Morishita M, Fukuda H, et al. Evaluating the efficacy of leading large language models in the Japanese national dental hygienist examination: a comparative analysis of ChatGPT, Bard, and Bing Chat. *J Dent Sci* 2024;19(4):2262–7. doi: [10.1016/j.jds.2024.06.022](https://doi.org/10.1016/j.jds.2024.06.022).
56. Yunsun J, Chae YK, Sun KM, Hyo-Seol L, Sung CC, Hyung NO. Evaluating the accuracy of artificial intelligence-based chatbots on pediatric dentistry questions in the Korean National Dental Board Exam. *J Korean Acad Pediatr Dent* 2024;51(3):299–309. doi: [10.5933/JKAPD.2024.51.3.299](https://doi.org/10.5933/JKAPD.2024.51.3.299).
57. Park YJ, Pillai A, Deng J, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Med Inform Decis Mak* 2024;24:72. doi: [10.1186/s12911-024-02438-9](https://doi.org/10.1186/s12911-024-02438-9).
58. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci* 2025;32(1):45. doi: [10.1186/s12929-025-01134-x](https://doi.org/10.1186/s12929-025-01134-x).
59. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. *JMIR Med Educ* 2024;10:e51391. doi: [10.2196/51391](https://doi.org/10.2196/51391).
60. Demir GB, Süküt Y, Duran GS, Topsakal KG, Görgülü S. Enhancing systematic reviews in orthodontics: a comparative examination of GPT-3.5 and GPT-4 for generating PICO-based queries with tailored prompts and configurations. *Eur J Orthod* 2024;46(2) cjae011. doi: [10.1093/ejo/cjae011](https://doi.org/10.1093/ejo/cjae011).
61. Shan G, Chen X, Wang C, et al. Comparing diagnostic accuracy of clinical professionals and large language models: systematic review and meta-analysis. *JMIR Med Inform* 2025;13:e64963. doi: [10.2196/64963](https://doi.org/10.2196/64963).
62. Wang L, Li J, Zhuang B, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. *J Med Internet Res* 2025;27:e64486. doi: [10.2196/64486](https://doi.org/10.2196/64486).
63. Dinc MT, Bardak AE, Bahar F, Noronha C. Comparative analysis of large language models in clinical diagnosis: performance evaluation across common and complex medical cases. *JAMIA Open* 2025;8(3) ooaf055. doi: [10.1093/jamiaopen/ooaf055](https://doi.org/10.1093/jamiaopen/ooaf055).
64. Bagde H, Dhopte A, Alam MK, Basri R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. *Heliyon* 2023;9(12):e23050. doi: [10.1016/j.heliyon.2023.e23050](https://doi.org/10.1016/j.heliyon.2023.e23050).
65. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform* 2024;151:104620. doi: [10.1016/j.jbi.2024.104620](https://doi.org/10.1016/j.jbi.2024.104620).
66. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613–22. doi: [10.1038/s41591-024-02993-w](https://doi.org/10.1038/s41591-024-02993-w).
67. Dashti M, Londono J, Ghasemi S, Moghaddasi N. How much can we rely on artificial intelligence chatbots such as the ChatGPT software program to assist with scientific writing? *J Prosthet Dent* 2025;133(4):1082–8. doi: [10.1016/j.prosdent.2024.11.025](https://doi.org/10.1016/j.prosdent.2024.11.025).
68. Dashti M, Ghasemi S, Khurshid Z. Role of artificial intelligence in oral diagnosis and dental treatment. *Eur J Gen Dent* 2023;12(3):135–7. doi: [10.1055/s-0043-1764408](https://doi.org/10.1055/s-0043-1764408).
69. Liu M, Okuhara T, Huang W, et al. Large language models in dental licensing examinations: systematic review and meta-analysis. *Int Dent J* 2025;75(1):213–22. doi: [10.1016/j.identj.2024.10.014](https://doi.org/10.1016/j.identj.2024.10.014).
70. Liu M, Okuhara T, Chang X, et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. *J Med Internet Res* 2024;26:e60807. doi: [10.2196/60807](https://doi.org/10.2196/60807).

71. Jin HK, Lee HE, Kim EY. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. *BMC Med Educ* 2024;24(1):1013. doi: [10.1186/s12909-024-05944-8](https://doi.org/10.1186/s12909-024-05944-8).
72. Rokhshad R, Tichy A, Ducret M, et al. The ethics and governance of large language models in dentistry: a framework for research and clinical implementation. *J Dent* 2025;164:106187. doi: [10.1016/j.jdent.2025.106187](https://doi.org/10.1016/j.jdent.2025.106187).
73. Ashery AF, Aiello LM, Baronchelli A. Emergent social conventions and collective bias in LLM populations. *Sci Adv* 2025;11(20):eadu9368. doi: [10.1126/sciadv.adu9368](https://doi.org/10.1126/sciadv.adu9368).
74. Hu T, Kyrychenko Y, Rathje S, Collier N, van der Linden S, Roozenbeek J. Generative language models exhibit social identity biases. *Nat Comput Sci* 2025;5(1):65–75. doi: [10.1038/s44256-024-00003-y](https://doi.org/10.1038/s44256-024-00003-y).
75. Li R, Wu T. Delving into the practical applications and pitfalls of large language models in medical education: narrative review. *Adv Med Educ Pract* 2025;16:625–36. doi: [10.2147/AMEP.S465844](https://doi.org/10.2147/AMEP.S465844).
76. Dashti M, Khosraviani F, Azimi T, et al. Assessing ChatGPT-4's performance on the US prosthodontic exam: impact of fine-tuning and contextual prompting vs. base knowledge, a cross-sectional study. *BMC Med Educ* 2025;25(1):761. doi: [10.1186/s12909-025-06948-z](https://doi.org/10.1186/s12909-025-06948-z).
77. Gertych A, Faust O. AI explainability and bias propagation in medical decision support. *Comput Methods Programs Biomed* 2024;257:108465. doi: [10.1016/j.cmpb.2024.108465](https://doi.org/10.1016/j.cmpb.2024.108465).
78. Migliavaca CB, Stein C, Colpani V, et al. Meta-analysis of prevalence: I2 statistic and how to deal with heterogeneity. *Res Synth Methods* 2022;13(3):363–7. doi: [10.1002/jrsm.1564](https://doi.org/10.1002/jrsm.1564).