Journal of Dentomaxillofacial Radiology, Pathology and Surgery

Summer 2025, Volume 14, Number 3

Review Paper: Performance of Large Language Models on Dental Board and Academic Examinations: Updated Narrative Review

Soheil Vafaeian[1*], Pedram Hajibagheri[1]

1. Department of Oral and Maxillofacial Medicine, Dental Sciences Research Center, School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran.

Use your device to scan and read the article online

Citation Vafaeian S, Hajibagheri P. Performance of Large Language Models on Dental Board and Academic Examinations: Updated Narrative Review. Journal of Dentomaxillofacial Radiology, Pathology and Surgery. 2025; 14(3):1-7. http://dx.doi.org/10.32598/3dj.14.3.1

http://dx.doi.org/10.32598/3dj.14.3.1

Article info:
Received: 10 July 2025
Accepted: 09 Aug 2025
Available Online: 30 Aug 2025

Keywords:
Artificial Intelligence, Dental Education, Large Language Models (LLMs)

ABSTRACT

Large language models (LLMs) are transforming dental education and practice by supporting clinical decision-making, administrative automation, and academic assessments. This review synthesizes 12 studies (May 2024–June 2025) evaluating LLMs, including ChatGPT, Gemini, and Claude, on dental board and academic examinations using a modified population, intervention, comparison, outcome (PICO) framework to assess accuracy, reliability, comprehensiveness, and reasoning quality. A narrative review of the literature was conducted, identifying relevant articles from PubMed, Scopus, Google Scholar, and arXiv. LLMs achieved acceptable accuracy on multiple-choice questions, often surpassing human benchmarks, though performance varied by model, question type, and language. They excel in factual recall and exam preparation, particularly in resource-limited settings, but struggle with clinical reasoning and text-based formats. LLMs show potential for enhancing dental education, especially in standardized assessments, but require standardized evaluation frameworks, diverse question formats, and ethical guidelines to address limitations in practical and visual applications for effective integration into dental curricula.

1. Introduction

Large language models (LLMs) are advanced artificial intelligence algorithms adept at processing and generating human-like text. These models are trained on vast datasets, allowing them to perform a variety of natural language processing tasks, which include summarization, question-answering, and applications involving logical reasoning and contextual understanding (1, 2).

In recent years, the application of LLMs in dentistry has garnered attention for their potential to enhance various facets of dental practice, including diagnosis, treatment planning, patient management, and education (3, 4).

One key area where LLMs are applied in dentistry is in clinical decision support. Generative AI models, such as ChatGPT, can assist dental practitioners in developing preliminary assessment protocols and management plans, particularly when clinical information is sparse or ambiguous. However, concerns about the "hallucinations" phenomenon, where LLMs may provide inac-

*** Corresponding Author:**
Soheil Vafaeian, DDS.
Address: *Department of Oral and Maxillofacial Medicine, Dental Sciences Research Center, School of Dentistry, Guilan University of Medical Sciences, Rasht, Iran.*
E-mail: *soheylvafa@gmail.com*

Journal of Dentomaxillofacial Radiology, Pathology and Surgery

curate or misleading information, necessitate cautious integration into clinical workflows (5, 6). Researchers have noted that these models can significantly improve diagnosis rates and enhance patient education by providing tailored information (7, 8).

Additionally, LLMs can automate administrative tasks like appointment scheduling and follow-up communications, enhancing practice efficiency and allowing dental professionals to focus more on patient care (6, 9). LLMs also contribute to educational strategies within dentistry. They can generate quizzes, summaries, and practice questions aligned with dental curricula, supporting medical students and residents in their learning (7, 10). Additionally, the potential for multilingual communication enabled by LLMs opens avenues for global outreach in dental health training programs (7).

The integration of LLMs in dentistry is not without challenges. Issues related to data privacy, quality of the generated content, and the need for continuous oversight to mitigate bias and ensure reliable information dissemination are pressing concerns (6, 11). Establishing ethical frameworks is essential to guide the deployment of these technologies in clinical settings, maximizing benefits while minimizing risks (11, 12).

The performance of LLMs on dental board and academic examinations has become a key research focus, underscoring their potential in medical education and licensure assessments. Studies have systematically evaluated the accuracy and capabilities of popular LLMs—such as ChatGPT (including ChatGPT-3.5 and ChatGPT-4o) and Google Bard—in the context of medical exams, including dental licensure tests (13). These advancements highlight significant opportunities for innovation in medical education.

However, while these findings are promising, researchers emphasize the need for further exploration into the integration of LLMs into formal educational settings. The current literature calls for standardized evaluation frameworks to ensure LLM responses are reliable, reproducible, and clinically relevant (14, 15). Given the rapid advancements in artificial intelligence and machine learning, ongoing assessments are crucial to gauge the effectiveness of these tools in real-world academic and clinical scenarios. This review synthesizes recent evidence on the performance of LLMs on dental board and academic examinations, while addressing gaps in validation and their potential role in shaping future dental education.

2. Materials and Methods

This study, designed as a narrative review, evaluated the performance of LLMs on dental board and academic examinations. A mixed-methods approach combined quantitative metrics—accuracy, reliability, and comprehensiveness—with qualitative assessments of reasoning and response quality to examine LLMs in the context of dental education and certification. The methodology was designed to elucidate how LLMs managed specialized knowledge and clinical reasoning in dentistry, updating findings from a prior systematic review whose database search was completed on May 1, 2024, by incorporating new evidence published since that date (16).

Data sources consisted of compiled studies, including peer-reviewed articles and preprints, on LLM performance in medical or dental contexts. These were sourced from PubMed, Scopus, Google Scholar, and arXiv. Preprints were included to capture recent advancements in AI applications for dentistry, with their non-peer-reviewed status noted for transparency. Studies were selected based on specific inclusion and exclusion criteria to ensure relevance.

Two independent reviewers evaluated the titles, abstracts, and study designs of all identified articles. To minimize bias, reviewers conducted their assessments independently, unaware of each other's decisions, ensuring objective evaluations. When disagreements occurred regarding the inclusion or exclusion of an article, reviewers discussed the points of contention and reached a consensus based on the study's inclusion criteria. This process ensured the accuracy and integrity of the study selection.

The search strategy comprised a literature review using targeted keywords and Boolean operators: ("large language model" OR "LLM" OR "artificial intelligence" OR "AI" OR "ChatGPT" OR "GPT-4" OR "GPT-4o" OR "Gemini" OR "Claude") AND ("dental board" OR "dental examination" OR " dental license" OR "dental education" OR "academic assessment") AND ("performance" OR "accuracy" OR "evaluation"). The search was limited to English-language publications from May 2024 to June 2025, with the English-only restriction and selected databases chosen for practicality but potentially limiting the scope of findings. Manual searches of reference lists from key articles supplemented the electronic search to enhance coverage (Figure 1).

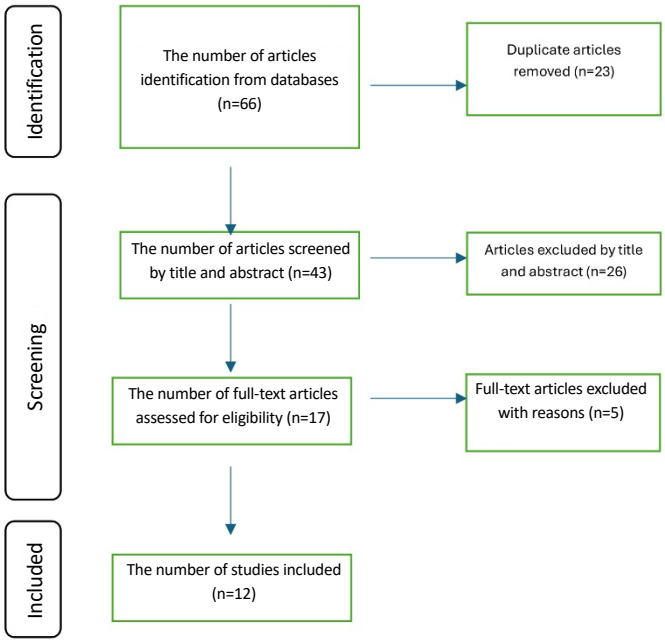

Figure 1. Literature selection flowchart

Studies were selected based on their relevance to evaluating LLMs (e.g. ChatGPT-4o, Gemini, Claude) on dental board examinations (e.g. NBDE, INBDE) or academic dental assessments. Inclusion criteria required that studies: Evaluate LLMs (e.g. ChatGPT-4o, Gemini, Claude) on dental board examinations (e.g. NBDE, INBDE) or academic dental assessments; report quantitative metrics (accuracy, reliability, or comprehensiveness) or qualitative insights (e.g. reasoning quality or response limitations); be published between May 2024 and June 2025; and provide sufficient methodological detail to assess study quality. Exclusion criteria eliminated studies that: focused exclusively on non-dental medical examinations; were not in English; lacked clear performance metrics or qualitative findings; were unpublished or inaccessible; or neither evaluated LLMs on dental board examinations nor on comprehensive academic dental assessments.

Data collection extracted performance metrics (accuracy, reliability, comprehensiveness) and qualitative insights (e.g. reasoning quality, limitations in handling complex questions) from selected studies. Accuracy was measured as the percentage of correct answers, reliability as response consistency across trials, and comprehensiveness as the completeness and relevance of responses. Qualitative data focused on LLMs' ability to address complex or ambiguous questions and their limitations in clinical reasoning.

3. Results

Sixty-six articles were initially identified from various databases. After removing 23 duplicate articles, 43 unique articles remained and were screened by title and abstract. Of these, 26 were excluded, leaving 17 full-text articles to be assessed for eligibility. Ultimately, 5 full-text articles were excluded because they either didn't focus on formal or academic dental examinations or lacked clear performance metrics or qualitative findings, resulting in a final total of 12 studies included in the analysis (Table 1).

Jaworski et al. in 2024 found that ChatGPT-4o accurately answered multiple-choice questions, including clinical case-based and factual questions, in a study involving 199 participants (17). Similarly, Kinikoglu in 2025 reported that ChatGPT-4o, ChatGPT-o1, Gemini 1.5 Pro, and Gemini 2.0 Advanced performed reliably on multiple-choice questions covering basic and clinical sciences with 238 participants (18).

Hu et al. in 2024 observed that ChatGPT 3.5, ChatGPT-4o, and New Bing effectively handled multiple-choice questions across various dental subjects in a study of 324 examinees (19). Expanding on this, Uehara et al. in 2025 noted that ChatGPT-3.5 and ChatGPT-4o achieved consistent performance on text-based multiple-choice questions in dental subjects, testing 1,399 participants (20). Similarly, Fujimoto et al. in 2024 evaluated ChatGPT-4o, Claude 3 Opus, and Gemini 1.0, finding strong performance on multiple-choice questions in physiology, anesthesia, and other subgroups with 295 participants (21).

Table 1. Main characteristics of selected studies

Author(s), Year	LLM Model (s)	Sample Size	Question's Type	Answer's Type
Jaworski et al. (2024) (17)	ChatGPT-4o	199	Multiple-choice (clinical case-based and factual)	Single correct answer
Kinikoglu (2025) (18)	ChatGPT-4o, ChatGPT-o1, Gemini 1.5 Pro, Gemini 2.0 Advanced	238	Multiple-choice (basic and clinical sciences)	Single correct answer
Hu et al. (2024) (19)	ChatGPT 3.5, ChatGPT-4o New Bing	324	Multiple-choice (various dental subjects)	Single correct answer
Uehara et al. (2025) (20)	ChatGPT-3.5, ChatGPT-4o	1399	Multiple-choice (text-based, various dental subjects)	Single correct answer
Fujimoto et al. (2024) (21)	ChatGPT-4o, Claude 3 Opus, Gemini 1.0	295	Multiple-choice (subgroups: Physiology; anesthesia, etc.)	Single correct answer
Sismanoglu & Capan (2025) (22)	ChatGPT-4o, Gemini Advanced	240	Multiple-choice (basic and clinical sciences)	Single correct answer
Xiong et al. (2025) (23)	ChatGPT-4o, Doubao-pro 32k, Qwen2-72b, ChatGLM-4	200	Likert-scale questions	Single correct answer
Kim et al. (2025) (24)	ChatGPT-3.5, GPT-4, Claude3-Opus	1777	Multiple-choice (various dental subjects)	Single correct answer
Sabri et al. (2025) (25)	ChatGPT-3.5, GPT-4, Google Gemini	1312	Multiple-choice (periodontology)	Single correct answer
Chan-Chia Lin et al. (2025) (26)	ChatGPT-3.5, Claude2, Gemini	2699	Multiple-choice (basic and clinical dentistry)	Single correct answer
Temiz & Güzel (2025) (27)	ChatGPT-4o	720	Multiple-choice (basic and clinical sciences)	Single correct answer
Wójcik et al. (2024) (28)	ChatGPT-4o, Gemini, Claude	198	Multiple-choice (various dental subjects)	Single correct answer

Journal of Dentomaxillofacial Radiology, Pathology and Surgery

Further reinforcing these findings, Sismanoglu and Capan in 2025 reported that ChatGPT-4o and Gemini Advanced successfully answered multiple-choice questions in basic and clinical sciences for 240 participants (22). Beyond traditional formats, Xiong et al. in 2025 found that ChatGPT-4o, Doubao-pro 32k, Qwen2-72b, and ChatGLM-4 performed well on Likert-scale questions with single correct answers in a study of 200 participants (23). Additionally, Kim et al. in 2025 observed that ChatGPT-3.5, ChatGPT-4o, and Claude 3 Opus demonstrated high accuracy on multiple-choice questions across various dental subjects, involving 1,777 test cases (24).

Specialized applications were also explored, such as Sabri et al. in 2025, who focused on periodontology, finding that ChatGPT-3.5, ChatGPT-4o, and Google Gemini provided reliable responses to multiple-choice questions for 1,312 participants (25). Broadening the scope, Chan-Chia Lin et al. in 2025 reported that ChatGPT-3.5, Claude 2, and Gemini excelled in multiple-choice questions covering basic and clinical dentistry with 2,699 examinees (26).

Supporting these results, Temiz and Güzel in 2025 noted that ChatGPT-4o achieved high accuracy on multiple-choice questions in basic and clinical sciences for 720 participants (27). Finally, Wójcik et al in 2024 found that ChatGPT-4o, Gemini, and Claude performed consistently on multiple-choice questions across various dental subjects with 198 participants (28).

4. Discussion

This narrative review synthesizes findings from 12 studies evaluating LLMs on dental board and academic examinations, organizing insights into three key themes: performance on standardized dental examinations, effectiveness in specialized dental fields, and comparative model performance. By comparing similarities and divergences across studies, this discussion highlights LLMs' potential as educational tools in dentistry while noting common limitations, such as reliance on text-based multiple-choice questions and limited testing of clinical reasoning, to provide a balanced perspective.

Several studies assessed LLMs on standardized dental licensing and academic examinations, demonstrating their potential as study aids. Kinikoglu in 2025 evaluated ChatGPT-4o, ChatGPT-o1, Gemini 1.5 Pro, and Gemini 2.0 Advanced on 238 multiple-choice questions from the Turkish dental specialization exam, finding ChatGPT-o1 achieved 97.46% accuracy, surpassing ChatGPT-4o's 88.66% (18). Uehara et al. in 2024 tested ChatGPT-3.5 and ChatGPT-4o on 1,399 multiple-choice questions from the Japanese National Dental Examination, with ChatGPT-4o reaching 84.63% accuracy compared to ChatGPT-3.5's 45.46% (20). Sismanoglu and Capan in 2025 (22) and Temiz and Güzel in 2025 (27) examined ChatGPT-4o and Gemini Advanced on Turkish DUS exams, reporting ChatGPT-4o's accuracy at 80.50%-83.30%, often outperforming human benchmarks) . Kim et al. in 2025 found Claude 3 Opus achieved 85.40% of human performance on 1,777 multiple-choice questions from the Korean dental licensing examination (24). Jaworski et al. in 2024 tested ChatGPT-4o on 199 multiple-choice questions from the Polish final dentistry examination, finding 70.85% overall accuracy but only 36.36% on clinical case-based questions compared to 72.87% on factual ones. These studies show that newer LLMs, like ChatGPT-4o and Claude 3 Opus, consistently excel in standardized multiple-choice exams, particularly in factual questions, suggesting their utility for exam preparation. However, a common limitation is the small question sample in some studies (17, 18), which may limit generalizability to broader examination contexts.

Studies focusing on specialized dental domains revealed LLMs' strengths in fact-based questions but challenges in clinical reasoning. Sabri et al. in 2024 evaluated ChatGPT-3.5, GPT-4, and Google Gemini on 1,312 periodontology multiple-choice questions, with GPT-4 achieving 78.80%-80.98% accuracy, surpassing human performance (25). Fujimoto et al. in 2024 assessed ChatGPT-4o, Claude 3 Opus, and Gemini 1.0 on 295 multiple-choice questions from the Japanese Dental Society of Anesthesiology board certification exam, noting ChatGPT-4o's moderate 51.20% accuracy (21). These mixed results suggest that while LLMs can effectively handle certain knowledge-based tasks in dentistry, they still struggle with the nuanced problem-solving required for complex clinical scenarios. Further research is needed to understand the specific limitations of these models and develop strategies to improve their performance in areas requiring critical thinking and clinical judgment.

Studies comparing multiple LLMs revealed variations in model effectiveness. Hu et al. in 2024 tested ChatGPT, GPT-4, and New Bing on 324 multiple-choice questions from the Chinese national dental licensing examination, with New Bing achieving 72.50% accuracy, surpassing GPT-4's 63.00% and ChatGPT's 42.60% (19). Xiong et al. in 2025 evaluated GPT-4, Doubao-pro 32k, Qwen2-72b, and ChatGLM-4 on 200 questions from the Chinese dental licensing examination, with Doubao-pro 32k leading at 81.00% accuracy. Chan-Chia Lin et al. in 2025 found Claude 2 outperformed ChatGPT-3.5 and Gemini on 2,699 multiple-choice questions from Taiwan's dental licensing exams, achieving 54.89% accuracy (26). Wójcik et al. in 2025 noted Claude outperformed ChatGPT-4o and Gemini in most areas except prosthodontics on 198 multiple-choice questions from the Polish LDEK (28). These studies suggest that while ChatGPT variants are widely used, alternative models like Claude, New Bing, and Doubao-pro 32k can outperform in specific contexts, possibly due to specialized training. A common limitation is the inconsistent performance on ambiguous or adversarial questions, indicating a need for further model refinement.

Across the 12 studies reviewed, common limitations in evaluating LLMs on dental board and academic examinations include a heavy reliance on multiple-choice questions, which primarily assess factual recall rather than clinical reasoning or practical skills. Most studies focused on text-based formats, with limited exploration of visual or case-based scenarios critical to dental practice, such as image interpretation or hands-on procedural assessments. Additionally, small sample sizes in some studies restrict generalizability. Specific gaps include insufficient evaluation of LLMs in dental specialties like prosthodontics, orthodontics, or oral surgery, where complex decision-making is essential. There is also a lack of standardized question formats beyond multiple-choice, such as open-ended or interactive case studies, and limited testing in multilingual or culturally diverse contexts. Further research is needed to develop diverse assessment formats, evaluate LLMs in underrepresented specialties, and create standardized evaluation frameworks to ensure clinical relevance and applicability.

5. Conclusions

This review demonstrates that advanced LLMs, such as ChatGPT-4o, Claude, and Doubao-pro 32k, show significant potential as educational tools in dental training, excelling in standardized assessments like multiple-choice and Likert-scale questions that evaluate factual knowledge and subjective opinions. They offer valuable support for exam preparation, particularly in resource-constrained settings, and show promise in specialized fields like periodontology. However, their limitations

in clinical reasoning and reliance on text-based formats highlight gaps in addressing the practical and visual aspects of dentistry. Variability in study designs and inconsistent reporting further challenge their broader application. To guide future work, we recommend: developing standardized question sets to ensure consistent evaluation across studies, evaluating LLMs in real-world dental examinations to assess their practical applicability, and integrating LLMs thoughtfully into curricula to balance technological benefits with the development of clinical competency.

Ethical Considerations

This review used publicly available data with proper citation and required no ethics approval, as no human subjects were involved.

Funding

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

Author's Contributions

Soheil Vafeaian: Conceptualization, Investigation, Writing - Original Draft, Writing - Review & Editing Pedram Hajibagheri: Investigation, Writing - Review & Editing.

Conflict of interest

The authors declared no conflicts of interest.

Availability of Data and Material

Not applicable.

Acknowledgements

The authors thank their institutional colleagues for providing valuable feedback during the preparation of this review.

References

1. Li J, Li W, Chen X, Deng X, Wen H, You M, et al. Are you asking GPT-4 medical questions properly?-prompt engineering in consistency and reliability with evidence-based guidelines for ChatGPT-4: A pilot study. 2023 [Unpiblished]. [DOI:10.21203/rs.3.rs-3336823/v1]

2. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023; 6(1):120. [DOI:10.1038/s41746-023-00873-0] [PMID]

3. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023; 35(7):1098-102. [DOI:10.1111/jerd.13046] [PMID]

4. Freitas MD, Lago-Méndez L, Posse JL, Dios PD. Challenging ChatGPT-4V for the diagnosis of oral diseases and conditions. Oral Dis. 2025; 31(2):701-6. [DOI:10.1111/odi.15169] [PMID]

5. Albagieh H, Alzeer ZO, Alasmari ON, Alkadhi AA, Naitah AN, Almasaad KF, et al. Comparing artificial intelligence and senior residents in oral lesion diagnosis: A comparative study. Cureus. 2024; 16(1):e51584. [DOI:10.7759/cureus.51584]

6. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. Int J Oral Sci. 2023; 15(1):29. [DOI:10.1038/s41368-023-00239-y] [PMID]

7. Nguyen HC, Dang H, Nguyen TL, Hoàng V, Anh NV. 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] [PMID]

8. Alhazmi N, Alshehri A, BaHammam F, Philip MR, Nadeem M, Khanagar SB. Can large language models serve as reliable tools for information in dentistry? A systematic review. Int Dent J. 2025; 75(4):100835. [DOI:10.1016/j.identj.2025.04.015] [PMID]

9. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK. ChatGPT in dentistry: A comprehensive review. Cureus. 2023; 15(4):e38317. [DOI:10.7759/cureus.38317] [PMID]

10. Giannakopoulos K, Kavadella A, Salim AA, 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 Int Res. 2023; 25:e51580. [DOI:10.2196/51580] [PMID]

11. Siluvai S, Narayanan V, Ramachandran VS, Lazar VR. Generative pre-trained transformer: Trends, applications, strengths and challenges in dentistry: A systematic review. Health Inform Res. 2025; 31(2):189-99. [DOI:10.4258/hir.2025.31.2.189] [PMID]

12. Shirani M. Trends and classification of artificial intelligence models utilized in dentistry: A bibliometric study. Cureus. 2025; 17(4):e81836. [DOI:10.7759/cureus.81836] [PMID]

13. 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]

14. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: A scoping review. BMC Med Inform Decis Mak. 2024; 24(1):366. [DOI:10.21203/rs.3.rs-3879872/v1]

15. Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. J Med Int Res. 2024; 26:e53297. [DOI:10.2196/53297] [PMID]

16. Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, et al. Large language models in dental licensing examinations: Systematic review and meta-analysis. Int Dent J. 2025; 75(1):213-222. [DOI:10.1016/j.identj.2024.10.014] [PMID]

17. Jaworski A, Jasiński D, Sławińska B, Błecha Z, Jaworski W, Kruplewicz M, 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]

18. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. Cureus. 2025; 17(1):e77292. [DOI:10.7759/cureus.77292]

19. Hu Z, Xu Z, Shi P, Zhang D, Yue Q, Zhang J, et al. Performance of large language models in the national dental licensing examination in China: A comparative analysis of ChatGPT, GPT-4, and New Bing. Int J Comput Dent. 2024; 27(4):401-11. [Link]

20. Uehara O, Morikawa T, Harada F, Sugiyama N, Matsuki Y, Hiraki D, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese national dental examination. J Dent Educ. 2025; 89(4):459-66. [DOI:10.1002/jdd.13766] [PMID]

21. Fujimoto M, Kuroda H, Katayama T, Yamaguchi A, Katagiri N, Kagawa K, et al. Evaluating large language models in dental anesthesiology: A comparative analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese dental society of anesthesiology board certification exam. Cureus. 2024; 16(9):e70302. [DOI:10.7759/cureus.70302]

22. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: Can ChatGPT-4.0 and Gemini advanced achieve comparable results to humans? BMC Med Educ. 2025; 25(1):214. [DOI:10.1186/s12909-024-06389-9] [PMID]

23. Xiong YT, Zhan ZZ, Zhong CL, Zeng W, Guo JX, Tang W, et al. Evaluating the performance of large language models (LLMs) in answering and analysing the Chinese dental licensing examination. Eur J Dent Educ. 2025; 29(2):332-40. [DOI:10.1111/eje.13073] [PMID]

24. Kim W, Kim BC, Yeom H-G. Performance of large language models on the Korean dental licensing examination: A comparative study. Int Dent J. 2025; 75(1):176-84. [DOI:10.1016/j.identj.2024.09.002] [PMID]

25. Sabri H, Saleh MH, Hazrati P, Merchant K, Misch J, Kumar PS, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. J Periodontal Res. 2025; 60(2):121-33. [DOI:10.1111/jre.13323] [PMID]

26. Chan-Chia Lin C, Sun JS, Chang CH, Chang YH, Zwei-Chieng Chang J. Performance of artificial intelligence chatbots in national dental licensing examination. J Dent Sci. 2025; 20(4):2307-14. [DOI:10.1016/j.jds.2025.05.012] [PMID]

27. Temiz M, Güzel C. Assessing the performance of ChatGPT on dentistry specialization exam questions: A comparative study with DUS examinees. Med Rec. 2025; 7(1):162-6. [DOI:10.37990/medr.1567242]

28. Wójcik D, Adamiak O, Czerepak G, Tokarczuk O, Szalewski L. A comparative analysis of the performance of chatGPT4, Gemini and Claude for the Polish medical final diploma exam and medical-dental verification exam. MedRxiv. 2024:2024. [Unpublished]. [DOI:10.1101/2024.07.29.24311077]