Journal of Dental Sciences 21 (2026) 96–102

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Evaluation of GPT-4o and Gemini Advanced on the Korean National Dental Licensing Examination: Accuracy, consistency, and question generation

Eun Sun Song [a], Ga Hyeon Kim [b], Seung-Pyo Lee [a*]

[a] *Department of Oral Anatomy, Dental Research Institute, School of Dentistry, Seoul National University, Seoul, Republic of Korea*
[b] *Department of Dental Hygiene, Jeonju Vision University, Jeonju, Republic of Korea*

Received 22 May 2025; Final revision received 21 July 2025
Available online 6 August 2025

KEYWORDS
Gemini Advanced;
GPT-4o;
Korean National Dental Examination;
Large language models

Abstract *Background/purpose:* Large language models (LLMs), such as GPT-4o and Gemini Advanced, have performed strongly on global medical examinations. However, their capabilities in non-English, dentistry-specific licensing contexts remain unclear. Thus, this study aimed to compare the performance, consistency, and question-generation abilities of GPT-4o and Gemini Advanced in the Korean National Dental Licensing Examination (KNDLE).
Materials and methods: This study used 1,401 text-based KNDLE questions from 2019 to 2023 in Korean. Each model responded to the questions in three separate runs. Accuracy and consistency were compared with human answers. The models generated new questions in four subject areas and attempted to solve each other's generated items. Paired t-tests and chi-square tests were conducted.
Results: GPT-4o achieved significantly higher average accuracy than Gemini Advanced (81.1 % vs. 76.6 %, $P = 0.013$) and showed greater consistency across attempts. Both models performed better in basic sciences than in clinical subjects, such as prosthodontics. In cross-solving tasks, GPT-4o's performance notably declined in Gemini-generated oral biology questions, indicating interpretation differences. However, the consistency difference between models was not significant ($P = 0.578$).
Conclusion: GPT-4o outperformed Gemini Advanced in accuracy, consistency, and alignment with its generated content. However, challenges remain in clinical domains and cross-model understanding, highlighting the potential of LLMs as supportive tools for non-English dental education and question generation while emphasizing the persistent need for expert oversight and domain-specific refinement.

* Corresponding author. Department of Oral Anatomy, Dental Research Institute, School of Dentistry, Seoul National University, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea.
E-mail address: orana9@snu.ac.kr (S.-P. Lee).

https://doi.org/10.1016/j.jds.2025.07.020

Introduction

Recent advancements in artificial intelligence (AI) have led to significant progress in different fields. In medicine and dentistry, AI and deep learning have facilitated applications ranging from clinical decision support to educational assessment and medical imaging.[1—5] Among these innovations, large language models (LLMs), such as OpenAI's GPT-4o and Google's Gemini Advanced, have demonstrated notable capabilities in understanding complex medical information and assisting in knowledge-based tasks.[6—9] They support applications in clinical decision-making, medical image interpretation, and patient education.[10—13]

LLMs are increasingly used in medical and dental education, clinical reasoning, patient communication, and even solving board examination questions. A review of 11 studies across eight countries revealed that GPT-4 achieved an average accuracy of 72 %, surpassing GPT-3.5 and Bard, and passed over half of the dental licensing examinations evaluated—even in non-English-speaking contexts.[14] In the USA, GPT-4 exceeded the passing threshold in the United States Medical Licensing Examination and outperformed earlier models, indicating its potential in clinical assessment.[15] Japan and China studies have reported similar results; however, limitations remained in specialized clinical areas, such as prosthodontics and surgery.[16,17] LLMs can also generate dental board-style questions at near-human quality, further expanding their utility in education.[18,19]

Previous studies have evaluated GPT-3.5, GPT-4, and Gemini on the Korean National Dental Hygienist Examination. Although GPT-4 showed the highest performance, it exhibited lower accuracy in Korean-language domains, such as health and medical law.[20] Thus, newer-generation LLMs must be evaluated in more linguistically complex and clinically demanding environments. This is particularly relevant given that most LLMs are primarily trained on English-language corpora, raising concerns about their performance in low-resource languages like Korean.[21,22]

This study focused on two of the most advanced LLMs: GPT-4o and Gemini Advanced. GPT-4o is the latest version of OpenAI's GPT series, offering improvements in reasoning and multimodal processing over GPT-4. Although previous research has evaluated GPT-4, the present study extends this work by analyzing GPT-4o's performance in a more demanding context. Similarly, Gemini Advanced is the successor to Google Bard, designed with enhanced capabilities for contextual understanding and response generation. Throughout this manuscript, the term "LLMs" is used broadly to refer to both models unless otherwise specified.

To address the aforementioned gaps, this study examined the performance of GPT-4o and Gemini Advanced in the Korean National Dental Licensing Examination (KNDLE), a more comprehensive and clinically rigorous test than the hygienist examination. Both models were prompted to generate new multiple-choice questions and solve each other's items. This dual evaluation aimed to assess subject-level accuracy and consistency, internal coherence, and mutual interpretability of the models. Accordingly, this study sought to evaluate the feasibility and limitations of using LLMs in AI-assisted content development, formative testing, and self-directed learning in dental education. This study aimed to evaluate and compare the accuracy, consistency, and question-generation capabilities of GPT-4o and Gemini Advanced on the Korean National Dental Licensing Examination to assess their utility in non-English dental education.

Materials and methods

LLMs

This study utilized two advanced LLMs: GPT-4o (OpenAI) and Gemini Advanced (Google), both accessed through subscription versions in February—March 2025. To evaluate their accuracy and response consistency, each KNDLE question was presented three times in Korean to both models. Model performance was compared against pass rates of human examinees for contextual interpretation.

This repeated-trial approach allowed for a robust evaluation of model performance in a non-English, high-stakes setting and examined LLM's feasibility in dental education. Fig. 1 shows the overall study design.

Dataset

Publicly available KNDLE items and answer keys from 2019 to 2023, provided by the Korea Health Personnel Licensing Examination Institute, were used. Questions containing images or diagrams were excluded.

The KNDLE includes five-option multiple-choice items covering 13 clinical subjects, such as oral medicine, prosthodontics, and radiology. Examinations from 2019 to 2021 included 364 questions per year, whereas those from 2022 to 2023 had 321 questions per year. Overall, 1,734 questions were collected, with the analysis including 1,401 after excluding non-text items.

To minimize memory or context effects, each question was submitted in a newly reset chat session using a consistent prompt: "You are a student taking the Korean National Dental Licensing Examination. This is a multiple-choice exam, and you must select only the most appropriate answer. Which of the following options best represents the correct answer?"

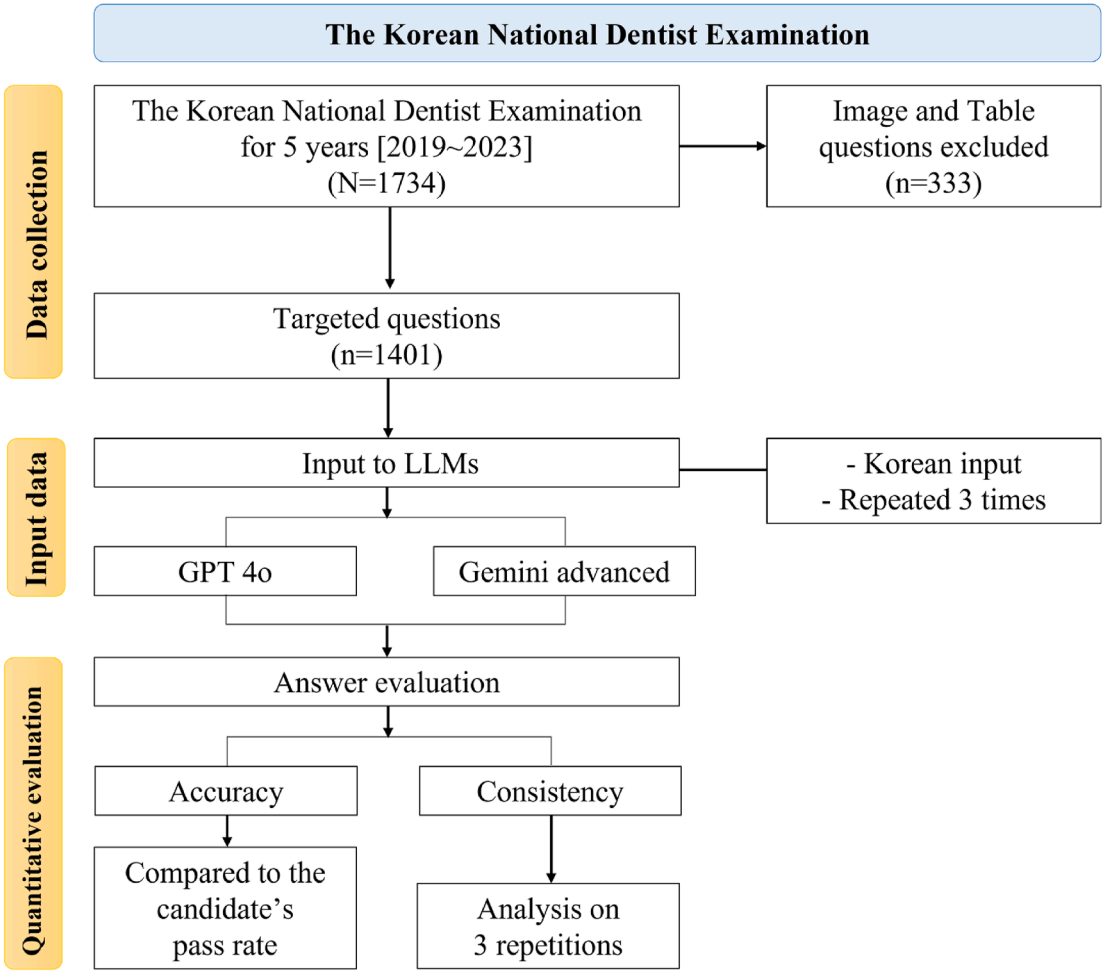

Figure 1 Study design for evaluating GPT-4o and Gemini advanced large language models' performance on correctly answering the exam questions from the Korean National Dental Licensing Examination (2019–2023).

To ensure content clarity and appropriateness for LLM analysis, the authors manually reviewed all text-based questions before inclusion.

Problem generation and cross-solving

In a secondary experiment, both models generated 10 original questions for each subject (orthodontics, prosthodontics, oral pathology, and oral biology) based on previous KNDLE. Each item included one correct answer and a brief explanation. To evaluate mutual interpretability and generation quality, the models solved each other's generated questions.

Each model generated 10 original multiple-choice questions each for orthodontics, prosthodontics, oral pathology, and oral biology, and the other model attempted to solve them. The cross-solving task was conducted as a single trial.

Statistical analysis

Data from GPT-4o, Gemini Advanced, and human examinees (2019–2023) were analyzed using Microsoft Excel and SPSS (SPSS Inc., Chicago, IL, USA). Descriptive statistics (mean, SD) summarized annual and subject-level accuracy. Performances between models and with human examinees were compared by paired t-tests. Repeated-measures analysis of variance was employed to evaluate within-model variability across the three trials. A Pearson chi-square test of independence assessed intramodel consistency based on categorical accuracy patterns (3/3, 1–2/3, and 0/3 correct), and assumption validity was confirmed by expected cell counts. A separate chi-square test compared the distribution of correct/incorrect responses between models in the cross-solving task. A P-value <0.05 was considered significant. Weighted means and SDs were calculated based on the annual number of test items (2019–2021, n = 364; 2022–2023, n = 321) (see Table 1).

Results

Table 2 presents the yearly accuracy of GPT-4o, Gemini Advanced, and human examinees from 2019 to 2023. Arithmetic means, weighted means, and weighted SDs were calculated to account for yearly variations in the number of

Table 1 Distribution of questions by subject in the Korean National Dental Licensing Examination.

Subject	Number of questions (n)
Oral medicine	13
Prosthodontics	35
Pediatric dentistry	23
Orthodontics	29
Oral pathology	12
Oral biology	43
Oral and maxillofacial radiology	23
Periodontology	23
Oral and maxillofacial surgery	35
Conservative dentistry	35
Oral health science	17
Dental materials	13
Health and medical law	20
Total	321

Table 2 Comparison of the yearly accuracy rates of GPT-4o, Gemini Advanced, and student examinees (2019–2023).

Year	GPT-4o (%)	Gemini Advanced (%)	Students	Passing score
2019	80.8	76.27	97.3	60
2020	81.47	73.47	97.3	60
2021	84.87	83.6	92.3	60
2022	82.13	77.8	94.8	60
2023	76.23	71.83	93.3	60
Average	81.1	76.59	95	60
Weighted mean ± SD	81.20 ± 2.76	76.68 ± 4.10	95.05 ± 2.07	–

test items. The weighted values were based on the number of exam items per year (364 and 321 items for 2019–2021 and 2022–2023, respectively). GPT-4o achieved a weighted mean accuracy of 81.20 % (SD, 2.76), significantly outperforming Gemini Advanced (76.68 %; SD, 4.10). Human examinees demonstrated the highest performance (weighted mean, 95.05 %; SD, 2.07).

A paired t-test showed a significant difference between GPT-4o and Gemini (mean difference, 4.51 %; SD, 2.38; $P = 0.013$), favoring GPT-4o. Compared with students, GPT-4o and Gemini scored significantly lower (GPT-4o, −13.90 %, $P = 0.0015$; Gemini, −18.41 %, $P = 0.002$). Correlation analysis revealed strong agreement between GPT-4o and Gemini ($r = 0.87$) but not with human performance.

As shown in Table 3, GPT-4o outperformed Gemini in 11 of 13 subjects. The largest margin was in periodontology (79.97 % vs. 64.68 %, Δ = 15.29 %), followed by oral medicine (+7.26 %), conservative dentistry (+5.78 %), pediatric dentistry (+5.58 %), and oral pathology (+5.22 %). Gemini outperformed GPT-4o in oral and maxillofacial radiology (90.97 % vs. 87.80 %) and health and medical law (72.00 % vs. 68.67 %).

Table 3 Comparison of the subject-wise accuracy rates between GPT-4o and Gemini Advanced.

Subject	GPT-4o (%)	Gemini Advanced (%)	Δ accuracy
Oral medicine	85.67	78.41	7.26
Prosthodontics	66.75	62.8	3.95
Pediatric dentistry	82.51	76.93	5.58
Orthodontics	71.34	68.51	2.83
Oral pathology	93.05	87.83	5.22
Oral biology	91.17	87.11	4.06
Oral and maxillofacial radiology	87.8	90.97	−3.17
Periodontology	79.97	64.68	15.29
Oral and maxillofacial surgery	87.47	82.59	4.88
Conservative dentistry	81.35	75.57	5.78
Oral health science	77.35	72.9	4.45
Dental materials	89.41	85.95	3.46
Health and medical law	68.67	72	−3.33

A paired t-test confirmed a significant difference in subject-wise performance, favoring GPT-4o ($P = 0.005$).

As shown in Table 4, GPT-4o consistently scored 81.1 %, 81.7 %, and 80.9 % (average, 81.23 %; SD, 0.42 %), whereas Gemini had greater variability, scoring 72.0 %, 76.9 %, and 77.2 % (average, 75.37 %; SD, 2.92 %). Although the average performance gap (5.87 %) did not reach significance ($P = 0.071$), a repeated-measures analysis of variance (ANOVA) indicated that GPT-4o demonstrated lower within-model variability across trials, supporting its greater consistency in repeated responses.

Table 5 presents intramodel consistency. GPT-4o had a higher identical response rate (81.8 %) than Gemini (78.0 %) and more consistently correct answers (3/3 correct, 73.2 % vs. 66.5 %). Gemini showed a slightly higher proportion of consistently incorrect answers (15.8 % vs. 11.5 %). Categorical response patterns (i.e., 3/3 correct, 1–2/3 correct,

Table 4 Comparison of accuracy rates across three runs for GPT-4o and Gemini Advanced.

Evaluation (run)	GPT-4o (%)	Gemini Advanced (%)
1st	81.1	72
2nd	81.7	76.9
3rd	80.9	77.2
Average	81.23	75.37

Table 5 Analysis of response consistency across three runs for GPT-4o and Gemini Advanced.

Metric	GPT-4o (%)	Gemini Advanced (%)
Identical response rate	81.8	78
Consistently correct (3/3)	73.2	66.5
Partially correct (1–2/3)	15.3	17.6
Consistently incorrect (0/3)	11.5	15.8

Table 6 Accuracy rates of GPT-4o and Gemini Advanced on self-generated questions by subject.

Subject	GPT-4o (%)	Gemini advanced (%)
Orthodontics	100	100
Prosthodontics	100	80
Oral pathology	90	100
Oral biology	80	100
Average	**92.5**	**95**

and 0/3 correct) were compared between models by Pearson chi-square test of independence, showing no significant difference ($\chi^2(2, N = 201) = 1.10$, $P = 0.578$). All expected cell frequencies met the minimum requirement (expected count ≥ 5), satisfying the test assumption.

In the cross-solving task (Table 7), Gemini Advanced showed higher overall accuracy than GPT-4o (87.5 % vs. 80.0 %), with oral biology showing the largest performance gap. To assess the significance of this difference, a Pearson chi-square test was conducted to compare the distribution of correct and incorrect responses between the two models. However, the result was not significant ($\chi^2(1) = 0.37$, $P = 0.544$), indicating that the models demonstrated comparable capabilities in solving each other's generated questions. Given that this task involved a single 40-item trial, measures of variability such as SDs were not applicable.

Discussion

This study evaluated the performance of GPT-4o and Gemini Advanced in KNDLE, particularly on subject-level accuracy and intramodel consistency across repeated trials.[20] Studies involving the Japanese national dental examinations have demonstrated that GPT-4-based models showed superior performance to Gemini, aligning with the outcomes of the present study.[23,24]

This study extends previously published research on GPT-3.5, GPT-4, and Gemini, which revealed strong overall performance but lower accuracy in Korean-language and legally oriented items in the National Dental Hygienist Examination. These earlier findings emphasized the need to test newer-generation LLMs under more linguistically complex and clinically demanding conditions.

Accordingly, GPT-4o and Gemini Advanced were subjected to a more comprehensive and rigorous evaluation: the KNDLE. GPT-4o both outperformed Gemini Advanced in overall accuracy and showed broader domain-level competence and greater response stability across repeated trials.

Table 7 Accuracy rates of GPT-4o and Gemini Advanced on cross-generated questions (cross-solving task).Each model attempted to answer 40 questions (10 per subject) generated by the other model in a single trial.

Subject	GPT-4o (%)	Gemini advanced (%)
Orthodontics	80	100
Prosthodontics	90	70
Oral pathology	90	90
Oral biology	60	90
Average	**80**	**87.5**

These findings highlight the growing capabilities of LLMs in non-English dental education and the need for multi-run evaluations in assessing model reliability in high-stakes testing. The use of three independent runs for each question provides a more robust evaluation framework than previous single-pass approaches.

GPT-4o consistently outperformed Gemini Advanced across 5 years (2019—2023) of KNDLE data, achieving an average accuracy of 81.1 %, compared with 76.59 % for Gemini (Table 2). Although both models surpassed the minimum passing score of 60 %, they remained below the average performance of actual student examinees (95.0 %). Thus, although LLMs demonstrate impressive competence, they do not yet match the depth of clinical knowledge expected of licensed practitioners.[25,26] The significant difference in accuracy between GPT-4o and Gemini ($P = 0.013$) supports the superiority of GPT-4o in handling complex dental knowledge. To enhance accuracy and reduce potential bias, weighted averages and SDs were used to summarize overall performance across the 5-year dataset. This method accounts for year-to-year differences in the question volume and provides a more balanced assessment of model performance. The relatively low weighted SD of GPT-4o (2.76) indicates greater consistency across years compared with Gemini Advanced (4.10) that showed more fluctuations in yearly accuracy.

Subject-specific accuracy results (Table 3) reveal important trends regarding domain sensitivity in both models. GPT-4o outperformed Gemini Advanced in 11 of the 13 dental subjects, with the highest accuracy observed in periodontology (+15.29 %), oral medicine (+7.26 %), and conservative dentistry (+5.78 %). However, both models showed relatively lower accuracy in prosthodontics (GPT-4o, 66.75 %; Gemini, 62.8 %)—a subject that requires nuanced clinical reasoning and procedural understanding. These findings are consistent with the results of previous studies, indicating that LLMs tend to perform less accurately in clinical disciplines requiring multi-step interpretation or treatment planning than in knowledge-based domains, such as oral biology or dental materials.[27,28] Notably, Gemini Advanced performed slightly better than GPT-4o in oral and maxillofacial radiology and health and medical law; however, these were the only exceptions.

Both models consistently showed lower accuracy in prosthodontics and orthodontics, which require integrated clinical reasoning and procedural decision-making. This trend aligns with the results of previous studies that LLMs tend to underperform in domains that require multistep logic, treatment planning, and nuanced judgment. Thus, the observed errors are more likely attributable to domain-specific cognitive demands rather than with linguistic comprehension issues.

Conversely, the relatively low accuracy in health and medical law may stem from the limited exposure to Korean-specific legal terminology and healthcare regulations, which are typically underrepresented in general LLM training corpora. Thus, culturally and linguistically relevant data must be incorporated when evaluating LLMs in localized educational contexts.

Table 4 demonstrates the consistent accuracy of GPT-4o across three independent evaluation runs (average, 81.23 %, SD = 0.42) compared with more variable performance by Gemini (average, 75.37 %, SD = 2.92 %). Although the difference in means was not significant ($P = 0.071$), the trend underscores GPT-4o's stability, a critical factor when considering LLM integration in educational or assessment settings. This consistency was further supported by repeated-measures ANOVA, which did not find significant differences among the three trials for GPT-4o or Gemini Advanced ($P > 0.5$ for both), underscoring their internal stability across repeated runs.

Table 5 presents the intramodel consistency analysis. GPT-4o exhibited a higher identical response rate (81.8 %) than Gemini Advanced (78.0 %) and a greater proportion of consistently correct responses (3/3 runs correct; 73.2 % vs. 66.5 %). Although Gemini had a slightly higher proportion of consistently incorrect answers, the differences did not reach significance ($P = 0.578$).

In addition, both models were evaluated for their ability to generate new questions and solve each other's items. Both models demonstrated high accuracy when solving their generated questions (Tables 6 and 7); however, GPT-4o showed a notable drop in cross-solving—particularly on Gemini's oral biology items (60 %), despite previously strong performance on official oral biology items (91.2 %, Table 3).

Thus, variations in phrasing or conceptual framing between models can significantly affect interpretability—particularly in knowledge-dense domains such as oral biology, where nuanced biomedical terminology may vary between models and affect comprehension. These findings highlight the need to incorporate expert review or standardized prompt protocols when using LLM-generated items across platforms.

However, both models demonstrated relatively lower accuracy in procedure-intensive clinical subjects, such as prosthodontics (Table 3), aligning with previous findings that current LLMs may struggle with context-specific treatment planning or multistep clinical reasoning.[27,28] While LLMs may serve as useful adjuncts in clinical education, further refinement is necessary before deploying them in real-world clinical settings. In the near term, GPT-4o may be best positioned as a clinical education adjunct, for example, by generating formative quiz, offering initial diagnostic suggestions for discussion, or simulating clinical scenarios in a safe learning environment. LLMs also provide valuable opportunities for self-directed learning. Through these models, enabling students to create personalized quiz items, explore clinical case variations, and receive immediate feedback can promote active engagement and individualized learning, particularly in preclinical and revision-focused settings. With further domain-specific fine-tuning and integration of multimodal data, LLMs, like GPT-4o, could be adapted for more advanced clinical decision support in dental practice.[10,29—31]

Although LLMs, such as ChatGPT, can generate dental board-style questions with near-human quality, our review of self-generated items revealed that incorrect answers often stemmed from ambiguous phrasing, poorly constructed distractors, or factual errors, indicative of hallucination.[18,19] These issues underscore the critical role of expert oversight in ensuring the reliability of AI-generated educational content. LLMs occasionally failed to answer questions they generated, with analysis indicating that these errors resulted from item flaws and hallucination. This observation is consistent with the results of previous studies reporting similar limitations in LLM-based question generation.

Several study limitations must be acknowledged. First, the study utilized only text-based questions, excluding image-based or diagrammatic content that is critical in dental assessments. Second, the evaluation focused solely on multiple-choice questions, limiting insight into open-ended reasoning capabilities. Furthermore, reasoning quality, clinical justification, and explanation generation, which are essential in real-world educational settings, were not systematically evaluated.

Thus, future studies should include image-based and open-ended item formats and expert-led qualitative evaluations to analyze the reasoning and explanation quality of LLMs. Cross-linguistic comparisons using equivalent questions in Korean and English can provide insights into the influence of language and cultural context on model performance. Given that LLM performance is sensitive to prompt phrasing, standardized question-design guidelines ensure fairness and reproducibility. Finally, expanding the range of evaluated models—including open-source and multimodal architectures—will improve the generalizability of findings across platforms.

This study focused on two leading commercial LLMs—GPT-4o and Gemini Advanced. Although this allowed for a structured and in-depth comparison, future studies should explore a wider array of models, including open-source alternatives and emerging multimodal architectures, to improve the generalizability and applicability of findings across different LLM platforms.

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

Acknowledgements

The authors received no specific funding or institutional support for this study.

References

1. Guan H, Yap P-T, Bozoki A, Liu M. Federated learning for medical image analysis: a survey. *Pattern Recogn* 2024;151:110424.
2. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30—6.
3. Carrillo-Perez F, Pecho OE, Morales JC, et al. Applications of artificial intelligence in dentistry: a comprehensive review. *J Esthetic Restor Dent* 2022;34:259—80.
4. Jiang L, Wu Z, Xu X, et al. Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies. *J Int Med Res* 2021;49. 03000605211000157.

5. Zarei M, Mamaghani HE, Abbasi A, Hosseini MS. Application of artificial intelligence in medical education: a review of benefits, challenges, and solutions. *Med Clin Pract* 2024;7:100422.
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29:1930—40.
7. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *npj Digit Med* 2023;6:210.
8. Mu Y, He D. The potential applications and challenges of chatgpt in the medical field. *Int J Gen Med* 2024:817—26.
9. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. *Health Care Sci* 2023;2:255—63.
10. Huang H, Zheng O, Wang D, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *Int J Oral Sci* 2023;15:29.
11. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as chatgpt for dental medicine. *J Esthetic Restor Dent* 2023;35:1098—102.
12. Claman D, Sezgin E. Artificial intelligence in dental education: opportunities and challenges of large language models and multimodal foundation models. *JMIR Med Educ* 2024;10: e52346.
13. Thurzo A, Strunga M, Urban R, Surovková J, Afrashtehfar KI. Impact of artificial intelligence on dental education: a review and guide for curriculum update. *Educ Sci* 2023;13:150.
14. Liu M, Okuhara T, Huang W, et al. Large language models in dental licensing examinations: Systematic review and meta-analysis. *Int Dent J* 2025;75:213—22.
15. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep* 2023; 13:16492.
16. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002.
17. Hu Z, Xu Z, Shi P, 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;27:401—11.
18. Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu CT. The quality of AI-generated dental caries multiple choice questions: a comparative analysis of ChatGPT and Google Bard language models. *Heliyon* 2024;10:e28198.
19. Kim HS, Kim GT. Can a large language model create acceptable dental board-style examination questions? A cross-sectional prospective study. *J Dent Sci* 2025;20:895—900.
20. 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* 2025;23:267—76.
21. Patil R, Gudivada V. A review of current trends, techniques, and challenges in large language models (llms). *Appl Sci* 2024; 14:2074.
22. Chi Z, Huang H, Liu L, Bai Y, Gao X, Mao XL. Can pretrained English language models benefit non-English NLP systems in low-resource scenarios?. In: *IEEE/ACM Trans Audio Speech Lang Process*, 32; 2023:1061—74.
23. 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;19:2262—7.
24. Fukuda H, Morishita M, Muraoka K, et al. Evaluating the image recognition capabilities of GTP-4v and Gemini Pro in the Japanese national dental examination. *J Dent Sci* 2025;20: 368—72.
25. Wan N, Jin Q, Chan J, et al. Humans continue to outperform large language models in complex clinical decision-making: a study with medical calculators. *arXiv Prepr arXiv:241105897* 2024.
26. Liu J, Zhou P, Hua Y, et al. Benchmarking large language models on cmexam-a comprehensive Chinese medical exam dataset. *Adv Neural Inf Process Syst* 2023;36:52430—52.
27. Chau RCW, Thu KM, Yu OY, Hsung RTC, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. *Int Dent J* 2024;74:616—21.
28. Uehara O, Morikawa T, Harada F, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese national dental examination. *J Dent Educ* 2025;89:459—66.
29. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. *arXiv Prepr arXiv:240413066* 2024.
30. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. *Medicina* 2025;13:603.
31. Rutledge GW. Diagnostic accuracy of gpt-4 on common clinical scenarios and challenging cases. *Learn Health Syst* 2024;8: e10438.