Journal of Dental Sciences 21 (2026) 198–205

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Analysis of multimodal large language models on visually-based questions in the Japanese National Examination for Dental Hygienists: A preliminary comparative study

Yoshino Kaneyasu [a], Yuichi Mine [b,c*], Yoshie Niitani [d], Tsuyoshi Taji [e], Saori Takeda [b,c], Ryohei Tokinaga [c], Hideo Shigeishi [a], Toshinobu Takemoto [d], Naoya Kakimoto [f], Takeshi Murayama [b,c], Kouji Ohta [a]

[a] *Department of Public Oral Health, Graduate School of Biomedical and Health Sciences, Hiroshima, Japan*
[b] *Project Research Center for Integrating Digital Dentistry, Hiroshima University, Hiroshima, Japan*
[c] *Department of Medical Systems Engineering, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan*
[d] *Department of Oral Health Management, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan*
[e] *Department of Oral Biology & Engineering, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan*
[f] *Department of Oral and Maxillofacial Radiology, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan*

Received 1 June 2025; Final revision received 7 June 2025
Available online 20 June 2025

KEYWORDS
Dental hygienist licensing examination;
OpenAI o3-mini-high;
ChatGPT-4.5 Preview;
Gemini 2.0 Flash Thinking

Abstract *Background/purpose:* The importance of oral health is globally recognized, which has increased the demand for qualified dental hygienists. This study assessed the performance of multimodal large language models (LLMs), on the Japanese National Examination for Dental Hygienists, focusing on their ability to answer visually-based questions and evaluating image-recognition capabilities.
Materials and methods: The 34th Japanese National Examination for Dental Hygienists (March 2025) supplied 213 multiple-choice questions (74 text-only, 139 visually-based). Five multimodal LLMs were tested: OpenAI o3-mini-high (o3-mh), ChatGPT-4.5 Preview (GPT-4.5), Gemini

* Corresponding author. Department of Medical Systems Engineering, Graduate School of Biomedical and Health Sciences, Hiroshima University, 1-2-3 Kasumi Minami-ku, Hiroshima 734-8553, Japan.
E-mail address: mine@hiroshima-u.ac.jp (Y. Mine).

https://doi.org/10.1016/j.jds.2025.06.005

Experimental; Gemini 2.5 Pro Experimental; Claude 3.7 Sonnet

2.0 Flash Thinking Experimental (Gemini 2.0), Gemini 2.5 Pro Experimental (Gemini 2.5), and Claude 3.7 Sonnet (Claude 3.7). Performance was evaluated by comparing LLM answers to official correct answers. Cochran's Q test and McNemar's tests with Bonferroni correction were used for statistical analysis.

Results: Gemini 2.5 achieved the highest overall correct response rate (85.0 %), followed by Claude 3.7 (77.5 %), o3-mh (77.0 %), GPT-4.5 (76.1 %), and Gemini 2.0 (75.1 %). For text-only questions, Claude 3.7 (91.9 %) performed best. On visually-based questions, Gemini 2.5 was superior (82.0 %), while other models scored around 70–73 %. Gemini 2.5 significantly outperformed, GPT-4.5 and Gemini 2.0 overall, and GPT-4.5 and Claude 3.7 on visually-based questions.

Conclusion: Multimodal LLMs, particularly Gemini 2.5, demonstrate significant proficiency on the Japanese National Examination for Dental Hygienists, including questions with visual elements. These findings suggest a growing potential for LLMs as educational tools in dental hygiene. However, current limitations in accuracy and reliability necessitate further refinement and cautious integration into educational and clinical settings.

Introduction

Adopted at the 75th World Health Assembly in 2022, the Landmark Global Strategy on Oral Health includes six strategic objectives: oral health governance, oral health promotion and disease prevention, the health workforce, oral healthcare, oral health information systems, and oral health research agendas.[1] The importance of oral health has increased worldwide in recent years. Consequently, there is a demand for personnel who can provide proper oral care, such as dental hygienists, to help people maintain healthy oral hygiene.[2] Formally established in 34 countries worldwide with defined core competencies for practice,[3] dental hygienists serve as preventive oral health professionals. Their work significantly contributes to population health through a focus on disease prevention and health promotion, encompassing the provision of preventive oral care, alongside education and instruction in oral health. Recently, technology has been approved for dental hygienist education, as well as for educating patients about oral health. Technology is predicted to receive more attention in the future.[4,5]

However, the acceptance and scope of the profession varies widely from region to region.[6] Educational pathways range from shorter vocational programs to university-level degrees, reflecting different national standards and requirements.[7] Workforce distribution issues further complicate the global landscape of dental hygiene. Marked disparities exist in the supply of dental hygienists between regions: high-income countries generally enjoy a high density of hygienists, whereas many low-income countries have few or none, creating large gaps in access to preventive oral health.[2] Workforce analyses indicate that a complex interplay of factors, including the expansion of educational program capacity, evolving scope of practice regulations, and economic fluctuations, has created regional disparities in the availability of dental hygiene services.[8] These imbalances have direct implications for oral health outcomes, as insufficient workforce supply can limit access to preventive care.

Large language models (LLMs) have garnered attention as a novel support mechanism in health care.[9,10] LLMs are sophisticated machine learning models trained on vast corpora of text, enabling them to understand natural language and generate human-like responses. These models, exemplified by generative transformers such as OpenAI's GPT series, can serve a variety of functions from answering questions and summarizing information to aiding in complex problem-solving. A key advancement in this domain is the advent of multimodal LLMs, which can process not only text but also other data types (such as images, audio, or numerical data).[11] This is especially relevant for dentistry and oral health, where diagnostic information is inherently multimodal (*i.e.*, clinicians rely on written notes, radiographic images, photographs, and even patient videos).

Understanding how these models interpret professional knowledge, especially visual elements critical to dental situations, is essential for evaluating their potential applications in dental hygiene education and clinical practice. Licensing exam questions provide an objective standard to evaluate LLMs' capabilities in healthcare domains.[12–14] Previous studies have primarily used text-based inputs to assess LLMs; however, with the emergence of multimodal LLMs, researchers have begun to evaluate performance on questions that include visual elements such as photographs, radiographs, and diagrams.[15–18] Therefore, this study investigated the correct response rates of multimodal LLMs on the Japanese National Examination for Dental Hygienists. In particular, we aimed to provide new insights into the image recognition capabilities of these models, since more than half of the exam questions incorporate visual elements alongside text. The findings may contribute to developing effective educational tools and resources for dental hygiene education and practice.

Materials and methods

Dataset

This study used the 34th Japanese National Examination for Dental Hygienists from March 2025 as a dataset. This exam consists of 220 multiple-choice questions, each of which requires the selection of a certain number of correct answers from four options. According to the Ordinance for Enforcement of the Dental Hygienists Act,[19] the examination subjects are divided into nine areas: structure and function of the human body excluding teeth and oral cavity, structure and function of teeth and oral cavity, pathology and principles of recovery, human and social systems related to dental and oral health promotion and prevention, introduction to dental hygiene, clinical dentistry, theory of dental preventive treatment, theory of dental health guidance, and theory of dental assistance. The exam includes visual materials such as clinical photographs, photographs of dental instruments, photographs of dental prostheses, radiographs, and illustrations.

Seven questions that the Ministry of Health, Labour and Welfare of Japan (MHLW) had officially withdrawn from scoring due to validity concerns were excluded from the dataset.[20] As a result, a total of 213 questions were scored, including 74 text-only questions and 139 visually-based questions. The specialties of the questions were determined by two researchers (Y.K. and Y.N.).

No ethical approval or institutional review was required for this study, as it relied on exam questions published by the MHLW.

Multimodal large language models and prompting

To examine their performance on the Japanese National Examination for Dental Hygienists, we employed five multimodal LLMs in this study. The chosen models were OpenAI o3-mini-high (o3-mh; OpenAI Global, San Francisco, CA, USA, launched January 31, 2025), ChatGPT-4.5 Preview (GPT-4.5; OpenAI, launched February 27, 2025), Gemini 2.0 Flash Thinking Experimental (Gemini 2.0; Google, Mountain View, CA, USA, updated January 21, 2025), Gemini 2.5 Pro Experimental (Gemini 2.5; Google, Mountain View, launched March 25, 2025), and Claude 3.7 Sonnet (Claude 3.7; Anthropic, San Francisco, CA, USA, launched February 24, 2025). Each LLM can process both textual and visual data at the same time.

No prompt engineering or special instructions for guiding the models were provided; a zero-shot approach was used.[21] Only the original Japanese text of each question and the corresponding answer choices were entered directly into the prompt window. If a question included figures, tables, and/or images, those materials were presented to the multimodal LLMs in their original form, without any additional explanation. o3-mh, GPT-4.5, and Claude 3.7 used the official web interface, while Gemini was accessed via Google AI Studio. Each time an answer to one question was output, the session was reset, and the next question was entered anew. To obtain the correct response rate, the answers to the questions officially announced by the MHLW were compared with the results output by each LLM. Claude 3.7 adopted the extended thinking mode. For Gemini models, where parameters like temperature were adjustable, all queries were sent with the temperature set to zero.

Statistical analysis

Data analysis was carried out using IBM SPSS Statistics version 27 (IBM SPSS, Inc., Armonk, NY, USA). An initial assessment using Cochran's Q test was conducted to evaluate whether the proportion of correct responses differed significantly among the five multimodal LLMs. Following a significant result from Cochran's Q test, McNemar's tests were employed for post-hoc pairwise comparisons between each pair of models. For these ten pairwise comparisons, *P* values were Bonferroni-adjusted (multiplied by 10); adjusted *P*-values <0.05 were considered statistically significant.

Results

Table 1 shows the correct response rates for the five multimodal LLMs on the Japanese National Examination for Dental Hygienists. Gemini 2.5 achieved the highest correct response rate of 85.0 % (95 % CI: 79.5—89.5) for all 213 questions, followed by Claude 3.7 with 77.5 % (95 % CI: 71.3—82.9), o3-mh with 77.0 % (95 % CI: 70.8—82.5), GPT-4.5 with 76.1 % (95 % CI: 69.7—81.6), and Gemini 2.0 with 75.1 % (95 % CI: 68.8—80.8). When analyzing the 74 text-only questions, all models demonstrated higher correct response rates than their overall performance. Claude 3.7 performed best with a score of 91.9 % (95 % CI: 83.2—97.0), followed closely by Gemini 2.5 with a score of 90.5 % (95 % CI: 81.5—96.1). GPT-4.5, o3-mh, and Gemini 2.0 achieved 87.8 % (95 % CI: 78.2—94.3), 85.1 % (95 % CI: 75.0—92.3), and 82.4 % (95 % CI: 71.8—90.3), respectively. For the 139 visually-based questions that included images, figures, or

Table 1 Correct response rates (%) and 95 % CIs of the five LLMs.

	o3-mh	GPT-4.5	Gemini 2.0	Gemini 2.5	Claude 3.7
All questions	77.0 (70.8—82.5)	76.1 (69.7—81.6)	75.1 (68.8—80.8)	85.0 (79.5—89.5)	77.5 (71.3—82.9)
Text-only questions	85.1 (75.0—92.3)	87.8 (78.2—94.3)	82.4 (71.8—90.3)	90.5 (81.5—96.1)	91.9 (83.2—97.0)
Visually-based questions[a]	72.7 (64.5—79.9)	69.8 (61.4—77.3)	71.2 (62.9—78.6)	82.0 (74.6—88.0)	69.8 (61.4—77.3)

CI, Confidence interval; LLMs, Large language models; o3-mh, OpenAI o3-mini-high; GPT-4.5, ChatGPT-4.5 Preview; Gemini 2.0, Gemini 2.0 Flash Thinking Experimental; Gemini 2.5, Gemini 2.5 Pro Experimental; Claude 3.7, Claude 3.7 Sonnet.

[a] Includes one or more images, figures, or tables.

tables, the correct response rates were lower across all models. Gemini 2.5 performed best with a rate of 82.0 % (95 % CI: 74.6–88.0). o3-mh followed with a rate of 72.7 % (95 % CI: 64.5–79.9). Gemini 2.0 followed with a rate of 71.2 % (95 % CI: 62.9–78.6). GPT-4.5 and Claude 3.7 both achieved a rate of 69.8 % (95 % CI: 61.4–77.3).

Table 2 shows the statistical comparison of correct response rates between the models using McNemar's tests with Bonferroni correction. Gemini 2.5 significantly outperformed GPT-4.5 ($P = 0.029$) and Gemini 2.0 ($P = 0.010$) for all questions combined, while other pairwise comparisons did not reach statistical significance ($P > 0.05$). Cochran's Q test revealed no significant differences among the five models for text-only questions, suggesting comparable performance in processing text-based content. For visually-based questions, Gemini 2.5 significantly outperformed GPT-4.5 and Claude 3.7 (both $P = 0.027$), while other comparisons were not statistically significant.

Table 3 provides a detailed analysis of the correct response rates for each specialty in the examination. The highest correct response rates were observed in the "Pathology and principles of recovery" category for all five models, with 100 % accuracy on all 12 questions. Conversely, the models generally performed less well in the "Theory of dental preventive treatment" and "Theory of dental assistance" categories. When analyzing text-only questions by specialty, all models demonstrated 100 % accuracy in the "Structure and function of the human body excluding teeth and oral cavity" and "Pathology and principles of recovery" categories. Performance varied in other specialties. In the "Theory of dental preventive treatment" category, notable variation in performance was observed. o3-mh achieved 40.0 % accuracy, while Claude 3.7 and Gemini 2.5 achieved 100 %. Performance varied more for visually-based questions by specialty. In the "Introduction to dental hygiene" category with visual elements, o3-mh, Gemini 2.0, and Claude 3.7 achieved 0 % accuracy on the single question in this category. In contrast, GPT-4.5 and Gemini 2.5 achieved 100 % accuracy on the single question in this category. In the "Clinical dentistry" specialty, which had the most questions with visual elements (51), Gemini 2.5 had the highest accuracy (84.3 %), and Claude 3.7 had the lowest (68.6 %).

Fig. 1 presents an UpSet plot showing the distribution of correct and incorrect responses among the five multimodal LLMs. Of the 213 questions, 119 (55.9 %) were correctly answered by all five models. These questions included 53 text-only questions and 66 visually-based questions. Conversely, all models missed 14 questions (6.6 %), including three text-only questions and 11 visually-based questions.

Discussion

Our results show that Gemini 2.5 outperformed other multimodal LLMs by a significant margin on the 34th Japanese National Examination for Dental Hygienists. Gemini 2.5 achieved an overall correct response rate of approximately 85 %, compared to 75–78 % for the other models (o3-mh, GPT-4.5, Gemini 2.0, and Claude 3.7). These results are consistent with prior studies that have shown more advanced LLMs tend to perform better on dental board exams.[22,23] For instance, Yamaguchi et al.[22] reported that GPT-4 scored approximately 75 % on the text-only part of the 32nd Japanese National Examination for Dental Hygienists. This score was higher than the approximately 63 % achieved by GPT-3.5. Similarly, GPT-4 outperformed GPT-3.5 across question categories in U.S. dental exams.[23] Our study extends these findings by including visually-based questions, where performance gaps became even more severe. Gemini 2.5 performed well on visually-based questions (~82 % correct), while the other models answered ~73 % correctly. In contrast, a previous study that excluded visual materials reported no significant

Table 2 P-value between overall correct response rates of five LLMs.

		o3-mh	GPT-4.5	Gemini 2.0	Gemini 2.5	Claude 3.7
All questions	o3-mh	–	$P = 1.000$	$P = 1.000$	$P = 0.076$	$P = 1.000$
	GPT-4.5	–	–	$P = 1.000$	$P = 0.029$	$P = 1.000$
	Gemini 2.0	–	–	–	$P = 0.010$	$P = 1.000$
	Gemini 2.5	–	–	–	–	$P = 0.120$
	Claude 3.7	–	–	–	–	–
Text-only questions	o3-mh	–	N.S.	N.S.	N.S.	N.S.
	GPT-4.5	–	–	N.S.	N.S.	N.S.
	Gemini 2.0	–	–	–	N.S.	N.S.
	Gemini 2.5	–	–	–	–	N.S.
	Claude 3.7	–	–	–	–	–
Visually-based questions[a]	o3-mh	–	$P = 1.000$	$P = 1.000$	$P = 0.220$	$P = 1.000$
	GPT-4.5	–	–	$P = 1.000$	$P = 0.027$	$P = 1.000$
	Gemini 2.0	–	–	–	$P = 0.082$	$P = 1.000$
	Gemini 2.5	–	–	–	–	$P = 0.027$
	Claude 3.7	–	–	–	–	–

N.S., Not significant (Cochran's Q test); LLMs, Large language models; o3-mh, OpenAI o3-mini-high; GPT-4.5, ChatGPT-4.5 Preview; Gemini 2.0, Gemini 2.0 Flash Thinking Experimental; Gemini 2.5, Gemini 2.5 Pro Experimental; Claude 3.7, Claude 3.7 Sonnet
Because Cochran's Q test did not indicate a significant difference among the samples, no multiple comparisons were performed.
[a] Includes one or more images, figures, or tables

Table 3 Comparing correct response rates (%) of five LLMs in different specialties on all questions.

Specialty	Questions (n)	o3-mh	GPT-4.5	Gemini 2.0	Gemini 2.5	Claude 3.7
All questions	213	77.0	76.1	75.1	85.0	77.5
Structure and function of the human body excluding teeth and oral cavity	6	83.3	66.7	100	100	83.3
Structure and function of teeth and oral cavity	8	87.5	75.0	75.0	100	100
Pathology and principles of recovery	12	100	100	100	100	100
Human and social systems related to dental and oral health promotion and prevention	34	76.5	76.5	73.5	85.3	85.3
Introduction to dental hygiene	5	80.0	100	60.0	100	80.0
Clinical dentistry	64	79.7	81.3	79.7	84.4	75.0
Theory of dental preventive treatment	23	60.9	56.5	65.2	82.6	69.6
Theory of dental health guidance	19	89.5	79.0	68.4	89.5	73.7
Theory of dental assistance	42	66.7	69.1	69.1	73.8	69.1
Text-only questions	74	85.1	87.8	82.4	90.5	91.9
Structure and function of the human body excluding teeth and oral cavity	2	100	100	100	100	100
Structure and function of teeth and oral cavity	3	66.7	66.7	100	100	100
Pathology and principles of recovery	9	100	100	100	100	100
Human and social systems related to dental and oral health promotion and prevention	21	76.2	76.2	71.4	85.7	85.7
Introduction to dental hygiene	4	100	100	75.0	100	100
Clinical dentistry	13	92.3	92.3	100	84.6	100
Theory of dental preventive treatment	5	40.0	80.0	60.0	100	100
Theory of dental health guidance	6	100	83.3	66.7	83.3	83.3
Theory of dental assistance	11	90.9	100	81.8	90.9	81.8
Visually-based questions[a]	139	72.7	69.8	71.2	82.0	69.8
Structure and function of the human body excluding teeth and oral cavity	4	75.0	50.0	100	100	75.0
Structure and function of teeth and oral cavity	5	100	80.0	60.0	100	100
Pathology and principles of recovery	3	100	100	100	100	100
Human and social systems related to dental and oral health promotion and prevention	13	76.9	76.9	76.9	84.6	84.6
Introduction to dental hygiene	1	0	100	0	100	0
Clinical dentistry	51	76.5	78.4	74.5	84.3	68.6
Theory of dental preventive treatment	18	66.7	50.0	66.7	77.8	61.1
Theory of dental health guidance	13	84.6	76.9	69.2	92.3	69.2
Theory of dental assistance	31	58.1	58.1	64.5	67.7	64.5

LLMs, Large language models; o3-mh, OpenAI o3-mini-high; GPT-4.5, ChatGPT-4.5 Preview; Gemini 2.0, Gemini 2.0 Flash Thinking Experimental; Gemini 2.5, Gemini 2.5 Pro Experimental; Claude 3.7, Claude 3.7 Sonnet.

[a] Includes one or more images, figures, or tables.

differences among text-only models.[22] These results suggest that advanced image understanding is a new differentiator.

The strong performance of certain LLMs suggests their potential as educational support tools in dentistry.[24] For example, an LLM could ask students board-style questions, explain the rationale behind correct and incorrect answers, and help them identify areas that need more study. Indeed, multiple studies have suggested using LLM chatbots to supplement traditional learning methods due to their ability to provide interactive, on-demand explanations.[24–26] our results show that the latest models not only answer correctly, but also often provide detailed reasoning. This can benefit learners by modeling clinical reasoning processes, though caution is needed to ensure the reasoning is accurate. Overall, the educational utility of LLMs is expected to grow as these models improve. In our study, even when the models provided incorrect answers, they often offered instructive feedback. This observation is consistent with the findings of others who have noted that LLMs sometimes provide more detailed explanations for incorrect answers than for correct ones.[14] With careful integration into curricula, LLMs could enhance exam preparation and continuous education, serving as knowledgeable, tireless assistants to instructors and students alike.

A unique feature of this study was the inclusion of visual information in the questions for the assessment of multimodal LLMs. Dental licensing exams often include images, such as radiographs, clinical photos, and diagrams, that test visual diagnostic skills. However, most prior studies of LLMs in dentistry have focused solely on text-based questions.[16,17] In this study, Gemini 2.5 achieved a visually-

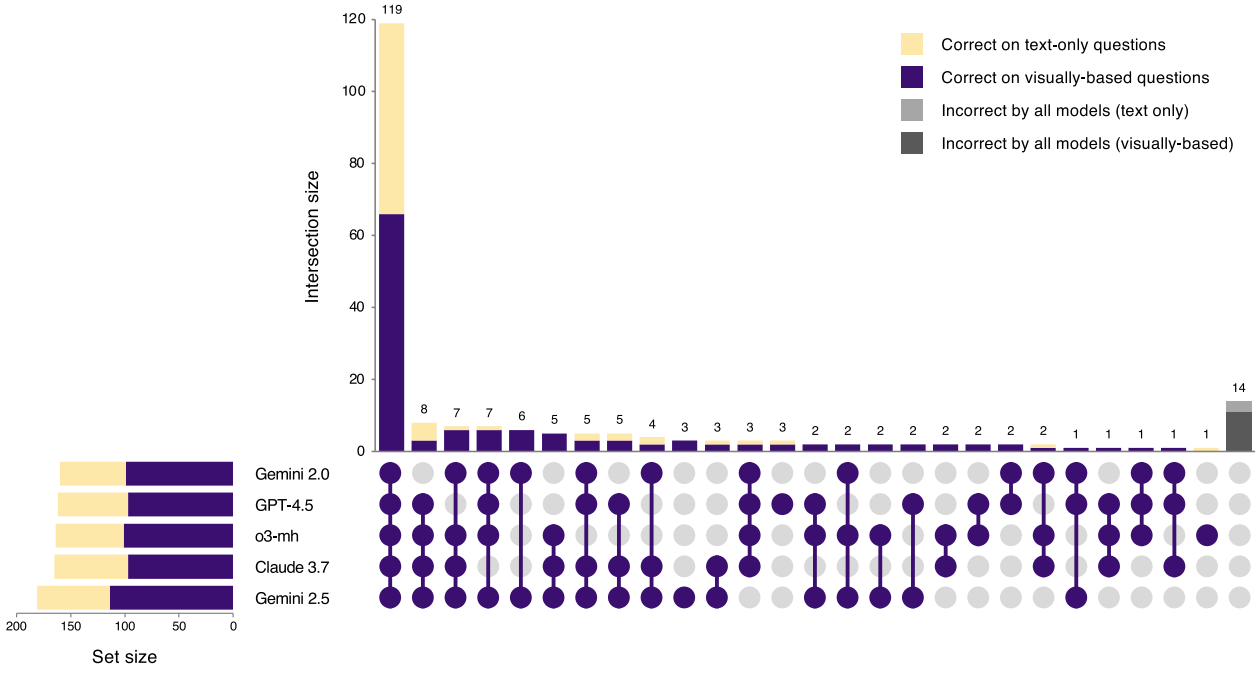

Figure 1 The intersection patterns of correct and incorrect responses among the five multimodal LLMs. The bar graph shows the number of questions correctly responded exclusively by text-only (light yellow) or visual-based (purple) questions, and the number of questions incorrectly responded by all models for text-only (light gray) and visually-based (dark gray) questions. The UpSet plot shows the intersection of correct responses among the five models: Gemini 2.5, Claude 3.7, o3-mh, GPT-4.5, and Gemini 2.0. The horizontal bars on the left represent the total set size (the number of correct responses) for each model. Connected dots indicate which models share correct responses for specific questions. The corresponding bar heights indicate the size of each intersection. LLMs, Large language models; Gemini 2.5, Gemini 2.5 Pro Experimental; Claude 3.7, Claude 3.7 Sonnet; o3-mh, OpenAI o3-mini-high; GPT-4.5, ChatGPT-4.5 Preview; Gemini 2.0, Gemini 2.0 Flash Thinking Experimental.

based correct response rate of 82.0 %, outperforming other models, which averaged around 70 %. These results suggest progress in vision-language integration, a capability that was previously a major limitation of LLM applications in dentistry. A recent study[17] using the 117th Japanese National Dental Examination provides additional insight into the challenges that multimodal LLMs face when interpreting clinical visuals. The study found that 73 % of the questions that all four models missed were visually-based and clustered in specialties requiring complex visual elements, such as orthodontics and crown and bridge prosthodontics. These error types reveal the need for improved visual reasoning architectures and access to training data specific to dentistry. Ongoing evaluation in real-world, image-inclusive tasks, such as those used in national licensing exams, will continue to serve as a critical benchmark for progress in this domain.

Although they show promise, current LLMs have important limitations that restrict their use in high-stakes settings. First, their accuracy is imperfect. In this study, Gemini 2.5 answered about 15 % of questions incorrectly, and other models had error rates around 25 %. A recent meta-analysis revealed that, despite its relative strength, GPT-4's accuracy still falls below the threshold required for clinical application in dentistry.[14] In practice, an error rate of 15–25 % on exam-level questions is too high for unsupervised use in patient care or certification exams. Furthermore, LLMs can exhibit unpredictable knowledge gaps. Our data and previous studies demonstrate that these models perform poorly on certain question types, such as mathematical calculations[23] or specialized regional topics. This indicates that LLMs may not be fully reliable across the entire spectrum of dental knowledge. Language and training-data biases are another concern. For example, GPT-3.5 has been found to perform significantly better on English exam questions than on non-English versions.[14] Similarly, Song and Lee's study[27] on the Korean National Dental Hygienist Examination revealed that all models achieved significantly higher accuracy rates when answering questions in English compared to Korean, with GPT-3.5 showing a remarkable 23.6 % performance gap (61.3 % in English vs. 37.7 % in Korean), highlighting substantial linguistic bias in these models. While our study could handle Japanese content, subtle translation or interpretation issues may have affected their understanding of nuanced clinical terms. Another critical limitation is the tendency toward hallucinations and a lack of trustworthiness. LLMs sometimes generate information that sounds plausible but is incorrect, which can be dangerous in an educational or clinical context. In a comparison of answers to oral pathology case questions, GPT-4o frequently provided fake literature references (50 out of 62 were fake).[28] While we did not specifically test for fabricated outputs in our exam answers, any tendency to present false facts or citations with confidence would undermine the credibility of LLMs as study tools.

Looking ahead, future research should address these gaps. First, more comprehensive evaluations are needed. Our study was limited to one year's exam and a specific set of models. As LLMs evolve rapidly, continual benchmarking using updated exam questions, including image-based, multi-step clinical scenarios, and open-ended items, will be important.[25] This will help track progress and identify persistent weak spots. Second, exploring improvements in model training is crucial. Fine-tuning large models on dental curricula or leveraging retrieval-augmented generation, where the LLM consults external databases, could boost accuracy and reduce hallucinations. Collaboration between dental educators, clinicians, and Artificial Intelligence (AI) developers could facilitate the development of custom models balancing broad linguistic capability and deep dental knowledge. Additionally, techniques that increase the interpretability and transparency of LLM responses would build user trust. For example, one could prompt models to cite sources and implement verification mechanisms for those citations or prompt them to clarify uncertainty when an answer is not well-supported. Finally, studying the impact of LLM assistance on learning outcomes will be valuable. Early adoption in medical education shows promise, but there are mixed results regarding whether AI hints actually improve human problem-solving.[29,30] Rigorous trials in dental education could determine whether students who use AI tutors perform better or become overly reliant on them. Understanding how learners interact with these models will allow us to develop best practices that maximize benefits, such as improved understanding, engagement, and retention, while mitigating risks, such as the propagation of errors or superficial learning.

In conclusion, our study adds to the growing body of evidence indicating that LLMs, especially those with multimodal capabilities, are reaching a level of proficiency suitable for integration into dental education and practice. However, as many studies have cautioned, further refinement and caution are needed. Continued research and careful integration could establish LLMs as reliable tools in training the next generation of dental professionals, improving the diagnostic process, and enhancing patient care.

Declaration of competing interest

All authors declare no conflicts of interest.

Acknowledgments

This study was partially supported by SmaSo-X Challenge Project Young Researchers Research Grant from the Graduate School of Innovation and Practice for Smart Society, Hiroshima University to Y.M., and Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan to Y.M., N.K. and T.M. [25K15961].

References

1. World Health Organization. *Landmark global strategy on oral health adopted at World Health Assembly 75*. Available at: https://www.who.int/news-room/feature-stories/detail/landmark-global-strategy-on-oral-health-adopted-at-world-health-assembly-75. [Accessed 20 May 2025].
2. Gallagher JE, Mattos Savage GC, Crummey SC, Sabbah W, Makino Y, Varenne B. Health workforce for oral health inequity: opportunity for action. *PLoS One* 2024;19:e0292549.
3. Lavigne SE. Dental hygiene's century-long journey to the world stage: professional pride. *Can J Dent Hyg* 2023;57:143—4.
4. Takenouchi A, Otani E, Sunaga M, et al. Development and evaluation of e-learning materials for dental hygiene students in six schools: using smartphones to learn dental treatment procedures. *Int J Dent Hyg* 2020;18:413—21.
5. Kaneyasu Y, Shigeishi H, Sugiyama M, Ohta K. Development and evaluation of the "toothbrushing timer with information on toothbrushes" application: a prospective cohort pilot study. *Clin Exp Dent Res* 2023;9:1206—13.
6. Rederiene G, Bol-van den Hil E, Pajak-Lysek E, Eaton KA. The employment of dental hygienists in European countries: report of a European Dental Hygienists Federation/European Association of Dental Public Health survey in 2021. *Int J Dent Hyg* 2024;22:814—24.
7. Inukai J, Sakurai M, Nakagaki H, et al. Comparison of clinical practice education in dental hygiene schools in eight countries. *Int Dent J* 2012;62:122—6.
8. Dobrow MJ, Valela A, Bruce E, Simpson K, Pettifer G. Identification and assessment of factors that impact the demand for and supply of dental hygienists amidst an evolving workforce context: a scoping review. *BMC Oral Health* 2024;24:631.
9. Büttner M, Leser U, Schneider L, Schwendicke F. Natural language processing: chances and challenges in dentistry. *J Dent* 2024;141:104796.
10. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. *J Med Internet Res* 2025;27:e59069.
11. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res* 2024;26:e59505.
12. Waldock WJ, Zhang J, Guni A, Nabeel A, Darzi A, Ashrafian H. The accuracy and capability of artificial intelligence solutions in health care examinations and certificates: systematic review and meta-analysis. *J Med Internet Res* 2024;26:e56532.
13. 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.
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. Kim W, Kim BC, Yeom HG. Performance of large language models on the Korean dental licensing examination: a comparative study. *Int Dent J* 2025;75:176—84.
16. Mine Y, Taji T, Okazaki S, et al. Analyzing the performance of multimodal large language models on visually-based questions in the Japanese National Examination for Dental Technicians. *J Dent Sci* 2025 (in press).
17. Mine Y, Okazaki S, Taji T, Kawaguchi H, Kakimoto N, Murayama T. Benchmarking multimodal large language models on the dental licensing examination: challenges with clinical image interpretation. *J Dent Sci* 2025 (in press).
18. Wu YH, Tso KY, Chiang CP. Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations. *J Dent Sci* 2025 (in press).
19. *e-Gov portal, Dental Hygienists Act*. Available at: https://laws.e-gov.go.jp/law/401M50000100046. [Accessed 6 April 2025].

20. The Ministry of Health Labour and Welfare of Japan. *Announcement of successful candidates for the 34th Japanese National Examination for Dental Hygienists*. Available at: https://www.mhlw.go.jp/general/sikaku/successlist/2025/siken19/about.html. [Accessed 6 April 2025].
21. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;33:1877—901.
22. 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.
23. 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;54:271—5.
24. Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. *Eur J Dent Educ* 2024;28:865—76.
25. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health* 2025;25:573.
26. Shahzad T, Mazhar T, Tariq MU, Ahmad W, Ouahada K, Hamam H. A comprehensive review of large language models: issues and solutions in learning environments. *Discov Sustain* 2025;6:27.
27. 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.
28. Kaygisiz ÖF, Teke MT. Can deepseek and ChatGPT be used in the diagnosis of oral pathologies? *BMC Oral Health* 2025;25:638.
29. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ* 2024;58:1276—85.
30. Arain SA, Akhund SA, Barakzai MA, Meo SA. Transforming medical education: leveraging large language models to enhance PBL-a proof-of-concept study. *Adv Physiol Educ* 2025;49:398—404.