Journal of Dental Sciences 20 (2025) 2427—2435

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Benchmarking multimodal large language models on the dental licensing examination: Challenges with clinical image interpretation

Yuichi Mine [a,b*], Shota Okazaki [a,b], Tsuyoshi Taji [c], Hiroyuki Kawaguchi [d], Naoya Kakimoto [e], Takeshi Murayama [a,b]

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

Received 6 March 2025; Final revision received 14 March 2025
Available online 26 March 2025

KEYWORDS
Multimodal large language models;
Image recognition;
ChatGPT-4o;
OpenAI o1;
Claude 3.5 sonnet;
Gemini 2.0 flash thinking experimental

Abstract *Background:* /purpose: Large language models (LLMs) have been studied in text-based healthcare tasks, but their performance in multimodal dental applications has not yet been fully explored. This study evaluated the performance of four multimodal LLMs on dental licensing examination questions with both text-only and visually-based components.

Materials and methods: Four multimodal LLMs, ChatGPT-4o (4o), OpenAI o1 (o1), Claude 3.5 Sonnet (Sonnet), and Gemini 2.0 Flash Thinking Experimental (Gemini), were tested on 353 questions from the 2024 Japanese National Dental Examination, including 204 text-only and 149 visually-based questions spanning 17 dental specialties. A zero-shot approach was used without prompt engineering. Performance was analyzed using Cochran's Q test and McNemar's test with Bonferroni correction.

Results: o1 achieved the highest overall correct response rate (81.9 %), followed by Sonnet (71.7 %), Gemini (66.6 %), and 4o (65.7 %). All models performed significantly better on text-only questions (79.9—92.2 %) than on visually-based questions (45.6—67.8 %). Performance varied by specialty, with highest scores in basic medical sciences (Dental pharmacology: 100 %; Oral physiology: 86.7—100 %) and lower scores in clinical specialties requiring visual interpretation (Orthodontics: 36.4—66.7 %).

[*] 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.2025.03.018

Conclusion: Multimodal LLMs demonstrate promising performance on dental examination questions, particularly in text-based scenarios, but significant challenges remain in complex visual interpretation. The remarkable zero-shot performance of newer models such as o1 suggests potential applications in dental education and certain aspects of clinical decision support, although further advances are needed before reliable application in visually complex diagnostic workflows.

Introduction

Digital technologies have progressively reshaped modern dentistry, resulting in new diagnostic tools, treatment modalities, and patient-education strategies.[1] Cone-beam computed tomography (CBCT), intraoral scanners, and Computer-Aided Design/Computer-Aided Manufacturing systems have become an integral part of routine procedures, enabling greater precision and efficiency.[2] Building on these advances, artificial intelligence (AI) has received increasing attention as a means to further enhance clinical workflows, including automated radiographic analysis and predictive algorithms for treatment outcomes.[3,4] While early AI applications focused on image processing for caries detection[5,6] or orthodontic measurements,[7,8] attention has recently turned to large language models (LLMs) that can generate, summarize, and reason about textual information in large knowledge domains.[9,10]

LLMs offer the potential to improve communication with patients by providing accessible explanations of treatment options and oral health guidelines.[11] In an educational context, LLMs can provide personalized tutoring, simulate case-based learning, or help students review question banks for licensing exams.[12,13] They also show potential for chairside clinical decision support, suggesting diagnostic or therapeutic considerations for specific patient presentations.[14] However, despite their growing interest in medical and dental research, much of the current literature on LLM performance remains limited to text-only tasks. In contrast, dentistry relies heavily on image-based evidence such as radiographs, clinical photographs, and histology slides to diagnose pathology, design restorations, and plan treatment. In fact, many dental licensing exams include visuals to reflect the real-world decision-making scenarios; however, LLM assessments often remove these visual materials.[12,15–19] Addressing this gap is essential to advancing AI applications that can truly improve clinical care, rather than simply generating text-based summaries. Therefore, it remains important to determine whether advanced multimodal LLMs can process both textual and visual data in realistic dental situations, which could facilitate more robust image-integrated AI solutions in modern dentistry.

In this study, we evaluated four multimodal LLMs on a set of text-only and visually-based questions derived from a dental licensing exam. Unlike many previous LLM evaluations that omit images or reduce visual content, we intentionally included all items, such as radiographs, clinical photographs, diagrams, and textual prompts, to assess each model's capacity for accurate visual-text integration. Our goal was not simply to see if LLMs could "pass" a licensing exam, but to identify the potential and limitations of multimodal AI in dentistry. By using a real-world, image-inclusive question set and examining the results at the specialty level, we sought to provide a more detailed perspective on where these models succeed, where they fail, and what steps might move dental AI toward true clinical use.

Materials and methods

Dataset

This study used the 117th Japanese National Dental Examination (JNDE-2024)[20] from February 2024 as a comprehensive benchmark for assessing LLM performance. The JNDE-2024 consists of 360 multiple-choice questions, each requiring the selection of a certain number of correct answers from five options (one, two, three, four or all correct). The questions cover a wide range of dental knowledge, including basic medical sciences, general dentistry, epidemiology, and clinical decision-making scenarios. The exam includes a variety of visual materials, such as clinical photographs, radiographs, histopathology slides, diagnostic diagrams, and statistical visualizations. In this study, "figure" refers to graphical information and diagnostic diagrams, while "image" refers to intraoral photographs, radiographs, and similar materials.

We excluded from our analysis the seven questions that the Ministry of Health, Labour and Welfare of Japan (MHLW) had officially withdrawn from scoring due to validity concerns. As a result, a total of 353 questions were scored, including 204 text-only questions and 149 visually-based questions. The specialties of the questions were determined by two researchers (Y.M. and T.T.) based on the Explanatory Guide for the 117th National Dental Examination Questions published by Azabu Dental Academy (Tokyo, Japan).

All figures, tables, and images were provided by the MHLW in PDF format and were converted to JPEG format for use in this study.

Multimodal large language models and prompting

In this study, four multimodal LLMs were employed to evaluate their performance on the JNDE-2024, including questions containing figures, tables and images. ChatGPT-

4o (4o; OpenAI Global, San Francisco, CA, USA, released on May 13, 2024), OpenAI o1 (o1; OpenAI, released on December 5, 2024), Claude 3.5 Sonnet (Sonnet; Anthropic, San Francisco, CA, USA, updated on October 22, 2024), and Gemini 2.0 Flash Thinking Experimental (Gemini; Google, Mountain View, CA, USA, released on December 19, 2024) were selected. These models are capable of processing both textual and visual data, allowing them to address a broad range of question types.

A zero-shot approach[21] was used, with no special prompt engineering or additional instructions provided to guide the models. The original Japanese text of each question and its corresponding answer choices were input directly into the prompt window. For 4o, o1, and Sonnet, the official web interfaces were used, while Gemini was accessed through Google AI Studio. In most cases, each question was presented individually, and a new conversation was initiated for each question to avoid carry-over context. However, for sections of the exam where two consecutive questions were based on the same image, both questions were input together to the LLMs to replicate the original test format. For questions that included figures, tables, and/or images, these visual materials were provided directly to the multimodal LLM without any additional textual explanation. In cases where the answer choices involved tooth notation, the options were provided as JPEG images. Gemini allowed for parameter adjustments; hence, all queries for Gemini were submitted with the temperature set to zero. This setup ensured that each model received the exact information from the actual exam questions, without any customized prompts or clarifications.

Statistical analysis

Statistical analyses were performed using IBM SPSS Statistics 27 (IBM SPSS, Inc., Armonk, NY, USA). Cochran's Q test was used to statistically compare correct response rates among the four models. If Cochran's Q test indicated a significant difference between the four models, pairwise comparisons were then performed using McNemar's test. Because multiple comparisons were performed (six pairwise tests in total), the Bonferroni correction was applied to maintain an overall family-wise error rate of 0.05.

Results

Overall performance on all questions

A total of 353 questions were administered, of which 149 contained at least one figure, table, or image (visually-based questions) and 204 were text-only. As shown in Table 1, o1 achieved the highest overall correct response rate with 81.9 % [95 % CI (77.4—85.7)]. Sonnet followed with 71.7 % [95 % CI (66.7—76.3)], then Gemini with 66.6 % [95 % CI (61.4—71.5)], and finally 4o with 65.7 % [95 % CI (60.5—70.7)]. Table 2 showed that o1 significantly outperformed all other models ($P < 0.001$ vs 4o, Sonnet and Gemini). In contrast, no significant differences were found among the other three models (4o vs Sonnet: $P = 0.107$; 4o vs Gemini: $P = 1.000$; Sonnet vs Gemini: $P = 0.253$).

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

		4o	o1	Sonnet	Gemini
All questions	4o	—	$P < 0.001$	$P = 0.107$	$P = 1.000$
	o1	—	—	$P < 0.001$	$P < 0.001$
	Sonnet	—	—	—	$P = 0.253$
	Gemini	—	—	—	—
Text-only questions	4o	—	$P < 0.001$	$P = 1.000$	$P = 1.000$
	o1	—	—	$P = 0.004$	$P < 0.001$
	Sonnet	—	—	—	$P = 1.000$
	Gemini	—	—	—	—
Visually-based questions[a]	4o	—	$P < 0.001$	$P = 0.211$	$P = 1.000$
	o1	—	—	$P = 0.069$	$P < 0.001$
	Sonnet	—	—	—	$P = 0.734$
	Gemini	—	—	—	—

LLMs, Large language models; 4o, ChatGPT-4o; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental.

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

Text-only vs visually-based questions

When the questions were subdivided into text-only and visually-based questions, all four models performed significantly better on text-only questions than on visually-based questions. For text-only questions (n = 204), o1 led with a correct response rate of 92.2 % [95 % CI (87.6—95.5)]. Sonnet achieved 83.3 % [95 % CI (77.5—88.2)], 4o 80.4 % [95 % CI (74.3—85.6)], and Gemini 79.9 % [95 % CI (73.7—85.2)]. Statistical analysis revealed significant differences (o1 vs 4o: $P < 0.001$; o1 vs Gemini: $P < 0.001$; o1 vs Sonnet: $P = 0.004$), while no significant differences were found between Sonnet and 4o ($P = 1.000$), Sonnet and Gemini ($P = 1.000$), or 4o and Gemini ($P = 1.000$).

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

	4o	o1	Sonnet	Gemini
All questions	65.7 (60.5—70.7)	81.9 (77.4—85.7)	71.7 (66.7—76.3)	66.6 (61.4—71.5)
Text-only questions	80.4 (74.3—85.6)	92.2 (87.6—95.5)	83.3 (77.5—88.2)	79.9 (73.7—85.2)
Visually-based questions[a]	45.6 (37.5—54.0)	67.8 (59.6—75.2)	55.7 (47.3—63.8)	48.3 (40.1—56.6)

CI, Confidence interval; LLMs, Large language models; 4o, ChatGPT-4o; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental.

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

In contrast, performance declined for all models on the visually-based questions (n = 149). o1 remained the highest-scoring model with 67.8 % [95 % CI (59.6–75.2)], followed by Sonnet with 55.7 % [95 % CI (47.3–63.8)], Gemini with 48.3 % [95 % CI (40.1–56.6)], and 4o with 45.6 % [95 % CI (37.5–54.0)]. Statistical analysis showed a significant difference between o1 and 4o ($P < 0.001$), as well as between o1 and Gemini ($P < 0.001$). However, the difference between o1 and Sonnet was not statistically significant ($P = 0.069$), and the performance of Gemini was not significantly different from that of Sonnet ($P = 0.734$).

Performance by specialty

Detailed performance for the 17 dental specialties is shown in Fig. 1 and Tables 3–5. For all questions, performance varied considerably by specialty, with the highest scores observed in basic medical science areas such as Dental pharmacology (100 % for all models), Oral physiology (86.7–100 %), and Oral pathology (80–100 %) (Table 3). Clinical specialties showed more variation, with particularly challenging areas including Orthodontics (36.4–66.7 %) and Pediatric dentistry (45.7–77.1 %).

When examining text-only questions, all models showed strong performance in basic medical science subjects (Table 4). All four models scored 100 % correct response in Dental pharmacology and nearly all models scored 100 % correct response in Oral physiology and Oral pathology. o1 demonstrated high performance in clinical subjects, achieving over 90 % accuracy in several areas including Oral surgery (96.4 %), Orthodontics (91.7 %), and Oral health (95.2 %).

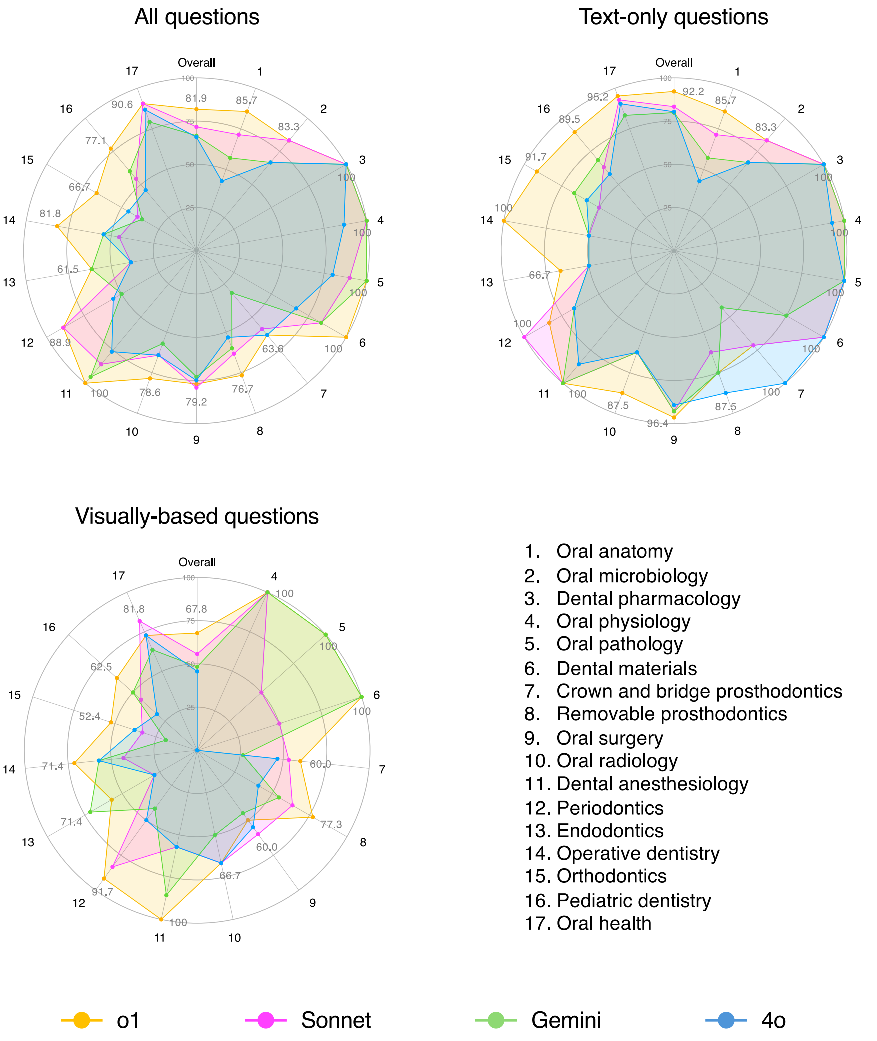

Figure 1 Comparing correct response rates (%) of four LLMs in different specialties in JNDE-2024. Correct response rates of four LLMs (o1, yellow line; Sonnet, pink line; Gemini, green line; and 4o, blue line) were evaluated across 17 dental specialties using three different question types (Results from all questions combined, performance on text-only questions, and performance on visually-based questions). Numbers on radar charts (1–17) correspond to dental specialities: Oral anatomy (1), Oral microbiology (2), Dental pharmacology (3), Oral physiology (4), Oral pathology (5), Dental materials (6), Crown and bridge prosthodontics (7), Removable prosthodontics (8), Oral surgery (9), Oral radiology (10), Dental anesthesiology (11), Periodontics (12), Endodontics (13), Operative dentistry (14), Orthodontics (15), Pediatric dentistry (16), and Oral health (17). LLMs, Large language models; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental; 4o, ChatGPT-4o.

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

Specialty	Questions (n)	4o	o1	Sonnet	Gemini
All questions	353	65.7	81.9	71.7	66.6
Oral anatomy	7	42.9	85.7	71.4	57.1
Oral microbiology	6	66.7	83.3	83.3	66.7
Dental pharmacology	11	100	100	100	100
Oral physiology	15	86.7	100	100	100
Oral pathology	10	80	100	90	100
Dental materials	6	66.7	100	83.3	83.3
Crown and bridge prosthodontics	22	63.6	63.6	59.1	31.8
Removable prosthodontics	30	53.3	76.7	63.3	60
Oral surgery	48	75	77.1	79.2	72.9
Oral radiology	14	64.3	78.6	64.3	57.1
Dental anesthesiology	21	76.2	100	85.7	95.2
Periodontics	18	55.6	88.9	88.9	50
Endodontics	13	38.5	61.5	38.5	61.5
Operative dentistry	11	54.5	81.8	45.5	54.5
Orthodontics	33	45.5	66.7	39.4	36.4
Pediatric dentistry	35	45.7	77.1	54.3	60
Oral health	53	86.8	90.6	90.6	79.2

LLMs, Large language models; 4o, ChatGPT-4o; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental.

Table 4 Comparing correct response rates (%) of four LLMs in different specialties on text-only questions.

Specialty	Questions (n)	4o	o1	Sonnet	Gemini
Text-only questions	204	80.4	92.2	83.3	79.9
Oral anatomy	7	42.9	85.7	71.4	57.1
Oral microbiology	6	66.7	83.3	83.3	66.7
Dental pharmacology	11	100	100	100	100
Oral physiology	14	92.9	100	100	100
Oral pathology	8	100	100	100	100
Dental materials	4	100	100	100	75
Crown and bridge prosthodontics	7	100	71.4	71.4	42.9
Removable prosthodontics	8	87.5	75	62.5	75
Oral surgery	28	89.3	96.4	92.9	92.9
Oral radiology	8	62.5	87.5	62.5	62.5
Dental anesthesiology	14	85.7	100	100	100
Periodontics	6	66.7	83.3	100	66.7
Endodontics	6	50	66.7	50	50
Operative dentistry	4	50	100	50	50
Orthodontics	12	58.3	91.7	50	66.7
Pediatric dentistry	19	57.9	89.5	63.2	68.4
Oral health	42	90.5	95.2	92.9	83.3

LLMs, Large language models; 4o, ChatGPT-4o; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental.

Table 5 Comparing correct response rates (%) of four LLMs in different specialties on visually-based questions.

Specialty	Questions (n)	4o	o1	Sonnet	Gemini
Visually-based questions*	149	45.6	67.8	55.7	48.3
Oral physiology	1	0	100	100	100
Oral pathology	2	0	100	50	100
Dental materials	2	0	100	50	100
Crown and bridge prosthodontics	15	46.7	60	53.3	26.7
Removable prosthodontics	22	40.9	77.3	63.6	54.5
Oral surgery	20	55	50	60	45
Oral radiology	6	66.7	66.7	66.7	50
Dental anesthesiology	7	57.1	100	57.1	85.7
Periodontics	12	50	91.7	83.3	41.7
Endodontics	7	28.6	57.1	28.6	71.4
Operative dentistry	7	57.1	71.4	42.9	57.1
Orthodontics	21	38.1	52.4	33.3	19
Pediatric dentistry	16	31.3	62.5	43.8	50
Oral health	11	72.7	72.7	81.8	63.6

LLMs, Large language models; 4o, ChatGPT-4o; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental.

The introduction of visual elements had a significant impact on performance across all models (Table 5). For visually-based questions, performance varied by specialty. o1 maintained higher performance in some areas, particularly Dental anesthesiology (100 %), Periodontics (91.7 %), and Removable prosthodontics (77.3 %). All models showed lower accuracy in visually complex specialties such as Orthodontics (o1: 52.4 %, Sonnet: 33.3 %, 4o: 38.1 %, Gemini: 19.0 %) and Crown and bridge prosthodontics (ranging from 26.7 % to 60.0 %).

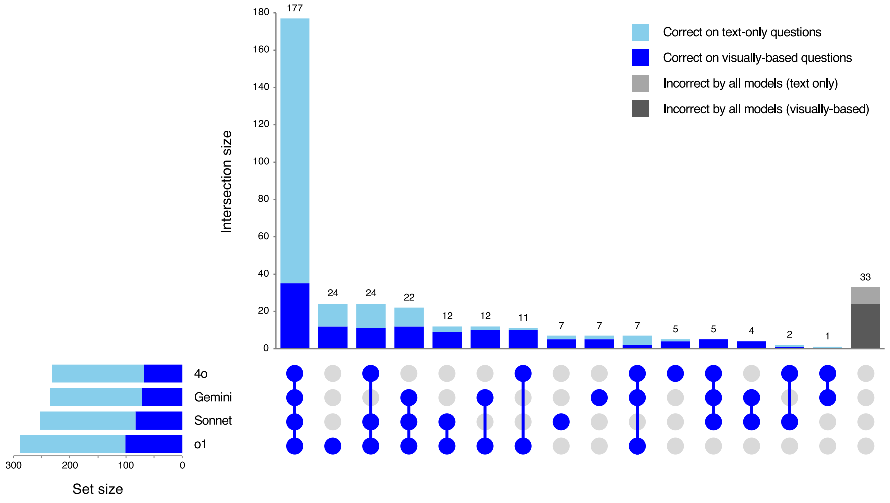

Figure 2 Distribution of questions responded correctly and incorrectly among LLMs. Top: Bar graph shows the number of questions correctly responded exclusively by text-only (light blue) or visual-based (dark blue) questions, and the number of questions incorrectly responded by all models for text-only (light gray) and visually-based (dark gray) questions. Bottom: UpSet plot showing the intersection of correct responses among models (o1, Sonnet, Gemini, and 4o). The horizontal bars on the left represent the total set size (number of correct responses) for each model, while the connected dots indicate which models share correct responses for specific questions, with the corresponding bar heights indicating the size of each intersection. LLMs, Large language models; o1, OpenAI o1; Sonnet, Claude 3.5 Sonnet; Gemini, Gemini 2.0 Flash Thinking Experimental; 4o, ChatGPT-4o.

Model agreement analysis

The UpSet plot revealed both shared successes and model-specific performance patterns (Fig. 2). The largest intersection consisted of 177 questions correctly responded by all four models, with the majority of which were text-only questions. Notably, o1 solved a significant number of questions that the other models could not correctly respond, demonstrating its superior problem-solving capabilities in certain cases. Each model had some unique correct answers, although the frequency varied considerably between models. At the other end of the spectrum, 33 questions (mostly visually-based) proved challenging for all models, with none providing correct answers.

The 33 questions that were not responded correctly by all models were concentrated in specific specialties (Table 6): Orthodontics (7 questions, 21.2 % of all questions in this speciality), Crown and Bridge Prosthodontics (4 questions, 18.2 %), and Oral Surgery (6 questions, 12.5 %). The majority of these universally challenging questions were visually-based, particularly in clinical specialties such as Crown and bridge prosthodontics, Periodontics, and Operative dentistry, where all of the most frequently missed questions contained visual elements. In contrast, basic medical science subjects such as Oral microbiology and Oral health showed fewer universally challenging questions, with most of their difficult questions being text-based rather than visual.

Discussion

In this study, we evaluated four multimodal LLMs, 4o, o1, Sonnet, and Gemini, using text-only and visually-based questions derived from a dental licensing exam.

Our result showed the superior performance of o1 (81.9 % overall correct response rate) without the need for prompt engineering or additional instructions. Some previous studies have examined prompt engineering approaches[18,22,23] or attempted to evaluate image-based questions,[24–26] but often reported unsatisfactory results. o1 processed the dental licensing exam questions in Japanese under zero-shot conditions by directly inputting the exam text and images. According to OpenAI,[27] o1 naturally employs a chain-of-thought[28] approach, much like human test takers who pause to consider challenging questions, and shows encouraging results even in zero-shot scenarios. This performance not only exceeds previous response scores, but is also remarkable for its ability to handle complex dental content in a non-English language. From a practical perspective, such zero-shot robustness indicates that multimodal LLMs can operate in specialized medical or dental domains without prompt tuning or domain-specific instruction sets. At the same time, our results also revealed that visual interpretation remains a critical limitation in complex diagnostic and technical workflows. Therefore, while the text-based performance of zero-shot is promising, further improvements are needed for LLMs to reliably interpret complex dental images and provide accurate, context-sensitive responses in all specialties of dentistry.

Our results, when considered with the recent review of natural language processing in dentistry by Büttner et al.,[10] suggest several promising application areas, but also reveal significant implementation challenges. While our results highlight the robust performance of the models in the basic medical sciences (*e.g.*, Dental pharmacology: 100 %, Oral physiology: 86.7—100 %), the global survey by Uribe et al.[13] also suggests that educators see value in AI for knowledge acquisition (74.3 %) and research support (68.5 %). Taken

Table 6 Analysis of incorrect responses across dental specialties by question type.

Specialty with questions responded incorrectly by all models	Questions (n)	Incorrect by all models within the same specialty (n, %)	Incorrect by all models (text-only; n, %)	Incorrect by all models (visually-based; n, %)	Types of visual materials
Oral microbiology	6	1 (16.7)	1 (100)	0 (0)	–
Crown and bridge prosthodontics	22	4 (18.2)	0 (0)	4 (100)	Intraoral photograph, Prosthesis photograph, Dental technical work photograph,
Removable prosthodontics	30	1 (3.3)	0 (0)	1 (100)	Intraoral photograph, complete denture photograph
Oral surgery	48	6 (12.5)	1 (16.7)	5 (83.3)	Intraoral photograph, Extraoral photograph, Dental radiograph, Panoramic radiograph, Cephalogram, CT, 3D-CT image, MR image, Surgical instrument
Oral radiology	14	2 (14.3)	1 (50)	1 (50)	Panoramic radiograph, CT
Periodontics	18	1 (5.6)	0 (0)	1 (100)	Intraoral photograph, Dental radiograph
Endodontics	13	3 (23.1)	2 (66.7)	1(33.3)	Intraoral photograph, Dental radiograph
Operative dentistry	11	1 (9.1)	0 (0)	1 (100)	Intraoral photograph, Dental radiograph
Orthodontics	33	7 (21.2)	1 (14.3)	6 (85.7)	Intraoral photograph, Extraoral photograph, Panoramic radiograph, technical work photograph, Craniofacial polygon analysis
Pediatric dentistry	35	5 (14.3)	1 (20)	4 (80)	Intraoral photograph, Dental radiograph, Panoramic radiograph
Oral health	53	2 (3.8)	2 (100)	0 (0)	–
Total	283	33 (11.7)[a]	9 (27.3)	24 (72.7)	–

CT, Computed tomography; MR, magnetic resonance.

[a] The incorrect response rate is 11.7 % for questions from listed specialties (283 questions) and 9.4 % for questions from all dental specialties (353 questions total).

together, these points suggest that modern LLMs-especially when operating with minimal prompt engineering (*e.g.*, o1 with 81.9 % overall correct response rate)-could be readily used for factual or theoretical components of dental education. However, the performance gap between text-only questions (92.2 % correct response rate) and visually-based questions (67.8 %) reveals a challenge and is consistent with the report by Uribe et al.[13] that only 38.8 % of educators recognized the promise of AI in clinical skills training. This disparity was most pronounced in specialties such as Orthodontics, Endodontics, and Prosthodontics, suggesting that LLMs may still have difficulty interpreting radiographs, soft tissue contours, or tooth angles. These findings, in turn, validate educators' cautious stance on how AI might-or might not-enhance advanced clinical competencies.

The UpSet plot analysis provides valuable insight into these challenges, revealing that while 177 questions (mostly text-based) were accurately responded by all four LLMs, 33 questions (mostly involving complex visual elements) proved universally challenging. This pattern suggests specific areas where current multimodal systems need improvement, particularly in handling complex clinical imaging scenarios. The improvement in visual task performance compared to previous studies (from 53.7 % to 67.1 % in the current study, albeit with different exams in a previous study[25,26]) indicates progress, but also highlights the significant work still needed to achieve reliable clinical application. A previous study analyzing the Japanese National Examination for Dental Technicians using 4o, o1, and Sonnet found that o1 produced significantly higher percentages of correct responses than 4o on text-only questions, but no significant differences were found between the three LLMs used on the visually-based questions.[29] This may be one reason why the amount of training data related to dental technology in dentistry is so small. The limitations of visual processing revealed in our study indicate the need for significant advances in image interpretation capabilities, particularly for complex clinical images such as radiographs and 3D scans. This challenge can be explained by data availability and quality issues, as Büttner et al. noted a lack of high-quality dental training data and privacy concerns that limit data sharing.[10]

Several limitations of our study should be considered. Our analysis focused on a single country's dental licensing examination, and although the JNDE-2024 is comprehensive, variations in question style or content between different national examinations may result in different performance patterns. The rapid development of LLM technology also means that our results reflect a point in time, and performance characteristics may change with future model updates. In addition, our scoring methodology for multiple-answer items, which is based on an all-or-nothing approach, may not fully capture nuanced performance differences that alternative scoring methods might reveal. Moreover, real-world dental practice often involves dynamic or 3D data (*e.g.*, CBCT scans, continuous patient monitoring) that are beyond the scope of typical LLM image input systems. However, technological innovation is progressing rapidly. In December 2024, OpenAI implemented real-time video mode, screen sharing, and image uploading in its mobile application, in addition to voice chat function.[30] These features, which use native multimodal models, enable more real-time and natural interactions while capturing nonverbal cues. This technological advance suggests possibilities for future dynamic support in dental practice.

Looking forward, the successful integration of LLMs into dental practice will require careful attention to implementation frameworks and quality assurance mechanisms. As both our findings and recent literature emphasize, technical advances must be accompanied by thoughtful consideration of institutional policies, professional education requirements, and ethical boundaries. The development of standardized evaluation metrics and validation protocols will be critical to ensuring the safe and effective use of these technologies in clinical settings.

By benchmarking four advanced multimodal LLMs (4o, o1, Sonnet, and Gemini) on a real JNDE-2024 that included both text-only and visually-based questions, we have demonstrated the models' growing strengths and persistent weaknesses in the dental domain. Although o1 achieved remarkably high overall correct response rate, particularly for text-based questions, performance declined for visually-based questions, demonstrating the challenges of image interpretation. These results demonstrate the potential of LLMs to support educational initiatives and certain aspects of clinical decision-making, but they also reveal the need for significant advances before LLMs can reliably handle the visual demands unique to dentistry.

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.

References

1. Watanabe H, Fellows C, An H. Digital technologies for restorative dentistry. *Dent Clin* 2022;66:567—90.
2. Schierz O, Hirsch C, Krey KF, Ganss C, Kämmerer PW, Schlenz MA. Digital dentistry and its impact on oral health-related quality of life. *J Evid Base Dent Pract* 2024;24:101946.
3. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res* 2020;99:769—74.
4. Hung KF, Yeung AWK, Bornstein MM, Schwendicke F. Personalized dental medicine, artificial intelligence, and their relevance for dentomaxillofacial imaging. *Dentomaxillofacial Radiol* 2023;52:20220335.
5. Mohammad-Rahimi H, Motamedian SR, Rohban MH, et al. Deep learning for caries detection: a systematic review. *J Dent* 2022;122:104115.
6. Moharrami M, Farmer J, Singhal S, et al. Detecting dental caries on oral photographs using artificial intelligence: a systematic review. *Oral Dis* 2024;30:1765—83.
7. Wang CW, Huang CT, Hsieh MC, et al. Evaluation and comparison of anatomical landmark detection methods for cephalometric X-Ray images: a grand challenge. *IEEE Trans Med Imag* 2015;34:1890—900.
8. Khanagar SB, Al-Ehaideb A, Vishwanathaiah S, et al. Scope and performance of artificial intelligence technology in orthodontic diagnosis, treatment planning, and clinical decision-making - a systematic review. *J Dent Sci* 2021;16:482—92.
9. 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.
10. Büttner M, Leser U, Schneider L, Schwendicke F. Natural language processing: chances and challenges in dentistry. *J Dent* 2024;141:104796.
11. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. *Front Med* 2024;11:1477898.
12. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
13. 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.
14. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. *J Med Internet Res* 2023;25:e51580.
15. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. *Int Dent J* 2024;74:616—21.
16. Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. *J Periodontal Res* 2025;60:121—33.
17. 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* 2024 (in press).
18. Morishita M, Fukuda H, Yamaguchi S, et al. An exploratory assessment of GPT-4o and GPT-4 performance on the Japanese national dental examination. *Saudi Dent J* 2024;36:1577—81.
19. 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.
20. The Ministry of Health Labour and Welfare of Japan. *Questions and correct answers for the 117th National dental examination.*

https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp240424-02.html. [Accessed 11 December 2024].

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. Morishita M, Fukuda H, Muraoka K, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. *J Dent Sci* 2024;19:1595—600.
23. Chen Y, Huang X, Yang F, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. *BMC Med Educ* 2024;24:1372.
24. Ishida K, Arisaka N, Fujii K. Analysis of responses of GPT-4 V to the Japanese national clinical engineer licensing examination. *J Med Syst* 2024;48:83.
25. Morishita M, Fukuda H, Muraoka K, et al. Evaluating the image recognition capabilities of GPT-4V and Gemini Pro in the Japanese national dental examination. *J Dent Sci* 2025;20:368—72.
26. 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.
27. OpenAI. *Learning to reason with LLMs*. https://openai.com/index/learning-to-reason-with-llms/. [Accessed 25 December 2024].
28. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *arxiv* 2022;2201:11903v6.
29. 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;20:2460—6.
30. OpenAI. *Santa mode & video in advanced voice—12 Days of OpenAI: day 6*. https://www.youtube.com/live/NIQDnWlwYyQ?si=ewh9_hWY92Yogi5h. [Accessed 25 December 2024].