Performance evaluation of text- and image-based questions by large language model and large multimodal model chatbots in oral and maxillofacial radiology

Jeong, Hui

Department of Dentistry

Graduate School

Yonsei University

Performance evaluation of text- and image-based questions by large language model and large multimodal model chatbots in oral and maxillofacial radiology

Advisor Han, Sang-Sun

**A Dissertation Submitted
to the Department of Dentistry
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
Requirements for the Degree of
Doctor of philosophy in Dental Science**

Jeong, Hui

June 2025

Performance evaluation of text- and image-based questions by large language model and large multimodal model chatbots in oral and maxillofacial radiology

This certifies that the Dissertation of Jeong, Hui is approved

Committee Chair	**Lee, Chena**
Committee Member	**Han, Sang-Sun**
Committee Member	**Jeon, Kug Jin**
Committee Member	**Choi, Yoon Joo**
Committee Member	**Jo, Gyu-Dong**

Department of Dentistry
Graduate School
Yonsei University
June 2025

감사의 글

긴 대학원 생활의 결실을 맺는 이 순간을 맞이하기까지 함께 해 주신 분들께 감사의 인사를 전하기 위하여 이 자리를 빌려 글을 올립니다.

부족한 저에게 학계에 발을 들일 기회를 주시고 성장할 수 있도록 이끌어 주신 한상선 지도교수님과 전국진 교수님께 깊은 감사의 말씀을 드립니다. 더딘 걸음에도 끝까지 인내로 기다려주시고, 늘 따뜻한 응원과 격려의 말씀으로 세심히 지도해 주신 덕분에 단 한 번의 슬럼프 없이 학위과정을 마칠 수 있었습니다. 연구년으로 타국에서 바쁘신 와중에도 학위논문 심사위원장을 맡아 주신 이채나 교수님, 미처 발견하지 못한 연구의 오류를 함께 고민해 주신 최윤주 교수님과 조규동 교수님, 지금은 다른 곳에 계시지만 늘 가까이에서 따뜻하게 연구원들과 소통해 주셨던 김학선 교수님께도 깊이 감사드립니다.

연구실을 떠난 뒤에도 끊임없는 관심과 애정으로 후배들을 아껴 주신 영현 선생님, 아리 선생님, 은규 선생님께 감사드립니다. 같은 길을 먼저 걸어가신 선생님들의 모습을 동경하며 삶의 나침반을 그려 나갔고, 연구실 생활에서 마주하는 크고 작은 문제들에도 선생님들의 본보기를 통해 해답을 찾을 수 있었습니다. 가장 많은 시간을 함께 울고 웃었던 한승 선생님, 멋진 선배이자 든든한 친구로 곁에 있어 주셔서 약 5년 간의 연구실 생활이 인생에 손꼽는 행복한 시간으로 기억될 것 같습니다. 힘들고 바쁠 때면 기꺼이 손과 발이 되어 도와주셨던 지윤 선생님과 예림 선생님, 차분하고 현명한 모습으로 큰 귀감이 되어 주신 유진 선생님, 늘 환한 미소로 교실 분위기를 밝혀 주신 혜원 선생님, 궂은 일도 마다하지 않고 도와주신 강인식 선생님과 문상현 선생님, 각종 행정 업무에 큰 도움을 주신 혜리 선생님께도 감사 인사를 전합니다.

비록 전공은 다르지만 비슷한 시기에 대학원에 입학해 같은 공간에 있다는

것만으로도 큰 의지가 되었던 친구 지영이와 성윤이에게 감사의 마음을 전하며, 재료학교실 실습생 시절부터 졸업을 앞둔 지금까지 아낌없는 조언과 응원을 보내 주신 류정현 박사님께도 감사드립니다.

한참을 방황하던 스무 살 새내기에게 박사학위라는 꿈을 심어 주신 정성균 교수님과 이경희 교수님께 감사드립니다. 시간이 흐를수록, 마음 깊이 존경하는 은사님이 계시다는 사실이 삶의 큰 원동력이 됨을 더욱 절실히 느낍니다.

매사에 고생스러운 길을 자처해도 묵묵히 믿어 주신 부모님, 사랑받은 어린 날들이 모두 모여 오늘의 제가 되었습니다. 지치지 않도록 곁에서 힘을 준 동생 인구에게도 감사합니다. 학위논문 심사 통과 소식에 누구보다 기뻐해 주신 시부모님과, 물심양면으로 지원을 아끼지 않았던 사랑하는 남편에게도 깊은 감사와 존경의 마음을 전합니다. 마지막으로, 학위논문 심사 준비 기간을 함께해 준 소중한 아기 부다에게도 훗날 꼭 고마운 마음을 전하고 싶습니다.

"밀밭에 부는 바람은 오월이 제격이라 나는 무던히 오월을 기다리는 사람이 되었다"는 어느 책 속의 문장을 좋아합니다. 오늘의 결실이 많은 분들의 도움과 애정으로 이루어졌음을 잊지 않고, 모든 무게를 겸허히 짊어지고 다시 사회에 나아가 언제나 그 자리에서 무던히 최선을 다하는 사람이 되겠습니다.

2025년 6월 저자 씀

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Performance evaluation of text- and image-based questions by large language model and large multimodal model chatbots in oral and maxillofacial radiology

Purpose: This study aimed to conduct a comprehensive evaluation of general-purpose large language model (LLM) and large multimodal model (LMM) chatbots in oral and maxillofacial radiology (OMFR) by comparing their performance with dental students, and assessing performance changes from LLM to LMM chatbots.

Materials and Methods: A total of 90 text- and image-based examination questions were extracted from OMFR curriculum in a Korean dental school and categorized into six educational content categories and two question types. Four LLM chatbots (ChatGPT, ChatGPT Plus, Bard, Bing Chat) generated a single response per question, while two LMM chatbots (ChatGPT-4o, Gemini 2.0 Flash) produced ten responses per item. Accuracy was assessed using the first response from each chatbot and compared to student scores. For LMM chatbots, response consistency across repeated outputs was analyzed using Fleiss' kappa coefficient. Hallucination was evaluated by two oral and maxillofacial radiologists using a five-point Global Quality Scales, with mean and standard deviation, and the effect of zero-shot chain-of-thought (ZS-CoT) prompting was examined.

Results: LMM chatbots demonstrated higher accuracy than LLM chatbots on text-based items and outperformed dental students in certain domains. However, their performance remained limited in image-based diagnostic tasks, with frequent variability and hallucinations observed in complex image interpretation and short-answer formats. ZS-CoT prompting did not produce meaningful improvement in accuracy.

Conclusions: This is the first study to compare chatbot performance with student scores using an OMFR questions that includes both textual and image components, while also examining longitudinal performance changes from LLM to LMM chatbots. These findings offer timely insight into the current strengths and limitations of general-purpose AI chatbots. Future work incorporating more diverse clinical images and case scenarios, combined with model customization and advanced prompting strategies, may help enable safer and more effective application of AI chatbots in dental education, patient communication, and clinical practice.

Key words : Oral and maxillofacial radiology, Large language model, Large multimodal model, Artificial intelligence, Chatbot, Performance evaluation

1. INTRODUCTION

Recent advances in artificial intelligence (AI) have been significantly accelerated by the emergence of large language models (LLMs), which are sophisticated systems that use mathematical and statistical methods to understand and generate human-like language (Shanahan, 2024). LLMs learn linguistic patterns and contextual relationships across a wide range of topics by analyzing massive volumes of text data sourced from various literature, academic writing, and online content (Thirunavukarasu et al., 2023). Given an input prompt, an LLM generates coherent and contextually appropriate responses by predicting the most probable next word based on learned probability distributions. These models are built upon the transformer architecture, first introduced by Vaswani et al. in 2017, which replaces traditional recurrence and convolutional layers with a novel attention mechanism known as multi-headed self-attention. This design enables the model to capture dependencies between tokens (i.e., the smallest units of text used by the model to process language) regardless of their relative distance within the input or output sequence (Vaswani et al., 2017). These foundational advancements have led to the development of highly influential LLMs, such as GPT (Generative Pre-trained Transformer) (Floridi & Chiriatti, 2020; Radford, Narasimhan, Salimans, & Sutskever, 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee, & Toutanova, 2019). Along with their subsequent iterations, LLMs have demonstrated exceptional performance of natural language processing (NLP) tasks, including summarization, translation, question answering, and logical reasoning.

The launch of ChatGPT – a conversational chatbot powered by the GPT-3.5 model – introduced to the general public in November 2022 and marked a major paradigm shift in the accessibility and application of generative AI: while it took Facebook 10 months to reach one million users, ChatGPT achieved this milestone in just five days. In addition, ChatGPT's unexpected success in passing the United States Medical Licensing

Examination (USMLE) without any domain-specific training captured global attention in January 2023 (Kung et al., 2023). In February 2023, OpenAI released ChatGPT Plus based on GPT-4 (Achiam et al., 2023), which provide improved response speed and enhanced performance through more extensive and up-to-date training data. In line with this trend, Microsoft introduced Bing Chat in February 2023, leveraging Prometheus model, built upon GPT-4 and integrating real-time web search capabilities. Google followed in March 2023 with Bard, initially based on its Pathways Language Model (PaLM) (Chowdhery et al., 2023) and the Language Model for Dialogue Applications (LaMDA) architecture (Thoppilan et al., 2022).

Large multimodal models (LMMs) are designed to process and integrate information from multiple modalities – including text, images, audio, and video – within a unified framework (Huang, Yan, Li, & Peng, 2024). They are the next evolutionary step of LLM, which only works with text input and output. This shift reflects an ambition to create AI systems that perceive and reason more like humans by simultaneously analyzing linguistic and visual cues. LMMs retain transformer-based architecture of their predecessors but are further trained on large-scale image-text pairs or multimodal datasets (Li et al., 2024; Qi et al., 2020). Notable milestones include Google's Gemini series (Team et al., 2023), which began incorporating LMM capabilities with the release of Gemini 1.0 in December 2023. Another key advancement was OpenAI's GPT-4o (with "o" standing for "omni") launched in May 2024, which introduced a truly unified multimodal model capable of natively processing text, images, video, and audio within a single architecture (Islam & Moushi, 2024). Furthermore, these LMMs adopt mixture-of-experts (MoE) models – an architecture that has been widely applied in tasks such as classification, clustering, and regression (Nguyen & Chamroukhi, 2018). MoE selectively activates only the most relevant expert subnetworks per each query, significantly reducing computational overhead while maintaining high-quality responses in complex multimodal tasks (Kim, Lee, & Kim, 2024).

Response consistency is an important factor in evaluating chatbot performance. LLMs and LMMs are inherently probabilistic in nature. Rather than applying fixed rules, these models generate responses by sampling from probability distributions over possible tokens (Kim et al., 2024). As a result, identical prompts submitted at different times can yield varying outputs, even within the same version of model. In practice, medical researchers often submit the same prompts multiple times – typically three to five iterations – to assess the consistency of model outputs (Kuşcu, Pamuk, Sütay Süslü, & Hosal, 2023; Wu et al., 2024).

Hallucination refers to the generation of information that appears contextually appropriate but is factually incorrect. Such inaccuracies may stem from imbalanced or incomplete training data, restricted access to up-to-date information, or intrinsic limitations in generating responses that are both logically accurate and contextually appropriate (Rawte, Sheth, & Das, 2023). Especially, recent study reported that LMMs may generate severe ungrounded or inaccurate outputs that are not properly aligned with the provided visual context. This issue is often attributed to the imbalance in the amount and quality of multimodal training data compared to text data (Sun et al., 2023). The problem becomes more prominent in tasks that require precise integration of image and text, such as complex image reasoning or medical image interpretation.

Researchers have explored prompt engineering as a practical method to enhance the performance of LLMs and LMMs. This approach guides the model toward producing more accurate and reliable responses – for example, by explicitly specifying the model's assumed role (e.g., "Answer as if you were an oral and maxillofacial radiologist with 20 years of experience") or by providing clear and detailed instructions. Several structured prompt engineering strategies have also been developed to enhance model reasoning. A common classification includes zero-shot, one-shot, and few-shot prompting, depending on how many examples are provided in the prompt to guide the model's understanding of the expected context and response structure. Generally, the more examples provided, the better the model's performance (Brown et al., 2020). Another widely studied method is the chain-of-thought (CoT) prompting strategy, which encourages the model to reason step by step through explicit instructions that prompt it to articulate its thought process.

Zero-shot chain-of-thought (ZS-CoT), proposed by Kojima et al. in 2022, is a prompt engineering technique that improves the reasoning of chatbots. (Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022). Simply prepending the phrase "Let's think step by step" to a prompt has been shown to dramatically enhance performance in reasoning tasks. For example, in arithmetic problems, model accuracy improved from 17.7% to 78.7% using ZS-CoT without the need for any examples. Although ZS-CoT did not outperform few-shot CoT in the previous study (Kojima et al., 2022), it offers a compelling balance between effectiveness and simplicity. Given that few-shot CoT often requires careful task alignment and manually designed examples, ZS-CoT remains a highly efficient and practical strategy for improving model reasoning with minimal prompt engineering. However, its application in dentistry remains relatively underexplored, highlighting the need for further investigation in domain-specific contexts.

LLMs and LMMs are now widely accessible to the public through general-purpose, web-based chatbot platforms. As their educational and clinical applications continue to expand, there is growing interest within the dental community regarding their potential utility. In dental education, chatbots can function as interactive platforms for clinical practice simulation, knowledge reinforcement, and personalized competency assessment. By enabling continuous access to educational content and delivering immediate, adaptive feedback, they contribute to enhancing students' learning processes (Fang et al., 2024). Clinically, chatbots hold potential to improve patient education by offering tailored health information, facilitating remote consultations, and supporting multilingual communication, thereby promoting better patient comprehension and engagement in care (Helvacioglu-Yigit et al., 2025). Oral and maxillofacial radiology (OMFR), which involves both text-based and image-based tasks, offers opportunities for chatbots to support experts by addressing language-related challenges, assisting in the generation and standardization of radiology reports, and aiding in the interpretation of dental radiographic images for the diagnosis of various head and neck conditions (Kim et al., 2024). To ensure their safe and effective integration into these domains, systemic evaluation of their current performance is essential.

Several recent studies have attempted to evaluate the performance of publicly available LLM and/or LMM chatbots using OMFR examination questions (Mine et al., 2025; Tassoker, 2025; Uehara et al., 2025). However, studies were limited by small datasets with fewer than 50 questions, focused exclusively on multiple-choice question formats, or primarily utilized text-only input. In the study by Mine et al., image-based questions were included but were limited to only six items (Mine et al., 2025). In addition, previous studies primarily assessed chatbot accuracy or reliability without conduction expert-based hallucination evaluation or applying prompt engineering strategies for performance enhancement (Mine et al., 2025; Tassoker, 2025; Uehara et al., 2025).

This study aimed to conduct a comprehensive evaluation of general-purpose LLM and LMM chatbots across four key dimensions – accuracy, response consistency, hallucination, and the effect of ZS-CoT – using OMFR questions that incorporate both text and image-based items. It offers timely insight into both the current capabilities and critical limitations of chatbots in OMFR, and contributes a novel methodological perspective for future research and practical applications in complex clinical fields. It also provides a longitudinal assessment of the evolving capabilities of AI chatbots in domain-specific fields by presenting performance changes from LLM to LMM chatbots.

2. MATERIALS AND METHODS

The overall workflow is presented in Fig. 1.

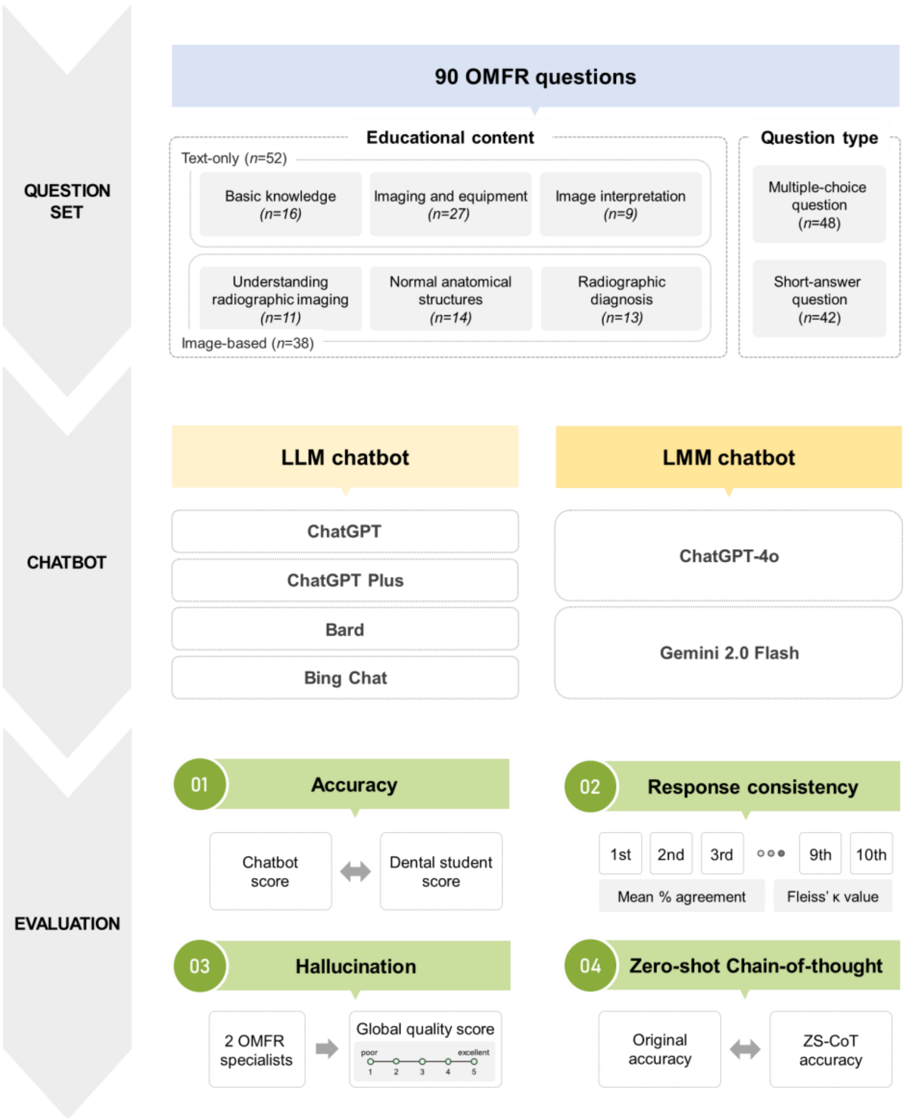


Fig. 1 Overall workflow of this study.

2.1. Question preparation and categorization

The study utilized a total of 90 examination questions from the oral and maxillofacial radiology (OMFR) curriculum at Yonsei University College of Dentistry. The questions were selected from mid- and end-of-semester examinations administered in April and June 2023 and were developed by experienced oral and maxillofacial radiologists. All questions were originally written in Korean, the native language of the students. Essay-format questions that lacked objective scoring criteria were excluded from the question set.

The questions were categorized into two groups and further subdivided by specific educational content as follows:

i. **Text-only questions (n=52)**: These items did not contain any visual elements and consisted entirely of text.
   - Basic knowledge (n=16): Understanding of X-ray generation and measurement units, radiation biology, exposure and protection principles.
   - Imaging and equipment (n=27): Understanding of panoramic radiography, periapical radiography, cone-beam computed tomography (CBCT), magnetic resonance imaging (MRI), and digital imaging systems.
   - Image interpretation (n=9): Interpretation of radiographic features associated with cysts, trauma, fractures, soft tissue calcifications, and systemic diseases involving the oral and maxillofacial region.

ii. **Image-based questions (n=38)**: These items included visual elements such as dental radiographic images, illustrations, schematic diagrams, or graphs.
   - Understanding radiographic imaging (n=11): Comprehension of digital image characteristics (e.g., bit depth, grayscale levels, window width and window center), identification of imaging artifacts or acquisition errors, and schematic understanding of X-ray power supply and generation systems.
   - Normal anatomical structures (n=14): Identification of normal hard and soft

tissue anatomical structures as seen on panoramic and periapical radiography, computed tomography (CT), and magnetic resonance imaging (MRI).

- Radiographic diagnosis (n=13): Recognition of radiographic manifestations of cysts, tumors, fractures, soft tissue calcifications, other bone diseases, and systemic conditions on various dental radiographs.

The questions were also classified by question format:

i. **Multiple-choice questions (MCQs; n=48)**: These items required to select only one correct answer among multiple options.

ii. **Short-answer questions (SAQs; n=42)**: These items required a clear and concise response that involved no inference or subjective judgment.

2.2. Educational background of dental students

A total of 120 dental students – 58 in their third-year and 62 in their fifth-year – were enrolled in the examinations, which formed an integral component of the structured dental curriculum. These students served as the reference group for evaluating the performance of LLM and LMM chatbots on identical test items. Each set of examinations was independently designed by the oral and maxillofacial radiologists leading the respective course. As the assessments were conducted as part of routine academic instruction, individual consent from the students or institutional review board (IRB) approval was not required.

Third-year students received 32 hours of conventional classroom instruction, delivered by a radiologist with 29 years of clinical and teaching experience. They were assessed on basic knowledge, imaging and equipment, understanding radiographic imaging, and normal anatomical structures. Fifth-year students underwent 16 hours of instruction provided by a radiologist with 26 years of experience. Their curriculum focused on advanced diagnostic skills, and their examination primarily covered the domains of image interpretation and radiographic diagnosis.

2.3. Model descriptions and input strategies

2.3.1. Large language model chatbots

Four text-based large language model (LLM) chatbots – ChatGPT, ChatGPT Plus, Bard, and Bing Chat – were evaluated using 52 text-only questions, as these models did not support image input at the time. The assessments were conducted between July and September 2023.

ChatGPT and ChatGPT Plus (OpenAI, San Francisco, California, USA) are based on OpenAI's GPT-3.5 and GPT-4 architectures, respectively. ChatGPT did not support real-time internet access, and its responses were generated solely from pre-trained data (cutoff: September 2021). ChatGPT Plus also relied primarily on pre-trained data (cutoff: January 2023), but real-time web access was available through the optional use of the "web browsing" setting.

Bard (Google, Mountain view, California, USA) is built on Google's Pathways Language Model (PaLM) and the Language Model for Dialogue Applications (LaMDA), and it supported real-time web access through integration with Google Search. Bing Chat (Microsoft, Redmond, Washington, USA) is powered by Microsoft's Prometheus model, built upon GPT-4, and also provided real-time internet search capabilities. The training data for Bard was current up to April 2023, and Bing Chat's was up to March 2023 at the time of evaluation.

2.3.2. Large multimodal model chatbots

Two large multimodal models (LMM) chatbots – ChatGPT-4o and Gemini 2.0 Flash – were evaluated using all 90 questions, which included 52 text-only and 38 image-based items. The assessments were conducted between February and March 2025.

ChatGPT-4o (OpenAI, San Francisco, California, USA) is based on the GPT-4 architecture and integrates multimodal capabilities, including text, image, and audio input. While it includes a web browsing feature, ChatGPT-4o does not access real-time information by default; unless the user manually enables the browsing function before submitting a prompt, responses are generated solely from its internal knowledge base, last updated in October 2023.

Gemini 2.0 Flash (Google DeepMind, London, England) is a lightweight variant of the Gemini 2.0 architecture, which combines large language modeling with multimodal pre-training. Unlike ChatGPT-4o, Gemini 2.0 Flash support real-time internet access by default through integration with Google Search. Its training data was current up to December 2023 at the time of evaluation.

2.3.3. Prompt formatting and input strategies

All questions were entered into each chatbot in Korean, as they were originally administered in the student examinations. To maintain consistency and ensure objective evaluation, input queries were standardized across models.

All items were reformatted to require a single response. Detailed instructions such as "Select the most accurate/inaccurate statement" or "Write the most appropriate response" were included, depending on the question type. For SAQs requiring the identification of specific term described in a statement, a blank space was inserted within the sentence to prompt the chatbot to generate a direct and contextually appropriate answer.

In the LLM chatbots evaluation, a single-input strategy was employed (Lin, Chan, Hsu, & Kao, 2024), in which each question was entered only once per model without any rephrasing, repetition, or follow-up attempts. This approach was intended to mirror the conditions under which students completed their examinations, ensuring a fair and realistic comparison between human and model performance.

For LMM chatbots, each question was submitted ten times in independent chat sessions to evaluate response consistency, considering the potential variability of multimodal processing. However, in order to evaluate response accuracy under conditions consistent with the LLM protocol, only the first response of LMM chatbots was used for scoring.

To enhance transparency and methodological rigor, this study referenced the MInimum reporting items for CLear Evaluation of Accuracy Reports of Large Language Models in healthcare (MI-CLEAR-LLM) checklist (Park, Suh, Lee, Kahn Jr, & Moy, 2024). A detailed account of how each checklist item was addressed is provided in Table 1.

Table 1. Checklist of evaluation strategies and input conditions (Park et al., 2024).

Checklist item	Application in this study
Handling of stochasticity	
• Number of query attempts	• One attempt per question for LLM chatbots; ten independent attempts for LMM chatbots.
• Response synthesis method and rationale	• To mirror real student testing conditions, only the first response generated by the model was used for accuracy evaluation.
• Response consistency analysis	• Response consistency was assessed by analyzing the agreement among ten responses for each question.
• Technical parameter settings	• No hyperparameters (e.g., temperature) were modified; default settings were used.
Exact prompt wording and syntax	
• Prompt formatting details	• Prompts maintained consistent and precise use of spelling, symbols, punctuation, and spacing.
Prompt application procedure	
• Query session structure	• Multiple items were entered sequentially within a single chat session per model.
• Query input sequence	• Items were submitted one at a time over multiple chat rounds to isolate responses.

Prompt testing and optimization	
• Prompt development steps	• All items included a directive to select the most appropriate responses; SAQs were further reformatted with blank spaces to elicit concise answers.
• Prompt wording rationale	• Terminology was drawn from standard oral and maxillofacial radiology textbooks to reflect academic appropriateness.
Dataset independence	
• Use of test data in model training or prompt tuning	• None of the items were used in the training or prompt tuning of any chatbot models.
• Data source URL disclosure	• Not applicable.

LLM, large language model; LMM, large multimodal model; SAQ, short-answer question.

2.4. Data analysis

2.4.1. Accuracy

Accuracy was assessed using a single response from each LLM chatbot and the first of ten responses from each LMM chatbot. Accuracy was calculated as the percentage of correct responses out of the total number of questions. Analyses were conducted across three dimensions – overall performance, educational content category, and question type – using student scores as a reference standard for comparison.

To ensure consistency and objectivity in the evaluation of both chatbot-generated and student responses, test items were independently developed by course instructors and thoroughly reviewed through three iterative rounds to minimize potential grading inconsistencies. MCQ responses were classified as either correct of incorrect to eliminate evaluator bias. SAQs were designed to elicit concise, fact-based answers requiring the identification of specific key terms rather than interpretive reasoning, thereby minimizing the need for subjective judgment. Any response that deviated from the reference answer in format or spelling was marked as incorrect.

2.4.2. Response consistency

Response consistency was assessed for LMM chatbots due to the inherent variability of multimodal input processing. Each question was submitted ten times in separate chat sessions, and response set was considered consistent if the responses were identical in content, regardless of correctness. For short-answer responses, a rule-based normalization process was applied prior to comparison to account for minor lexical variations that retained equivalent meaning. Expressions with negligible differences – such as spacing, inflectional suffixes, or commonly interchangeable terminology – were considered equivalent and treated as consistent responses during this assessment.

Response consistency was evaluated by calculating the mean percentage agreement across ten repeated outputs. To quantify agreement beyond chance, Fleiss' kappa coefficient was calculated using SPSS software (version 26.0; IBM Corp., Armonk, NY, USA), with a two-tailed significance level set at 0.05.

2.4.3. Hallucination

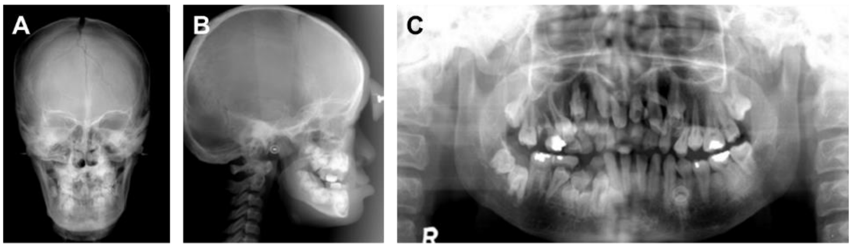
LMM chatbots, which are more susceptible to hallucinations due to their multimodal nature, were evaluated based on the first response out of ten generated for each question, in accordance with the protocol used for accuracy evaluation. Two oral and maxillofacial radiologists who had administered the test questions independently rated the responses using the modified Global Quality Scales (GQS) (Bernard et al., 2007), a five-point scale ranging from 1 (poor) to 5 (excellent) (Table 2). The evaluation considered not only the factual correctness of each answer but also the validity and appropriateness of the underlying reasoning and supporting evidence (Fig. 2). The results were analyzed using descriptive statistics (mean and standard deviation).

Table 2. Modified Global Quality Scales (GQS) (Bernard et al., 2007).

Score	Quality	Description
1	Poor	Disorganized flow of the information, most key information missing, not useful for patients or education.
2	Generally poor	Poorly structured flow of the information, many key information missing, very limited usefulness for patients or education.
3	Moderate	Suboptimal flow of the information, some key information missing or incomplete, limited usefulness for patients or education.
4	Good	Well-organized flow or the information, most key information included, generally useful for patients or education.
5	Excellent	Clear and logical flow of the information, all key information included, highly useful for patients or education.

Question

Write the most appropriate diagnosis based on the following radiographs (A. posteroanterior cephalometric radiograph, B. lateral cephalometric radiograph, C. panoramic radiograph).



Response (answer + supporting evidence)

Most likely diagnosis:

✅ **Rickets**

🔍 **Analysis of radiographic findings and their association with rickets**

1️⃣ **Posteroanterior cephalometric radiograph**

- The skull is abnormally wide and sutures may be prominent.
- Cranial growth abnormalities are possible.

2️⃣ **Lateral cephalometric radiograph**

- Delayed growth of the maxilla and mandible.
- Abnormal tooth alignment, overbite or malocclusion may be observed.
- The mandible may be smaller than normal or deformed.

3️⃣ **Panoramic radiograph**

- Multiple teeth are delayed in eruption and amelogenesis imperfecta is observed.
- Abnormal tooth alignment, overbite or malocclusion may be observed.
- Teeth are not aligned normally, and skeletal growth abnormalities are accompanied.

Fig. 2 An example of a chatbot response. Hallucination was evaluated based on both the chatbots' answer and the validity of its supporting evidence. (The input of the question and the output of the response were both in Korean, but an English translation is provided in this figure for the reader's convenience.)

2.4.4. Zero-shot chain-of-thought

ZS-CoT prompting was applied to evaluate whether stepwise reasoning could enhance model performance. This strategy was implemented for the LMM chatbots across all questions, including both text-only and image-based items. A single sentence – “Let’s think step by step” – was added in Korean at the beginning of each standardized input query, with no further modifications to the original formatting.

Model performance under the ZS-CoT condition was evaluated by calculating accuracy as the percentage of correct responses out of the total number of questions. These results were compared to the model’s original performance without ZS-CoT prompting to assess the effectiveness of the strategy.

3. RESULTS

3.1. Accuracy

Table 3 presents the overall accuracy rates of the six chatbots evaluated in this study. Among the LLM chatbots, ChatGPT Plus achieved the highest accuracy (65.4%), followed by Bing Chat (63.5%). ChatGPT and Bard recorded the lowest accuracy, both at 50.0%. In the LMM group, ChatGPT-4o and Gemini 2.0 Flash achieved accuracy rates of 61.1% and 58.9%, respectively.

In comparison to the dental student reference scores (81.2% for text-only items and 77.7% for text- and image-based items), all chatbot models showed lower accuracy than the students' scores.

Table 3. Accuracy rates and comparison with reference standard (dental student scores).

Chatbot classification	Name	Accuracy	Ref.
LLM	ChatGPT	50.0	81.2
	ChatGPT Plus	65.4	
	Bard	50.0	
	Bing Chat	63.5	
LMM	ChatGPT-4o	61.1	77.7
	Gemini 2.0 Flash	58.9	

LLM, large language model; LMM, large multimodal model; Ref., reference standard (dental student scores).

3.1.1. Accuracy based on educational content

Table 4 presents the accuracy of chatbot responses across different educational content categories. Among the LLM chatbots evaluated on text-only questions, ChatGPT Plus achieved the highest accuracy in basic knowledge (93.8%), surpassing the dental student reference score of 78.7%. In the imaging and equipment category, Bing Chat scored 74.1%, which was higher than other LLM chatbots but still below the student score of 83.5%. For image interpretation, LLM chatbots showed accuracy ranging from 22.2% (ChatGPT Plus) to 33.3% (ChatGPT, Bard, and Bing Chat).

For the LMM chatbots in text-only questions, ChatGPT-4o and Gemini 2.0 Flash achieved accuracy rates of 86.5% and 82.7%, respectively, both exceeding the student score of 81.2%. Gemini 2.0 Flash achieved 100.0% accuracy in the basic knowledge category. Compared to ChatGPT Plus – an earlier model from the same developer – ChatGPT-4o showed an improvement of 25.9 percentage points in imaging and equipment, and 55.6 points in image interpretation (Fig. 3A). Similarly, Gemini 2.0 Flash outperformed its predecessor Bard by 50.0 points in basic knowledge, 18.5 points in imaging and equipment, and 44.5 points in image interpretation, as illustrated in Fig. 3B.

Image-based questions were evaluated using LMM chatbots only. Both ChatGPT-4o and Gemini 2.0 Flash scored 26.3%, below the student reference score of 72.9%. The categories of interpreting normal anatomy (ChatGPT-4o: 14.3%, Gemini 2.0 Flash: 21.4%) and radiographic diagnosis (15.4% for both chatbots) yielded the lowest accuracy rates across all content categories.

Table 4. Accuracy rates based on educational content and comparison with reference standard (dental student scores).

		LLM				LMM		Ref.
		Chat GPT	ChatGPT Plus	Bard	Bing Chat	ChatGPT-4o	Gemini 2.0 Flash	
Text-only (n=52)	Basic knowledge (n=16)	31.3	93.8	50.0	68.8	87.5	100.0	78.7
	Imaging and equipment (n=27)	66.7	63.0	55.6	74.1	88.9	74.1	83.5
	Image interpretation (n=9)	33.3	22.2	33.3	33.3	77.8	77.8	78.3
	TOTAL (n=52)	**50.0**	**65.4**	**50.0**	**63.5**	**86.5**	**82.7**	**81.2**
Image-based (n=38)	Understanding radiographic imaging (n=11)	-	-	-	-	54.5	45.5	75.4
	Interpreting normal anatomy (n=14)	-	-	-	-	14.3	21.4	64.8
	Radiographic diagnosis (n=13)	-	-	-	-	15.4	15.4	79.5
	TOTAL (n=38)	**-**	**-**	**-**	**-**	**26.3**	**26.3**	**72.9**

LLM, large language model; LMM, large multimodal model; Ref., reference standard (dental student scores).

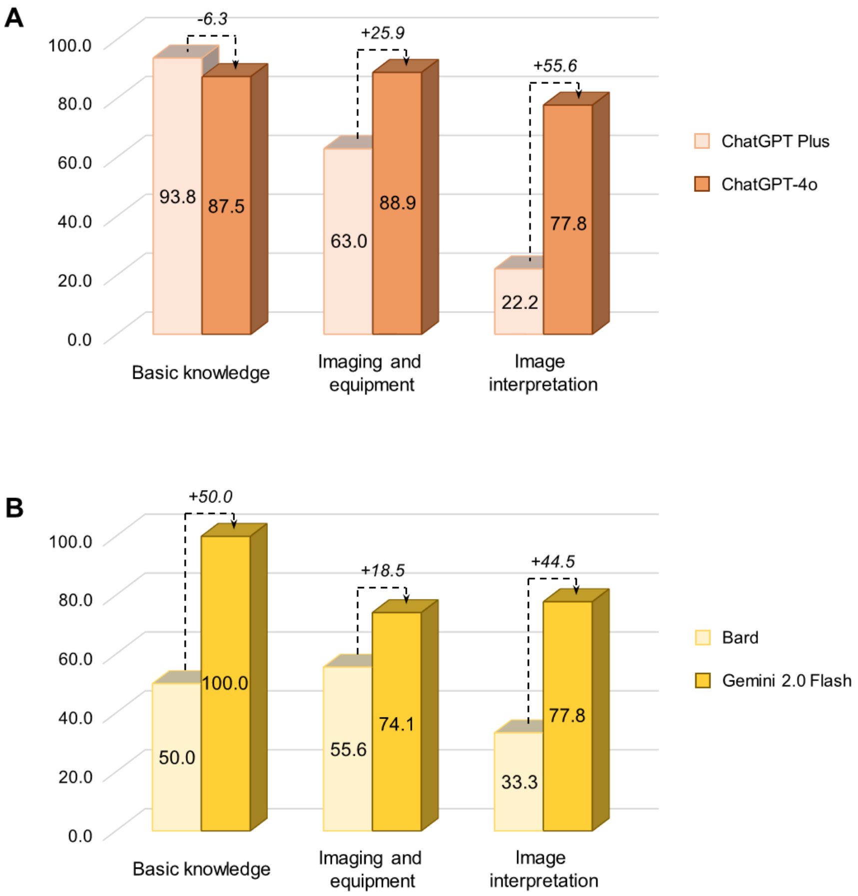


Fig. 3 Accuracy comparison between LLM and LMM chatbots across educational content categories in text-only questions. (A) Performance comparison between ChatGPT Plus (LLM) and ChatGPT-4o (LMM), (B) Performance comparison between Bard (LLM) and Gemini 2.0 Flash (LMM).

3.1.2. Accuracy based on question type

Table 5 summarizes chatbot performance based on question type. For text-only questions, all LLM chatbots showed higher accuracy on SAQs than on MCQs. ChatGPT and ChatGPT Plus each achieved 85.7% accuracy on SAQs, while their MCQ scores were lower at 36.8% and 57.9%, respectively. Bard showed the smallest performance gap between the two question types, with 47.4% accuracy on MCQs and 57.1% on SAQs – a difference of 9.7 percentage points.

Among the LMM chatbots, ChatGPT-4o achieved an accuracy of 86.8% on MCQs and 85.7% on SAQs, representing a 28.9 percentage point increase in MCQ performance compared to ChatGPT Plus (Fig. 4A). Gemini 2.0 Flash scored 78.9% on MCQs and 92.9% on SAQs, showing the same pattern observed in LLM chatbots, where SAQ performance exceeded that of MCQs. Compared to Bard, Gemini 2.0 Flash showed improvements of 31.5 points on MCQs and 35.8 points on SAQs (Fig. 4B).

In image-based questions, both ChatGPT-4o and Gemini 2.0 Flash recorded higher accuracy on MCQs (60.0% and 50.0%, respectively) than on SAQs (14.3% and 17.9%, respectively). Neither model exceeded the student reference scores, which were 76.7% for MCQs and 71.5% for SAQs.

Table 5. Accuracy rates based on question type and comparison with reference standard (dental student scores).

		LLM				LMM		Ref.
		ChatGPT	ChatGPT Plus	Bard	Bing Chat	ChatGPT-4o	Gemini 2.0 Flash	
Text-only (n=52)	Multiple-choice question (n=38)	36.8	57.9	47.4	57.9	86.8	78.9	80.5
	Short-answer question (n=14)	85.7	85.7	57.1	78.6	85.7	92.9	82.9
	TOTAL (n=52)	**50.0**	**65.4**	**50.0**	**63.5**	**86.5**	**82.7**	**81.2**
Image-based (n=38)	Multiple-choice question (n=10)	-	-	-	-	60.0	50.0	76.7
	Short-answer question (n=28)	-	-	-	-	14.3	17.9	71.5
	TOTAL (n=38)	**-**	**-**	**-**	**-**	**26.3**	**26.3**	**72.9**

LLM, large language model; LMM, large multimodal model; Ref., reference standard (dental student scores).

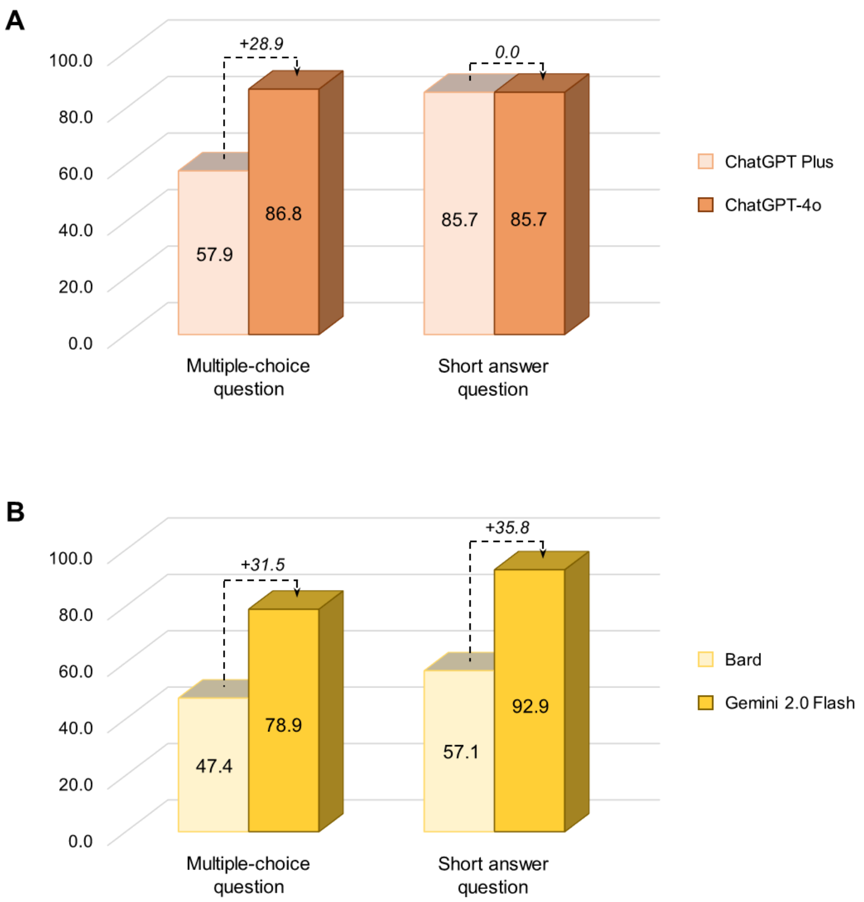


Fig. 4 Accuracy comparison between LLM and LMM chatbots across question type in text-only questions. (A) Performance comparison between ChatGPT Plus (LLM) and ChatGPT-4o (LMM), (B) Performance comparison between Bard (LLM) and Gemini 2.0 Flash (LMM).

3.2. Response consistency

Response consistency was evaluated by comparing the consistency of ten independently generated responses for each question, regardless of correctness. Agreement was evaluated using both percentage agreement and Fleiss' kappa coefficient. According to established interpretation guidelines (Landis & Koch, 1977), a kappa value between 0.00-0.20 indicates "Slight" agreement, 0.21-0.40 indicates "Fair" agreement, 0.41-0.60 indicates "Moderate" agreement, 0.61-0.80 indicates "Substantial" agreement, and values above 0.80 represent "Almost perfect" agreement.

Overall results of the response consistency assessment are shown in Table 6. ChatGPT-4o demonstrated a mean percentage agreement of 80.9% across ten responses per question, and Gemini 2.0 Flash showed 81.4% of agreement. Fleiss' kappa values were 0.709 and 0.722 for ChatGPT-4o and Gemini 2.0 Flash, respectively, both interpreted as indicating "Substantial" agreement.

Table 6. Response consistency evaluation and interpretation.

	Mean % agree	κ value	Interp.
ChatGPT-4o	80.9	0.709	Substantial
Gemini 2.0 Flash	81.4	0.722	Substantial

Mean % agree, average percentage of the most frequent response among ten repeated outputs; κ value, Fleiss' kappa coefficient; Interp., interpretation of κ value based on standard agreement levels.

3.2.1. Response consistency based on educational content

Table 7 presents the response consistency of the LMM chatbots across different educational content categories, based on ten repeated outputs per question. For text-only questions, ChatGPT-4o achieved mean agreement rates of 95.6% for basic knowledge, 97.0% for imaging and equipment, and 94.4% for image interpretation, with corresponding Fleiss' kappa values ranging from 0.863 to 0.948, all interpreted as "Almost perfect." Gemini 2.0 Flash showed mean agreement rates of 96.3%, 97.8%, and 95.6%, and κ values between 0.900 and 0.962, also categorized as "Almost perfect." The overall consistency for text-only questions was 96.2% (κ = 0.926) for ChatGPT-4o and 96.9% (κ = 0.945) for Gemini 2.0 Flash.

For image-based questions, ChatGPT-4o demonstrated mean agreement ranged from 49.2% to 80.0%, and κ values from 0.267 to 0.617. Gemini 2.0 Flash showed agreement rates between 48.5% and 85.5%, and κ values ranging from 0.271 to 0.719. The highest consistency was observed in the category of understanding radiographic imaging, where both chatbots achieved "Substantial" agreement. However, in interpreting normal anatomy and radiographic diagnosis, agreement levels dropped to the "Fair" range. The total consistency for image-based questions was 60.0% (κ = 0.427) for ChatGPT-4o and 60.3% (κ = 0.433) for Gemini 2.0 Flash, both interpreted as "Moderate."

Table 7. Response consistency evaluation and interpretation based on educational content.

		ChatGPT-4o			Gemini 2.0 Flash		
		Mean % agree	**κ value**	**Interp.**	**Mean % agree**	**κ value**	**Interp.**
Text -only (n=52)	Basic knowledge (n=16)	95.6	0.910	Almost perfect	96.3	0.928	Almost perfect
	Imaging and equipment (n=27)	97.0	0.948	Almost perfect	97.8	0.962	Almost perfect
	Image interpretation (n=9)	94.4	0.863	Almost perfect	95.6	0.900	Almost perfect
	TOTAL (n=52)	**96.2**	**0.926**	**Almost perfect**	**96.9**	**0.945**	**Almost perfect**
Image -based (n=38)	Understanding radiographic imaging (n=11)	80.0	0.617	Substantial	85.5	0.719	Substantial
	Interpreting normal anatomy (n=14)	54.3	0.338	Fair	51.4	0.304	Fair
	Radiographic diagnosis (n=13)	49.2	0.267	Fair	48.5	0.271	Fair
	TOTAL (n=38)	**60.0**	**0.427**	**Moderate**	**60.3**	**0.433**	**Moderate**

Mean % agree, average percentage of the most frequent response among ten repeated outputs; κ value, Fleiss' kappa coefficient; Interp., interpretation of κ value based on standard agreement levels.

3.2.2. Response consistency based on question type

Table 8 summarizes the response consistency of LMM chatbots for each question type, based on ten repeated outputs per item. For text-only questions, ChatGPT-4o showed a mean agreement of 96.3% for MCQs and 95.7% for SAQs, with Fleiss' kappa values of 0.923 and 0.916, respectively. These values fall within the "Almost perfect" category of agreement. Gemini 2.0 Flash demonstrated a mean agreement rates of 97.1% ($\kappa = 0.939$) for MCQs and 96.4% ($\kappa = 0.944$) for SAQs, also interpreted as "Almost perfect."

For image-based questions, ChatGPT-4o achieved a mean agreement of 82.0% ($\kappa = 0.635$) for MCQs, corresponding to "Substantial" agreement, and 52.1% ($\kappa = 0.324$) for SAQs, interpreted as "Fair." Gemini 2.0 Flash showed 51.1% agreement ($\kappa = 0.312$) for image-based SAQs, categorized as "Fair."

Table 8. Response consistency evaluation and interpretation based on question type.

		ChatGPT-4o			Gemini 2.0 Flash		
		Mean % agree	**κ value**	**Interp.**	**Mean % agree**	**κ value**	**Interp.**
Text -only (n=52)	Multiple-choice question (n=38)	96.3	0.923	Almost perfect	97.1	0.939	Almost perfect
	Short-answer question (n=14)	95.7	0.916	Almost perfect	96.4	0.944	Almost perfect
	TOTAL (n=52)	**96.2**	**0.926**	**Almost perfect**	**96.9**	**0.945**	**Almost perfect**
Image -based (n=38)	Multiple-choice question (n=10)	82.0	0.635	Substantial	86.0	0.722	Substantial
	Short-answer question (n=28)	52.1	0.324	Fair	51.1	0.312	Fair
	TOTAL (n=38)	**60.0**	**0.427**	**Moderate**	**60.3**	**0.433**	**Moderate**

Mean % agree, average percentage of the most frequent response among ten repeated outputs; κ value, Fleiss' kappa coefficient; Interp., interpretation of κ value based on standard agreement levels.

3.3. Hallucination

Table 9 shows the hallucination evaluation results of the two LMM chatbots, by two oral and maxillofacial radiologists, based on the modified Global Quality Scales (GQS) (Bernard et al., 2007). To facilitate standardized interpretation of the GQS score, the mean scores were categorized into five levels of quality as follows: scores between 0.00 and 1.00 were interpreted as “Poor,” 1.01 to 2.00 as “Generally poor,” 2.01 to 3.00 as “Moderate,” 3.01 to 4.00 as “Good,” and 4.01 to 5.00 as “Excellent.”

The mean GQS score was 3.37 (SD = 1.77) for ChatGPT-4o and 3.41 (SD = 1.79) for Gemini 2.0 Flash. According to the predefined interpretation criteria, both chatbots were classified as “Good.”

Table 9. Hallucination evaluation and interpretation.

	M ± SD	**Interp.**
ChatGPT-4o	3.37 ± 1.77	Good
Gemini 2.0 Flash	3.41 ± 1.79	Good

M, mean; SD, standard deviation; Interp., interpretation of mean Global Quality Score values based on predefined score ranges.

3.3.1. Hallucination evaluation based on educational content

Table 10 presents the hallucination evaluation results by educational content categories. For text-only questions, ChatGPT-4o showed mean GQS score of 4.41 (SD = 1.16) in basic knowledge, 4.41 (SD = 1.10) in imaging and equipment, and 3.56 (SD = 1.81) in image interpretation. Gemini 2.0 Flash recorded 4.50 (SD = 1.02), 4.22 (SD = 1.27), and 4.11 (SD = 1.76) in the respective categories. According to the predefined interpretation criteria, both chatbots were classified as “Excellent” in basic knowledge and imaging and equipment. In image interpretation, ChatGPT-4o was classified as “Good,” and Gemini 2.0 Flash was classified as “Excellent.”

For image-based questions, mean GQS score for ChatGPT-4o were 3.55 (SD = 1.77) in understanding radiographic imaging, 1.82 (SD = 1.49) in interpreting normal anatomy, and 1.31 (SD = 0.75) in radiographic diagnosis. Gemini 2.0 Flash showed corresponding scores of 3.55 (SD = 1.81), 1.61 (SD = 1.36), and 1.69 (SD = 1.30). Both chatbots were classified as “Good” for understanding radiographic imaging and as “Generally poor” for both interpreting normal anatomy and radiographic diagnosis.

Table 10. Hallucination evaluation and interpretation based on educational content.

		ChatGPT-4o		Gemini 2.0 Flash	
		M ± SD	**Interp.**	**M ± SD**	**Interp.**
Text -only (n=52)	Basic knowledge (n=16)	4.41 ± 1.16	Excellent	4.50 ± 1.02	Excellent
	Imaging and equipment (n=27)	4.41 ± 1.10	Excellent	4.22 ± 1.27	Excellent
	Image interpretation (n=9)	3.56 ± 1.81	Good	4.11 ± 1.76	Excellent
	TOTAL (n=52)	**4.26 ± 1.28**	**Excellent**	**4.29 ± 1.28**	**Excellent**
Image -based (n=38)	Understanding radiographic imaging (n=11)	3.55 ± 1.77	Good	3.55 ± 1.81	Good
	Interpreting normal anatomy (n=14)	1.82 ± 1.49	Generally poor	1.61 ± 1.36	Generally poor
	Radiographic diagnosis (n=13)	1.31 ± 0.75	Generally poor	1.69 ± 1.30	Generally poor
	TOTAL (n=38)	**2.14 ± 1.64**	**Moderate**	**2.20 ± 1.69**	**Moderate**

M, mean; SD, standard deviation; Interp., interpretation of mean Global Quality Score values based on predefined score ranges.

3.3.2. Hallucination evaluation based on question type

Table 11 summarizes the hallucination evaluation results by question type. For text-only MCQs, ChatGPT-4o recorded a mean score of 4.24 (SD = 1.27) and Gemini 2.0 Flash recorded 4.13 (SD = 1.33). In text-only SAQs, the mean GQS score was 4.32 (SD = 1.34) for ChatGPT-4o and 4.71 (SD = 1.07) for Gemini 2.0 Flash. All of these scores were classified as “Excellent” based on the predefined interpretation criteria.

For image-based MCQs, both ChatGPT-4o and Gemini 2.0 Flash recorded identical mean scores of 3.60, with standard deviations of 1.85 and 1.90, respectively. These were classified as “Good.” In image-based SAQs, mean GQS score were 1.63 (SD = 1.21) for ChatGPT-4o and 1.70 (SD = 1.31) for Gemini 2.0 Flash, corresponding to “Generally poor.”

Table 11. Hallucination evaluation and interpretation based on question type.

		ChatGPT-4o		Gemini 2.0 Flash	
		M ± SD	**Interp.**	**M ± SD**	**Interp.**
Text -only (n=52)	Multiple-choice question (n=38)	4.24 ± 1.27	Excellent	4.13 ± 1.33	Excellent
	Short-answer question (n=14)	4.32 ± 1.34	Excellent	4.71 ± 1.07	Excellent
	TOTAL (n=52)	**4.26 ± 1.28**	**Excellent**	**4.29 ± 1.28**	**Excellent**
Image -based (n=38)	Multiple-choice question (n=10)	3.60 ± 1.85	Good	3.60 ± 1.90	Good
	Short-answer question (n=28)	1.63 ± 1.21	Generally poor	1.70 ± 1.31	Generally poor
	TOTAL (n=38)	**2.14 ± 1.64**	**Moderate**	**2.20 ± 1.69**	**Moderate**

M, mean; SD, standard deviation; Interp., interpretation of mean Global Quality Score values based on predefined score ranges.

3.4. Zero-shot chain-of-thought

Table 12 presents the change in accuracy for the two LMM chatbots when ZS-CoT prompting was applied. For ChatGPT-4o, accuracy decreased from 61.1% to 52.2%, showing a performance decline of 8.9 percentage points under the ZS-CoT condition, and Gemini 2.0 Flash showed 1.1 percentage points improvement, with accuracy increasing from 58.9% to 60.0%. Both models underperformed compared to the reference standard score of 77.7%.

Table 12. Accuracy changes with zero-shot chain-of-thought and comparison with reference standard (dental student scores).

	Original	ZS-CoT	Ref.
ChatGPT-4o	61.1	52.2	77.7
Gemini 2.0 Flash	58.9	60.0	

Ref., reference standard (dental student scores); ZS-CoT, zero-shot chain-of-thought; Diff., difference in accuracy between original and zero-shot chain-of-thought conditions.

3.4.1. Changes in accuracy rates after applying zero-shot chain-of-thought based on educational content

Table 13 shows the accuracy of LMM chatbots under original and ZS-CoT prompting conditions, categorized by educational content. For ChatGPT-4o, no improvement observed in any content category. The largest decline occurred in image interpretation, where accuracy dropped from 77.8% to 33.3% (-44.5 points). Basic knowledge remained unchanged (87.5%), imaging and equipment decreased by 3.7 points (from 88.9% to 85.2%), and all image-based subcategories also showed declines ranging from -7.2 to -9.0 points.

For Gemini 2.0 Flash, the largest improvement was observed in radiographic diagnosis, where accuracy increased from 15.4% to 38.5% (+23.1 points). In other categories, basic knowledge decreased by 6.2 points (from 100.0% to 93.8%), and image interpretation decreased by 11.1 points (from 77.8% to 66.7%). No change was observed in imaging and equipment (74.1%), understanding radiographic imaging (45.5%), and interpreting normal anatomy (21.4%).

Table 13. Accuracy changes with zero-shot chain-of-thought based on educational content, and comparison with reference standard (dental student scores).

		ChatGPT-4o		Gemini 2.0 Flash		Ref.
		Original	ZS-CoT	Original	ZS-CoT	
Text -only (n=52)	Basic knowledge (n=16)	87.5	87.5	100.0	93.8	78.7
	Imaging and equipment (n=27)	88.9	85.2	74.1	74.1	83.5
	Image interpretation (n=9)	77.8	33.3	77.8	66.7	78.3
	TOTAL (n=52)	**86.5**	**76.9**	**82.7**	**78.8**	**81.2**
Image -based (n=38)	Understanding radiographic imaging (n=11)	54.5	45.5	45.5	45.5	75.4
	Interpreting normal anatomy (n=14)	14.3	7.1	21.4	21.4	64.8
	Radiographic diagnosis (n=13)	15.4	7.7	15.4	38.5	79.5
	TOTAL (n=38)	**26.3**	**18.4**	**26.3**	**34.2**	**72.9**

ZS-CoT, zero-shot chain-of-thought; Ref., reference standard (dental student scores).

3.4.2. Changes in accuracy rates after applying zero-shot chain-of-thought based on question type

Table 14 presents the accuracy of LMM chatbots under original and ZS-CoT prompting conditions, categorized by question type. For ChatGPT-4o, the largest decrease was observed in MCQs for text-only questions, with accuracy declining from 86.8% to 71.1% (-15.7 points). SAQs in the same category increased from 85.7% to 92.9% (+7.2 points), surpassing the student score of 82.9%. For image-based questions, both MCQs and SAQs declined – from 50.0% to 40.0% (-10.0 points) and from 14.3% to 7.1% (-7.2 points), respectively.

For Gemini 2.0 Flash, MCQ accuracy declined from 78.9% to 73.7% (-5.2 points), and SAQ performance unchanged at 92.9%, exceeding the student score of 82.9% in text-only questions. For image-based questions, MCQ accuracy remained at 50.0%. The accuracy of SAQ increased from 17.9% to 28.6% (+10.7 points), but still fell short of the student reference score of 71.5%.

Table 14. Accuracy changes with zero-shot chain-of-thought based on question type, and comparison with reference standard (dental student scores).

		ChatGPT-4o		Gemini 2.0 Flash		Ref.
		Original	ZS-CoT	Original	ZS-CoT	
Text -only (n=52)	Multiple-choice question (n=38)	86.8	71.1	78.9	73.7	80.5
	Short-answer question (n=14)	85.7	92.9	92.9	92.9	82.9
	TOTAL (n=52)	**86.5**	**76.9**	**82.7**	**78.8**	**81.2**
Image -based (n=38)	Multiple-choice question (n=10)	50.0	40.0	50.0	50.0	76.7
	Short-answer question (n=28)	14.3	7.1	17.9	28.6	71.5
	TOTAL (n=38)	**23.7**	**15.8**	**26.3**	**34.2**	**72.9**

ZS-CoT, zero-shot chain-of-thought; Ref., reference standard (dental student scores).

4. DISCUSSION

LLMs are transformer-based AI systems trained on extensive textual datasets to generate coherent and contextually relevant responses. Extending this architecture, LMMs can simultaneously process and integrate information from diverse sources, including images, audio, and video. Chatbots based on these models are now widely accessible and have garnered increasing interest for their potential applications in dental education and clinical decision support. However, the performance of LLM and LMM chatbots has not been fully studied in specialized fields such as OMFR, where both factual knowledge and image interpretation are essential. To address this gap, this study evaluated the performance of these models using standardized examination items and provided insights into their current utility and limitations in the OMFR field.

This study evaluated the performance of four LLM chatbots (ChatGPT, ChatGPT Plus, Bard, Bing Chat) and two LMM chatbots (ChatGPT-4o, Gemini 2.0 Flash) using 90 examination questions from the OMFR curriculum, comprising 52 text-only and 38 image-based items. All questions were entered into the chatbots in Korean and were slightly reformatted to ensure a consistent prompt structure. For further analysis, the items were grouped into six educational content categories and two question types. As no official passing threshold was defined for this examination, the performance of dental students who had previously completed the same test was used as the reference standard.

Across all six chatbots evaluated in this study, none outperformed the dental student reference score. ChatGPT Plus demonstrated the highest overall accuracy among the LLM chatbots at 65.4%, showing a 15.8 percentage point gap compared to student performance. Although this trend is consistent with prior studies (Ali et al., 2023; Danesh, Pazouki, Danesh, Danesh, & Vardar-Sengul, 2024; Ohta & Ohta, 2023; Toyama et al., 2024) the performance of ChatGPT Plus in this study did not reach the highest accuracy levels reported for the top-performing LLMs in other medical and dental domains – such as 82.6% in neurosurgery (Ali et al., 2023), 87.11% in radiology (Patil, Huang, van der Pol, & Larocque, 2024), 70.8% in respiratory medicine (Rahsepar et al., 2023)), and 73.6% in periodontology (Danesh et al., 2024). ChatGPT-4o and Gemini 2.0 Flash also failed to achieve the dental student score. This result aligns with findings from a recent evaluation using the 2024 Japanese National Dental Examination (Mine et al., 2025), in which ChatGPT-4o scored 64.3% and Gemini 2.0 Flash scored 57.1% in the OMFR domain – both performances comparable to those observed in the present study.

In text-based questions, LMM chatbots – ChatGPT-4o and Gemini 2.0 Flash – demonstrated clear improvements over their LLM-based predecessors and outperformed the dental student reference score of 81.2%. Notably, the most pronounced gains were observed in the image interpretation category, where the LLM chatbots had previously shown poor performance (22.2% for ChatGPT Plus and 33.3% for Bard); in contrast, both LMMs achieved 77.8% accuracy in the same category. These results are consistent with findings from a recent study by Tassoker (2025), which evaluated chatbot performance on 123 multiple-choice questions in the OMFR domain (Tassoker, 2025). In that study, the LMM-based ChatGPT-4o achieved the highest accuracy (86.1%), followed by the LLM chatbots Bard (61.8%), ChatGPT (43.9%), and Microsoft's Copilot (41.5%) – a later version of Bing Chat. The superior performance of LMMs on text-based tasks may not be solely attributed to model size or recentness. Rather, their multimodal training process likely contributes to more comprehensive understanding and adaptive reasoning by enabling the development of broader and more flexible conceptual frameworks – an advantage in acquiring diverse domain-specific knowledge in highly specialized fields such as OMFR.

Despite notable advances in text-based reasoning, both LMM chatbots exhibited substantial limitations in image-based interpretation. This finding indicates considerable difficulty in recognizing key visual patterns essential for accurate radiographic assessment. Particularly low accuracy was observed in the identification of normal anatomical structures. For example, in a panoramic radiograph where the ear lobe was to be identified (Fig. 5) – a task correctly answered by 93.1% of students – ChatGPT-4o misidentified it as the styloid process, while Gemini 2.0 Flash answered the mandibular condyle. In another item requiring localization of a missing floor of the maxillary sinus (Fig. 6), students achieved a high accuracy of 87.1%, while ChatGPT-4o and Gemini 2.0 Flash incorrectly responded with “mandibular canal” and “mandibular condyle,” respectively. One possible explanation lies in the nature of the training data used for LMMs. While these models are pre-trained on large-scale image-text pairs (Li et al., 2024; Qi et al., 2020), much of the data emphasizes general visual understanding rather than the subtle identification of normal anatomical landmarks. Furthermore, many medical or dental image datasets used during pre-training are heavily weighted toward pathological cases, potentially biasing model attention away from normal structures.

The diagnostic performance on radiographic images was also limited. In a case that presented multiple diagnostic cues – including posterior-anterior and lateral cephalograph alongside a panoramic radiograph – students unanimously (100.0%) identified the condition as cleidocranial dysplasia (Fig. 7). ChatGPT-4o, however, incorrectly suggested a possibility of rickets, failing to recognize key radiographic features such as delayed closure of cranial sutures and fontanelles, underdeveloped maxilla, prolonged retention of primary teeth, and multiple unerupted supernumerary teeth. In another example, a panoramic radiograph showing a right mandibular condyle fracture (Fig. 8) was correctly interpreted by 98.4% of students. Both LMM chatbots, however, failed to localize the lesion, instead misidentifying it as a left mandibular body fracture, also overlooking the reversed left-right orientation typical of dental radiographic images. These findings suggest that current LMMs, despite their multimodal capabilities, lack the domain-specific radiographic interpretive accuracy required for clinical application in specialized fields such as OMFR.

Analysis results by question type, all LLM chatbots consistently demonstrated higher accuracy on SAQs than on MCQs. This trend may be attributed to the relative simplicity of the SAQs used in this study, which required concise, keyword-based responses, in contrast to MCQs that demanded comprehensive evaluation of all given options and a deeper understanding of OMFR knowledge. However, this pattern was reversed in the LLM chatbots, which exhibited higher accuracy on MCQs than on SAQs. These findings are consistent with those of a previous study by Mine et al. (2025), which reported similar MCQ accuracy scores in the OMFR domain – 66.7% for ChatGPT-4o and 50.0% for Gemini 2.0 Flash – although image-based SAQs were not included in their evaluation (Mine et al., 2025). This suggests that the enhanced architecture of LMMs may better support structured decision-making tasks such as MCQs, despite their continued limitations in generating accurate responses for complex dental imaging interpretation.

Response consistency was generally high in text-based and multiple-choice formats but decreased in image-based and short-answer tasks. These discrepancies may stem from fundamental structural differences between response types. MCQs in this study required selecting a single answer from a fixed set of choices while SAQs allow open-ended responses, resulting in broader variability in phrasing. The stochastic nature of autoregressive language models, which generate text by sampling from probability distributions, may further contribute to inconsistencies in complex image-based tasks (Kim et al., 2024). Despite the use of a rule-based normalization process to address minor lexical variation, agreement remained low – especially in image-based items, where models frequently generated entirely different interpretations. For instance, when interpreting panoramic image of a patient with "hyperparathyroidism" (Fig. 9), the chatbots generated a wide range of diagnoses, including osteogenesis imperfecta, nevoid basal cell carcinoma syndrome, cleidocranial dysplasia. Although only the first responses were used for accuracy scoring in this study, such variability underscores an important concern: even advanced LMMs may generate inconsistent outputs for the same task. This emphasizes that blind reliance on chatbot outputs, especially by individuals without domain expertise, may lead to serious errors in judgment.

The hallucination analysis demonstrated that both LMM chatbots generally produced responses of acceptable quality when evaluated by expert radiologists, particularly for text-based questions. Performance was consistently higher in multiple-choice formats than in short-answer formats, and for text-only items compared to image-based items. But hallucination increased substantially in image-based tasks, particularly for SAQs. These findings suggest that the integration of visual information introduces additional complexity that can compromise the models' ability to generate well-grounded, evidence-based responses (Sun et al., 2023). In clinical applications where multimodal input is essential, such as radiologic diagnosis, this vulnerability remains a critical challenge. Ongoing improvements in multimodal alignment, domain-specific training, and reasoning capabilities will likely be necessary to mitigate hallucination risks in complex diagnostic contexts.

ZS-CoT prompting was applied to evaluate whether structured reasoning could improve chatbot performance by inserting a simple instruction in Korean (“Let’s think step by step”) at the beginning of each query. Although this technique has previously demonstrated substantial improvements in accuracy across various reasoning tasks in natural language processing (NLP) (Kojima et al., 2022), it failed to make meaningful improvements not only on text-based tasks where chatbots already perform well, but also on image interpretation, where chatbots struggle significantly. These results suggest that the benefits of ZS-CoT prompting in facilitating robust reasoning effects may not generalize to highly domain-specific fields such as OMFR. Many OMFR tasks require factual recall rather than multi-step reasoning, thereby limiting the utility of stepwise prompting strategies. Similar findings have been reported in prior biomedical (Nagar et al., 2024) and oncology (Chen et al., 2024) studies, highlighting the need for further research to establish the reliability of prompt engineering approaches in enhancing the accuracy of multimodal chatbots.

This study has several limitations. First, the evaluation was conducted using general-purpose chatbot interfaces without API-based access or additional domain-specific training. It remains unclear to what extent the models have been exposed to OMFR content during their pretraining, limiting the ability to interpret the source and depth of their domain knowledge. Second, all experiments were performed using default settings without any adjustment of parameters such as temperature, response length, or prompt structure. The performance observed in this study may therefore not fully reflect the chatbots’ potential under optimized conditions.

Q. Write the most appropriate anatomical structure indicated by the yellow arrows in the following panoramic radiograph.

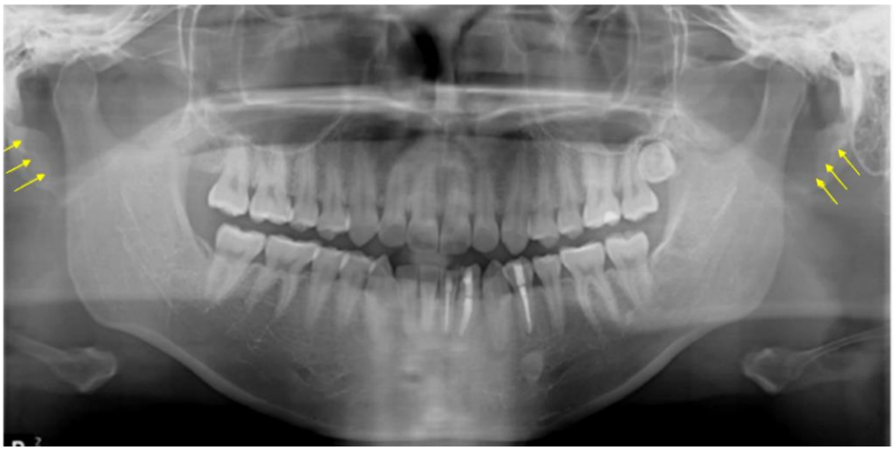


Fig. 5 Panoramic radiograph presented as an example of identifying a normal anatomical structure (ear lobe).

Q. Write the most appropriate normal anatomical structure missing in the following panoramic radiograph.

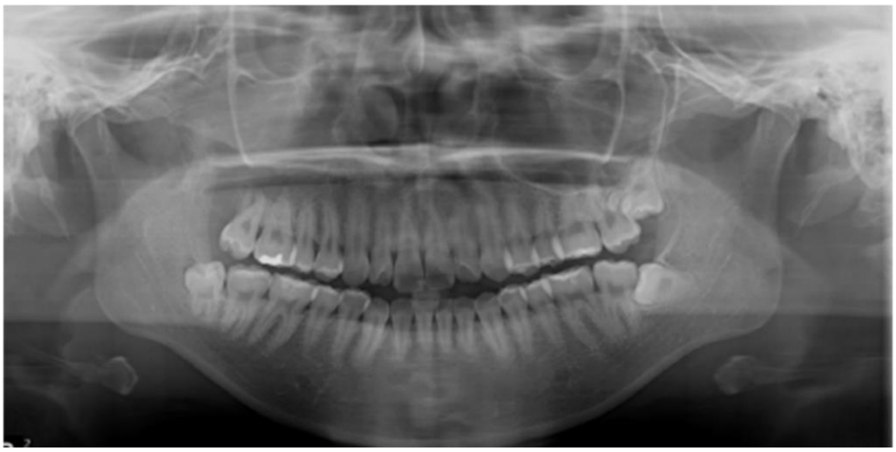


Fig. 6 Panoramic radiograph presented as an example of identifying a normal anatomical structure (missing floor of the maxillary sinus).

Q. Write the most appropriate diagnosis based on the following radiographs (A. posteroanterior cephalometric radiograph, B. lateral cephalometric radiograph, C. panoramic radiograph).

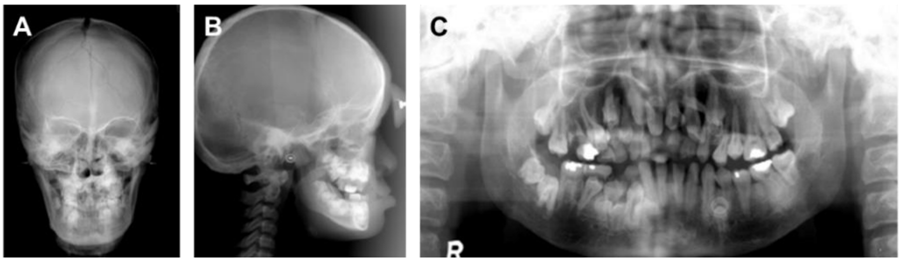


Fig. 7 Various radiographs presented as an example requiring radiographic diagnosis (cleidocranial dysplasia).

Q. Write the most appropriate anatomical structure in which a fracture is observed in the following panoramic radiograph.

Fig. 8 Panoramic radiograph presented as an example of identifying a fracture of an anatomical structure (condylar head).

Q. Write the most suspicious systemic disease from the following panoramic radiograph.

Fig. 9 Panoramic radiographs presented as an example requiring radiographic diagnosis (hyperparathyroidism).

5. CONCLUSION

This is the first study to present a comprehensive evaluation comparing general-purpose chatbot performance to actual student outcomes in the OMFR domain using a multidimensional dataset encompassing both text- and image-based questions. Additionally, this study provides a longitudinal assessment of performance progression from LLM to LMM chatbots, offering insight into the evolving capabilities of AI chatbot within a specialized dental domain.

The performance of four LLM and two LMM chatbots was evaluated using text- and image-based examination questions covering six educational content categories and two question types. Key performance indicators–including accuracy, hallucination, ZS-CoT prompting, and response consistency–were systematically analyzed to provide a comprehensive assessment of chatbot performance in this highly specialized dental domain.

LMM chatbots demonstrated superior accuracy and response quality compared to LLM chatbots in text-based tasks, and outperformed students in some areas. However, their performance remained limited in image-based diagnostic tasks. A high degree of inconsistency and hallucination was observed, particularly in complex visual interpretation and short-answer formats. ZS-CoT did not result in meaningful improvements in response accuracy.

Future research should include datasets with diverse clinical images and case scenarios to better evaluate multimodal reasoning. Model customization, including hyperparameter tuning and advanced prompting, may help reduce hallucination and improve performance in complex diagnostic tasks. These efforts are essential for the safe and effective application of AI chatbots in dental education and clinical practice.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., . . . Anadkat, S. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L. Z., . . . Telfeian, A. E. (2023). Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. *Neurosurgery, 93*(5), 1090-1098.

Bernard, A., Langille, M., Hughes, S., Rose, C., Leddin, D., & Van Zanten, S. V. (2007). A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Official journal of the American College of Gastroenterology| ACG, 102*(9), 2070-2077.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Chen, D., Huang, R. S., Jomy, J., Wong, P., Yan, M., Croke, J., . . . Raman, S. (2024). Performance of Multimodal Artificial Intelligence Chatbots Evaluated on Clinical Oncology Cases. *JAMA Network Open, 7*(10), e2437711-e2437711.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., . . . Gehrmann, S. (2023). Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research, 24*(240), 1-113.

Danesh, A., Pazouki, H., Danesh, F., Danesh, A., & Vardar-Sengul, S. (2024). Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. *Journal of Periodontology, 95*(7), 682-687.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep* *bidirectional transformers for language understanding.* Paper presented at the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers).

Fang, Q., Reynaldi, R., Araminta, A. S., Kamal, I., Saini, P., Afshari, F. S., . . . Sukotjo, C. (2024). Artificial Intelligence (AI)-driven dental education: Exploring the role of chatbots in a clinical learning environment. *The Journal of Prosthetic Dentistry*.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*, 681-694.

Helvacioglu-Yigit, D., Demirturk, H., Ali, K., Tamimi, D., Koenig, L., & Almashraqi, A. (2025). Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*.

Huang, D., Yan, C., Li, Q., & Peng, X. (2024). From large language models to large multimodal models: A literature review. *Applied Sciences, 14*(12), 5068.

Islam, R., & Moushi, O. M. (2024). Gpt-4o: The cutting-edge advancement in multimodal llm. *Authorea Preprints*.

Kim, S., Lee, C.-k., & Kim, S.-s. (2024). Large Language Models: A Comprehensive Guide for Radiologists. *Journal of the Korean Society of Radiology, 85*(5), 861-882.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems, 35*, 22199-22213.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., . . . Maningo, J. (2023). Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health, 2*(2), e0000198.

Kuşcu, O., Pamuk, A. E., Sütay Süslü, N., & Hosal, S. (2023). Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Frontiers in Oncology, 13*, 1256459.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., . . . Bai, X. (2024). *Monkey: Image resolution and text label are important things for large multi-modal models.* Paper presented at the proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Lin, S.-Y., Chan, P. K., Hsu, W.-H., & Kao, C.-H. (2024). Exploring the proficiency of ChatGPT-4: an evaluation of its performance in the Taiwan advanced medical licensing examination. *Digital Health, 10*, 20552076241237678.

Mine, Y., Okazaki, S., Taji, T., Kawaguchi, H., Kakimoto, N., & Murayama, T. (2025). Benchmarking multimodal large language models on the dental licensing examination: Challenges with clinical image interpretation. *Journal of Dental Sciences*.

Nagar, A., Schlegel, V., Nguyen, T.-T., Li, H., Wu, Y., Binici, K., & Winkler, S. (2024). LLMs are not Zero-Shot Reasoners for Biomedical Information Extraction. *arXiv preprint arXiv:2408.12249*.

Nguyen, H. D., & Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), e1246.

Ohta, K., & Ohta, S. (2023). The performance of GPT-3.5, GPT-4, and Bard on the Japanese

National Dentist Examination: a comparison study. *Cureus, 15*(12).

Park, S. H., Suh, C. H., Lee, J. H., Kahn Jr, C. E., & Moy, L. (2024). Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean Journal of Radiology, 25*(10), 865.

Patil, N. S., Huang, R. S., van der Pol, C. B., & Larocque, N. (2024). Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. *Canadian Association of Radiologists Journal, 75*(2), 344-350.

Qi, D., Su, L., Song, J., Cui, E., Bharti, T., & Sacheti, A. (2020). Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Rahsepar, A. A., Tavakoli, N., Kim, G. H. J., Hassani, C., Abtin, F., & Bedayat, A. (2023). How AI responds to common lung cancer questions: ChatGPT versus Google Bard. *Radiology, 307*(5), e230922.

Rawte, V., Sheth, A., & Das, A. (2023). A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 68-79.

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., . . . Yang, Y. (2023). Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Tassoker, M. (2025). ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health, 25*(1), 173.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., . . . Millican, K. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine, 29*(8), 1930-1940.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., . . . Du, Y. (2022). Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Toyama, Y., Harigai, A., Abe, M., Nagano, M., Kawabata, M., Seki, Y., & Takase, K. (2024). Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. *Japanese journal of radiology, 42*(2), 201-207.

Uehara, O., Morikawa, T., Harada, F., Sugiyama, N., Matsuki, Y., Hiraki, D., . . . Murata, Y. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National Dental Examination. *Journal of Dental Education*.

Uehara, O., Morikawa, T., Harada, F., Sugiyama, N., Matsuki, Y., Hiraki, D., . . . Murata, Y. (2025). Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National Dental Examination. *Journal of Dental Education, 89*(4), 459-466.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Wu, S.-H., Tong, W.-J., Li, M.-D., Hu, H.-T., Lu, X.-Z., Huang, Z.-R., . . . Chen, L.-D. (2024). Collaborative enhancement of consistency and accuracy in US diagnosis of thyroid nodules using large language models. *Radiology, 310*(3), e232255.

ABSTRACT IN KOREAN

텍스트 및 이미지 기반 문제를 활용한 거대 언어 모델 및 거대 다중모달 모델 인공지능 챗봇의 성능 평가

목적: 본 연구는 영상치의학에서 범용 거대 언어 모델 및 거대 다중모달 모델 기반 인공지능 챗봇의 성능을 평가하여 실제 치과대학생의 성적과 비교하고, 거대 언어 모델에서 거대 다중모달 모델 챗봇으로의 종단적 성능 변화를 분석함으로써 다차원 평가를 수행하는 것을 목표로 한다.

재료 및 방법: 국내 치과대학의 영상치의학 교육과정에서 추출한 90 개의 텍스트 및 영상 기반 시험 문항을 6 개의 교육 내용과 2 개의 문제 유형으로 분류하였다. 4 개의 거대 언어 모델 챗봇(ChatGPT, ChatGPT Plus, Bard, Bing Chat)은 각 문항에 대해 1 회, 2 개의 거대 다중모달 모델 챗봇(ChatGPT-4o, Gemini 2.0 Flash)은 10 회씩 응답을 수집하였다. 모든 챗봇의 첫 회차 응답을 기준으로 정확도를 산출하여 실제 치과대학생 성적과 비교하였다. 거대 다중모달 모델 챗봇에 한하여 10 회 반복 응답의 일관성을 Fleiss' kappa 계수로 평가하였고, 2 명의 영상치의학 전문의가 Global Quality Scales 지표의 5 점 척도에 따라 환각 정도를 평가하여 평균 및 표준편차를 계산하였다. 마지막으로는 단계별 추론을 유도하는 제로샷 생각의 사슬 프롬프트의 적용 효과를 확인하였다.

결과: 거대 다중모달 모델 챗봇은 텍스트 기반의 문항에서 거대 언어 모델 챗봇보다 높은 정확도를 보였고, 일부 영역에서는 치과대학생의 성적을 상회하는 성과를 나타냈다. 그러나 영상 기반 문제에서는 매우 제한적인 성능을 보였으며, 복잡한 영상 해석 및 단답형 문항에서 높은 수준의 변동성과 환각이 관찰되었다. 제로샷 생각의 사슬 프롬프트 적용은 챗봇의 정확도 향상에 유의미한 효과를 보이지 않았다.

결론: 본 연구는 텍스트와 이미지를 모두 포함하는 영상치의학 관련 시험 문항을 활용하여 챗봇의 성능을 학생 성적과 비교하는 동시에 거대 언어 모델 챗봇에서 거대 다중모달 모델 챗봇으로의 종단적 성능 변화를 조사한 최초의 연구로서, 현시점에서 범용성 인공지능 챗봇의 역량과 한계를 규명하는 데에 시의적절한 통찰을 제공한다. 향후 연구에서 다양한 임상 영상과 사례를 포함한 특화된 데이터셋을 활용하고, 환각 감소를 위해 모델 맞춤화와 고급 프롬프트 전략을 적용한다면 학생 및 환자 교육과 임상 실무에서 인공지능 챗봇을 보다 안전하고 효과적으로 활용할 수 있을 것이다.

핵심되는 말 : 영상치의학, 거대 언어 모델, 거대 다중모달 모델, 인공지능, 챗봇, 성능 평가