

Original Article

Evaluating the Accuracy of Artificial Intelligence-Based Chatbots on Pediatric Dentistry Questions in the Korean National Dental Board Exam

Yun Sun Jung¹, Yong Kwon Chae¹, Mi Sun Kim^{2,3}, Hyo-Seol Lee^{1,3}, Sung Chul Choi^{1,3}, Ok Hyung Nam^{1,3}

¹Department of Pediatric Dentistry, Kyung Hee University College of Dentistry, Kyung Hee University Medical Center, Seoul, Republic of Korea

²Department of Pediatric Dentistry, Kyung Hee University Dental Hospital at Gangdong, Seoul, Republic of Korea

³Department of Pediatric Dentistry, School of Dentistry, Kyung Hee University, Seoul, Republic of Korea

Abstract

This study aimed to assess the competency of artificial intelligence (AI) in pediatric dentistry and compare it with that of dentists. We used open-source data obtained from the Korea Health Personnel Licensing Examination Institute. A total of 32 item multiple-choice pediatric dentistry exam questions were included. Two AI-based chatbots (ChatGPT 3.5 and Gemini) were evaluated. Each chatbot received the same questions seven times in separate chat sessions initiated on April 25, 2024. The accuracy was assessed by measuring the percentage of correct answers, and consistency was evaluated using Cronbach's alpha coefficient. Both ChatGPT 3.5 and Gemini demonstrated similar accuracy, with no significant differences observed between them. However, neither chatbot achieved the minimum passing score set by the Pediatric Dentistry National Examination. However, both chatbots exhibited acceptable consistency in their responses. Within the limits of this study, both AI-based chatbots did not sufficiently answer the pediatric dentistry exam questions. This finding suggests that pediatric dentists should be aware of the advantages and limitations of this new tool and effectively utilize it to promote patient health. [J Korean Acad Pediatr Dent 2024;51(3):299-309]

Keywords

Artificial Intelligence, Machine Intelligence, Pediatric Dentistry, Dentist

ORCID

Yun Sun Jung

<https://orcid.org/0009-0001-5352-0160>

Yong Kwon Chae

<https://orcid.org/0000-0001-8059-9305>

Mi Sun Kim

<https://orcid.org/0000-0001-8338-1838>

Hyo-Seol Lee

<https://orcid.org/0000-0001-7287-5082>

Sung Chul Choi

<https://orcid.org/0000-0001-7221-2000>

Ok Hyung Nam

<https://orcid.org/0000-0002-6386-803X>

Article history

Received July 10, 2024

Revised August 10, 2024

Accepted August 12, 2024

© 2024 Korean Academy of Pediatric Dentistry

© This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Artificial intelligence (AI) is a branch of computer science that deals with the creation of intelligent agents, which are systems that can reason, learn, and

Corresponding author: Ok Hyung Nam

Department of Pediatric Dentistry, Kyung Hee University School of Dentistry, 26 Kyungheedae-ro, Dongdaemun-gu, Seoul 02447, Republic of Korea.

Tel: +82-2-958-9371 / Fax: +82-2-958-9478 / E-mail: pedokhyung@gmail.com

act autonomously. AI has been used in a wide variety of fields, and its importance in healthcare is growing rapidly[1,2]. The main types of AI chatbots available include rule-based chatbots, machine learning chatbots, natural language processing (NLP) chatbots, generative AI chatbots, hybrid chatbots, and domain-specific chatbots. Large language models (LLMs), such as the Chat-Generative Pretrained Transformer (ChatGPT), are NLP tools capable of understanding and generating human-like text[3]. Conventional LLMs are simpler models that predict sequences based on fixed windows of words, while ChatGPT uses vast datasets to generate human-like conversations.

Most AI applications in medicine rely on machine learning technology, particularly supervised learning. In this approach, machines learn from pairs of data and annotated labels, often provided by human experts (for example, “this radiograph contains a carious lesion”)[4]. The machine iteratively learns the statistical patterns inherent in these data-label pairs, allowing it to make predictions on unseen and unlabeled data over time. Typically, this prediction occurs on a test set that is separate and independent of the training dataset or occurs later in real-world clinical applications.

Failure of machine learning systems occurs when they fail to consistently reproduce intended behaviors, leading to unpleasant outcomes in contexts that differ from their intended behavior[5]. This risk is particularly high in the application of machine learning to AI and medical healthcare (MLH) because algorithmic discoveries can directly affect human health management. Unfortunately, several factors affect the reproducibility of MLH applications. These factors are related to the availability, quality, and consistency of clinical or biomedical data. When trained on a general corpus of text data, ChatGPT may lack specialized medical knowledge or terminologies[6]. Consequently, their ability to answer specific medical questions or provide accurate advice related to medical conditions may be limited.

ChatGPT could represent the first in a new line of models that may better represent the combination of clinical knowledge and dialogic interaction[7]. Artificial

intelligence (AI) has progressively become an essential component in medical education, revolutionizing access to information and learning processes through advanced tools such as computer-based models, virtual reality simulations, and personalized learning platforms. The literature effectively showcases improvements in the accuracy of the responses provided by ChatGPT across various medical education contexts[8-10]. However, the reliability of ChatGPT to repeated queries remains uncertain[6,11]. To ensure that AI tools used in medical education are both accurate and reliable, it is crucial to assess their performance against established benchmarks. One study examined the reliability of references generated by ChatGPT language models in the Head and Neck field. ChatGPT 4.0, outperformed version 3.5, in terms of reference reliability. However, both versions tended to provide erroneous or nonexistent references[12]. On the contrary, previous studies have tasked AI with taking the United States Medical Licensing Exam (USMLE), which yielded different results. This study reported the increasing accuracy of ChatGPT, reaching or surpassing the passing threshold for USMLE, and the potential of AI to generate fresh insights that could aid human learners in a medical education environment[11]. Previous research in various medical fields, such as orthopedics [13]stock ownership, equity interest, patent/licensing arrangements, etc., has highlighted the importance of these evaluations by comparing AI performance to that of human practitioners using standardized exams. This method provides a rigorous and relevant assessment framework.

Limited availability and accessibility of medical and dental data, owing to concerns regarding data protection and organizational hurdles, contribute to the scarcity of papers on AI applications in dentistry[14]. Additionally, the insufficient replicability and robustness of dental AI research, along with the limited applicability of AI outcomes in addressing the complex decision-making processes required in clinical care, further contribute to the limited number of studies in this area.

Research on the application of AI in dentistry is limited, particularly in pediatric dentistry. Pediatric dentistry requires specialized expertise and knowledge, making

it challenging for machine learning algorithms to fully understand this complexity. In our study, we aimed to evaluate the performance of AI chatbots in the context of pediatric dentistry by utilizing the Korean National Dental Board Examination. This exam was selected due to its comprehensive nature, which ensures that dental professionals possess the necessary knowledge and skills to practice safely and effectively. Therefore, the purpose of this study was to evaluate whether AI possesses competencies comparable to those of dentists. The null hypothesis of this study was that “artificial intelligence can provide accurate answers in the pediatric dentistry national board examination at a level equivalent to that of a dentist.”

Materials and Methods

1. Establishment of a Multiple-Choice Questionnaire Based on the Korean National Dental Board Examination

This study used open-source data from the Korea Health Personnel Licensing Examination Institute (<https://www.kuksiwon.or.kr/index.do>). Briefly, the dental examination consisted of 321 questions, with 23 questions specifically related to pediatric dentistry. Questions regarding pediatric dentistry were selected from the Korean National Dental Board Examination conducted in 2023 and 2024. Questions containing images or diagrams were excluded owing to copyright restrictions. Finally, a 32-item multiple-choice questionnaire was created (Table 1). The questionnaire was structured around the following topics: development and morphology, radiographic techniques, behavioral guidance, restorative dentistry, pulp therapy, occlusion and orthodontics, local anesthesia, and trauma.

2. Evaluating the Accuracy of AI-Based Chatbot Systems

Two AI-based chatbots were evaluated: (i) ChatGPT 3.5 (OpenAI, San Francisco, USA) and (ii) Gemini (Google DeepMind, London, United Kingdom) (Table 2). All the questions were asked on the same day (April 25th, 2024).

The questionnaire was administered seven times to both AI-based chatbots. A new chat session was initiated each time (Fig. 1). This method aims to prevent potential learning biases and performance improvements that may occur if the same set of questions is repeatedly presented in a single session[4].

In the actual national dental examination for dentists, a score of 40% or higher in each subject was determined based on the following categories: Pediatric Dentistry and Orthodontics as one subject; Oral Radiology, Oral Medicine, and Oral Pathology as one subject; Periodontics and Oral Health as one subject; and Dental Materials and Oral Biology as one subject. In our study, we excluded orthodontic examination questions and set the failing score for pediatric dentistry at 40%.

To assess accuracy, the percentages of correct answers and consistency in each chatbot were measured. The answers from the chatbots were based on the corresponding answers that were officially provided by the Korean National Dental Board Examination. The correct answer rate was determined by averaging the results from each of these seven attempts.

3. Statistical Analysis


Descriptive statistics, including means, medians, and percentages, were calculated. Data were analyzed using IBM SPSS Statistics 20 (SPSS Inc., Chicago, IL, USA). After confirming normality using the Kolmogorov-Smirnov test, data that followed normality were analyzed using an independent t-test, while data that did not follow normality underwent the Mann-Whitney test (with a significance level set at 0.05). Consistency was assessed using the Cronbach's alpha coefficient, with p-values and confidence intervals (CI) calculated from the two-way mixed effects model. The consistency levels were categorized as follows: excellent (greater than 0.9), good (from 0.8 to 0.9), acceptable (from 0.7 to 0.8), and questionable (less than 0.7)[15]. Consistency in responses from artificial intelligence-based chatbots was evaluated overall, as well as per topic.

Table 1. A 32-item multiple-choice questionnaire used in this study


No.	Questions	Topics
1	Which organ is being described below?	Development and Morphology
2	Which statement about tooth eruption before shedding is correct?	Development and Morphology
3	Which statement about the morphology of the upper first permanent molar is correct?	Development and Morphology
4	How to minimize radiation exposure in children during intraoral radiography?	Radiographic Technique
5	Which psychological-social stage of Erikson corresponds to the stage where children can think abstractly beyond the immediate physical world according to Piaget's cognitive development stages?	Behavior Guidance
6	What is the term for redirecting a child's behavior from watching a favorite video that they were told to stop by a dentist during treatment?	Behavior Guidance
7	What is correct about moderate sedation?	Behavior Guidance
8	Instructions for a 7-year-old child's rubber dam moisture control method are given. What are the correct numbers to fill in for (A), (B), and (C)?	Restorative Dentistry
9	A child's lower first primary tooth has extensive decay but no cavitation. What treatment is appropriate?	Restorative Dentistry
10	What is correct when restoring a lower first permanent molar with a stainless steel crown?	Restorative Dentistry
11	What is correct about evaluating the condition of dental pulps?	Pulp Therapy
12	What are the indications for an indirect pulp treatment?	Pulp Therapy
13	What appliance is used to correct a crossbite in the dental arch?	Occlusion and Orthodontics
14	What common factors contribute to early mixed dentition phase issues in children?	Occlusion and Orthodontics
15	Which stage of childhood development does the following correspond to?	Development and Morphology
16	Which statement regarding children's physical growth and development related to body weight and surface area is correct?	Development and Morphology
17	Which statement about the growth and development of the mandible is correct?	Development and Morphology
18	What is common among Down syndrome, cleidocranial dysostosis, hypopituitarism, and fibrous dysplasia?	Development and Morphology
19	Which statement regarding radiation protection for children is correct?	Radiographic Technique
20	Which stage in Erikson's psychosocial development corresponds to preschool age, characterized by efforts to satisfy curiosity while needing defined boundaries and guidelines?	Behavior Guidance
21	The description refers to children's behavior categorized by Wright as a potential cooperative group. Which behavior does it describe?	Behavior Guidance
22	In behavior modification, which category does the sham operation technique belong to?	Behavior Guidance
23	When is nitrous oxide-oxygen sedation appropriate?	Behavior Guidance
24	What is the appropriate response when a rubber dam clamp causes severe pain due to excessive pressure on the gingiva during placement?	Restorative Dentistry
25	What should be considered when restoring immature permanent teeth?	Restorative Dentistry
26	Which statement regarding zirconia crown restorations for primary teeth is correct?	Restorative Dentistry
27	The following description pertains to the anatomical structure of a primary tooth's root canals. Which tooth does it describe?	Pulp Therapy
28	Which component of paste-form calcium hydroxide used in pediatric tooth pulp treatment enhances infection control and increases radiopacity?	Pulp Therapy
29	Information about inferior alveolar nerve block anesthesia for children is provided. Arrange the correct sequence in the parentheses:	Local Anesthesia
30	A 7-year-old child presents two days after trauma with a crown fracture exposing about 1 mm of dentin on a maxillary central incisor. What treatment is appropriate?	Trauma
31	A 3-year-old child presents two hours after trauma with the following symptoms on a maxillary primary central incisor: 2 mm intrusion, buccal displacement of the crown, slight gingival bleeding, and no evidence of pulp or alveolar bone fracture. What treatment is indicated?	Trauma
32	Which disorder characterized by defective formation of the mesodermal system prior to type I collagen formation leads to skeletal, skin, tendon, and ligament abnormalities, along with dentinogenesis imperfecta?	Development and Morphology


Table 2. Comparison of artificial intelligence-based chatbot systems used in this study


Features	ChatGPT 3.5	Gemini
Specification	Input: Natural language text Output: Text responses Large pre-trained model (e.g., GPT-3)	Input: Multi-modal data (text, images, audio) Output: Multi-modal responses Tailored for multi-modal interactions
Algorithm	Transformer architecture Language model fine-tuning Focus on natural language understanding and generation	Hybrid architecture incorporating vision, language, and decision-making modules Emphasis on multi-modal fusion and reasoning
Training Data	Text-based corpora (e.g., books, articles, internet text) Large-scale, diverse datasets	Multi-modal datasets (text, images, audio) Incorporation of domain-specific data for decision-making
Advantages	Strong in language comprehension and generation tasks Versatile for text-based applications Extensive pre-trained knowledge base	Capable of handling multi-modal inputs and generating multi-modal outputs Suited for tasks requiring integration of different data types Potential for complex decision-making tasks



A  **You**
49.The organ described in the following is responsible for:
Influencing growth, development, protein synthesis, and metabolic activities.
Its dysfunction can lead to conditions such as cretinism and myxedema, delayed eruption and exfoliation of deciduous teeth, and delayed development of permanent teeth.



- ① Adrenal gland
- ② Thymus gland
- ③ Thyroid gland
- ④ Reproductive gland
- ⑤ Parathyroid gland




 **ChatGPT**
The correct answer is:
③ Thyroid gland

B  49.The organ described in the following is responsible for:
Influencing growth, development, protein synthesis, and metabolic activities.
Its dysfunction can lead to conditions such as cretinism and myxedema, delayed eruption and exfoliation of deciduous teeth, and delayed development of ...

 The answer is ③ Thyroid gland.

Here's why:

- The passage describes an organ influencing growth, development, protein synthesis, and metabolic activities.
- The thyroid gland is the primary regulator of these functions in the body.
- Dysfunction of the thyroid gland can cause cretinism and myxedema, delayed development of teeth, which aligns with the information in the passage.

Fig. 1. An example of multiple-choice questions given to artificial intelligence-based chatbot systems. (A) ChatGPT 3.5, (B) Gemini. To help readers understand how the interaction took place, only the images in the figures were translated into English.

Results

1. Percentage Correct Answers

Regarding the overall percentage of correct answers, there were no significant differences between ChatGPT 3.5 (35.3 ± 5.6%) and Gemini (33.0 ± 4.0%). Regarding the percentage of correct answers by topic, ChatGPT 3.5 and Gemini showed varying levels of performance (Fig. 2). ChatGPT 3.5 demonstrated the following rates according to the topics: Development and Morphology (33.9 ± 9.4%), Radiographic Technique (42.8 ± 18.9%), Behavior Guidance (26.5 ± 15.3%), Restorative Dentistry (31.0 ± 11.5%), Pulp therapy (53.5 ± 22.5%), Occlusion and Orthodontics (28.5 ± 26.7%), Trauma (50.0 ± 28.9%), and Local anesthesia (28.6 ± 48.8%). In contrast, Gemini

demonstrated the following rates: Development and Morphology (51.8 ± 4.7%), Radiographic Technique (0.0 ± 0.0%), Behavior Guidance (26.5 ± 9.9%), Restorative Dentistry (21.4 ± 8.1%), Pulp therapy (46.4 ± 9.4%), Occlusion and Orthodontics (50.0 ± 0.0%), Trauma (0.0 ± 0.0%), and Local anesthesia (42.8 ± 53.4%). ChatGPT 3.5 showed significantly higher rates in Radiographic Technique ($p = 0.001$) and Trauma ($p = 0.002$) than Gemini. Gemini showed significantly higher rates of Development and Morphology ($p = 0.001$) than ChatGPT 3.5.

2. Consistency in responses from AI-based chatbots

The Cronbach's alpha coefficient scores for ChatGPT 3.5 and Gemini were 0.833 and 0.969, respectively (Table 3). These scores indicated a good level of response consistency.

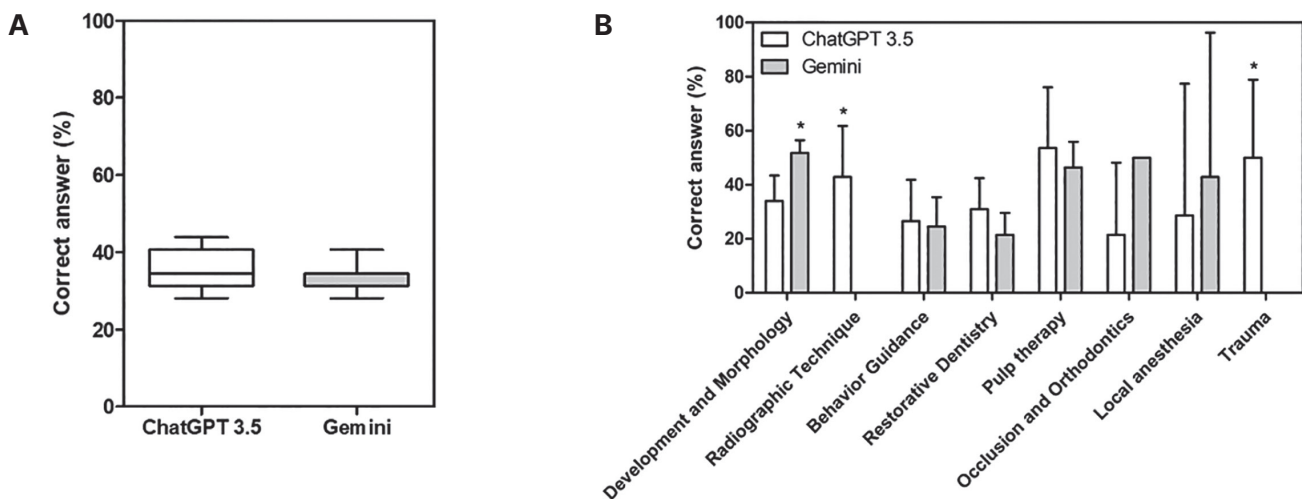


Fig. 2. Percentage of correct answers by artificial intelligence-based chatbots. Data shown are the mean percentage of correctly answered questions ± standard deviation (A) Comparison of the overall percentage of correct answers. There were no significant differences between ChatGPT 3.5 and Gemini ($p = 0.864$). Line: median, Box: Interquartile range, Whiskers: Min.-max. (B) Comparison of the percentage of correct answers per topic. Statistical analyses for the Development and Morphology, Radiographic Technique, Behavior Guidance, and Restorative Dentistry categories were conducted using an independent t-test, whereas the others were conducted using the Mann-Whitney test. * $p < 0.05$

Table 3. Consistency in responses from artificial intelligence-based chatbots

Group	Cronbach's alpha	95% CI	p-value
ChatGPT 3.5	0.833	0.726 - 0.908	< 0.0001
Gemini	0.969	0.949 - 0.983	< 0.0001

p-values and confidence intervals (CI) were calculated from the two-way mixed effects model.

tency for ChatGPT 3.5 and an excellent level for Gemini.

Both chatbots demonstrated acceptable levels of consistency for most topics (Table 4). However, ChatGPT 3.5 exhibited a very low level in the “Behavior Guidance” topic. Interestingly, Gemini incorrectly answered all questions on Radiographic Technique in all seven trials, resulting in a variance of zero on the scale.

Discussion

This study aimed to investigate whether AI chatbots can be applied in the field of pediatric dentistry by having them undergo a national pediatric dentistry examination. Both ChatGPT 3.5 ($35.3 \pm 5.6\%$) and Gemini ($33.0 \pm 4.0\%$) did not pass the pediatric dentistry section of the national examination based on their respective average scores. Therefore, the null hypothesis was rejected. An analysis of the 2023 National Dental Examination for Dentists revealed that none of the 754 candidates failed the pediatric dentistry course[16]. The average number of correct answers to the 23 pediatric dentistry questions was 18.6 (SD = 2.3). This outcome indicates that dental license exam candidates performed significantly better than AI candidates in the exam, as both ChatGPT 3.5 and Gemini scored below the passing threshold, and there was no significant difference between the two groups. Based on these results, AI is currently insufficient to re-

place pediatric dentistry experts. These results are consistent with the findings of a previous study[11], which showed that ChatGPT 3.5 was unable to attain a passing score for Step 1, which evaluates foundational medical knowledge, and Step 2, which assesses clinical knowledge examinations of the USMLE, scoring 55.8% and 59.1%, respectively. The results of a contrasting study[17] showed that a chatbot outperformed first- and second-year medical and physician assistant students in clinical reasoning examinations. This suggests that while the chatbot may excel for less experienced students, it might not perform as well in more advanced exams, such as graduation exams. According to prior research[18], an in-depth analysis of the question style revealed that single-choice questions were associated with a significantly higher rate ($p < 0.001$) of correct responses ($n = 1313$; 63%) than multiple-choice questions ($n = 162$; 34%). Based on these findings, one might consider that AI may still lack the ability to answer multiple-choice exam questions effectively.

In the case of Gemini, it got all the questions wrong in the Radiographic Technique and Trauma sections. Additionally, ChatGPT 3.5 performed better than Gemini in all items within these categories. The differences in performance between ChatGPT and Gemini, especially in the areas of trauma and radiographic techniques, could be attributed to the nature of the questions. Both

Table 4. Consistency in responses from artificial intelligence-based chatbots per topic

Group	ChatGPT 3.5			Gemini		
	ICC	95% CI	<i>p</i> -value	ICC	95% CI	<i>p</i> -value
Development and Morphology	0.902	0.745 - 0.977	<0.0001	0.993	0.983 - 0.998	<0.0001
Radiographic Technique	0.984	0.862 - 1.000	<0.0001	N/A*	N/A	N/A
Behavior Guidance	0.206	-1.121 - 0.842	0.300	0.958	0.883 - 0.992	<0.0001
Restorative Dentistry	0.7636	0.284 - 0.962	0.005	0.927	0.780 - 0.988	<0.0001
Pulp therapy	0.857	0.434 - 0.990	0.003	0.907	0.632 - 0.993	<0.0001
Occlusion and Orthodontics	0.910	0.207 - 1.00	0.016	1.00	1.00 - 1.00	N/A
Local anesthesia	N/A**	N/A	N/A	N/A	N/A	N/A
Trauma	0.843	-0.387 - 1.00	0.045	0.778	-0.958 - 1.00	0.078

Consistency was assessed using the intraclass correlation coefficient (ICC), with *p*-values and confidence intervals (CI) calculated from the two-way mixed effects model.

*Gemini’s radiographic technique value is N/A because of zero variance on the scale.

**Insufficient cases for analysis; both ChatGPT 3.5 and Gemini are marked as N/A.

trauma and radiographic technique questions often require not just theoretical knowledge but also the ability to assess situations and make diagnoses. Specifically, in the areas of trauma and radiographic techniques, the ICC value between the two chatbots was -1.988 (p -value = 0.996), which indicated a very low level of agreement between the two chatbots. This suggests that Gemini has a weakness in assessing situations and making diagnoses, whereas ChatGPT 3.5 appears to be better at diagnosis.

To trust and use AI in the field of healthcare, it is important for AI to provide consistent responses. Both ChatGPT and Gemini exhibited excellent consistency. These findings align with well-known characteristics of AI [19]. However, there are still gaps that must be filled before relying on these findings. Despite these advantages, a notable drawback of AI is its potential to generate seemingly reliable and highly plausible but incorrect answers[19,20]. A major concern with AI in healthcare is the risk of delivering inaccurate medical advice[21,22]. Because AI-generated content relies on extensive internet data, the information it provides can sometimes be misleading or entirely incorrect. A study examining the reliability of references generated by ChatGPT language reveals a significant problem known as “hallucination” or “stochastic parroting”[23]. This phenomenon describes the generation of convincing yet false information by AI systems such as ChatGPT and is recognized as an issue in various natural language processing models[19]. In pediatrics, traditional rule-based clinical decision support (CDS) systems are routinely used to improve patient care, but they are often limited by poor model specificity, leading to frequent false positive alerts[24]. AI models have the potential to revolutionize the healthcare sector, it is crucial to recognize and address the associated risks[25]. Each AI-generated reference should be cross-verified against primary sources and trusted academic databases. This might involve updating the AI’s training data, refining the algorithms, or using human oversight to ensure accuracy. Moreover, to fully utilize AI chatbots, ensuring accurate translation of medical terminology is essential. Ensuring the safe and effective integration of AI technologies into the healthcare system requires proper

oversight and regulation[25]. An important limitation for pediatric research, from the standpoint of model development and prediction, is the lack of large datasets[24]. Constructing predictive models requires large datasets. If these issues were resolved, AI chatbots hold promising potential in various areas of medicine, such as patient education, appointment scheduling, and mental health support.

Ethical, social, and legal considerations regarding AI are additional factors to consider. Ethical controls and limitations must be established in AI language models. Many studies have pointed out the ethical and legal concerns posed by ChatGPT and other AI tools in medicine [26-30]. The World Health Organization (WHO) has recently established a new digital health department and issued updated guidelines on digital health[31]. When using ChatGPT, it is essential to implement security measures to protect patient information. These measures include encryption, access control, secure data storage, and adherence to privacy regulations. The patient data used for training and fine-tuning ChatGPT should be anonymized to safeguard privacy. Obtaining patient consent for data use during the development of ChatGPT is critically important[20].

This study had several limitations. First, the exclusion of image-related questions limited the assessment of the AI’s ability to solve various types of problems, especially in diagnostics. In a previous study assessing the usability of information generated by ChatGPT in oral and maxillofacial surgery, two types of questions were posed[32]. The results revealed that ChatGPT provided reasonably accurate and helpful responses to patient-oriented questions but did not perform as well in responding to advanced technical questions. Additionally, a prior study[33] on the 643 Congress of Neurological Surgeons Self-Assessment Neurosurgery Exam (SANS) board-style questions found that out of these, 477 were text-based and 166 contained images. GPT-4 demonstrated a 79.0% accuracy rate for text-based questions and a 66.6% accuracy rate for image-based questions that it deemed appropriate to answer. Open AI has noted future capabilities for image input processing, which should be further

investigated once available to the public. AI's role in dentistry should extend beyond merely handling knowledge-based test questions. It should also include addressing patients' common inquiries, performing image diagnoses, incorporating advanced technologies, and more.

Second, this study utilized the free version of ChatGPT 3.5 instead of ChatGPT 4.0. To ensure the study's findings were applicable to a wider user base, this study tried to compare the easily accessible AI chatbot systems at the stage of study design. At the time of the study design, ChatGPT 4.0 was not freely available and had limited access, which could restrict the applicability of the study's findings. On the other hand, ChatGPT 3.5 and Gemini were both freely accessible to users, making them more representative of the tools available to the general public. In a previous study in which medical examinations like USMLE were administered, ChatGPT version 4 showed a significant improvement in performance[11]. A previous study showed a significant increase in the percentage of correct answers when ChatGPT 4.0 was used compared with those in ChatGPT 3.5.

Third, given the low correct answer rate of chatbots observed in this study, it is crucial to review and analyze both the correct and incorrect responses. A previous study showed that formatting questions into three variants can help avoid systematic errors introduced by rigid wording[11]. Considering this finding, future research comparing and analyzing response patterns by varying the question formats will be necessary. Additionally, it is acknowledged that the AI chatbot might perform differently in various languages due to the amount of training data available in each language. In this study, questions were asked in Korean to preserve the nuances and context of the original exam questions. It would be important to consider linguistic aspects by conducting studies that involve asking questions in English, as well as other languages, to account for these differences in proficiency. This approach could provide a deeper understanding of the chatbot's capabilities and limitations.

Conclusion

This study aimed to assess whether AI can match the skills of pediatric dentists. Neither ChatGPT 3.5 nor Gemini demonstrated sufficient proficiency in passing examinations in the pediatric dentistry domain. This implies that AI is not yet capable of replacing dentists in pediatric dentistry. Pediatric dentists should be aware of the advantages and limitations of this new tool and should effectively utilize it to promote patient health.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

References

1. Arif TB, Munaf U, Ul-Haque I : The future of medical education and research: Is ChatGPT a blessing or blight in disguise? *Med Educ Online*, 28:2181052, 2023.
2. Manickam P, Mariappan SA, Murugesan SM, Hansda S, Kaushik A, Shinde R, Thipperudraswamy SP : Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare. *Biosensors (Basel)*, 12:562, 2022.
3. Lee JW, Yoo IS, Kim JH, Kim WT, Jeon HJ, Yoo HS, Shin JG, Kim GH, Hwang S, Park S, Kim YJ : Development of AI-generated medical responses using the ChatGPT for cancer patients. *Comput Methods Programs Biomed*, 254:108302, 2024.
4. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, Uribe SE, Krois J : Artificial intelligence in dental research: Checklist for authors, reviewers, readers. *J Dent*, 107:103610, 2021.
5. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M : Reproducibility in machine learning for health research: Still a ways to go. *Sci Transl Med*, 13:eabb1655, 2021.
6. Vaishya R, Misra A, Vaish A : ChatGPT: Is this version good for healthcare and research? *Diabetes Metab*

- Syndr*, 17:102744, 2023.
7. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D : How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*, 9:E45312, 2023.
 8. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succu MD : Evaluating ChatGPT as an adjunct for radiologic decision-making. *MedRxiv*, 2023 Feb 7:2023.02.02.23285399, 2023.
 9. Potapenko I, Boberg - Ans LC, Stormly Hansen M, Klefter ON, van Dijk EH, Subhi Y : Artificial intelligence - based chatbot patient information on common retinal diseases using ChatGPT. *Acta Ophthalmol*, 101:829-831, 2023.
 10. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A : Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*, 29:721-732, 2023.
 11. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V : Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*, 2:E0000198, 2023.
 12. Frosolini A, Franz L, Benedetti S, Vaira LA, de Filipis C, Gennaro P, Marioni G, Gabriele G : Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol*, 280:5129-5133, 2023.
 13. Lum ZC : Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. *Clin Orthop Relat Res*, 481:1623-1630, 2023.
 14. Schwendicke F, Samek W, Krois J : Artificial Intelligence in Dentistry: Chances and Challenges. *J Dent Res*, 99:769-774, 2020.
 15. Taber KS : The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res Sci Educ*, 48:1273-1296, 2018.
 16. Korea Health Personnel Licensing Examination Institute : Results of the analysis of the 75th national examination of dentists in 2023. Available from URL: https://www.kuksiwon.or.kr/analysis/brd/m_91/view.do?seq=330&srchFr=&srchTo=&srchWord=&srchTtp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm= (Accessed on August 14, 2024).
 17. Strong E, DiGiammarino A, Weng Y, Kumar A, Hosamani P, Hom J, Chen JH : Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. *JAMA Intern Med*, 183:1028-1030, 2023.
 18. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, Cotofana S, Alfertshofer M : ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol*, 280:4271-4278, 2023.
 19. Alhaidry HM, Fatani B, Alrayes JO, Almanan AM, Alhaed NK : ChatGPT in Dentistry: A Comprehensive Review. *Cureus*, 15:E38317, 2023.
 20. Liu J, Wang C, Liu S : Utility of ChatGPT in Clinical Practice. *J Med Internet Res*, 25:E48568, 2023.
 21. Sallam M : ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)*, 11:887, 2023.
 22. Haupt CE, Marks M : AI-generated medical advice - GPT and beyond. *JAMA*, 329:1349-1350, 2023.
 23. Athaluri SA, Manthana SV, Kesapragada V, Yarlagadda V, Dave T, Duddumpudi RTS : Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, 15:E37432, 2023.
 24. Ramgopal S, Sanchez-Pinto LN, Horvat CM, Carroll MS, Luo Y, Florin TA : Artificial intelligence-based clinical decision support in pediatrics. *Pediatr Res*, 93:334-341, 2023.
 25. Liu Z, Zhang L, Wu Z, Yu X, Cao C, Dai H, Liu N, Liu J, Liu W, Li Q, Shen D, Li X, Zhu D, Liu T : Surviving

- ChatGPT in healthcare. *Front Radiol*, 3:1224682, 2024.
26. Dave T, Athaluri SA, Singh S : ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*, 6:1169595, 2023.
27. Stokel-Walker C : ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, 613:620-621, 2023.
28. Chatterjee J, Dethlefs N : This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy. *Patterns (N Y)*, 4:100676, 2023.
29. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, Moy L : ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*, 307:E230163, 2023.
30. Biswas S : ChatGPT and the Future of Medical Writing. *Radiology*, 307:E223312, 2023.
31. WHO : WHO guideline Recommendations on Digital Interventions for Health System Strengthening. World Health Organization, Geneva, 2019.
32. Balel Y : Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg*, 124:101471, 2023.
33. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, Hopkins BS, Dallas J, Pangal DJ, Cheok S, Nguyen VN, Mack WJ, Zada G : GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and Neurosurgery Residents on Neurosurgery Written Board-Like Questions. *World Neurosurg*, 179:E160-E165, 2023.