Journal of Dental Sciences

Volume 21 | Issue 2 Article 19

2026

Comparison of the performances of different updated generative artificial intelligence models on the Japanese National Dental Examination

Shuma Hamaguchi

Masakazu Hamada

Shunya Ikeda

Satoru Kusaka

Ryota Nomura

Journal of Dental Sciences 21 (2026) 833–837

Available online at https://jds.ads.org.tw/journal/

Digital Commons

journal homepage: https://jds.ads.org.tw/journal/

Original Article

Comparison of the performances of different updated generative artificial intelligence models on the Japanese National Dental Examination

Shuma Hamaguchi [a], Masakazu Hamada [b], Shunya Ikeda [a], Satoru Kusaka [c], Tatsuya Akitomo [a*], Ryota Nomura [a]

[a] *Department of Pediatric Dentistry, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan*
[b] *Department of Oral & Maxillofacial Oncology and Surgery, Graduate School of Dentistry, The University of Osaka, Suita, Osaka, Japan*
[c] *Department of Pediatric Dentistry, Hiroshima University Hospital, Hiroshima, Japan*

Received 3 September 2025; Final revision received 2 October 2025
Available online 1 April 2026

KEYWORDS
Artificial intelligence;
Dental education;
National dental examination

Abstract *Background/purpose:* Artificial intelligence (AI) has become widely used and applied in various fields. Although several studies have been conducted using generative AI for various qualification exams, to the best of our knowledge, none has focused on performance changes over time.
Materials and methods: In August 2025, ChatGPT 5, Gemini 2.5, Microsoft Copilot, and MediSearch were asked to answer compulsory questions from five years of the Japanese National Dental Examination. In 2024, we also conducted similar tests on other ChatGPT series and Gemini, and the scores were compared.
Results: In 2025, Copilot, Gemini, and MediSearch scored 80 % or higher, which was the passing standard, for all five years. Although ChatGPT 3.5 did not meet the passing standard for any of the five years, ChatGPT 4o mini and ChatGPT 5 exceeded it for two and three years, respectively. In addition, both Chat GPT's and Gemini's scores substantially improved over time and with each update.
Conclusion: This report suggests that generative AI is improving annually and adapting to the National Dental Examination. Although each AI model is suited to different fields, the trends may change over time. It is necessary to continue comparing and analyzing AI models and provide users with the latest information.

* Corresponding author. Department of Pediatric Dentistry, Graduate School of Biomedical and Health Sciences, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8553, Japan.
E-mail address: takitomo@hiroshima-u.ac.jp (T. Akitomo).

https://doi.org/10.1016/j.jds.2025.10.003

Introduction

In almost all fields, artificial intelligence (AI) is an emerging sector and mainly concentrates on how computers analyze data and mimic the human thought process.[1] The pace of AI integration into healthcare has accelerated, with rapid advancements in generative AI.[2] In dentistry, generative AI is mainly used for medical consultation and assistance, study assistance, and confirmation tests.[3] OpenAI developed the Chatbot Generative Pre-Trained Transformer (ChatGPT), an AI-based natural-language-processing (NLP) tool, which was launched in November 2022 and has been regularly updated ever since.[3]

Japanese National Dental Examination is an exam held once a year to obtain a national qualification as a dentist, and more than 3000 people take the exam every year. It consists of required basic topics called "compulsory questions", general dentistry, and each topic of dentistry. The compulsory questions are considered to constitute the basic knowledge and skills necessary to become a dentist, and the pass mark for compulsory questions is set at 80 % every year. The compulsory questions are classified into 13 basic topics and are asked in accordance with the prescribed question proportions, which were changed to 12 topics from 2023. Although we previously reported the performances of three generative AI models on the Japanese National Dental Examination, two did not meet the eligibility criteria.[4]

In August 2025, OpenAI released ChatGPT 5.[5] It is the best model yet at answering medical and healthcare questions, achieving a significantly higher score than any previous version on HealthBench.[5] This benchmark utilizes realistic scenarios and doctor-defined evaluation criteria. Additionally, chatbots such as MediSearch, which generate evidence-based answers from medical data, have also been developed and are gaining popularity.[6] Generative AI is continually improving, and our previous report noted that the upgraded version of generative AI has enhanced its adaptability to the National Examination.[4] However, there were few reports in the literature focusing on secular changes. To the best of our knowledge, this is the first report on not only the use of ChatGPT 5 for answering questions during a dental examination, but also examines one year of change in generative AI. The aim of this study was to clarify the usefulness of and changes in generative AI in dentistry by presenting the National Dental Examination to various generative AI models at different times and comparing their performance differences.

Materials and methods

Ethical approval was waived because this study investigated the usefulness of AI and did not involve human or animal subjects or patients' data. Many generative AIs still require a paid version to read images, and the paid version has different systems depending on the price; therefore, it is difficult to compare. In this study, we decided to exclude image problems and use free versions of generative AI to compare data from many generative AI models. This study was performed using slightly modified previously described methods.[4] Briefly, we used the official website of the Ministry of Health, Labour, and Welfare to extract the 400 compulsory questions from the National Dental Examinations from 2020 to 2024.[7] Of the 400 questions, the AI models attempted to answer 363, excluding questions that were difficult to reproduce in text because they contained diagrams and charts. All the AI models were free versions, and the questions were presented to ChatGPT 3.5 (OpenAI Global, San Francisco, CA, USA) and Gemini 1.5 (Google, Mountain View, CA, USA) in July 2024; ChatGPT 4o mini (OpenAI Global, San Francisco, CA, USA) in September 2024; and ChatGPT 5 (OpenAI Global, San Francisco, CA, USA), Gemini 2.5 (Google, Mountain View, CA, USA), Microsoft Copilot (Microsoft, Redmond, Washington, D.C., USA), and MediSearch in August 2025.[8]

All questions were presented to the AI in Japanese sentences, in accordance with the Japanese National Examination, and the first response was tallied as the answer. To ensure fairness, only questions that were answered by all the AI models were included in this study. For each model, scores were calculated as follows: (the total number of correct answers)/(the total number of questions included in this study) $\times$ 100 (%)). In addition, the AI models' performances were evaluated in each of the 13 categories of compulsory questions.[4] Statistical analyses were conducted using GraphPad Prism 9 (GraphPad Software Inc., La Jolla, CA, USA). For the five-year comparison and the 13 basic categories of the Japanese National Dental Examination, the Bonferroni correction was used to adjust for multiple comparisons between years and categories, respectively. Differences were considered as statistically significant at $P < 0.05$.

Results

Of the 363 questions, all seven AI models could evaluate 346, and the results are included in this study. Each score is shown in Table 1. In 2025, Copilot, Gemini 2.5, and MediSearch scored over 80 %, the passing standard, for all five years. Although ChatGPT 3.5 failed to meet the passing standard for all five years, ChatGPT 4o mini and ChatGPT 5 exceeded it for two and three years, respectively. Over the five years, the scores of the four AI models in 2025 consistently outperformed those of the three AI models in 2024. The average scores of all three AI models in 2024 were below 80 %, while those of all four AI models in 2025 were

Table 1 AI models' scores for five years on the Japanese National Dental Examination.

	Mean score (%)								
Year	ChatGPT 3.5	Gemini 1.5	ChatGPT 4o mini	**Average 2024**	ChatGPT 5	Copilot	Gemini 2.5	MediSearch	**Average 2025**
2020	62.1	72.7	75.8	**70.2 ± 7.2**	89.4	93.9	92.4	90.9	**91.7 ± 1.9***
2021	62.1	74.2	71.2	**69.2 ± 6.3**	75.8	80.3	83.3	84.8	**81.1 ± 4.0**
2022	65.3	76.0	85.3	**75.6 ± 10.0**	88.0	93.3	88.0	89.3	**89.7 ± 2.5**
2023	55.9	80.9	83.8	**73.5 ± 15.3**	86.8	91.2	92.6	86.8	**89.3 ± 3.0**
2024	59.2	69.0	74.6	**67.6 ± 7.8**	76.1	87.3	93.0	85.9	**85.6 ± 5.0**
Total	61.0	74.6	78.3	**71.3 ± 9.1**	83.2	89.3	89.9	87.6	**87.5 ± 3.0***

Average was expressed as the mean ± standard deviation. Bonferroni correction was used to adjust for multiple comparisons over the years: *$P < 0.05$ *versus* the average in 2024.

Table 2 Comparison of same AI models for five years on the Japanese National Dental Examination.

Year	ChatGPT 3.5	ChatGPT 5	**Difference**	ChatGPT 4o mini	ChatGPT 5	**Difference**	Gemini 1.5	Gemini 2.5	**Difference**
2020	62.1	89.4	**27.3**	75.8	89.4	**13.6**	72.7	92.4	**19.7**
2021	62.1	75.8	**13.7**	71.2	75.8	**4.6**	74.2	83.3	**9.1**
2022	65.3	88.0	**22.7**	85.3	88.0	**2.7**	76.0	88.0	**12.0**
2023	55.9	86.8	**30.9**	83.8	86.8	**3.0**	80.9	92.6	**11.7**
2024	59.2	76.1	**16.9**	74.6	76.1	**1.5**	69.0	93.0	**24.0**
Total	61.0	83.2	**22.2**	78.3	83.2	**4.9**	74.6	89.9	**15.3**

Difference was calculated as follows: (the score of AI model in 2025) - (the score of AI model in 2024).

above 80 % and showed significantly difference in the 2020 examination ($P < 0.05$).

Table 2 shows a comparison of the same AI models between 2024 and 2025. Comparing ChatGPT 3.5 and ChatGPT 5, the score of ChatGPT 5 showed an increase of 13.7 %–30.9 % over ChatGPT 3.5, with scores increasing by more than 10 % in all examinations. Compared to ChatGPT 4o mini, it also showed an increase of 1.5 %–13.6 %. On average over the five years, the score of ChatGPT 5 increased by 22.2 % compared to ChatGPT 3.5 and 4.9 % compared to ChatGPT 4o mini. In addition, Gemini also showed the improvement of 9.1 %–24.0 % between the Gemini 1.5 and 2.5 updates, with an average score of five years increase of 15.3 %.

Table 3 compares the performances of all the AI models on the 13 topics of compulsory questions. In all 13 categories, the average performances in 2025 were higher than those in 2024. All the AI models answered the "general education" questions correctly. Two categories showed statistically significant differences: "4. Prevention and health management/promotion" ($P < 0.05$) and "6. Human body development, growth, and aging" ($P < 0.01$).

In addition, in 2024, although the models scored over 90 % in only one category, they scored over 90 % in six categories in 2025. Most AI models have significantly improved their scores over time in many topics. However, even in 2025, all the models' scores remained at approximately 80 % in the "2. Society and dentistry", "4. Prevention and health management/promotion", "6. Human body development, growth, and aging" and "7. Etiology and pathogenesis of major diseases and disorders" categories.

Compared to ChatGPT 3.5, ChatGPT 5 demonstrated significant improvements in 12 categories (Table 4). Although the score of ChatGPT 5.0 was also higher than that of ChatGPT 4o mini in 7 basic topics, it decreased in 4 topics, such as "2. Society and dentistry", "5. Normal structure and function of the human body", "7. Etiology and pathogenesis of major diseases and disorders", and "10. Basics of examination and clinical diagnosis". In Gemini, it outperformed 2024's model in 11 topics.

Discussion

Currently, generative AI models, such as ChatGPT, are being used in various fields, including medicine. Previous studies on the National Dental Examination have shown that ChatGPT-4 outperformed ChatGPT-3.5 and ChatGPT-3.[3] In our previous study, although ChatGPT 4o mini achieved the highest overall score among the three generative AI models, it only met the passing criteria twice in five years.[4] In this study, we compared the performances of ChatGPT 5 and the other generative AI models to evaluate how the latest versions of the models adapted to the Japanese National Dental Examination.

Although the average scores for all three AI models administered in 2024 were below 80 %, those for all four AI models administered in 2025 were above 80 %, achieving the passing standard. In the 2020 examination, the average in 2025 shows statical significance compared to those in 2024 ($P < 0.05$). In addition, both Chat GPT's and Gemini's scores substantially improved over time and with each update, suggesting that the generative AI models' performances drastically improved, even in just one year.

Across the 13 categories, all seven AI models possessed a 100 % accuracy rate in "general education". In the

Table 3 AI models' scores for the 13 basic categories of the Japanese National Dental Examination.

Basic categories	Mean score (%)								
	ChatGPT 3.5	Gemini 1.5	ChatGPT 4o mini	**Average 2024**	ChatGPT 5	Copilot	Gemini 2.5	MediSearch	**Average 2025**
1. Medical ethics and dental professionalism	63.6	81.8	90.9	**78.8 ± 13.9**	90.9	100.0	100.0	100.0	**97.7 ± 4.6**
2. Society and dentistry	58.7	73.9	80.4	**71.0 ± 11.1**	78.3	78.3	82.6	76.1	**78.8 ± 2.7**
3. Team medical care	77.8	77.8	88.9	**81.5 ± 6.4**	100.0	100.0	88.9	100.0	**97.2 ± 5.6**
4. Prevention and health management/promotion	65.5	62.1	69.0	**65.5 ± 3.5**	79.3	86.2	79.3	86.2	**82.8 ± 4.0***
5. Normal structure and function of the human body	63.8	76.6	87.2	**75.9 ± 11.7**	80.9	91.5	93.6	91.5	**89.4 ± 5.7**
6. Human body development, growth, and aging	53.3	53.3	63.3	**56.7 ± 5.8**	80.0	83.3	80.0	80.0	**80.8 ± 1.7****
7. Etiology and pathogenesis of major diseases and disorders	60.0	72.0	76.0	**69.3 ± 8.3**	74.0	82.0	84.0	84.0	**81.0 ± 4.8**
8. Cardinal signs	47.1	58.8	76.5	**60.8 ± 14.8**	82.4	100.0	100.0	94.1	**94.1 ± 8.3**
9. Basics of medical examination	57.9	73.7	68.4	**66.7 ± 8.0**	84.2	84.2	94.7	84.2	**86.8 ± 5.3**
10. Basics of examination and clinical diagnosis	75.0	85.0	95.0	**85.0 ± 10.0**	90.0	100.0	100.0	90.0	**95.0 ± 5.8**
11. Emergency care	68.2	95.5	90.9	**84.8 ± 14.6**	100.0	100.0	90.9	95.5	**96.6 ± 4.4**
12. Fundamentals of treatment and basic techniques	51.2	85.4	68.3	**68.3 ± 17.1**	82.9	87.8	90.2	90.2	**87.8 ± 3.4**
13. General education	100.0	100.0	100.0	**100.0 ± 0.0**	100.0	100.0	100.0	100.0	**100.0 ± 0.0**

Average was expressed as the mean ± standard deviation. Bonferroni correction was used to adjust for multiple comparisons for categories: *$P < 0.05$, and **$P < 0.01$ *versus* the average in 2024.

Table 4 Comparison of same AI models for the 13 basic categories of the Japanese National Dental Examination.

Basic categories	ChatGPT 3.5	ChatGPT 5	**Difference**	ChatGPT 4o mini	ChatGPT 5	**Difference**	Gemini 1.5	Gemini 2.5	**Difference**
1	63.6	90.9	**27.3**	90.9	90.9	**0.0**	81.8	100.0	**18.2**
2	58.7	78.3	**19.6**	80.4	78.3	**−2.1**	73.9	82.6	**8.7**
3	77.8	100.0	**22.2**	88.9	100.0	**11.1**	77.8	88.9	**11.1**
4	65.5	79.3	**13.8**	69.0	79.3	**10.3**	62.1	79.3	**17.2**
5	63.8	80.9	**17.1**	87.2	80.9	**−6.3**	76.6	93.6	**17.0**
6	53.3	80.0	**26.7**	63.3	80.0	**16.7**	53.3	80.0	**26.7**
7	60.0	74.0	**14.0**	76.0	74.0	**−2.0**	72.0	84.0	**12.0**
8	47.1	82.4	**35.3**	76.5	82.4	**5.9**	58.8	100.0	**41.2**
9	57.9	84.2	**26.3**	68.4	84.2	**15.8**	73.7	94.7	**21.0**
10	75.0	90.0	**15.0**	95.0	90.0	**−5.0**	85.0	100.0	**15.0**
11	68.2	100.0	**31.8**	90.9	100.0	**9.1**	95.5	90.9	**−4.6**
12	51.2	82.9	**31.7**	68.3	82.9	**14.6**	85.4	90.2	**4.8**
13	100.0	100.0	**0.0**	100.0	100.0	**0.0**	100.0	100.0	**0.0**

Difference was calculated as follows: (the score of AI model in 2025) - (the score of AI model in 2024).

remaining 12 categories, the models' average scores in 2025 were higher than those in 2024. There were significant differences in the scores for "4. Prevention and Health Management/Promotion" and "6. Human Development, Growth, and Aging" between the 2025 and 2024 models ($P < 0.05$). However, the 2025 models' scores in the "2. Society and dentistry", "4. Prevention and health management/promotion", "6. Human body development, growth, and aging", and "7. Etiology and pathogenesis of major diseases and disorders" remained at approximately 80 %, indicating an ongoing challenge for generative AI models. Chatbot-based answers to national examinations are being developed not only in dentistry, but also in many other fields, including those for doctors, pharmacists,

nurses, and dental hygienists.[3] ChatGPT has been noted to have both advantages and disadvantages in these exams,[3] and the differences in scores between categories in our study may be related to such characteristics. Still, compared with those in 2024, AI models have shown improvements in these topics, suggesting that new model versions may eventually overcome these weaknesses and that future updates may lead all issues to achieve the passing standards.

Morishita et al. investigated the performance of GPT-4V's performance on the Japanese National Dental Examination including images and reported that the correct response rates were 57.1 % for compulsory questions, which showed the highest percentage of correct answers among 3 questions; compulsory questions, general question, and clinical practical question.[9] In addition, the score on basic knowledge questions was also higher than that on general questions in the Japanese National Nurse Examinations without images.[10] These past reports showed that the basic questions are easier for the AI to answer. Our present study showed an improvement in the score for compulsory questions only one year later, which suggests that most of today's generative AI models would be able to achieve the passing standards in the compulsory questions without images.

In this study, because free versions of the models were used to combine conditions, image-containing questions were excluded. Therefore, to accurately approximate the actual conditions of the National Dental Examination, additional research is required on all the questions, including those containing diagrams. In addition, ChatGPT is constantly being updated with new features and machine learning content, and the results from one year later may differ.[9] Multiple inputs, even the same question or image, may produce different results.[9] Therefore, an examination of score fluctuations under various conditions is important.

There have been many past reports evaluating the adaptability of various AI models; however, the highly acclaimed AI models vary depending on the report. Das et al. investigated the responses to clinically relevant questions related to ocular malignancies of three AI models (ChatGPT-4o, DeepSeek v3, and Gemini 2.0) and reported that ChatGPT showed the most balanced performance.[11] On the other hand, we previously investigated the responses of generative AI tools in clinical pediatric dentistry, and Microsoft Copilot showed the highest score from pediatric dental specialists among ChatGPT 3.5, Microsoft Copilot, and Gemini.[12] In the present study, the highest average over five years was Gemini 2.5 among 4 AI models in 2025. However, compared to the score increase from ChatGPT 3.5 and Gemini 1.5 in July 2024, ChatGPT outperformed Gemini's 15.3 % at 22.2 %. Each AI model may be suited to different fields; therefore, it is important to use it appropriately. In addition, with the rapid development of AI, the trends may change on time. It is necessary to continue comparing and analyzing AI models and provide users with the latest information.

Generative AI has ethical concerns.[13] Although verification of the knowledge of generative AI models through questions is important, ethical issues must be considered when using generative AI in clinical applications involving patients' information. Although the Japanese National Dental Examination in the present study was available for free on the Website, there is a risk of patients' information leakage in a clinical situation. Medical professionals must understand these limitations and promote the appropriate use of generative AI.

Declaration of competing interests

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

Acknowledgments

This work was not supported by any organizations.

References

1. Shanbhogue MH, Thirumaleshwar S, Tegginamath PK, Somareddy HK. Artificial intelligence in pharmaceutical field - a critical review. *Curr Drug Deliv* 2021;18:1456–66.
2. Soroush A, Giuffrè M, Chung S, Shung DL. Generative artificial intelligence in clinical medicine and impact on gastroenterology. *Gastroenterology* 2025;169:502–17.
3. Hamada M, Kikuchi S, Akitomo T, Kusaka S, Iwamoto Y, Nomura R. Applications and potential of ChatGPT in dentistry: scoping review of research perspectives. *J Dent Sci* 2025 (in press).
4. Akitomo T, Hamada M, Tsuge Y, et al. Artificial intelligence's performance on the Japanese National Dental Examination. *Cureus* 2024;16:e73103.
5. OpenAI. *Introducing GPT-5*. Available at: https://openai.com/index/introducing-gpt-5/. [Accessed 27 August 2025].
6. Akpınar H. Comparison of responses from different artificial intelligence-powered chatbots regarding the all-on-four dental implant concept. *BMC Oral Health* 2025;25:922.
7. Ministry of Health, Labor and Welfare. *Ministry of Health, Labor and Welfare website*. Available at: https://www.mhlw.go.jp/index.html. [Accessed 27 August 2025].
8. *Medisearch*. Available at: https://medisearch.io/ja. [Accessed 26 September 2025].
9. Morishita M, Fukuda H, Muraoka K, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. *J Dent Sci* 2023;19:1595–600.
10. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan: evaluation study. *JMIR Nurs* 2023;6:e47305.
11. Das D, Narayan A, Mishra V, et al. AI Chatbots in answering questions related to ocular oncology: a comparative study between DeepSeek v3, ChatGPT-4o, and Gemini 2.0. *Cureus* 2025;17:e90773.
12. Kusaka S, Akitomo T, Hamada M, et al. Usefulness of generative artificial intelligence (AI) tools in pediatric dentistry. *Diagnostics* 2024;14:2818.
13. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res* 2023;25:e48009.