Journal of Dental Sciences 20 (2025) 895—900

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Journal of Dental Sciences

Original Article

Can a large language model create acceptable dental board-style examination questions? A cross-sectional prospective study

Hak-Sun Kim [a], Gyu-Tae Kim [b*]

[a] *Department of Oral and Maxillofacial Radiology, Kyung Hee University Dental Hospital, Seoul, Republic of Korea*
[b] *Department of Oral and Maxillofacial Radiology, College of Dentistry, Kyung Hee University, Seoul, Republic of Korea*

Received 16 July 2024; Final revision received 28 August 2024
Available online 11 September 2024

KEYWORDS
Dental education;
Examination questions;
Professional competence;
Artificial intelligence;
Natural language processing

Abstract *Background/purpose:* Numerous studies have shown that large language models (LLMs) can score above the passing grade on various board examinations. Therefore, this study aimed to evaluate national dental board-style examination questions created by an LLM versus those created by human experts using item analysis.
Materials and methods: This study was conducted in June 2024 and included senior dental students ($n = 30$) who participated voluntarily. An LLM, ChatGPT 4o, was used to generate 44 national dental board-style examination questions based on textbook content. Twenty questions for the LLM set were randomly selected after removing false questions. Two experts created another set of 20 questions based on the same content and in the same style as the LLM. Participating students simultaneously answered a total of 40 questions divided into two sets using Google Forms in the classroom. The responses were analyzed to assess difficulty, discrimination index, and distractor efficiency. Statistical comparisons were performed using the Wilcoxon signed rank test or linear-by-linear association test, with a confidence level of 95%.
Results: The response rate was 100%. The median difficulty indices of the LLM and human set were 55.00% and 50.00%, both within the range of ″excellent″ range. The median discrimination indices were 0.29 for the LLM set and 0.14 for the human set. Both sets had a median distractor efficiency of 80.00%. The differences in all criteria were not statistically significant ($P > 0.050$).

* Corresponding author. Department of Oral and Maxillofacial Radiology, College of Dentistry, Kyung Hee University, 26 Kyungheedae-ro, Dongdaemun-gu, Seoul, 02447, Republic of Korea.
E-mail address: latinum.omfr@khu.ac.kr (G.-T. Kim).

https://doi.org/10.1016/j.jds.2024.08.020

Conclusion: The LLM can create national board-style examination questions of equivalent quality to those created by human experts.

Introduction

Numerous artificial intelligence (AI) models have been developed using deep learning and studied in a wide range of fields.[1–5] Recently, the growth of large language models (LLMs) has significantly impacted daily life and research.[6–19] LLMs are specialized in processing natural language, including creating, editing, and summarizing text.[6–8] These models are continuously learned from new inputs provided by users worldwide. While LLMs are capable of answering knowledge-based questions, they sometimes hallucinate, providing incorrect information instead of facts when the models lack specific knowledge.[20]

In the field of medicine and dentistry, discussions about effective integration of AI into real practice and education have been ongoing.[1,2,4,8–17] Most commonly, deep learning models have been primarily developed and tested for the detection, classification, and prediction of pathologies or medical image generation.[1,2,4] In education, language models can serve as virtual patients or learning resources.[18,19,21,22] Moreover, several studies have been conducted to test the ability of LLMs to answer questions.[10–17] These studies have demonstrated that LLMs often score above the passing grade for board-style examination questions across various regions and fields.[10–17]

While LLMs' strengths include manipulating natural languages, many studies have examined their test-taking capabilities rather than their proficiency in generating test questions specifically in dentistry. Therefore, this study aimed to evaluate board-style examination questions generated by an LLM and compare them to those written by human experts using item analysis.

Materials and methods

This prospective cross-sectional study was conducted at the Department of Oral and Maxillofacial Radiology, Kyung Hee University Dental Hospital. Ethical approval was obtained from the Institutional Review Board (IRB) of the Kyung Hee University Dental Hospital (IRB No: KH-DT24002) on May 30, 2024. All participants completed the questionnaire independently and provided informed consent. The study procedure following ethical approval is shown in Fig. 1.

Board-style examination questions

Examination questions were based on a summarized version of the most frequently used oral and maxillofacial radiology textbook in the Republic of Korea. ChatGPT (ChatGPT-4o, available at chatgpt.com, OpenAI, San Francisco, CA, USA) was used to generate 44 examination questions from 22 chapters, formatted as multiple-choice questions with five options designed to simulate a national dental board examination (Table 1). The questions were evenly distributed across the textbook content. Questions generated by ChatGPT were screened for potential errors or inaccuracies, resulting in the elimination of three false questions. Subsequently, 20 questions were randomly selected for inclusion in the "LLM set."

Human experts, two oral and maxillofacial radiology specialists with more than 8 years of experience in dental student education, created 20 examination questions for the "human set" in the same format and based on the same content used for the ChatGPT-generated questions. These specialists were blinded to the questions generated by ChatGPT. The questions were cross-checked for errors and

Table 1 An example prompt used to generate board examination-style questions for the LLM set.

Item for LLM	Content
Prompt to generate questions	Create two multiple-choice questions with five options each, and one correct answer, based on the contents of the attached file, in a style similar to dental board-examinations. Show the correct answer at the end of each question.
Attached file	Chapter 31. Interpretation of craniofacial anomalies

LLM, large language model.

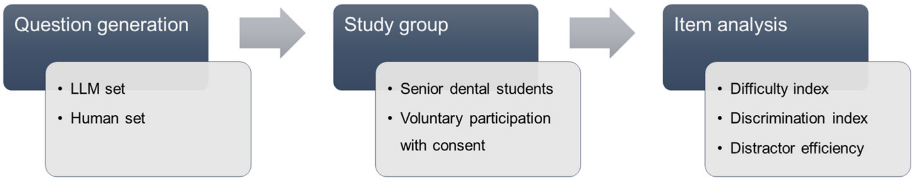


Figure 1 Schematic diagram of the overall process of this study. LLM, large language model.

corrected based on expert consensus. Examples of the questions from each set are shown in Fig. 2.

Study group

In total, 30 senior undergraduate dental students, who completed their first semester, voluntarily participated in the study. By the end of the first semester, the students had finished learning all the contents of the oral and maxillofacial radiology textbooks. In the second (final) semester, students would focus on reviewing and consolidating their knowledge in preparation for the national dental board examination. Participating in this study allowed students to review the content and answer new questions as part of their board examination preparation process.

Under the agreement of the participants, they were assembled in a classroom to answer 40 questions divided into two sets (20 each from the LLM and human sets) within a 30-min timeframe, an equivalent amount of time given during board examinations. Questions were distributed and answered using Google Forms (Alphabet Inc., Mountain View, CA, USA). Upon completing the question sets, students were asked to speculate which set was generated by ChatGPT and which by humans.

Item analysis

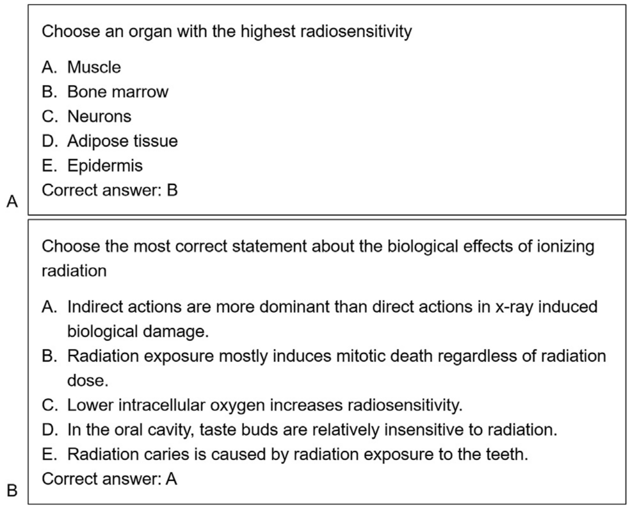
The results of both sets were analyzed using terms of item analysis, which included assessing the difficulty index, discrimination index, and distractor efficiency.[23,24] The difficulty index represents the proportion of students who chose the correct option. The discrimination index measures the difference in the number of students who chose the correct option between the upper 27% and the lower 27% of performers.

$$P = 100 * R/T \tag{1}$$

where P is the difficulty index of an item, R is the number of correct responses, and T is the total number of respondents.

$$D = (UG - LG)/n \tag{2}$$

where D is the discrimination index of an item, UG and LG are the number of correct responses to an item from upper and lower 27% students, respectively, and n is the number of the students in the larger group between upper and lower 27%.



A

Choose an organ with the highest radiosensitivity

A. Muscle
B. Bone marrow
C. Neurons
D. Adipose tissue
E. Epidermis
Correct answer: B

B

Choose the most correct statement about the biological effects of ionizing radiation

A. Indirect actions are more dominant than direct actions in x-ray induced biological damage.
B. Radiation exposure mostly induces mitotic death regardless of radiation dose.
C. Lower intracellular oxygen increases radiosensitivity.
D. In the oral cavity, taste buds are relatively insensitive to radiation.
E. Radiation caries is caused by radiation exposure to the teeth.
Correct answer: A

Figure 2 Example questions based on knowledge of the biological effects of ionizing radiation. (A) Large language model and (B) human sets.

Difficulty index values of 20% to 90% are considered good and acceptable, while those within the range of 40% to 60% are deemed excellent. Items with values less than 20% and above 90% need modification.[23] According to Ebel et al., discrimination indices can be categorized as follows: ≥0.40 indicates very good discrimination; between 0.30 and 0.39 suggests minimal or no modification needed; between 0.20 and 0.29 indicates marginal discrimination and needs modification; ≤0.19 indicates that the item must be modified or reconsidered.[25] Since this study evaluated the items only once, no further modifications were made to the questions in either set.

A non-functioning distractor refers to an option that was chosen by less than 5% of the students. Distractor efficiency is defined as the ratio of functioning distractors to the total number of distractors.

$$DE = 100 * (N - NFD)/N \tag{3}$$

where DE is the distractor efficiency of an item, N is the total number of options for each item, and NFD is the number of non-functioning distractors. Because all questions contained five options, N was 5.

Statistical analysis

The Wilcoxon signed-rank test was used to compare the difficulty index, discrimination index, and distractor efficiency values between the LLM and human sets. The presence of non-functioning distractors in both sets was assessed using a linear-by-linear association test. The significance level was set at 95% for the analyses. Data were analyzed using IBM SPSS Statistics for Windows, Version 29 (IBM Corp., Armonk, NY, USA).

Results

The median values for difficulty index, discrimination index, and distractor efficiency in the LLM and human sets were 55.00%, 0.29, and 80.00%, and 50.00%, 0.14, and 80.00%, respectively. None of these differences were statistically significant ($P > 0.050$). Detailed results of both sets and their statistical comparisons are presented in Table 2. Additionally, approximately 63.3% (19 out of 30) of the students correctly identified the origin of the question sets.

The median difficulty index values for both the LLM and the human sets were considered excellent. The relationship between the difficulty and discrimination indices for each item in both sets is shown in Fig. 3. Items with difficulty index values between 30% and 80% exhibited the highest discrimination index values. Conversely, items with a

Table 2 Median and quartile values of difficulty and discrimination index, and distraction efficiency of the human and LLM sets.

	LLM set	Human set	*P*-value
Difficulty index (%)	55.00 (35.83—68.33)	50.00 (9.17—63.33)	0.211
Discrimination index	0.29 (0.14—0.57)	0.14 (0.00—0.29)	0.064
Distractor efficiency (%)	80.00 (60.00—80.00)	80.00 (60.00—80.00)	0.776

LLM, large language model.

difficulty index below 30% or above 80% showed incrementally decreasing discrimination power. Among the questions in the LLM and human sets, two and eight questions, respectively, had difficulty indices below 20% or above 90%, thus requiring modification. Although the median discrimination index values of each LLM and human sets fell into different categories, both sets required modification for a significant number of questions (11 and 16 questions, respectively), and the differences were statistically insignificant. The distractor efficiency of the LLM and the human sets showed an equivalent level, and the number of non-functioning distractors also showed no statistical difference ($P > 0.050$), with both sets having a median of 1. The distribution is shown in Fig. 4.

Discussion

This study aimed to evaluate board-style examination questions fabricated by LLMs and compare them to those written by human experts using item analysis. The median difficulty and discrimination index values of the LLM and human sets showed slight differences that were not statistically significant. In addition, distractor efficiency was equivalent between the two sets. Of the 44 questions created by ChatGPT, three were based on hallucinations.

Although difficulty and discrimination index values did not differ significantly, the number of questions subjected to modification (difficulty index <20% or >90%; discrimination index <0.3) was greater for the human set. In terms of difficulty index values, the human set had a larger difference between the 25 percentile value (9.17%) and median (50.00%) than the LLM set (35.83% and 55.00%, respectively). For discrimination index values, the LLM set had a higher median value of 0.29 (vs 0.14) and a larger difference between the median and 75 percentile values than the human set (0.28 > 0.15). This resulted in a greater number of items in the LLM set with a discrimination index greater than 0.3.

Nevertheless, the distractor efficiency was almost identical between the LLM and human sets. Since more efficient distractors constitute difficult and discriminating items,[26] this finding may be construed by the range of difficulty and discrimination indices observed in both sets. The relationship between difficulty and discrimination indices revealed that the minimum and maximum values of both indices were very similar between the two sets, despite various combinations of these indices.

The present study assessed the test-making skills of LLM in dentistry, while most studies in dentistry and medicine have mainly focused on evaluating their test-taking skills.[10—17] The LLMs demonstrated passing scores in

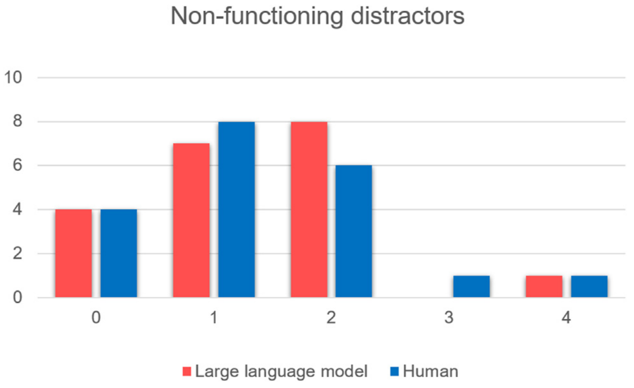


Figure 4 Number of non-functioning distractors in large language model and human sets.

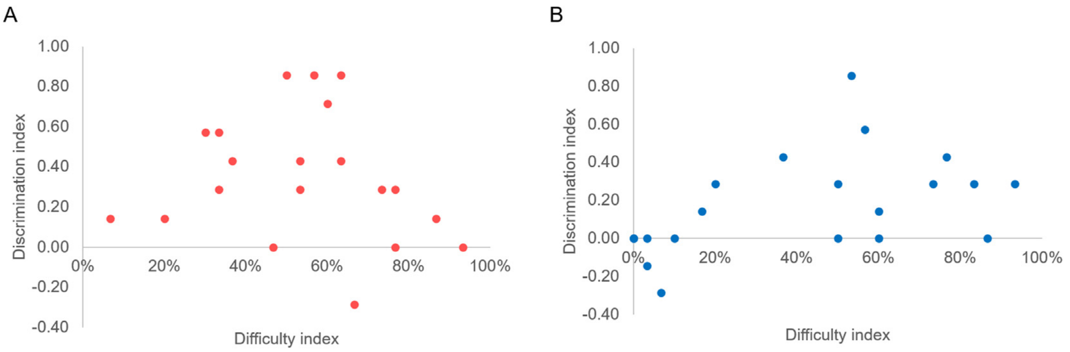


Figure 3 Plots of discrimination indices (Y axis) against difficulty indices (X-axis). (A) Large language model set and (B) human sets.

board-style examinations across various fields, including national medical boards, neurology, radiology, public health areas, and national dental and dental hygienist boards in various regions. In most studies comparing LLMs, the best result was observed with ChatGPT-4, which was also used in this study. One previous study also evaluated neurophysiology examination questions generated by ChatGPT and human experts.[9] The findings were consistent with those of this study, showing that questions generated by ChatGPT showed similar quality to those written by experts.[9] In that study, students were also able to correctly recognize the question writer slightly better than by chance (57%), which is comparable to the 63% accuracy observed in this study.[9]

Nevertheless, this study is unique in that it specifically explored the hallucinations produced by the LLM, despite existing literature with a similar objective.[9] Three questions were eliminated because what they inquired about, and the options provided, included false statements not mentioned in the given textbook summary. For example, one of those questions addressed characteristic radiologic features of malignant salivary gland tumors, which were not stated in the source material. Also, none of its options specified the imaging modality, such as ‘irregular borders and mixed radiolucency’ even though the textbook summary explicitly indicated the imaging features for each modality. ChatGPT falsely created statements and omitted crucial information, resulting in incorrect and unfounded questions.

Beyond the educational performance of artificial intelligence models in test-related tasks,[9—17] their practical application in dental and medical education have been thoroughly investigated.[18,19,21,22] Artificial intelligence chatbots have been implemented as virtual patients or clinical guidance for students and patients.[18,19,21,22,27,28] Students have generally satisfied with the chatbots,[19,21,22] but they also responded that chatbots could not replace human interaction.[22,29] In addition, the LLMs have been scrutinized for clinical application, for instance, answering patients’ clinical questions.[27,28] The results evaluated by dental experts indicated that the LLMs’ responses were acceptable, with the latest version found to be as reasonable as those provided by experts.[27,28]

This study has several limitations. First, the number of participants was relatively small.[9] Further studies with a larger number of students would yield more robust results. Second, the questions consisted only of text, without any accompanying images. For the LLM to create questions using radiographic images, the input would need to include sampled radiographic images. However, exposing patient radiographs to a third party raises a serious ethical issue. Despite this challenge, the absence of radiographic images remains a limitation, particularly for a study focused on oral and maxillofacial radiology. Additionally, while taking full advantage of LLMs to generate examination questions can be beneficial, it should be considered with caution. Although it is time-consuming for teachers to create new questions on the same content each year, the potential for hallucinations in LLMs-generated questions is inevitable.[20] Thus, LLMs’ examination questions require human surveillance and confirmation before practical use.

In conclusion, ChatGPT can generate dental board-style examination questions of equivalent quality to those generated by human experts. However, direct application in actual examinations should be carefully conducted, considering the potential for hallucinations.

Declaration of competing interest

The author has no conflicts of interest relevant to this article.

Acknowledgments

None.

References

1. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342—50.
2. Lee C, Ha EG, Choi YJ, Jeon KJ, Han SS. Synthesis of T2-weighted images from proton density images using a generative adversarial network in a temporomandibular joint magnetic resonance imaging protocol. *Imaging Sci Dent* 2022;52:393—8.
3. Lampinen AK, Dasgupta I, Chan SCY, et al. Language models show human-like content effects on reasoning tasks. *arXiv* 2022. 2207.07051.
4. Kim HS, Ha EG, Kim YH, Jeon KJ, Lee C, Han SS. Transfer learning in a deep convolutional neural network for implant fixture classification: a pilot study. *Imaging Sci Dent* 2022;52:219—24.
5. Jamwal A, Agrawal R, Sharma M. Deep learning for manufacturing sustainability: models, applications in Industry 4.0 and implications. *Int J Inf Manag Data Insights* 2022;2:100107.
6. Naveed H, Khan AU, Qiu S, et al. A comprehensive overview of large language models. *arXiv* 2023. 2307.06435.
7. Kasneci E, Sessler K, Kuchemann S, et al. ChatGPT for good: on opportunities and challenges of large language models for education. *Learn Indiv Differ* 2023;103:102274.
8. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology* 2023;307:e230163.
9. Laupichler MC, Rother JF, Kadow ICG, Ahmadi S, Raupach T. Large language models in medical education: comparing ChatGPT- to human-generated exam questions. *Acad Med* 2024;99:508—12.
10. Toyama Y, Harigai A, Abe M, et al. Performance evaluation of ChatGPT, GPT-4, and bard on the official board examination of the Japan radiology society. *Jpn J Radiol* 2024;42:201—7.
11. Güneş YC, Cesur T. Assessing the diagnostic performance of large language models with European Diploma in Musculoskeletal Radiology (EDiMSK) examination sample questions. *Jpn J Radiol* 2024;42:673—4.
12. Davies NP, Wilson R, Winder MS, et al. ChatGPT sits the DFPH exam: large language model performance and potential to support public health learning. *BMC Med Educ* 2024;24:57.
13. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the national board of medical examiners sample questions. *Cureus* 2024;16:e55991.
14. Javan R, Kim T, Mostaghni N, Sarin S. ChatGPT’s potential role in interventional radiology. *Cardiovasc Intervent Radiol* 2023;46:821—2.

15. Güneş YC, Cesur T. Diagnostic accuracy of large language models in the European Board of Interventional Radiology Examination (EBIR) sample questions. *Cardiovasc Intervent Radiol* 2024;47:836—7.
16. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? *Dentomaxillofacial Radiol* 2024:twae021.
17. Yamaguchi S, Morishita M, Rukuda H, et al. Evaluating the efficacy of leading large language modes in the Japanese national dental hygienist examination: a comparative analysis of ChatGPT, Bard, and Bing Chat. J Dent Sci (in press).
18. Li Y, Zeng C, Zhong J, Zhang R, Zhang M, Zou L. Leveraging large language model as simulated patients for clinical education. *arXiv* 2024. 2404.13066.
19. Borg A, Parodis I, Skantze G. Creating virtual patients using robots and large language models: a preliminary study with medical students. In: *In Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*; 2024:273—7.
20. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. *arXiv* 2024;2401:11817.
21. Wu D, Xian Y, Wu X, et al. Artificial intelligence-tutoring problem-based learning in ophthalmology clerkship. *Ann Transl Med* 2020;8:700.
22. Fang Q, Reynaldi R, Araminta AS, et al. Artificial intelligence (AI)-driven dental education: exploring the role of chatbots in a clinical learning environment. J Prosthet Dent (in press).
23. Quaigrain K, Arhin AK. Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Educ* 2017;4:1301013.
24. Iñarrairaegui M, Fernández-Ros N, Lucena F, et al. Evaluation of the quality of multiple-choice questions according to the students' academic level. *BMC Med Educ* 2022;22:779.
25. Ebel RL, Frisbie DA. *Essentials of educational measurement*, 5th ed. Englewood Cliffs: Prentice-Hall International Inc., 1991:220—40.
26. Rezigalla AA, Eleragi AMESA, Elhussein AB, et al. Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Med Educ* 2024;24:445.
27. Lv X, Zhang X, Li Y, Ding X, Lai H, Shi J. Leveraging large language models for improved patient access and self-management: assessor-blinded comparison between expert- and AI-generated content. *J Med Internet Res* 2024;26:e55847.
28. Batool I, Naved N, Kazmi SMR, Umer F. Leveraging large language models in the delivery of post-operative dental care: a comparison between an embedded GPT model and ChatGPT. *BDJ Open* 2024;10:48.
29. Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. *Eur J Dent Educ* 2024;00:1—12.