Journal of Dental Sciences and Education

Original Article

DOI: 10.51271/JDSE-0052

Performance comparison of artificial intelligence-based language models on oral pathology questions in dental education

Katibe Tuğçe Temur

Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Niğde Ömer Halisdemir University, Niğde, Turkiye

Cite this article as: Temur KT. Performance comparison of artificial intelligence-based language models on oral pathology questions in dental education. *J Dent Sci Educ.* 2025;3(2):31-34.

Received: 03.05.2025 • **Accepted**: 30.05.2025 • **Published**: 31.05.2025

ABSTRACT

Aims: This study aimed to comparatively evaluate the diagnostic accuracy levels of current large language models-including ChatGPT-3.5, ChatGPT-4, Gemini, and Deepseek-through multiple-choice questions related to oral pathology, a core subject in undergraduate dental education.

Methods: A total of 30 multiple-choice questions covering oral pathology topics within the dental curriculum were prepared by a domain expert using current textbooks and course materials. These questions were applied in a text-based format to four LLMs (ChatGPT-3.5, ChatGPT-4, Gemini, Deepseek), and only the first responses were recorded without any prompting or correction. Diagnostic accuracy was assessed based on the percentage of correct answers, and inter-model agreement was evaluated using the Kappa statistic. A p-value <0.05 was considered statistically significant.

Results: ChatGPT-3.5 and ChatGPT-4.0 each achieved an accuracy rate of 93.33%, with corresponding kappa values of κ=0.951 and κ=0.927, respectively. Gemini reached 86.67% accuracy (κ=0.879), while Deepseek achieved 80% accuracy (κ=0.855). All models showed statistically significant agreement with the reference answers (p=0.000).

Conclusion: This study demonstrated that ChatGPT-3.5 and ChatGPT-4.0 outperformed other models in terms of diagnostic accuracy and agreement in answering fundamental oral pathology questions. Although all models showed significant concordance, ChatGPT-based models stood out in terms of overall accuracy. However, considering potential error margins, these models should be used only as complementary tools in dental education and should not replace primary sources of knowledge acquisition for students.

Keywords: Artificial intelligence, dentistry, oral pathology, education

INTRODUCTION

Artificial intelligence (AI) has increasingly become a prominent field, particularly through applications such as large language models (LLMs) capable of generating human-like conversations. In recent years, AI technologies have rapidly expanded across various sectors worldwide, with their areas of application growing exponentially.[1,2]

AI systems especially those based on deep learning and artificial neural networks are driving transformations in many industries, most notably in healthcare. These systems are capable of processing large datasets, analyzing relationships between concepts, and generating well-reasoned, meaningful responses.[3]

One of the most notable advancements in AI is ChatGPT (Chat Generative Pre-trained Transformer), developed by OpenAI and publicly released on November 30, 2022.[4] ChatGPT-3.5 gained recognition as the first widely accessible large language model, followed by the emergence of other advanced models such as Claude (Anthropic, USA), Gemini (Google DeepMind, USA), Copilot (Microsoft in collaboration with OpenAI, USA), and DeepSeek (DeepSeek-Vision, China).

Among these, DeepSeek, version R1 stands out as an open-source, cost-effective alternative to models like OpenAI's, offering high reasoning performance and greater accessibility. Due to its promising capabilities and affordability, it has attracted growing interest among researchers and is expected to be more widely integrated into scientific workflows.[5]

In recent years, there has been a marked increase in interest regarding the potential contributions of AI technologies in dental education, leading to a growing body of scientific research in this domain. In particular, large language models such as ChatGPT have been the focus of numerous studies examining their utility in dental education. Similarly, other LLMs including Gemini and DeepSeek have been evaluated for their potential role as complementary tools in educational contexts, with increasing representation in the literature.[6-11] However, the number of studies that comparatively evaluate ChatGPT and similar AI-based models in the context of specific curriculum components particularly clinically significant subjects such as oral pathology remains limited in the current literature.[12,13]

Corresponding Author: Katibe Tuğçe Temur, tugce.uzmez@hotmail.com

Therefore, the aim of this study is to comparatively assess the diagnostic accuracy of current large language models namely ChatGPT-3.5, ChatGPT-4, Gemini, and DeepSeek using multiple choice questions related to oral pathology, a subject included in undergraduate dental curricula. This study further seeks to provide objective data on the potential applicability of these models in educational settings and to contribute a subject specific and domain focused perspective to the literature.

METHODS

Ethics Statement

This study does not involve any patient data or human participation. The questions were prepared based on course materials and were solely intended to assess knowledge level. Therefore, ethical approval was not required.[12]

Content and Topics of the Questions

In this study, a total of 30 multiple-choice questions were prepared, covering topics related to undergraduate-level oral pathology and oral mucosal lesions. The questions were designed to assess students' theoretical knowledge and were categorized under headings such as aphthous ulcers, mechanisms of white lesion formation, premalignant and malignant lesions, reactive lesions, tongue pathologies, HIV-related oral findings, histopathological concepts, infectious diseases, and pigmentation disorders related to systemic factors.

The question content was developed by a specialist in oral and maxillofacial radiology, based on current core textbooks and faculty lecture notes included in the dental curriculum. Designed with conceptual integrity that can be linked to clinical practice, the questions aimed not only to assess knowledge but also to evaluate clinical discernment, diagnostic classification skills, and critical thinking.

Testing and Comparative Evaluation of AI Models

In this study, the diagnostic performance of four advanced AI-based language models was comparatively evaluated using a standardized multiple-choice question set. All questions were presented in text-only format, and no visual content (e.g., radiographs, diagrams) was used.

The AI language models evaluated in this study were;

- ChatGPT-3.5 (OpenAI)
- ChatGPT-4 (OpenAI)
- Gemini (Google DeepMind)
- Deepseek (open-source large language model)

All models were assessed solely through text input to ensure a fair and standardized comparison.

The evaluation process was conducted as follows;

All questions were manually entered into the official web interfaces of each model by the same researcher using a single device.

Before the queries were initiated, cookies and browsing history were cleared, and all models were accessed on April 23, 2025.

Only the first response of each model was recorded, without any further input, prompting, or correction.

The responses of the models were compared against a predefined answer key, and evaluation was based solely on the accuracy of responses.

Statistical Analysis

Categorical variables were presented as frequencies (n) and percentages (%). The agreement between responses provided by different AI models was assessed using the Kappa test, and Kappa coefficients were reported. Statistical analyses were performed using IBM SPSS Statistics for Windows, Version 23.0 (IBM Corp., Released 2017). A p-value of <0.05 was considered statistically significant.

RESULTS

Following the evaluation comprising 30 multiple-choice questions, both ChatGPT-3.5 and ChatGPT-4.0 models demonstrated the highest diagnostic performance, each achieving 28 correct responses and 2 incorrect ones, corresponding to an accuracy rate of 93.33%. The Gemini model yielded 26 correct and 4 incorrect answers, with an overall accuracy of 86.67%, while the Deepseek model provided 24 correct and 6 incorrect responses, resulting in an accuracy rate of 80.00%. **Table 1** presents a comparative overview of the numerical and percentage-based distribution of correct responses across the different artificial intelligence-based language models in relation to undergraduate-level oral pathology questions.

Table 1. Frequency and percentage of predictions

Prediction tool	n	%
ChatGPT-3.5 (C/I)	28/2	93.33 /6.67
ChatGPT-4.0 (C/I)	28/2	93.33 /6.67
Gemini (C/I)	26/4	86.67/13.33
Deepseek (C/I)	24/6	80/20

C: Correct, I: Incorrect

ChatGPT-3.5 and ChatGPT-4.0 models achieved the highest performance with an accuracy rate of 93.33%. These models demonstrated statistically significant and high-level agreement, with kappa values of κ=0.951 and κ=0.927, respectively (p=0.000). The Gemini model reached an accuracy rate of 86.67% with a kappa value of κ = 0.879, indicating a high level of consistency with the reference answers (p=0.000). The Deepseek model exhibited the lowest accuracy rate at 80.00%, yet its kappa value was κ=0.855, demonstrating significant and high-level agreement with the reference responses (p=0.000). The calculated p-values for all AI models were 0.000, indicating that their agreement with the reference answers was statistically significant (**Table 2**).

Table 2. Comparison of artificial intelligence prediction models based on accuracy and consistency (kappa) metrics

Prediction tool	Accuracy (%)	Kappa value	p-value
ChatGPT-3.5	93.33	0.951	0.000
ChatGPT-4.0	93.33	0.927	0.000
Gemini	86.67	0.879	0.000
Deepseek	80.00	0.855	0.000

DISCUSSION

Although studies on the potential use of AI-based language models in health education are increasingly common,

comparative assessments targeting specific curricular subfields such as oral pathology are still limited. This study aims to contribute to the existing literature in this area.

In this context, our study demonstrated that the ChatGPT-3.5 and ChatGPT-4.o models achieved the highest accuracy rates (93.33%; κ=0.951 and 0.927; p = 0.000). Conversely, the Gemini model (86.67%; κ=0.879) and the Deepseek model (80.00%; κ=0.855) exhibited lower accuracy. Although all models showed statistically significant agreement with the reference answers, ChatGPT-based models demonstrated superior overall accuracy compared to others.

Similarly, Yılmaz et al.[12] compared eight different AI models, reporting that the ChatGPT-o.1 model achieved the highest accuracy (96%) for both knowledge-based and case-based questions. Models such as Claude 3.5, Gemini 2, and Deepseek showed significantly lower performance. These findings indicate that ChatGPT-based models perform more consistently, especially for clinical knowledge-based questions. On the other hand, another study reported that the ChatGPT-4.o model achieved the highest success, with an accuracy rate of 90.0%. In our current study, general accuracy rates were evaluated using undergraduate-level questions without clinical scenarios or specific categorization.

Additionally, Kınıkoğlu et al.,[6] analyzing data from the Dentistry Specialty Examination (DUS) for 2020-2021, found that ChatGPT-o.1 and Gemini 2.0 Advanced models achieved accuracy rates of 97.46% and 97.90%, respectively, significantly surpassing models such as ChatGPT-4.o and Gemini 1.5 Pro (p=0.0002). These findings highlight the importance of carefully evaluating performance differences among various model versions.

In contrast, Eraslan et al.,[9] in their comparative assessment within prosthodontics, found that the Copilot model had the highest accuracy (73%). This was followed by Gemini (63.5%), ChatGPT-3.5 (61.1%), Claude Pro (57.9%), and Perplexity (54.8%). Copilot significantly outperformed Perplexity (p=0.035), suggesting that model performance may vary depending on the subject area.

On the other hand, Wójcik et al.[10] reported that Claude could achieve high accuracy rates in general medical fields and maintained consistency across multilingual use; however, its moderate performance in this study indicates it may not be sufficiently optimized for specific fields such as dentistry.

Likewise, Tosun et al.[14] compared ChatGPT-3.5, ChatGPT-4, Gemini, Gemini Advanced, Claude AI, Microsoft Copilot, and Smodin AI in responding to DUS prosthodontics questions. The ChatGPT-4 model exhibited the highest accuracy (75.8%), while Gemini AI showed the lowest (46.1%). Although ChatGPT models were more successful in knowledge-based questions, all models showed limitations in case-based scenarios, suggesting the need for further optimization for complex clinical situations.

Lastly, a similar trend was observed in a study by Künz et al.[15] in restorative dentistry and endodontics. The ChatGPT-4.o model showed the highest accuracy (72%), particularly excelling in direct restoration and caries-related questions. This study also emphasized that GPT-4 and 3.5-based models provided higher accuracy levels. Additionally, another assessment with text-based case descriptions reported that the DeepSeek-R1 model reached a diagnostic accuracy of 91.6%, even surpassing ChatGPT-4.o and professional readers of the related journal.

On the other hand, Aygün et al.[11] reported high accuracy rates of AI models for anatomy questions but highlighted content inaccuracies in supplementary explanations, limiting their reliability as educational resources. Similarly, while all models evaluated in our study showed significant agreement with reference answers, incorrect or incomplete responses were observed in some questions. Thus, using these models as supplementary resources in dental education is advised, supported by supervised, multi-sourced, and critical learning approaches in fundamental knowledge acquisition processes.

Limitations

This study has several limitations. Firstly, the evaluation included only 30 multiple-choice questions, potentially restricting the assessment of general diagnostic performance of AI models. All questions were prepared in the same format (multiple-choice), excluding open-ended or case-based scenarios, limiting the ability to test clinical reasoning and problem-solving skills of the models. Additionally, as the study only covered oral pathology, findings may not be generalizable to other areas of dentistry. Finally, only initial responses from the models were evaluated, excluding interactive feedback or correction capabilities from analysis.

Future studies are recommended to include a larger number and variety of question formats, evaluate multiple dentistry fields, and examine the interactive capabilities of AI models. Such studies could provide more comprehensive assessments of model effectiveness in educational contexts.

CONCLUSION

This study revealed that ChatGPT-3.5 and ChatGPT-4.o models showed higher accuracy and statistical consistency in multiple-choice questions on fundamental oral pathology topics compared to other large language models. Although all models demonstrated significant agreement with reference responses, ChatGPT-based models exhibited superior accuracy. These results suggest these models may serve as supportive tools in dental education for specific subjects. However, considering the limitations and potential inaccuracies of AI models, their use should remain supplementary, not as primary resources for students' fundamental knowledge acquisition.

ETHICAL DECLARATIONS

Ethics Committee Approval

This study did not require ethical approval as it did not involve any human subjects or animal experiments.

Informed Consent

Because the study has no study with human and human participants, no written informed consent form was obtained.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Shorey S, Mattar C, Pereira TL, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. *Nurse Educ Today*. 2024;135:106121. doi:10.1016/j.nedt.2024.106121
2. Dashti M, Londono J, Ghasemi S, et al. Attitudes, knowledge, and perceptions of dentists and dental students toward artificial intelligence: a systematic review. *J Taibah Univ Med Sci*. 2024;19(2):327-337. doi:10.1016/j.jtumed.2023.12.010
3. Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. *Urol Pract*. 2023;10(4):409-415. doi:10.1097/UPJ.0000000000000406
4. OpenAI. ChatGPT: Optimizing language models for dialogue [Internet]. 2022. Available from: https://openai.com/blog/chatgpt.
5. Gibney E. Scientists flock to DeepSeek: how they're using the blockbuster AI model. *Nature*. 2025. doi:10.1038/d41586-025-00275-0
6. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish Dental Education. *Cureus*. 2025;17(1):e77292. doi:10.7759/cureus.77292
7. Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. *J Periodontal Res*. 2025;60(2):121-133. doi:10.1111/jre.13323
8. Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. *Eur J Dent Educ*. 2024;28(4):865-876. doi:10.1111/eje.13009
9. Eraslan R, Ayata M, Yagci F, Albayrak H. Exploring the potential of artificial intelligence chatbots in prosthodontics education. *BMC Med Educ*. 2025;25(1):321. doi:10.1186/s12909-025-06849-w
10. Wójcik D, Nowak M, Lewandowska M, et al. A comparative analysis of the performance of ChatGPT-4, Gemini and Claude for the Polish Medical Final Diploma Exam and Medical-Dental Verification Exam. *medRxiv*. 2024;2024.07.29.24311077. doi:10.1101/2024.07.29.24311077
11. Aygün T, Keskin M, Özkan T, et al. A performance of generative pre-trained transformers (GPT) in answering questions on anatomy in the Turkish dentistry specialization exam. *J Ilm Teknol Sist Inform*. 2024;5(4):188-192. doi:10.62527/jitsi.5.4.317
12. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health*. 2025;25(1):573. doi:10.1186/s12903-025-05926-2
13. Wu YH, Tso KY, Chiang CP. Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations. *J Dent Sci*. 2025. doi:10.1016/j.jds.2025.03.030
14. Tosun B, Yılmaz ZŞ. Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. *J Dent Sci*. 2025. doi:10.1016/j.jds.2025.01.025
15. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin Oral Investig*. 2024;28(11):575. doi:10.1007/s00784-024-05968-w