Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Diagnostic Performance of Large Language Models in Dental Image Interpretation: A Comparative Study with Undergraduate Dental Students

Burcu Yeliz KOLLAYAN

byevran@hotmail.com

Burdur Mehmet Akif Ersoy University

Gizem ALPAY

Burdur Mehmet Akif Ersoy University

Mehmet Egemen AYDEMİR

Burdur Mehmet Akif Ersoy University

Research Article

Keywords: large language models, dental education, radiographic interpretation, artificial intelligence, CBCT

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9193501/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

The rapid integration of artificial intelligence into healthcare and education has intensified interest in the diagnostic capabilities of large language models (LLMs), particularly with the emergence of multimodal systems capable of processing visual data. However, their ability to accurately interpret dental radiographic and clinical images remains uncertain.

Methods

This comparative observational study evaluated the diagnostic performance of two LLMs (ChatGPT and Google Gemini) against undergraduate dental students. A total of 129 fourth- and fifth-year students assessed 20 diagnostic cases based on panoramic, periapical, cone-beam computed tomography (CBCT), and clinical oral images. AI models were tested using standardized zero-shot prompts with identical inputs over five consecutive days to assess temporal variability. Diagnostic accuracy, response rates, and inter-day consistency were analyzed using chi-square tests, Cochran's Q test, and Cohen's kappa coefficient.

Results

Students demonstrated significantly higher diagnostic accuracy (81.9%) compared with Gemini (63%) and ChatGPT (60%) ($p = 0.015$). Across all imaging modalities, students consistently outperformed AI models ($p < 0.01$). ChatGPT achieved a 100% response rate but showed lower accuracy, whereas Gemini demonstrated higher accuracy among answered responses but failed to respond in 21% of cases. AI performance varied significantly across repeated prompts (45%–70%; $p = 0.036$), with only moderate inter-day agreement ($\kappa = 0.52$). Students were 1.78 times more likely to provide correct diagnoses than AI models ($p = 0.014$).

Conclusions

Despite recent advances in multimodal capabilities, current LLMs demonstrate limited diagnostic accuracy and inconsistent performance in image-based dental interpretation compared with trained human learners. While these systems may support educational processes, their variability, incomplete response behavior, and reduced diagnostic reliability restrict their use as independent clinical decision-making tools. Further refinement of multimodal AI systems is required before safe integration into dental diagnostics can be considered.

INTRODUCTION

Dental education requires the development of both theoretical knowledge and clinical diagnostic skills. In many countries, including Türkiye, dental education is structured as a five-year program integrating foundational sciences, preclinical training, and clinical practice. Preclinical education typically begins in the third year, whereas clinical training continues throughout the fourth and fifth years. During this period, students receive detailed instruction in dentomaxillofacial anatomy, including radiographic landmarks and the topographic anatomy of the oral mucosa. These courses aim to provide students with the essential knowledge required to recognize anatomical structures and pathological findings in clinical settings. However, students may experience difficulties in recognizing and retaining certain anatomical landmarks, which may complicate the accurate identification of these structures in clinical practice(1, 2).

In recent years, artificial intelligence (AI) technologies have rapidly advanced and have increasingly been integrated into healthcare and medical education(3). AI has demonstrated considerable potential in improving diagnostic accuracy, clinical decision-making, and educational support systems. Among recent AI developments, LLMs such as ChatGPT and Google Gemini have attracted significant attention due to their ability to generate human-like responses and process complex medical information. These models are based on transformer architectures and utilize reinforcement learning from human feedback to improve response quality and contextual understanding (4–6).

Recent advances in multimodal artificial intelligence systems have further expanded the capabilities of LLMs. These systems are now able to process not only textual information but also visual data, allowing them to analyze and interpret medical images in addition to generating text-based responses (7). This development has created new opportunities for applying AI tools in clinical education and diagnostic training.

Several studies have investigated the use of LLMs in medical and healthcare education. Previous research has shown that AI-based language models can perform well in answering medical examination questions and may serve as supportive tools in educational settings (7). However, concerns remain regarding the accuracy, reliability, and consistency of AI-generated responses, particularly in clinical contexts where diagnostic precision is essential.

Artificial intelligence applications have also begun to emerge in dentistry, particularly in areas such as radiographic interpretation, diagnostic support systems, and treatment planning (3, 8–14). Nevertheless, most existing studies in dental education have primarily focused on students' ability to identify specific anatomical landmarks on panoramic radiographs (1, 2, 15). Studies that evaluate diagnostic interpretation by considering both radiographic images and clinical oral findings simultaneously remain limited. However, the diagnostic performance of multimodal LLMs in image-based dental interpretation, particularly in comparison with dental students, remains insufficiently explored.

Understanding the diagnostic capabilities of these systems is particularly important as students increasingly rely on AI tools for learning and clinical guidance. Therefore, the aim of this study was to evaluate the diagnostic performance of large language models in interpreting dental radiographic and clinical images and to compare their performance with that of undergraduate dental students. In

addition, the study examined the temporal variability and reliability of AI responses across repeated prompts, providing further insight into the potential role and limitations of LLMs in dental education.

The null hypothesis of the study was that different AI chatbot models (ChatGPT and Google Gemini) would respond to diagnostic questions with accuracy comparable to dental students, and that no significant difference would exist between the chatbot models.

MATERIALS AND METHODS

Study design

This study was designed as a comparative observational study aimed at evaluating the diagnostic performance of LLMs in interpreting dental diagnostic images and comparing their performance with that of undergraduate dental students.

The study evaluated the interpretation of both dental radiographic images and clinical oral images and the diagnostic accuracy of LLMs was compared with that of dental students.

A total of 129 undergraduate dental students participated in the study. Participants were recruited from the fourth and fifth years of the dental program.

Demographic and educational information including gender, academic year, and completion of radiology training were recorded for each participant.

Participation was voluntary and all responses were collected anonymously.

Image dataset and diagnostic questions

The study consisted of 20 diagnostic questions based on dental diagnostic images.

The images were categorized into four groups:

- Panoramic radiographs (n = 5)
- Periapical radiographs (n = 5)
- Cone-beam computed tomography (CBCT) images (n = 5)
- Clinical oral images (n = 5)

The questionnaire used in this study was specifically developed by the authors and was not adapted from any previously published source. The diagnostic questions were designed based on clinical and radiographic scenarios and were reviewed by two oral and maxillofacial radiologists with more than 5 years of clinical experience (BYK & MEA) to ensure content validity, with a specific focus on anatomical structures and diagnostic scenarios that are frequently misinterpreted by undergraduate dental students,

thereby enhancing the educational relevance of the assessment; an English version of the questionnaire is provided as a supplementary file.

AI models evaluated

Two large language models were evaluated in the study:

- ChatGPT (Open AI, GPT-5-based model)
- Google Gemini (Google LLC)

Each diagnostic question was submitted to the models using identical prompts. The diagnostic images were directly uploaded to the AI models for evaluation.

To evaluate the temporal stability of AI responses, the same set of questions was asked once daily for five consecutive days, resulting in a total of 100 AI responses (20 questions × 5 days).

Responses that did not contain a clear answer were categorized as unanswered.

All questions were submitted to the AI models using the same standardized prompt format.

Each prompt included the diagnostic image and the following instruction:

"Please examine the image and select the most appropriate answer among the provided options."

The same prompts were used across all testing sessions to ensure consistency and reproducibility of AI responses.

Data collection

Student responses were collected using a structured questionnaire format. Each participant answered the same set of 20 diagnostic questions.

Responses were categorized as correct or incorrect according to the predetermined reference answers.

AI model responses were categorized as correct, incorrect, or unanswered.

Outcome measures

Outcome measures
The primary outcome of the study was diagnostic accuracy, defined as the proportion of correct answers relative to the total number of questions.

Secondary outcomes included:

- Response rate of AI models
- Diagnostic accuracy by imaging modality
- Comparison of diagnostic accuracy between students and AI models
- Temporal variability of AI responses
- Inter-day consistency of AI responses

Statistical analysis

All statistical analyses were performed using SPSS 23 software **(**IBM Corporation, Armonk, NY, USA**)**.

Descriptive statistics were calculated for participant characteristics and diagnostic accuracy.

Differences in diagnostic accuracy between groups were evaluated using the chi-square test.

Comparisons between two independent groups (such as academic year and radiology training status) were analyzed using the chi-square test.

The temporal variability of AI responses across days was assessed using the Cochran's Q test.

Inter-day consistency of AI responses was evaluated using Cohen's kappa coefficient.

A p-value < 0.05 was considered statistically significant.

Ethics statement

This study was approved by the Non-Interventional Clinical Research Ethics Committee of Mehmet Akif Ersoy University (Protocol number: GO 2026/2452) and designed in accordance with the Declaration of Helsinki. Participation in the study was voluntary and all data were collected anonymously.

RESULTS

Participant characteristics

A total of 129 dental students participated in the study. Among the participants, 80 (62.0%) were female and 49 (38.0%) were male. Regarding academic level, 76 students (58.9%) were fourth-year students and 53 (41.1%) were fifth-year students. Additionally, 85 students (65.9%) had completed radiology training, while 44 students (34.1%) had not yet received formal radiology training (Table 1).

Table 1
Participant characteristics of the study population (n = 129)

Variable	n	%
Female	80	62.0
Male	49	38.0
4th year students	76	58.9
5th year students	53	41.1
Radiology training completed	85	65.9
Radiology training not completed	44	34.1

Overall student diagnostic performance

The overall diagnostic performance of the students demonstrated a mean score of 16.39 ± 2.05 correct answers out of 20 questions, corresponding to a mean diagnostic accuracy of 81.9%. The lowest observed score among participants was 10, whereas the highest score reached 20, indicating variability in diagnostic performance across the cohort (Table 2).

Table 2
Overall diagnostic performance of dental students

Metric	Value
Number of students	129
Mean correct answers	16.39 ± 2.05
Mean accuracy	81.9%
Minimum score	10
Maximum score	20

Question-level diagnostic accuracy

Analysis of question-level performance revealed variability in diagnostic accuracy across the 20 questions. The accuracy ranged from 60.5% to 99.2%. The highest accuracy was observed for Question 8 (99.2%), followed by Question 6 (94.6%) and Question 7 (91.5%). In contrast, the lowest accuracy rates were observed for Question 14 (60.5%), Question 10 (61.2%), and Question 18 (65.9%), indicating that these items represented the most challenging diagnostic tasks for the students.

Diagnostic accuracy by imaging modality

When diagnostic performance was evaluated according to imaging modality, the highest accuracy was observed for periapical radiographs (87.0%), followed by CBCT images (81.0%) and panoramic

radiographs (81.1%). The lowest diagnostic accuracy was observed for clinical photographs (77.7%), suggesting that interpretation of clinical lesions may present greater diagnostic challenges for students (Table 3).

Table 3
Diagnostic accuracy of dental students according to imaging modality

Imaging modality	Student accuracy (%)
Panoramic	81.1
Periapical	87.0
CBCT	81.0
Clinical	77.7

AI model performance

Two large language models were evaluated for comparison with student performance. Gemini responded to 79 out of 100 prompts, corresponding to a response rate of 79%. Among the answered prompts, 63 responses were correct, yielding an accuracy of 79.7% among answered questions and an overall accuracy of 63%.

In contrast, ChatGPT responded to all prompts (response rate 100%), but demonstrated lower diagnostic accuracy, with 60 correct responses, corresponding to an overall accuracy of 60% (Table 4).

Table 4
Performance of AI models in diagnostic tasks

Model	Response rate	Accuracy (answered)	Overall accuracy
Gemini	79%	79.7%	63%
ChatGPT	100%	60%	60%
*Overall accuracy includes unanswered responses.			

Comparison between students and AI models

Comparative analysis demonstrated that students significantly outperformed AI models in diagnostic accuracy. The overall student accuracy was 81.9%, compared with 63% for Gemini and 60% for ChatGPT. Statistical analysis using the chi-square test revealed a significant difference between students and AI models ($p = 0.015$) (Table 5).

Table 5
Comparison of diagnostic accuracy between students and AI models

Group	Accuracy (%)
Students	81.9
Gemini	63
ChatGPT	60
*χ^2 test, p = 0.015	

Diagnostic accuracy by imaging modality: students vs AI

Further analysis showed that students consistently achieved higher diagnostic accuracy than AI models across all imaging modalities. Significant differences were observed for panoramic radiographs (χ^2 = 7.21, p = 0.007), periapical radiographs (χ^2 = 9.34, p = 0.002), CBCT images (χ^2 = 10.11, p = 0.001), and clinical photographs (χ^2 = 8.46, p = 0.004) (Table 6).

Table 6
Diagnostic accuracy by imaging modality: comparison between students and AI models

Imaging modality	Students (%)	AI (%)	χ^2	p
Panoramic	81	64	7.21	0.007
Periapical	87	68	9.34	0.002
CBCT	82	61	10.11	0.001
Clinical	78	56	8.46	0.004
*AI models represent combined performance of ChatGPT and Gemini.				

Students were significantly more likely to provide correct diagnoses than AI models (OR = 1.78, 95% CI: 1.12–2.83, p = 0.014), indicating a substantial performance advantage.

No statistically significant difference in diagnostic accuracy was observed between fourth-year and fifth-year students (p = 0.32). Similarly, completion of radiology training did not significantly affect diagnostic accuracy among students (p = 0.58).

Temporal variability of AI performance

Evaluation of AI responses across repeated prompts revealed variability in diagnostic performance across different days. Accuracy ranged from 45% to 70%, with the highest performance observed on Day 4 and the lowest on Day 5. Cochran's Q test demonstrated that this variation was statistically significant (Q = 10.24, p = 0.036) (Table 7).

Table 7
Temporal variability of AI diagnostic accuracy across repeated prompts

Day	Accuracy (%)	
Day 1	60	
Day 2	65	
Day 3	55	p = 0.036
Day 4	70	
Day 5	45	
*Cochran's Q test, p = 0.036		

Inter-day reliability analysis demonstrated moderate agreement in AI responses across repeated prompts, with a Cohen's kappa value of 0.52, indicating moderate consistency in AI-generated diagnostic outputs.

DISCUSSION

The findings of this study demonstrate a clear and consistent performance gap between undergraduate dental students and current LLMs in the interpretation of dental radiographic and clinical images. Students achieved significantly higher diagnostic accuracy than both evaluated AI models, indicating that current LLMs remain insufficient for complex diagnostic tasks requiring integration of visual information and clinical context.

This difference can be explained by the nature of diagnostic reasoning in dentistry. Accurate interpretation of dental images requires not only visual pattern recognition but also clinical reasoning, contextual integration, and decision-making under uncertainty. These competencies are developed through structured education and clinical experience. In contrast, LLMs generate responses based on probabilistic associations derived from large datasets and are primarily optimized for linguistic coherence rather than diagnostic accuracy. As a result, they may fail to capture subtle clinical cues and contextual relationships that are critical for correct diagnosis.

Previous studies have reported promising performance of LLMs in medical and dental education, particularly in knowledge-based tasks (5, 9, 13, 14, 22). Advanced models such as GPT-4 have demonstrated high accuracy in medical examinations, and ChatGPT has been shown to reach moderate performance levels in standardized tests such as the USMLE (5, 22). Similarly, studies in radiology have reported that LLMs can perform well in text-based, exam-style questions (16, 17). However, these findings primarily reflect text-based reasoning. The present study indicates that LLM performance is highly task-dependent and decreases substantially in image-based diagnostic tasks.

Evidence from dental radiology-specific studies further supports this observation. Öçbe and Demirdurak reported that AI models demonstrated lower accuracy than dental students in identifying anatomical structures on panoramic radiographs (1). Similarly, Kahalian et al. showed that ChatGPT-4 had limited accuracy in interpreting complex radiographic images (18). Additional studies have also reported that LLMs tend to generate generalized responses rather than detailed diagnostic interpretations in radiographic analysis (23, 24). Together with the present findings, these results indicate that LLMs remain limited in handling image-based diagnostic tasks despite recent advances in multimodal capabilities.

Image-based diagnostic interpretation requires spatial reasoning, visual discrimination, and integration of clinical context-processes that differ fundamentally from the text-based mechanisms of LLMs. Although multimodal AI systems have improved visual processing capabilities, their performance in complex diagnostic scenarios remains insufficient when compared with trained human learners.

In contrast, task-specific artificial intelligence systems, particularly those based on convolutional neural networks (CNNs), have demonstrated high accuracy in dental imaging tasks such as segmentation and lesion detection (19–21). These findings highlight a critical distinction between general-purpose language models and domain-specific AI systems, suggesting that task-specific models may be more suitable for complex visual diagnostic applications.

Another important finding of this study is the variability of AI responses across repeated prompts. Significant fluctuations in diagnostic accuracy were observed across different days, and inter-day agreement was only moderate. This variability raises critical concerns regarding the reproducibility and reliability of AI-generated outputs. In clinical settings, inconsistent responses to identical inputs may lead to diagnostic uncertainty and may pose potential risks to patient safety.

From an educational perspective, these findings suggest that LLMs may serve as supportive tools rather than replacements for traditional dental education. While AI systems can facilitate access to information and support learning processes, their limitations in diagnostic accuracy and consistency necessitate cautious use. Furthermore, overreliance on AI-generated outputs may impair the development of clinical reasoning skills and may lead to overconfidence or incorrect diagnostic approaches among students.

This study has several limitations. The number of diagnostic images and questions was limited, and only two LLMs were evaluated. Additionally, the use of a standardized prompt may have influenced AI responses. Future studies should include larger datasets, additional AI models, and alternative prompting strategies to further explore the role of LLMs in dental education and clinical diagnostics.

Taken together, these findings indicate that although LLMs represent a promising development in medical education, their current limitations in diagnostic accuracy, consistency, and contextual reasoning restrict their use in image-based diagnostic tasks. At present, LLMs should not be considered reliable standalone diagnostic tools in clinical settings. Finally, it emphasizes the importance of

maintaining human supervision, particularly in image-based clinical decision-making processes and AI-assisted diagnostic workflows.

CONCLUSION

In conclusion, large language models demonstrated limited diagnostic accuracy and moderate reliability in interpreting dental radiographic and clinical images compared with undergraduate dental students. Although these systems show potential as supportive educational tools, they currently lack the consistency and contextual understanding required for clinical decision-making. Therefore, LLMs should be integrated cautiously into dental education as complementary resources rather than replacements for clinical training and expert judgment. Future developments in multimodal AI may enhance their diagnostic capabilities and educational value.

Declarations

Acknowledgements

The authors would like to thank all undergraduate dental students who voluntarily participated in this study for their valuable contributions.

Author contributions

BYK: Conceptualization, Methodology, Investigation, Writing – Original Draft, Writing – Review & Editing.

GA: Conceptualization, Methodology, Investigation, Data Curation, Writing - Review & Editing.

MEA: Conceptualization, Methodology, Investigation, Writing – Original Draft, Writing – Review & Editing.

Funding statement

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Data availability statement

The datasets used and/or analysed during the current study are available from the corresponding author upon reasonable request.

Ethics Approval

This study was conducted using anonymized patient records and was approved by the Non-Interventional Clinical Research Ethics Committee of Mehmet Akif Ersoy University (Protocol number: **GO 2026/2452**). All procedures were carried out in accordance with the ethical principles of the Declaration of Helsinki.

Consent to Participate

Informed consent was obtained electronically from all participants prior to participation. The first page of the online questionnaire (Google Forms) included an informed consent statement, and only participants who provided consent were allowed to proceed with the study.

Consent for publication

Not applicable.

Competing interests

The author(s) declare no competing interests.

Clinical trial number:

Not applicable.

References

1. Öçbe M, Demirdurak Y. Dentistry Students vs. ChatGPT 4o: Assessing Knowledge of Panoramic Radiographic Anatomy. Aydın Dent J 31 Ağustos. 2025;11(2):139–46.
2. Maeda N, Hosoki H, Yoshida M, Suito H, Honda E. Dental students' levels of understanding normal panoramic anatomy. J Dent Sci Aralık. 2018;13(4):374–7. 10.1016/j.jds.2018.08. .002 PubMed PMID: 30895148; PubMed Central PMCID: PMC6388826.
3. Safı Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical Aspects of Developing Chatbots for Medical Applications: Scoping Review. J Med Internet Res 18 Aralık. 2020;22(12):e19127. doi:10.2196/19127 PubMed PMID: 33337337; PubMed Central PMCID: PMC7775817.
4. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: Use it with caution. Med Teach Mayıs. 2024;46(5):657–64. 2023.2271159 PubMed PMID: 37862566.
5. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C. vd. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health Şubat. 2023;2(2):e0000198. 10.1371/journal.pdig.0000198. PubMed PMID: 36812645; PubMed Central PMCID: PMC9931230.
6. Suta P, Lan X, Wu B, Mongkolnam P, Chan JH. An Overview of Machine Learning in Chatbots. Int J Mech Eng Robot Res. 2020;502 – 10. 10.18178/ijmerr.9.4.502-510
7. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW. vd. Large language models encode clinical knowledge. Nat Ağustos. 2023;620(7972):172–80. 10.1038/s41586-023-06291-2. PubMed PMID: 37438534; PubMed Central PMCID: PMC10396962.

8. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent Off Publ Am Acad Esthet Dent Al Ekim. 2023;35(7):1098–102. 10.1111/jerd.13046. PubMed PMID: 37017291.
9. Sabri H, Saleh MHA, Hazrati P, Merchant K, Misch J, Kumar PS. vd. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. J Periodontal Res Şubat. 2025;60(2):121–33. 10.1111/jre.13323. PubMed PMID: 39030766; PubMed Central PMCID: PMC11873669.
10. Fawaz P, Sayegh PE, Vannet BV. What is the current state of artificial intelligence applications in dentistry and orthodontics? J Stomatol Oral Maxillofac Surg Ekim. 2023;124(5):101524. 10.1016/j.jormas.2023.101524. PubMed PMID: 37270174.
11. Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y, vd. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofac Orthop Off Publ Am Assoc Orthod Its Const Soc Am Board Orthod. Haziran. 2024;165(6):652 – 62. 10.1016/j.ajodo.2024.01.012 PubMed PMID: 38493370.
12. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod 13 Nisan. 2024;48(1):cjae017. 10.1093/ejo/cjae017. PubMed PMID: 38613510; PubMed Central PMCID: PMC12810200.
13. Revilla-León M, Barmak BA, Sailer I, Kois JC, Att W. Performance of an Artificial Intelligence-Based Chatbot (ChatGPT) Answering the European Certification in Implant Dentistry Exam. Int J Prosthodont 22 Nisan. 2024;37(2):221–4. 10.11607/ijp.8852. PubMed PMID: 38270461.
14. Bulut AC, Bahadır HS, Ateş G. Artificial intelligence in dental education: can AI-based chatbots compete with general practitioners? BMC Med Educ 02 Ekim. 2025;25:1319. 10.1186/s12909-025-07880-7. PubMed PMID: 41039365; PubMed Central PMCID: PMC12492586.
15. Razmus TF, Williamson GF, Van Dis ML. Assessment of the knowledge of graduating American dental students about the panoramic image. Oral Surg Oral Med Oral Pathol Eylül. 1993;76(3):397–402. 10.1016/0030-4220. (93)90278-c PubMed PMID: 8378061.
16. Peker RB. Assessment of the Performance of Large Language Models ChatGPT, ChatGPT Plus, Gemini, and Microsoft Copilot in Responding to Oral and Maxillofacial Radiology Questions from the Turkish Dentistry Specialty Entrance Exam. Yeditepe Dent J. 2025;21(3):130–5. 10.5505/yeditepe.2025.93265.
17. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiol Haziran. 2023;307(5):e230582. 10.1148/radiol.230582. PubMed PMID: 37191485.
18. Kahalian S, Rajabzadeh M, Öçbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. Folia Med (Plovdiv) 31 Aralık. 2024;66(6):863–8. 10.3897/folmed.66.e135584.

19. Bayrakdar IS, Orhan K, Çelik Ö, Bilgir E, Sağlam H, Kaplan FA. vd. A U-Net Approach to Apical Lesion Segmentation on Panoramic Radiographs. BioMed Res Int. 2022;2022:7035367. 10.1155/2022/7035367. PubMed PMID: 35075428; PubMed Central PMCID: PMC8783705.
20. Kaygısız Yiğit M, Pınarbaşı A, Etöz M, Duman ŞB, Bayrakdar İŞ. Artificial intelligence-based fully automatic 3D paranasal sinus segmentation. Dento Maxillo Facial Radiol 01 Ocak. 2026;55(1):61–72. 10.1093/dmfr/twaf057. PubMed PMID: 40711942.
21. Gülşen İT, Kuran A, Evli C, Baydar O, Dinç Başar K, Bilgir E. vd. Deep learning model for automated segmentation of sphenoid sinus and middle skull base structures in CBCT volumes using nnU-Net v2. Oral Radiol Ocak. 2026;42(1):139–47. 10.1007/s11282-025-00848-9. PubMed PMID: 40748555.
22. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ 29 Haziran. 2023;9:e48002. doi:10.2196/48002 PubMed PMID: 37384388; PubMed Central PMCID: PMC10365615.
23. Silva TP, Andrade-Bortoletto MFS, Ocampo TSC, Alencar-Palha C, Bornstein MM, Oliveira-Santos C, vd. Performance of a commercially available Generative Pre-trained Transformer (GPT) in describing radiolucent lesions in panoramic radiographs and establishing differential diagnoses. Clin Oral Investig 09 Mart. 2024;28(3):204. 10.1007/s00784-024-05587-5. PubMed PMID: 38459362; PubMed Central PMCID: PMC10924032.
24. Hu Y, Hu Z, Liu W, Gao A, Wen S, Liu S. vd. Exploring the potential of ChatGPT as an adjunct for generating diagnosis based on chief complaint and cone beam CT radiologic findings. BMC Med Inf Decis Mak 19 Şubat. 2024;24(1):55. 10.1186/s12911-024-02445-. y PubMed PMID: 38374067; PubMed Central PMCID: PMC10875853.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFile.pdf