



Performance of Chat-GPT 5.1 in the Diagnostic Evaluation of Apical Lesions on Panoramic Radiographs

Ezgi Uzun^{1,a}, Burak Kerem Apaydin^{1,b,*}, İsmail Ongun^{2,c}

¹Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Pamukkale University, Denizli, Türkiye.

²Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Uşak University, Uşak, Türkiye.

*Corresponding author

Research Article

History

Received: 04/02/2026

Accepted: 24/02/2026

ABSTRACT

Objectives: The aim of this study was to evaluate the diagnostic performance of GPT-5.1, the latest version of ChatGPT, in determining the presence or absence of apical lesions on panoramic radiographs based on visual input and to analyze the obtained results on a jaw-specific basis.

Materials and Methods: A total of 207 anonymized panoramic radiographs were retrospectively analyzed. In each radiograph, the region in which an apical lesion was present was recorded as "lesion-present," whereas the contralateral jaw region without an apical lesion on the same radiograph was considered "lesion-absent." In this context, each lesion-present and lesion-absent region was treated as an independent unit of analysis. All evaluations were independently performed by GPT-5.1 using standardized and anatomically restricted prompts that clearly defined the jaw (maxilla/mandible), side (right/left), and anatomical region. Model outputs were classified as true positive, true negative, false positive, or false negative. Sensitivity, specificity, accuracy, and F1 score were calculated for overall performance and on a jaw-specific basis.

Results: Overall sensitivity, specificity, accuracy, and F1 score of GPT-5.1 were 67.15%, 60.87%, 64.01%, and 65.11%, respectively. Tooth-level sensitivity (the proportion of cases in which correctly detected the tooth with an apical lesion) was 67.6%. Mandibular performance was higher than maxillary performance (accuracy: 67.52% vs. 57.14%; tooth-level sensitivity: 69.89% vs. 63.04%).

Conclusions: GPT-5.1 demonstrated a moderate level of diagnostic performance in detecting apical lesions on panoramic radiographs. The findings indicate that the model is not suitable for use as a standalone reliable diagnostic tool.

Keywords: Diagnostic imaging, large language models, panoramic radiography

Panoramik Radyografide Apikal Lezyonların Tespiti için Chat-GPT 5.1'in Değerlendirilmesi

Araştırma Makalesi

Süreç

Gelis: 04/02/2026

Kabul: 24/02/2026

Copyright



This work is licensed under Creative Commons Attribution 4.0 International License

^a ezgiuzun096@hotmail.com

^c ismail.ongun@usak.edu.tr

ÖZ

Amaç: Bu çalışmanın amacı, görsel girdiyi dayanarak panoramik radyografilerde apikal lezyonların varlığını veya yokluğunu belirlemede ChatGPT'nin en son sürümü olan GPT-5.1'in tanılal performansını değerlendirmek ve elde edilen sonuçları çeneye özgü olarak analiz etmektir.

Gereç ve Yöntemler: Toplam 207 adet anonimleştirilmiş panoramik radyografi retrospektif olarak analiz edilmiştir. Her bir radyografide apikal lezyonun mevcut olduğu bölge "lezyonlu" olarak kaydedilirken, aynı radyografi üzerindeki apikal lezyon bulunmayan kontralateral çene bölgesi "lezyonsuz" olarak kabul edilmiştir. Bu kapsamda, her bir lezyonlu ve lezyonsuz bölge birbirinden bağımsız analiz birimleri olarak ele alınmıştır. Tüm değerlendirmeler; çene (maksilla/mandibula), taraf (sağ/sol) ve anatomik bölgeyi açıkça tanımlayan standartlaştırılmış ve anatomik olarak sınırlandırılmış istemler kullanılarak GPT-5.1 tarafından bağımsız olarak gerçekleştirilmiştir. Model çıktıları doğru pozitif, doğru negatif, yanlış pozitif ve yanlış negatif olarak sınıflandırılmıştır. Genel tanılal performans ve çeneye özgü performans için duyarlılık, özgüllük, doğruluk ve F1 skoru hesaplanmıştır.

Bulgular: GPT-5.1'in genel duyarlılık, özgüllük, doğruluk ve F1 skoru sırasıyla %67,15, %60,87, %64,01 ve %65,11 olarak bulunmuştur. Diş düzeyinde duyarlılık (apikal lezyon bulunan diş doğru tespit etme oranı) %67,6 olarak bulundu. Mandibuladaki performans maksillaya kıyasla daha yüksek bulunmuştur (doğruluk: %67,52'ye karşı %57,14; diş düzeyinde duyarlılık: %69,89'a karşı %63,04).

Sonuçlar: GPT-5.1, panoramik radyografilerde apikal lezyonların saptanmasında orta düzeyde bir tanılal performans sergilemiştir. Elde edilen bulgular, modelin tek başına güvenilir bir tanı aracı olarak kullanımına uygun olmadığını göstermektedir.

Anahtar Kelimeler: Büyük dil modelleri, panoramik radyografi, tanılal görüntüleme

^b 0000-0003-3198-8325

^d 0000-0003-1546-461X

^b drkeremapaydin@gmail.com

^d 0000-0003-2621-4704

How to Cite: Uzun E, Apaydin BK, Ongun İ. (2026) Performance of Chat-GPT 5.1 in the Diagnostic Evaluation of Apical Lesions on Panoramic Radiographs. Cumhuriyet Dental Journal, 29(1): 168-173.

Introduction

Large Language Models (LLM) are advanced artificial intelligence (AI) systems based on deep learning architectures, trained on extensive datasets comprising billions of data points, and capable of processing and generating human-like text.^{1,2} In recent years, these models have enabled substantial advancements across various domains of natural language processing and have begun to be widely employed in numerous fields.¹ One of the most prominent examples of this technology is Chatbot Generative Pre-trained Transformer (ChatGPT), developed by OpenAI (San Francisco, USA).³ This model, whose foundations were laid with the introduction of GPT-1 in November 2018, attracted substantial global attention following the public release of its enhanced GPT-3.5-based version on November 30, 2022, and rapidly reached a wide user base.⁴⁻⁶

In the following years, successive versions of ChatGPT were continuously improved, achieving significant enhancements in model capacity, language comprehension, contextual reasoning, and multimodal data processing capabilities.^{1,4,7} Notably, GPT-4, released in September 2023, enhanced the ability to process visual inputs and perform multimodal tasks.^{4,8} Subsequently, GPT-5, defined by its developers as a smart and efficient foundation model, was made available for use on August 7, 2025. This model provided significant improvements such as real-time guidance, enhanced reasoning capabilities, accuracy, and multimodal functionality.⁷

In conjunction with these advancements, numerous studies have been conducted with the aim of evaluating the performance of the model's versions and scientifically examining its inherent potential. Research conducted particularly in the fields of dentistry have comprehensively evaluated the model's usability as a patient information tool, its capacity to respond to clinical and expert-level questions, and its performance across various case scenarios.^{1,2,9-15} The findings indicated that ChatGPT is capable of generating clinically acceptable responses in fundamental clinical areas such as endodontic radiology and oral diseases, providing information that may be beneficial for both patients and dentists, and achieving high accuracy rates in certain specific clinical questions and expert-level evaluations.^{2,9,10,12,13,16}

Nevertheless, the model's visual data processing capability was introduced with ChatGPT-4, studies directly examining the diagnostic role of this feature in radiographic assessments remain limited; existing research indicates that the model's performance in this specific area has not yet reached an optimal level.^{8,17-20} On the other hand, based on the existing literature, there are no studies that directly evaluate the diagnostic accuracy of GPT-5-based models in dental radiography. In this context, this study aimed to evaluate the potential of ChatGPT in radiographic image analysis by examining the performance of GPT-5.1 in detecting the presence or absence of apical lesions on panoramic radiographs.

Materials and Methods

This study was conducted in accordance with the Declaration of Helsinki, and the necessary ethical approval was obtained from the Pamukkale University Non-Invasive Clinical Research Ethics Committee (E-60116787-020-797246).

Radiographic Dataset and Sample Size

In this study, panoramic radiographs from the archive of the Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Pamukkale University (Denizli, Turkey) were evaluated. The radiographs were obtained using the OP200D panoramic radiography device (Instrumentarium, Tuusula, Finland) according to standard protocols, with predefined exposure parameters (60 kV, 6.3 mA, 14.1 s scanning time, magnification ratio: 1.3). The radiographs were saved in JPEG format, and all patient-identifying information was removed to ensure anonymization.

In this study, the sample size was determined to support separate analyses for each subgroup (Mandible and Maxilla). Using G Power 3.1.9.7 software, a minimum sample size of $N = 49$ per subgroup was calculated based on an effect size of 0.25, a significance level of 5% ($\alpha = 0.05$), and 95% statistical power. A total of 207 panoramic radiographs were evaluated in accordance with the predefined inclusion and exclusion criteria; of these radiographs, 70 were classified as belonging to the maxilla and 137 as belonging to the mandible.

Inclusion Criteria

- Panoramic radiographs of individuals aged 18 years and older.
- Panoramic radiographs presenting a single apical lesion located in the jaw to be evaluated (either the maxilla or the mandible).
- Apical lesions with a maximum diameter of ≤ 5 mm, measured on panoramic radiographs, and presenting radiolucent margins that are clearly distinguishable from the surrounding bone.
- Radiographs with sufficient image quality and adequate diagnostic quality to allow clear visualization of the apical lesion.

Exclusion Criteria

- Panoramic radiographs with insufficient image quality, poor diagnostic quality, artifacts, or distortions.
- Radiographs in which missing teeth are present either in the region containing the apical lesion or in the contralateral region.

Diagnostic Evaluation Process with GPT-5.1

Panoramic radiographs were independently analyzed by two expert clinicians to establish a reference evaluation for apical lesions. In instances where consensus could not be reached between the evaluators, a final decision was determined using a consensus-based approach. Subsequently, for each radiograph, the apical lesion detected was recorded with respect to the jaw

(maxilla or mandible), side (right or left), and region (anterior, premolar, or molar).

The panoramic radiographs included in the study were uploaded as supplementary files in the chat section, and for each radiograph, the model was asked about the presence or absence of apical lesions. In cases where an apical lesion was present, the question specified the jaw and region where the lesion was located, and the model was asked: "Is there an apical lesion in the specified region? If so, indicate the tooth using the FDI tooth number on which tooth." The prompt structure is shown in Figure 1.

- A model response of "Lesion Present" was considered a True Positive (TP).
- A model response of "Lesion Absent" was considered a False Negative (FN).

In the same panoramic radiographs, the contralateral region without a lesion (for example, if a lesion was present in the right molar region, the left molar region) was presented to the model, and the same question was posed.

- A model response of "Lesion Absent" was considered a True Negative (TN).
- A model response of "Lesion Present" was considered a False Positive (FP).

This approach allowed standardized responses from the model regarding both the presence and absence of lesions in each radiograph. Additionally, all dental radiographs were evaluated in separate chat sessions to prevent contextual learning. The GPT model was operated with default settings, and its "Memory" feature was disabled to prevent them from recalling previously presented images and responses. This approach minimized potential biases arising from memory and ensured that each radiograph was analyzed independently.

Statistical Analysis

The data obtained from the study were analyzed using SPSS 25.0 (Statistical Package for the Social Sciences, Version 25). The distribution of apical lesions according to jaw and anatomical region was summarized using descriptive statistics. The model's responses were classified as TP, TN, FP, and FN. In addition, diagnostic performance measures were calculated to comprehensively evaluate the model's performance.

- **Sensitivity**, representing the model's ability to correctly detect true positive cases:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- **Specificity**, indicating the model's ability to correctly detect true negative cases:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- **Accuracy**, defined as the proportion of correctly predicted observations out of all predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**, indicates the proportion of correctly predicted positive cases among all cases predicted as positive:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **F1-Score**, the harmonic mean of Precision and Sensitivity:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

Results

A total of 207 panoramic radiographs were evaluated in the study. In each radiograph, the region where an apical lesion was present was recorded as lesion-present, whereas the contralateral jaw region on the same radiograph, where an apical lesion was absent, was considered lesion-absent. Accordingly, two separate evaluations were performed for each radiograph, resulting in a total of 414 assessments by ChatGPT.

Of the apical lesions, 137 (66.2%) were located in the mandible and 70 (33.8%) in the maxilla. Among the mandibular apical lesions, 70 (51.1%) were in the molar region, 51 (37.2%) in the premolar region, and 16 (11.7%) in the anterior region. In the maxilla, 29 apical lesions (41.4%) were in the molar region, 21 (30.0%) in the premolar region, and 20 (28.6%) in the anterior region. The distribution of cases without lesions was evaluated in the same manner, using the contralateral region as a reference.

In the evaluation of 207 cases with apical lesions, the GPT-5.1 model correctly detected the presence of apical lesions in 139 cases (TP = 139), while failing to detect apical lesions in 68 cases (FN = 68). In the assessment of 207 cases without apical lesions, the absence of apical lesions was correctly reported in 126 cases (TN = 126), whereas 81 cases were incorrectly recorded as having apical lesions (FP = 81) (Figure 2). Based on these results, the model's overall sensitivity, specificity, accuracy, and F1 score were calculated as 67.15%, 60.87%, 64.01%, and 65.11%, respectively (Table 1). Of the 139 correctly detected apical lesions, 94 were accurately localized to the corresponding tooth, indicating a tooth-level sensitivity of 67.6%.

In the maxilla, the GPT-5.1 model correctly detected apical lesions in 46 out of 70 cases (TP = 46) and failed to detect lesions in 24 cases (FN = 24). Among 70 maxillary cases without apical lesions, the absence of lesions was correctly reported in 34 cases (TN = 34), while 36 cases were incorrectly detected as having lesions (FP = 36). Accordingly, the overall sensitivity, specificity, accuracy, and F1 score for the maxilla were calculated as 65.71%, 48.57%, 57.14%, and 60.47%, respectively (Table 1). Among the 46 correctly detected apical lesions, 29 were accurately localized to the corresponding tooth, resulting in an overall tooth-level sensitivity of 63.04%.

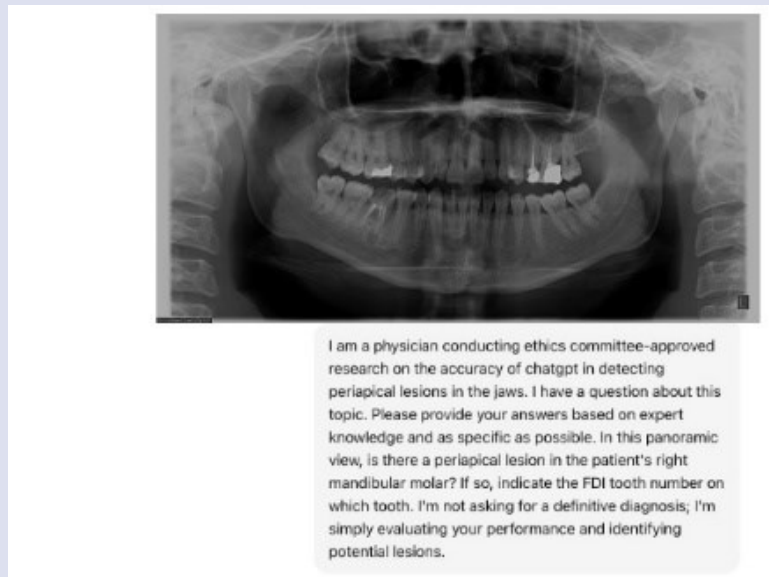


Figure 1. Example of the question posed to the model regarding apical lesion detection.

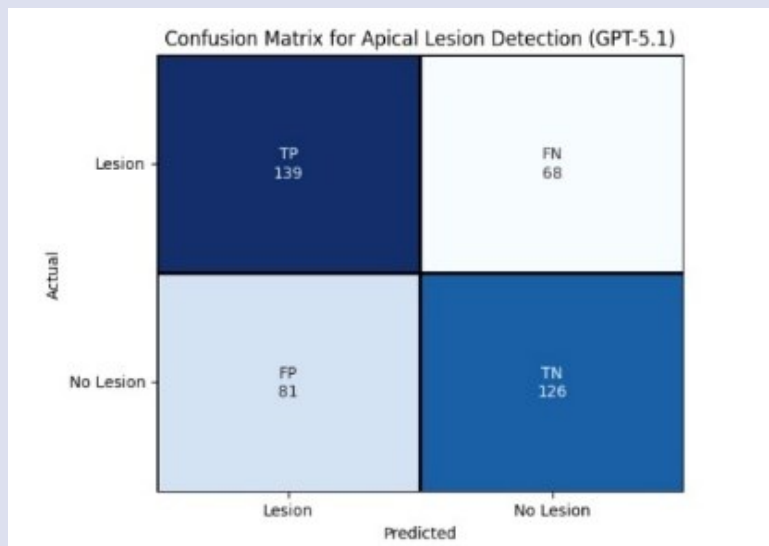


Figure 2. Confusion matrix of apical lesion detection by the GPT-5.1 model.

Table 1. Diagnostic performance metrics for apical lesion detection

	Sensitivity	Specificity	Accuracy	F1 Score
Total	67.15%	60.87%	64.01%	65.11%
Mandible	67.88%	67.15%	67.52%	67.62%
Maxilla	65.71%	48.57%	57.14%	60.47%

Discussion

ChatGPT has attracted significant attention in a short period of time owing to its user-friendly structure and interactive interface, reaching a broad user base.²¹ With this rapid expansion, it has begun to be widely utilized in dentistry for purposes such as supporting diagnostic processes, enhancing clinical decision-making, facilitating digital data recording, performing image analyses, contributing to professional education, and conducting research activities.^{1,15,21,22} The increasing integration of

AI-based applications into healthcare has necessitated the maintenance of high-quality standards in the information provided despite existing limitations; therefore, the reliability of the information generated by models such as ChatGPT, as well as their performance in responding to questions, has gained increasing importance to evaluate.^{14,21}

In this context, the study evaluated the apical lesion detection performance of the latest version of ChatGPT, GPT-5.1, based on visual analyses performed on panoramic radiographs. However, when radiographic

images are provided to the model, it is known to frequently generate responses stating that it cannot interpret medical images or establish a definitive diagnosis, instead directing users to consult a qualified healthcare professional such as a dentist or radiologist.⁶ Therefore, it was explicitly stated in the study that an ethics committee–approved investigation was being conducted to evaluate the accuracy of ChatGPT in detecting apical lesions in the jaw bones, and that no definitive diagnosis was requested from the model; rather, the aim was solely to assess its performance and to detect possible lesions.

In the literature, it has been reported that when ChatGPT is provided with images along with textual inputs, it demonstrates higher performance compared to tasks relying solely on image-based diagnostic assessments.^{15,20} For example, Ding et al.²⁰ reported in their study, in which they evaluated 129 different radiology cases using GPT-4, that diagnostic accuracy was lowest (19.90%) under the “image-only” condition, whereas it reached the highest accuracy (80.88%) when the image, clinical information, and diagnostic options were provided together. These findings have shown that the structured textual context provided to the model significantly affects its diagnostic performance.

In this study, however, the textual inputs provided to the model were deliberately restricted to exclude clinical information; they were limited solely to the anatomical localization of the lesions, and the model was asked to indicate the presence or absence of apical lesions in the relevant region along with the corresponding tooth number. This approach allowed for the assessment of ChatGPT’s performance in detecting tooth-associated apical lesions based on visual recognition alone, independent of guiding textual context.

In our study, GPT-5.1 detected apical lesions with 67.15% sensitivity, 60.87% specificity, 64.01% accuracy, and 65.11% F1 score, with tooth-level sensitivity of 67.60%. Suarez et al.,¹⁹ in their study evaluating mandibular third molars using GPT-4o, reported an overall accuracy of 38%, noting that the model was more reliable in detecting conditions such as implants, restorations, and orthodontic treatments, but limited in interpreting third molar positions. However, in their study, questions were not posed directly and were aimed at describing observations; in contrast, our study explicitly required the model to report the presence or absence of apical lesions. Similarly, in another study, no direct specific questions were asked; periapical radiographs were evaluated for conditions such as fillings, crowns, restorations, impacted teeth, and apical lesions, and the overall correct interpretation rate was found to be 11%. The sensitivity for apical lesions was 27.78%.²³ These findings indicate that GPT models demonstrate higher accuracy and sensitivity when asked directly about the presence or absence of a condition compared to open-ended observational questions. Furthermore, the provision of additional contextual cues in our study (e.g., the anatomical location of the lesion) may explain the higher

model performance compared to previous studies; for instance, another study reported that the pre-diagnosis rate of GPT-4.0 increased from 30.7% without guided prompts to 56.9% when guided prompts were provided.²⁴

Aşar et al.²⁵ evaluated the presence of supernumerary teeth on periapical radiographs using direct “presence/absence” questions and compared the performance of different GPT versions. The customized and specially trained cGPT-4V achieved the highest accuracy at 91%, GPT-4o demonstrated moderate performance with 77% accuracy, and GPT-4V showed the lowest performance with 63% accuracy and sensitivity. In our study, a lower performance was observed compared to the aforementioned study; this may be explained by the fact that panoramic radiographs contain more anatomical structures than periapical radiographs, thereby increasing the difficulty of detection. Similarly, in a study by Salmanpour and Akpınar¹⁸ using GPT-4, panoramic radiographs combined with specific questioning were employed to assess the labiolingual position of impacted maxillary canines and the presence of resorption, resulting in comparably low accuracy rates of 37.1% and 46.0%, respectively.

Furthermore, when evaluating apical lesion detection on a jaw-specific basis, the accuracy and tooth-level lesion sensitivity were observed to be higher in the mandible compared to the maxilla (mandible accuracy: 67.52%, tooth-level sensitivity: 69.89%; maxilla accuracy: 57.14%, tooth-level sensitivity: 63.04%). This can be attributed to the anatomical structure of mandibular teeth, which facilitates the visual detection of apical lesions more readily than in maxillary teeth. The literature also indicates that apical lesions in the mandible are more accurately and readily detected due to fewer surrounding anatomical structures (e.g., airways, nasal bone/cartilage, hard palate, etc.) and limited superimposition.^{26,27} The findings of this study support this observation, indicating that GPT-based models may provide a more reliable contribution to clinical practice in detecting mandibular apical lesions compared to the maxilla.

Nevertheless, the moderate levels of accuracy and sensitivity observed both overall and on a jaw-specific basis indicate that LLM-based models remain limited in their ability to serve as standalone reliable diagnostic tools. In particular, for models intended for use in healthcare, an accuracy value of typically above 85–90%, along with precision and recall rates exceeding 80%, is generally considered indicative of clinically robust and reliable performance.¹⁸ In light of these criteria, the current performance of GPT-5.1 suggests that it may be utilized as a supportive tool; however, its use as a sole diagnostic instrument would not be appropriate. Additionally, it should also be considered that the limited anatomical guidance provided to the model in this study (jaw laterality and regional localization), the single-center dataset, and the images obtained in JPEG format may have structured the evaluation process and influenced the results. Therefore, for LLM-based models to be employed effectively and safely in clinical practice, both the further

development of model training and the continuation of evaluations under expert human supervision are essential.

Conclusions

This study demonstrated that GPT-5.1 exhibited a limited–moderate level of diagnostic performance in detecting apical lesions on panoramic radiographs based solely on visual input. The findings indicate that, in its current form, GPT-5.1 is not suitable for use as an independent diagnostic tool in clinical practice; however, they also suggest that the model may have potential for clinical application with further development. Therefore, additional studies are needed before these results can be generalized to clinical practice. In this context, future research should focus on testing the model with diverse and large datasets and on conducting evaluations without providing anatomical guidance, in order to more clearly elucidate the true visual perception and diagnostic capabilities of GPT-based models.

Acknowledgements

No acknowledgements.

Conflicts of Interest Statement

The authors declare that they have no conflict of interest.

References

- Sezer B, Okutan AE. Evaluation of ChatGPT-4's performance on pediatric dentistry questions: accuracy and completeness analysis. *BMC Oral Health* 2025;25(1):1427.
- Durmazpinar PM, Ekmekci E. Comparing diagnostic skills in endodontic cases: dental students versus ChatGPT-4o. *BMC Oral Health* 2025;25(1):457.
- Tussie C, Starosta A. Comparing the dental knowledge of large language models. *Br Dent J* 2024.
- 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* 2026;21(1):1-8.
- Özdemir ÖT, Güven Y. ChatGPT usage areas and limitations in dentistry. *Selcuk Dent J* 2025;12(1):184-190.
- Puleio F, Lo Giudice G, Bellocchio AM, Boschetti CE, Lo Giudice R. clinical, research, and educational applications of ChatGPT in dentistry: a narrative review. *Appl Sci* 2024;14(23):10802.
- Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. *Int Dent Res* 2025;15(3):123-135.
- Atakır K, Işın K, Taş A, Önder H. Diagnostic accuracy and consistency of ChatGPT-4o in radiology: influence of image, clinical data, and answer options on performance. *Diagn Interv Radiol* 2025.
- Zhou X, Chen Y, Abdulghani EA, Zhang X, Zheng W, Li Y. Performance in answering orthodontic patients' frequently asked questions: Conversational artificial intelligence versus orthodontists. *J World Fed Orthod* 2025;14(4):202-207.
- Çekiç EC, Tavşan O. Evaluating large language models using national endodontic specialty examination questions: are they ready for real-world dentistry? *BMC Med Educ* 2025;25(1):1308.
- Yilmaz B, Kahraman EN, Brennan MT, Grewal AS, Aktas A. Accuracy of ChatGPT-4 Plus in providing information on oral cancer management. *Oral Dis* 2025.
- Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health* 2025;25(1):173.
- Jacobs T, Shaari A, Gazonas CB, Ziccardi VB. Is ChatGPT an accurate and readable patient aid for third molar extractions? *J Oral Maxillofac Surg* 2024;82(10):1239-1245.
- Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: assessment of accuracy and repeatability in answer generation. *J Prosthet Dent* 2024;131(4):659.e1-659.e6.
- Akkoca F, Özdede M, İlhan G, Koyuncu E, Ellidokuz H. Assessing the success of ChatGPT-4o in oral radiology education and practice: a pioneering research. *Cumhuriyet Dent J* 2025;28(2):210-215.
- Ekici Ö, Çalıřkan İ. Comparison of performance of leading large language models in answering medical pathology questions in dentistry specialization education entrance exams: a cross-sectional research. *Türkiye Klinikleri J Dental Sci* 2025.
- Makrygiannakis MA, Kaklamanos EG. Assessment of AI software's diagnostic accuracy in identifying impacted teeth in panoramic radiographs. *Eur J Orthod* 2025;47(5):cjaf085.
- Salmanpour F, Akpınar M. Performance of Chat Generative Pretrained Transformer-4.0 in determining labiolingual localization of maxillary impacted canine and presence of resorption in incisors through panoramic radiographs: a retrospective study. *Am J Orthod Dentofacial Orthop* 2025;168(2):220-231.
- Suárez A, Arena S, Herranz Calzada A, Castillo Varón AI, Diaz-Flores García V, Freire Y. Decoding wisdom: evaluating ChatGPT's accuracy and reproducibility in analyzing orthopantomographic images for third molar assessment. *Comput Struct Biotechnol J* 2025;28:141-147.
- Ding L, Fan L, Shen M, Wang Y, Sheng K, Zou Z, et al. Evaluating ChatGPT's diagnostic potential for pathology images. *Front Med (Lausanne)* 2024;11:1507203.
- Shrivastava PK, Rai A, Injety RJ, Singh S, Jain A, Mahuli AV et al. Performance of ChatGPT in dentistry: a cross-sectional, multi-specialty and multi-centric study. *Braz J Oral Sci* 2025;24:e254954.
- Achanur M, Bhatt S, Maniyar RN, et al. ChatGPT's emerging role in dentistry: a review. *J Pharm Bioallied Sci* 2025;17(Suppl 1):S99-S101.
- Bragazzi NL, Szarpak L, Piccotti F. Assessing ChatGPT's potential in endodontics: preliminary findings from a diagnostic accuracy study. *SSRN* 2023;4631017.
- Kahalian S, Rajabzadeh M, Öcbe M, Medisoglu MS. ChatGPT-4.0 in oral and maxillofacial radiology: prediction of anatomical and pathological conditions from radiographic images. *Folia Medica* 66(6): 863-868. 2024;66(6):863-868.
- Aşar EM, İpek İ, Bilge K. Customized GPT-4V(ision) for radiographic diagnosis: can large language model detect supernumerary teeth? *BMC Oral Health* 2025;25(1):756.
- Stera G, Giusti M, Magnini A, Calistri L, Izzetti R, Nardi C. Diagnostic accuracy of periapical radiography and panoramic radiography in the detection of apical periodontitis: a systematic review and meta-analysis. *Radiol Med* 2024;129(11):1682.
- Dhillon M, Raju SM, Verma S, et al. Positioning errors and quality assessment in panoramic radiography. *Imaging Sci Dent* 2012;42(4):207-212.