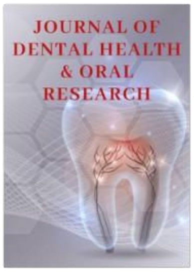

e-ISSN 3050-9432
Journal of Dental
Health and Oral
Research

ATHENAEUM
SCIENTIFIC PUBLISHERS

*Research Article*

# Evaluating Artificial Intelligence (AI) Tools in Clinical Pulpal Scenarios

**Emre Bayram[1], H Melike Bayram[1*]**
[1]Tokat Gaziosmanpasa University Faculty of Dentistry, Department of Endodontics, Turkey
***Correspondence author:** H Melike Bayram, Tokat Gaziosmanpasa University Faculty of Dentistry, Department of Endodontics, Turkey;
E-mail: melikealaca@yahoo.com

Citation: Bayram E, et al. Evaluating Artificial Intelligence (AI) Tools in Clinical Pulpal Scenarios. J Dental Health Oral Res. 2025;6(2):1-4.

https://doi.org/10.46889/JDHOR.2025.6202

Received Date: 30-04-2025
Accepted Date: 19-05-2025
Published Date: 26-05-2025

## Abstract

Background: This study aimed to compare the responses of various AI-based chatbots to different scenarios related to pulpal diseases and to analyze the reliability of these chatbots.
Materials and Method: The declaration of Helsinki conducted the study and did not require ethical approval. A total of 20 open-ended clinical case questions involving various pulpal diseases were created and submitted to ChatGPT, Google Gemini, DeepSeek and Microsoft Copilot. The responses were scored on a scale from 1 (very poor) to 5 (excellent) using the Modified Global Quality Scale (MGQS), based on expert evaluation. Data were analyzed using IBM SPSS Statistics version 22.0 and both the Friedman test and Pearson correlation analysis were performed.
Results: A strong and statistically significant correlation was found between ChatGPT and both DeepSeek ($p < .001$) and Microsoft ($p < .001$). A moderate but significant correlation was also observed between ChatGPT and Google ($p = .019$), as well as between Google and Microsoft ($p < .018$). In contrast, DeepSeek showed a weak and non-significant correlation with Google and Microsoft ($p = .089$). Descriptive statistics indicated that DeepSeek provided the highest and most consistent scores, whereas Google and Microsoft yielded more variable evaluations. However, according to the Friedman test results, there was no statistically significant difference in MGQS scores among the four AI systems ($p = .461$).
Conclusion: ChatGPT produced more consistent responses than the other AI systems, while DeepSeek provided the highest and most stable responses. However, no statistically significant differences were observed among the systems based on MGQS scores.

**Keywords:** Artificial Intelligence; Chatbot; Endodontics; Pulpal Diagnosis

## Introduction

Accurate diagnosis is one of the most critical determinants of treatment success in dentistry. In particular, the early and precise diagnosis of pulpal and periapical diseases plays a key role in determining appropriate treatment protocols [1]. However, interpreting clinical and radiographic findings often depends on the clinician's experience and subjective judgment, which can limit diagnostic accuracy [2].

Recent advances in Artificial Intelligence (AI) technologies have led to their adoption as supportive tools in diagnostic processes in healthcare and dentistry [3]. In particular, deep learning algorithms have achieved significant success in medical imaging due to their ability to recognize patterns within large datasets [4]. In the field of endodontics, AI systems have been utilized in various applications, including the assessment of root canal morphology, the detection of periapical lesions and the diagnosis and treatment planning of pulpal diseases [5]. The main advantages of these chatbots include reducing error rates, standardizing diagnostic processes and detecting microscopic changes that may be difficult for the human eye to perceive [6].

Several studies have reported that different AI models exhibit variations in performance regarding general medical diagnostic accuracy [7, 8]. However, no comparative studies have specifically evaluated the performance of large language models such as ChatGPT, Google Gemini, DeepSeek and Microsoft Copilot in diagnosing and treating pulpal diseases. In light of this information, this study aims to evaluate the diagnostic reliability of different AI-based chatbots in pulpal disease diagnosis and treatment and identify potential performance differences among these systems.

## Material and Method

The principles of the Declaration of Helsinki were followed when conducting this study. Ethical approval was not required since no direct human data were used and all case information was anonymized. To evaluate AI tools' diagnostic and treatment capabilities in pulpal diseases, 20 open-ended clinical case scenarios reflecting real-world dental practice were created. The scenarios were designed based on patients' clinical symptoms, vitality test results, radiographic findings and medical histories to assess diagnostic reasoning and proposed treatment plans. Each clinical scenario was submitted sequentially to four different AI-based language models:

- ChatGPT (based on OpenAI's GPT-4 architecture)
- Google Gemini
- DeepSeek
- Microsoft Copilot

Each AI system received the identical case input and the first response generated by the model was recorded without any modifications or follow-up prompts. The responses obtained from the AI systems were evaluated using the Modified Global Quality Scale (MGQS) by three independent and experienced endodontists.[9, 10]. The evaluation criteria included the following:

- Clinical diagnostic accuracy
- Appropriateness of the proposed treatment
- Completeness of information and clarity of expression

For each case, the mean score given by the three evaluators was calculated.

The scoring consistency among the three independent evaluators was assessed using Intraclass Correlation Coefficient (ICC) analysis. The ICC analysis revealed a high level of agreement among the evaluators. The responses were anonymized and presented to the evaluators to minimize potential bias without indicating the originating AI system.

### *Statistical Analysis*

Pearson correlation analysis was performed to assess the similarity of responses among the different AI chatbots. Pearson correlation and Friedman tests were performed to analyze AI chatbots' response consistency and score differences. A p-value of less than 0.05 was considered statistically significant.

## Results

Descriptive statistics of MGQS scores for the four AI systems are presented in Table 1,2. DeepSeek demonstrated the highest mean MGQS score, followed by ChatGPT, while Google and Microsoft Copilot shared identical mean scores. Friedman test results indicated no statistically significant difference in MGQS scores among the four AI systems (p = 0.461). Pearson correlation analysis revealed a strong, statistically significant correlation between ChatGPT and DeepSeek (p < 0.001), as well as between ChatGPT and Microsoft Copilot (p < 0.001). A moderate but significant correlation was observed between ChatGPT and Google (p = 0.019) and between Google and Microsoft Copilot (p = 0.018). In contrast, DeepSeek, Google and Microsoft Copilot correlations were weak and insignificant (p = 0.089).

| AI System | Mean | SD | Friedman $\chi^2$ | p-value |
|---|---|---|---|---|
| DeepSeek | 4.55 | 1.15 | 2.580 | 0.461 |
| ChatGPT | 4.25 | 1.25 | | |
| Google | 3.55 | 1.88 | | |
| Microsoft | 3.55 | 1.88 | | |

**Table 1:** MGQS scores for each AI system.

| | Google | Microsoft | DeepSeek |
|---|---|---|---|
| ChatGPT (r) | 0.521 | 0.700 | 0.817 |
| ChatGPT (p) | 0.019* | <0.001* | <0.001* |
| Google (r) | | 0.522 | 0.390 |
| Google (p) | | <0.018* | 0.089 |
| r: Pearson Correlation Coefficient<br>* p < 0.05 is considered statistically significant.<br>Interpretation of r values:<br>- r < 0.4: weak correlation<br>- r = 0.4-0.6: moderate correlation<br>- r > 0.6: strong correlation | | | |

**Table 2:** Pearson correlation coefficients between AI systems.

**Discussion**

The role of AI in healthcare has been expanding rapidly, with growing evidence supporting its utility in clinical decision support systems [3]. In dentistry, particularly in endodontics, there is a strong trend toward integrating AI-based systems into clinical workflows for tasks such as recognizing root canal morphology, detecting periapical lesions and assisting in treatment planning [5, 11]. In this context, case scenario-based evaluations, as employed in our study, offer a more realistic approach to assessing the extent to which AI systems can integrate into actual clinical reasoning processes [11]. In this context, case scenario-based evaluations, as employed in our study, offer a more realistic approach to assessing the extent to which AI systems can integrate into actual clinical reasoning processes [1,11]. In this study, the diagnostic reliability of four different AI systems-ChatGPT, Google Gemini, DeepSeek and Microsoft Copilot-was comparatively evaluated in the context of pulpal disease diagnosis and treatment. The results revealed that DeepSeek provided the highest and most consistent scores, while ChatGPT demonstrated strong correlations with the other systems. However, no statistically significant differences in MGQS scores were observed among the chatbot systems. The strong correlations observed between ChatGPT, DeepSeek and Microsoft Copilot in this study suggest that Large Language Model (LLM)-based systems may similarly process specific clinical reasoning patterns. The literature has reported that the clinical accuracy of language model-based AI systems can vary significantly depending on the scope and quality of their training data [7]. The adequacy of training data is a critical factor that directly influences the consistency and accuracy of a model’s responses to clinical scenarios [8]. The high performance of DeepSeek may also be attributed to the breadth of its training dataset and the richness of its embedded clinical knowledge. However, the weaker correlations between DeepSeek and Google and Microsoft Copilot suggest that algorithmic differences may influence clinical decision-making processes. As noted by Ma, et al., different AI models employ varying data processing and decision-making strategies, which can affect the consistency of clinical outcomes [12]. Therefore, complete agreement in outcomes across different platforms should not be expected. Notably, no statistically significant differences were found among the systems regarding MGQS scores in our study. This result suggests that the four AI systems evaluated in this study demonstrated comparable reliability in pulpal disease clinical scenarios. Consistent with our findings, Schwendicke, et al., also reported that different AI systems may exhibit similar overall accuracy, although performance differences can emerge in specific clinical situations [3]. Therefore, case-specific analyses are essential when interpreting the overall performance rates of AI systems. Topol similarly emphasized that AI applications in healthcare should be developed to support, rather than replace, clinicians in their decision-making processes [8]. Our findings are also consistent with those reported in the literature. In a study by Orhan, et al., AI systems achieved diagnostic accuracy comparable to human experts in detecting periapical lesions on CBCT images [5]. Similarly, Ma, et al., reported that a deep learning-based model achieved high diagnostic accuracy in identifying irreversible pulpitis in primary molars [12]. However, while these studies primarily focused on image-based analysis, our study was based on clinical scenario interpretation, which partially limits the direct comparability of the results. In another study conducted on a similar topic, Moura, et al., evaluated the diagnostic and treatment accuracy of ChatGPT, Bing (Copilot) and Google Bard (Gemini) in pulpal and periradicular cases [13]. In that study, each clinical scenario was presented to the AI systems in Portuguese and English and their responses' diagnostic accuracy and consistency were compared [13]. While the study by Moura, et al., evaluated diagnostic accuracy using a binary correct/incorrect classification, our study employed the Modified Global Quality Scale (MGQS) to assess broader aspects of response quality, including clinical coherence, clarity of recommendations and appropriateness. Additionally, our study utilized three independent expert evaluators and inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC). Both studies support the notion that AI systems have potential as clinical decision support tools rather than definitive diagnostic authorities; however, methodological differences limit the comparability of interpretation depth and findings.

Regarding the limitations of our study, the first is that the number of clinical questions was limited to 20, which may constrain the generalizability of the findings. Future studies incorporating larger and more diverse case sets could help reveal more pronounced differences between AI systems. Secondly, the lack of detailed information on each AI platform's algorithmic architecture and training datasets makes it impossible to determine definitively which parameters contributed to the observed performance differences.

## Conclusion

This study comparatively evaluated the performance of four different artificial intelligence systems in diagnosing and planning treatment for pulpal diseases using clinical scenarios. DeepSeek received the highest mean quality score and showed strong correlations with ChatGPT. However, no statistically significant differences in MGQS scores were found among the chatbots. The findings suggest that AI systems can be supportive tools in endodontic diagnostic processes, but should not be considered definitive decision-makers.

## Conflict of Interest

The authors declare no conflict of interest.

## Funding

None.

## Reference

1. Zanini M, Meyer E, Simon S. Pulp inflammation diagnosis from clinical to inflammatory mediators: A systematic review. J Endod. 2017;43(7):1033-51.
2. Bjørndal L, Ricucci D. Pulp inflammation: From the reversible pulpitis to pulp necrosis during caries progression. The Dental Pulp: Biology, Pathology and Regenerative Therapies. Berlin: Springer. 2014;125-39.
3. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: Chances and challenges. J Dent Res. 2020;99(7):769-74.
4. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60-88.
5. Orhan K, Bayrakdar IS, Ezhov M, Kravtsov A, Özyürek T. Evaluation of artificial intelligence for detecting periapical pathosis on cone-beam computed tomography scans. Int Endod J. 2020;53(5):680-89.
6. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-9.
7. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347-58.
8. Topol E. Deep medicine: How artificial intelligence can make healthcare human again. London: Hachette UK. 2019.
9. Bernard A, Langille M, Hughes S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102(9):2070-7.
10. Kumar N, Pandey A, Venkatraman A, Garg N. Are video sharing web sites a useful source of information on hypertension? J Am Soc Hypertens. 2014;8(7):481-90.
11. Aminoshariae A, Kulild J, Nagendrababu V. Artificial intelligence in endodontics: Current applications and future directions. J Endod. 2021;47(9):1352-7.
12. Ma T, Zhu J, Wang D. Deep learning-based detection of irreversible pulpitis in primary molars. Int J Paediatr Dent. 2025;35(1):57-67.
13. Mendonça de Moura JD, Fontana CE, Reis da Silva Lima VH. Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: A cross-sectional study. Comput Biol Med. 2024;183:109332.