**International Journal of Dental Research and Allied Sciences**
2025, Volume 5, Issue 2, Page No: 1-9

Available online at: www.tsdp.net

**ISSN: 3062-3502**

**Original Article**

# Investigation of Large Language Models' Capabilities in Answering Clinical Questions Related to Periodontal Furcation Therapy

**Dimitris Kounatidis[1*], Bhanu Raghunathan[2]**

[1] Department of Preventive Dentistry, Periodontology and Implant Biology, School of Dentistry, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece.
[2] Division of Periodontology, Department of Developmental and Surgical Sciences, School of Dentistry, University of Minnesota, Minneapolis, MN 55455, USA.

***E-mail** ✉ Dimitris.kounatidis.gr@gmail.com

**Received:** 09 April 2025; **Revised:** 01 July 2025; **Accepted:** 04 July 2025

## ABSTRACT

Large Language Models (LLMs) represent advanced artificial-intelligence (AI) tools capable of processing massive textual datasets and producing language that resembles human expression. Their emergence suggests new possibilities for retrieving clinically relevant knowledge within healthcare. This investigation sought to measure and compare how four separate LLMs responded to practical questions on the management and therapy of periodontal furcation defects, assessing the degree to which their answers aligned with scientific evidence. Four models—ChatGPT 4.0, Google Gemini, Google Gemini Advanced, and Microsoft Copilot—were each prompted with ten clinical questions on furcation-defect management. Their replies were benchmarked against a reference source drawn from the European Federation of Periodontology (EFP) S3 guidelines and recent systematic reviews. Two certified periodontists independently graded every response for depth, factual correctness, clarity, and clinical appropriateness, following a fixed rubric that assigned 0–10 points per criterion. Performance varied among systems. Google Gemini Advanced consistently achieved the highest averages, especially for breadth and readability, whereas Google Gemini and Microsoft Copilot tended to receive lower marks, most notably in relevance. A Kruskal–Wallis comparison, however, detected no statistically significant overall difference in mean scores. Inter-rater and intra-rater reliability were both strong. Although all LLMs demonstrated an ability to generate answers regarding furcation-defect care, their quality profiles diverged across domains of completeness, precision, transparency, and contextual fit. Clinicians should therefore recognize both the utility and constraints of such models when consulting them for professional insight.

**Keywords:** Artificial intelligence, ChatGPT, Google Gemini, Microsoft Copilot, Periodontology, Furcation

**How to Cite This Article:** Kounatidis D, Raghunathan B. Investigation of Large Language Models' Capabilities in Answering Clinical Questions Related to Periodontal Furcation Therapy. Int J Dent Res Allied Sci. 2025;5(2):1-9. https://doi.org/10.51847/pPUa3JWPaM

## Introduction

Recent years have witnessed a rapid expansion of artificial-intelligence (AI) applications across numerous disciplines. Among its most transformative branches are Large Language Models (LLMs)—systems trained on immense textual corpora to produce linguistically coherent and context-aware content [1]. These engines can markedly accelerate access to complex information compared with conventional retrieval methods, introducing new efficiencies in healthcare communication and data interpretation.

Within medicine, LLMs may support clinicians by extracting structured data from electronic health records, summarizing literature, simplifying technical phrasing, and automating administrative operations, thus improving both workflow and accuracy [2–4]. They have also been linked to progress in research analytics, educational programs, and quality-assurance

processes [2–4]. Growing evidence points to their promise in diagnostic assistance and predictive modeling [5].

The AI subfield of deep learning further extends these benefits by interpreting medical images for disease identification and individualized therapy planning, at times outperforming human assessment in precision and speed [6]. Moreover, AI systems can help detect individuals with increased disease susceptibility, promoting personalized prevention and targeted treatment protocols [7].

In dentistry, AI tools are emerging as integral aids in diagnosis, treatment planning, radiographic evaluation, outcome prediction, and clinical documentation, thereby optimizing efficiency and reliability [8]. Deep-learning frameworks already assist in early detection of pathologies such as caries and periapical periodontitis, refining decision-making and reducing chair time [9, 10].

Despite these advantages, questions persist regarding factual accuracy, bias, and ethical accountability when AI models deliver clinical information [11, 12]. Prior assessments of LLM responses to medical inquiries have shown inconsistent precision, partly due to heterogeneity in study design and reporting standards [13].

Barriers also remain to seamless clinical adoption. Transparency concerning the training data behind LLM outputs is limited, and models may produce fabricated or misleading statements ("hallucinations") when information gaps exist [14]. Access restrictions such as paywalls or subscription requirements further constrain the range of scholarly material available for model training, including that of ChatGPT [15]. Additionally, most systems rely on static knowledge cut-off dates—for instance, September 2021 for GPT-4—which limits incorporation of the latest evidence [16].

Clinicians often encounter major obstacles when managing molars affected by Class II and III furcation defects, as these teeth show a greater tendency toward loss [17, 18].

Treating molars with Class II or III furcation involvement remains one of the most demanding tasks in dentistry due to their higher susceptibility to tooth loss [17, 18]. The intricate internal architecture of furcation areas significantly complicates adequate cleaning, which represents the principal obstacle to successful therapy [19–21]. Non-surgical interventions rarely produce satisfactory outcomes, and a comprehensive review has indicated that surgical cleaning yields only limited clinical benefits [22–24]. A recently published meta-analysis compared regenerative periodontal procedures with conventional open flap debridement for these cases, also analyzing the effectiveness of different regenerative methods [25]. Results from 20 randomized controlled trials revealed that regenerative protocols consistently achieved better results than open flap debridement in improving furcation conditions, increasing both horizontal and vertical attachment levels, and reducing probing depths. Treatments incorporating bone graft materials showed the highest likelihood of optimal horizontal bone level recovery, while the combination of bone grafts with non-resorbable membranes ranked best for vertical attachment gain and reduction in pocket depth [25].

The ability of chatbots to accurately respond to clinically essential questions has been assessed across various dental specialties…

The capability of AI-driven chatbots to deliver precise answers to clinically important questions has been examined in numerous dental disciplines, including pediatric [26], operative [27], oral and maxillofacial radiology [28], orthodontics [29], community [30], endodontic [31], prosthodontic [32], oral pathology [33], dental trauma [34], periodontal [35], and implant-related dentistry [36]. Early investigations also explored how facial enhancement tools such as *FaceApp* could support orthodontic planning by digitally adjusting facial proportions. Researchers found that AI-enhanced images were generally rated as more attractive, with visible alterations in features like lip thickness, eye proportions, and lower facial height—implying potential use for individualized, soft-tissue-oriented orthodontic designs [37]. As large language models (LLMs) increasingly emerge as reference tools in dentistry, it is vital to examine their reliability and precision. Because furcation defect management requires nuanced decision-making and individualized planning, it is important to test whether LLMs can generate responses consistent with current scientific knowledge. This study therefore evaluated and compared four LLMs for their ability to provide accurate, evidence-based explanations to common clinical queries related to periodontal furcation treatment. The null hypothesis proposed no significant variation among models regarding accuracy, completeness, clarity, or scientific consistency with established guidelines.



## Materials and Methods

The investigation assessed how effectively four leading LLMs could formulate evidence-supported answers about the management of periodontal furcation lesions. Out of a broader set of ten periodontal questions **(Table 1)** derived from the European Federation of

Periodontology (EFP) S3 periodontitis guidelines [38], a subset focusing on furcation treatment was selected to reflect cases typically encountered by practicing dentists. Each model's output was compared with a benchmark "gold standard" compiled from EFP recommendations [38] and recent systematic reviews addressing furcation defect therapy [39–46]. The LLMs evaluated included ChatGPT (GPT-4.0), Google Gemini (2.0 Flash Experimental), Google Gemini Advanced (Gemini Ultra 1.0), and Microsoft Copilot (Free Version).

**Table 1.** Open-ended clinical questions answered using ChatGPT 4.0, Google Gemini, Google Gemini Advanced, and Microsoft Copilot

| Query ID | Question Summary |
|---|---|
| 1 | How to optimally treat molars with Class II and III furcation defects and persistent pockets? |
| 2 | What is the most effective approach for managing deep residual pockets linked to mandibular Class II furcation defects? |
| 3 | What is the ideal therapy for deep residual pockets associated with maxillary buccal Class II furcation defects? |
| 4 | Which regenerative biomaterials are best suited for treating persistent deep pockets in Class II mandibular and maxillary buccal furcation defects? |
| 5 | What is the preferred treatment for maxillary interdental Class II furcation defects? |
| 6 | What is the optimal strategy for addressing maxillary Class III furcation defects? |
| 7 | What is the most suitable treatment for mandibular Class III furcation defects? |
| 8 | Does the addition of local medications to subgingival scaling enhance outcomes for furcation defects? |
| 9 | Does the use of systemic antibiotics improve therapeutic results for furcation defects? |
| 10 | Which imaging method is most effective for evaluating furcation defects? |

Two periodontists certified by the American Board of Periodontology (G.S.C. and V.P.K.) independently reviewed the responses. They rated each LLM's answers according to a standardized 0–10 scale assessing clarity, depth, scientific precision, and relevance [29], comparing them against the established benchmark. Each question was submitted to each model once on December 13, 2024, with no follow-up prompts. The reviewers repeated the evaluation one month later to determine scoring consistency.

*Statistical analysis*

Data analysis involved descriptive statistics, correlation tests (Pearson's r and Spearman's ρ), Cronbach's α, and intraclass correlation coefficient (ICC) to assess inter-rater reliability and overall scoring consistency. Nonparametric tests, including Wilcoxon's and Friedman's, were used to identify significant differences ($p < 0.05$) among the LLMs' scores for furcation-related questions. Additional comparison across all four models was performed using the Kruskal–Wallis test. All statistical computations were conducted in SPSS version 29.0 (IBM, Armonk, NY, USA) with a significance level set at 0.05.

## Results

Responses to ten clinical questions about the management and treatment of periodontal furcation defects were generated by four LLMs: Microsoft Copilot, Google Gemini Advanced, ChatGPT 4.0 and Google Gemini. These outputs were compared against a guideline- and evidence-based reference, detailed in Supplementary Table S1. Each response was evaluated independently by two periodontology specialists on four metrics—comprehensiveness, scientific accuracy, clarity, and relevance—using a 0–10 scoring system. The assessments were performed twice, with a one-month interval between sessions. Descriptive statistics for these scores are summarized in Table 2. Overall, Google Gemini Advanced achieved the highest mean scores, whereas Google Gemini and Microsoft Copilot had the lowest.

**Table 2.** Summary of descriptive statistics for the evaluations of Google Gemini, Microsoft Copilot, ChatGPT 4.0 and Google Gemini Advanced across two scoring sessions by two evaluators

| Model | Rating 1 | Rating 2 |
|---|---|---|
| | ChatGPT 4.0 | Google Gemini |
| Assessor | A | B |
| Average Score | 6.0 | 6.0 |
| Std. Error | 0.8 | 0.8 |
| Midpoint Value | 6.0 | 6.0 |
| Lowest Score | 2.0 | 2.0 |
| Highest Score | 9.0 | 9.0 |
| Std. Deviation | 2.4 | 2.4 |
| Score Variance | 5.6 | 5.6 |

Correlation analyses using Pearson and Spearman's rho **(Table 3)** revealed strong agreement between the two evaluators across all LLMs and time points, indicating consistent scoring patterns [47, 48]. Inter-evaluator reliability was further confirmed through Cronbach's α and Intraclass Correlation Coefficient (ICC) analyses, as shown in **Table 4.** Additionally, Wilcoxon and Friedman tests **(Table 5)** showed no statistically significant differences between scores assigned by the two evaluators in either session or when pooled together.

**Table 3.** Pearson and Spearman correlations between the two evaluators' scores for all four LLMs at two time points

| AI Models [Assessors A–B] | Rating 1 | Rating 2 |
|---|---|---|
| | Pearson Coefficient | Spearman Rank |
| ChatGPT 4.0 | 1.000 ($p < 0.001$) | 1.000 (-) |
| Google Gemini | 1.000 ($p < 0.001$) | 1.000 (-) |
| Gemini Advanced | 0.985 ($p < 0.001$) | 0.975 ($p < 0.001$) |
| Microsoft CoPilot | 1.000 ($p < 0.001$) | 1.000 (-) |

**Table 4.** Cronbach α and ICC values demonstrating inter-evaluator reliability for scores given to Microsoft Copilot, ChatGPT 4.0, Google Gemini Advanced and Google Geminiacross the two sessions and combined data

| AI Systems | Rating 1 | Rating 2 | Combined Ratings 1 and 2 |
|---|---|---|---|
| | Cronbach's Alpha | Interclass Correlation (Single) | Interclass Correlation (Average) |
| ChatGPT 4.0 | 1.000 | 1.000 ($p < 0.001$) | 1.000 ($p < 0.001$) |
| Google Gemini | 1.000 | 1.000 ($p < 0.001$) | 1.000 ($p < 0.001$) |
| Gemini Advanced | 0.992 | 0.983 ($p < 0.001$) | 0.992 ($p < 0.001$) |
| Microsoft CoPilot | 1.000 | 1.000 ($p < 0.001$) | 1.000 ($p < 0.001$) |

An overall average score was then calculated for each model. Google Gemini Advanced scored the highest (6.80), while Google Gemini (5.70) and Microsoft Copilot (5.68) had the lowest mean scores **(Table 6)**. **Figure 1** illustrates the overall average scores, and **Figure 2** shows the scores for each individual question. Only Questions 1 and 10 consistently received scores above 7 across all LLMs. The Kruskal–Wallis test **(Table 7)** indicated no statistically significant differences among the models' average scores ($p > 0.05$), suggesting similar overall performance.

**Table 5.** Wilcoxon rank test assessing ratings assigned by two assessors to responses from ChatGPT 4.0, Google Gemini, Google Gemini Advanced, and Microsoft CoPilot at two distinct time points. Friedman rank test evaluating combined ratings from both assessors. Statistical analysis revealed no notable variations in the ratings provided by the two assessors for initial, follow-up, and aggregated assessments

| AI Systems [Assessors A–B] | Rating 1 | Rating 2 | Combined Ratings 1 and 2 |
|---|---|---|---|
| | Wilcoxon Rank Test | Friedman Rank Test | Combined Test Result |
| ChatGPT 4.0 | 1.000 | 0.157 | 1.000 |
| Google Gemini | 1.000 | 1.000 | 1.000 |
| Gemini Advanced | 0.157 | 0.157 | 1.000 |
| Microsoft CoPilot | 1.000 | 0.317 | 1.000 |

**Table 6.** Average scores of the four LLMs

| Metric | ChatGPT 4.0 | Google Gemini | Gemini Advanced | Microsoft CoPilot |
|---|---|---|---|---|
| Average Rating | 5.95 | 5.70 | 6.80 | 5.68 |
| Std. Error | 0.73 | 0.87 | 0.72 | 0.82 |
| Midpoint Score | 6.00 | 6.00 | 6.75 | 6.00 |
| Lowest Value | 2.00 | 2.00 | 3.00 | 2.00 |
| Highest Value | 8.75 | 10.00 | 10.00 | 9.00 |
| Std. Deviation | 2.30 | 2.75 | 2.29 | 2.60 |
| Score Variance | 5.29 | 7.57 | 5.23 | 6.78 |

**Table 7.** Kruskal–Wallis analysis comparing the mean scores of Microsoft Copilot, ChatGPT 4.0, Google Gemini Advanced and Google Gemini

| AI Systems [Mean Ratings] | Kruskal–Wallis Test (Bonferroni-Adjusted p-Value for Multiple Comparisons) |
|---|---|
| ChatGPT 4.0 vs. Google Gemini | 0.870 (1.000) |
| ChatGPT 4.0 vs. Gemini Advanced | 0.329 (1.000) |
| ChatGPT 4.0 vs. Microsoft CoPilot | 0.938 (1.000) |
| Google Gemini vs. Gemini Advanced | 0.254 (1.000) |
| Google Gemini vs. Microsoft CoPilot | 0.931 (1.000) |
| Gemini Advanced vs. Microsoft CoPilot | 0.292 (1.000) |

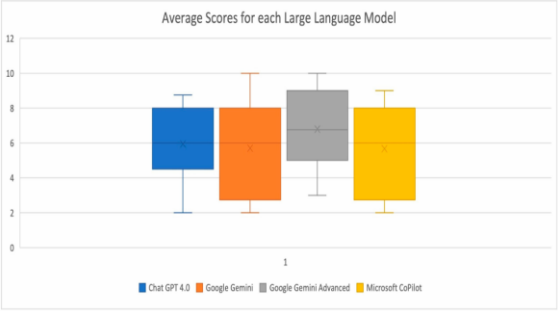

**Figure 1.** Overall average scores of each LLM

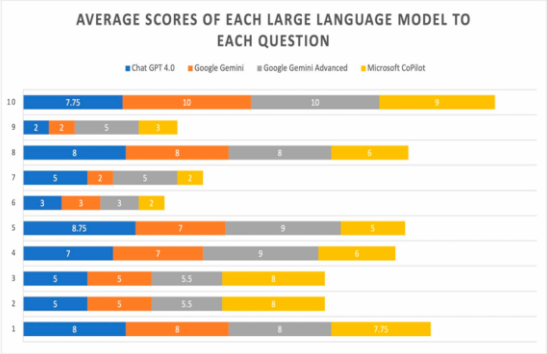


**Figure 2.** Average scores for individual questions across all LLMs

**Table 8** presents descriptive statistics for each LLM's performance across the four evaluation metrics. Google Gemini Advanced generally achieved higher mean scores in comprehensiveness and clarity, whereas Microsoft Copilot tended to score lower, particularly in relevance. The variability in scores (range from minimum to maximum) demonstrates differences in answer quality both within each LLM and across criteria. While no LLM outperformed all others consistently in every category, Google Gemini Advanced showed a trend toward higher comprehensiveness and clarity, and Microsoft Copilot tended toward lower relevance.

**Table 8.** Mean scores and variability for each LLM across comprehensiveness (1), scientific accuracy (2), clarity (3), and relevance (4)

| AI System | ChatGPT 4.0 | Google Gemini | Gemini Advanced | Microsoft CoPilot |
|---|---|---|---|---|
| Evaluated Aspects | A | B | C | D |
| Average Rating | 6.0 | 6.2 | 6.0 | 5.2 |
| Std. Error | 0.7 | 0.8 | 0.7 | 0.9 |
| Midpoint Value | 5.5 | 6.0 | 6.0 | 6.0 |
| Lowest Score | 2.0 | 1.5 | 2.0 | 2.0 |
| Highest Score | 9.0 | 9.0 | 9.0 | 8.5 |
| Std. Deviation | 2.2 | 2.5 | 2.3 | 2.9 |
| Rating Variance | 4.9 | 6.0 | 5.4 | 8.2 |

## Discussion

Artificial intelligence integration in healthcare offers notable advantages but also comes with challenges. This study examined the ability of four LLMs to respond accurately to common clinical questions regarding the treatment and management of periodontal furcation defects, comparing their outputs against the EFP S3 Clinical Practice Guidelines [38] and recent systematic reviews and meta-analyses [39–46].

Key observations from this study include:

Google Gemini Advanced consistently obtained the highest average scores, while Google Gemini and Microsoft Copilot were lower.

Statistical analysis using the Kruskal–Wallis test revealed no significant differences in average scores among the four LLMs.

Each model demonstrated unique strengths and limitations across the evaluated criteria, with no single LLM consistently outperforming the others in all areas.

Trends observed suggest that Google Gemini Advanced performed better in terms of comprehensiveness and clarity, while Microsoft Copilot showed lower relevance scores.

These findings indicate that although all LLMs can provide clinically relevant responses, their quality may differ depending on the evaluation metric. Understanding these differences is essential for dental professionals using AI tools for evidence-based clinical decision-making.

In this analysis, pairwise evaluations between the four LLMs did not reveal any statistically meaningful differences in their mean answer scores. The adjusted p-values were uniformly high (1.000), well above the standard 0.05 threshold, indicating comparable overall performance across the models. Among the LLMs, Google Gemini Advanced consistently achieved the highest mean and median ratings, suggesting that evaluators generally judged its outputs more favorably. Conversely, Google Gemini and Microsoft Copilot recorded the lowest mean values, reflecting lower perceived quality of their responses. Examination of standard deviations and variances demonstrated that score dispersion varied among the models. Google Gemini had the broadest spread (SD = 2.75; variance = 7.57), while Google Gemini Advanced displayed the narrowest variation (SD = 2.29; variance = 5.23). The minimum score recorded was 2.00 for all models except Google Gemini Advanced, which had a floor of 3.00. Maximum ratings ranged between 8.75 and 10.00.

Performance across the ten clinical questions was not uniform. Questions 1 and 10 tended to receive higher ratings across all models, whereas questions 2, 6, and 7 generally scored lower, indicating that the complexity and specifics of each query influenced model outputs.

Google Gemini Advanced repeatedly performed strongly, frequently earning the highest or near-highest scores within each question group. Microsoft Copilot showed the greatest inconsistency, scoring well on certain questions (e.g., Question 10) but falling behind on others. ChatGPT 4.0 exhibited a mix of strong and moderate scores, depending on the question, while Google Gemini typically scored in the middle-to-lower range, reflecting a performance gap relative to its advanced sibling. Overall, the findings suggest that simpler questions were handled more effectively by all LLMs, while nuanced or complex items resulted in lower performance.

Regarding comprehensiveness, Google Gemini Advanced produced the most detailed and extensive responses. Microsoft Copilot's answers were often less thorough, while ChatGPT 4.0 and Google Gemini fell in between. In the category of scientific accuracy, Google Gemini Advanced again scored highly, indicating a tendency to provide correct information, whereas Microsoft Copilot achieved the lowest average scores, pointing to occasional inaccuracies. Variations in standard deviations also reveal inconsistencies in accuracy across models.

Clarity was another area where Google Gemini Advanced excelled, consistently generating well-structured and understandable content. Microsoft Copilot remained reasonably clear but lagged slightly behind in comparison. All models showed moderate reliability in clarity. For relevance, Google Gemini Advanced again performed well, generally delivering content aligned with the questions asked. Microsoft Copilot had the lowest mean relevance scores, indicating that some of its answers were occasionally off-topic. Google Gemini also showed a relatively low mean for relevance.

A review of prior studies highlights a gap in research examining LLM accuracy for clinical questions on periodontal furcation defect management, particularly when benchmarked against a "gold standard." One previous investigation tested ChatGPT 4.0, Google Gemini, Google Gemini Advanced, and Microsoft Copilot on ten open-ended queries about peri-implant disease management, including peri-implantitis and mucositis [36]. In that study, Google Gemini Advanced outperformed the other models, while Google Gemini scored the lowest, mirroring trends observed here. Another study in periodontology used a comparable design, evaluating LLMs based on comprehensiveness, scientific accuracy, clarity, and relevance; the results indicated ChatGPT 4.0 performed best, while Google Gemini was least effective [35].

Performance across dental subfields varies by model. In endodontics, ChatGPT showed 57.3% accuracy and 85.4% consistency for binary-choice questions [31]. For oral and maxillofacial surgery, it scored 4.6 ± 0.8 out of 5 on patient-centered questions but only 3.1 ± 1.5 on technical queries [49]. On a board-style multiple-choice dental exam, ChatGPT answered 76.9% of questions correctly, reflecting competent knowledge acquisition [50]. Similarly, ChatGPT 4.0 performed well on open-ended head and neck surgery questions and delivered precise, up-to-date responses on common dental amalgam queries and removal practices [51].

Comparisons of large language models (LLMs) in dental contexts have produced inconsistent findings. One study demonstrated that ChatGPT-4 significantly outperformed ChatGPT-3.5, Bing Chat, and Bard [52]. In the field of pediatric dentistry, ChatGPT achieved the highest accuracy among several LLMs, with researchers suggesting that such models may support both dental education and patient information delivery [53]. Conversely, other research found no statistically meaningful difference between ChatGPT and Google Bard when generating queries about dental caries [9]. In oral and maxillofacial radiology assessments, ChatGPT, ChatGPT Plus, Bard, and Bing Chat generally performed below expectations, although ChatGPT Plus showed better accuracy on foundational knowledge questions [28]. Meanwhile, for clinically relevant orthodontic problems, Microsoft Bing Chat scored highest, surpassing both ChatGPT 3.5 and Google Bard [29]. In endodontics, GPT-3.5 generated more reliable responses than Google Bard or Bing [37]. Taken together, these results highlight that direct performance comparisons are difficult due to variability in methods and specialty areas.

Several methodological considerations should be noted. LLM performance can fluctuate depending on the phrasing of questions and the technical depth required in answers. This underscores the need for further studies investigating how question complexity affects output accuracy and relevance. To control for bias from factors such as question count, wording, and specificity, follow-up prompts were excluded. Each question was posed only once to standardize comparison, though this does not fully reflect real-world usage, where iterative questioning often occurs. Open-ended questions, while potentially producing incomplete or biased responses, more accurately represent the types of inquiries clinicians typically pose and enable a more realistic evaluation of LLM capabilities. Furthermore, the breadth and quality of an LLM's training data influence its ability to generate

precise answers, which may account for observed variability. Even when using a well-defined "gold standard" for evaluation, limitations such as training cutoffs and restricted access to paywalled literature remain important factors affecting response quality.

Responses were independently scored on two occasions by two periodontists certified by the American Board of Periodontology. Using a predefined benchmark, the evaluation achieved high inter- and intra-rater reliability, reducing the influence of individual subjectivity. Ten carefully chosen questions covering periodontal furcation defect management and treatment were used to assess the models. A structured scoring rubric rated each answer for comprehensiveness, scientific accuracy, clarity, and relevance, ensuring consistent and objective assessment. Open-ended formats allowed evaluation of not only factual correctness but also depth, clarity, and applicability.

The clinical relevance of LLM assessment is emphasized by the complexity of furcation defect management, where multiple treatment approaches exist and accurate information is essential for patient outcomes. This study offers a point-in-time evaluation of four widely available models (ChatGPT 4.0, Google Gemini, Google Gemini Advanced, and Microsoft Copilot), providing a baseline despite ongoing model updates. The findings illuminate current strengths and limitations, offering guidance for future model development, responsible clinical adoption, and the continued need for human oversight in AI-assisted dental decision-making.

Future research should explore a broader spectrum of clinical scenarios and question types. Efforts to refine and validate LLMs could improve information security, support clinical recommendation generation, and enhance patient care. While LLMs have potential as tools for dental professionals and patients, further work is needed to determine how they can optimally contribute to patient outcomes and experience.

## Conclusions

This investigation evaluated four LLMs' ability to answer clinical questions related to periodontal furcation defects. Results indicate that these models hold promise, though performance differs by model and evaluation metric. Google Gemini Advanced generally performed best, particularly in terms of comprehensiveness and clarity, whereas Google Gemini and Microsoft Copilot tended to score lower. Despite these differences, statistical analyses revealed no significant variation in mean scores across models, suggesting comparable overall performance. Variability in responses across individual questions highlights the influence of question complexity on answer quality.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| LLM | Large Language Model |

**Acknowledgments:** None

**Conflict of Interest:** None

**Financial Support:** None

**Ethics Statement:** None

## References

1. Orrù G, Piarulli A, Conversano C, Gemignani A. Human-like problem-solving abilities in large language models using ChatGPT. Front Artif Intell. 2023;6:1199350.
2. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5:194.
3. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform. 2023;25:bbad493.
4. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. Radiology. 2023;307:e230725.
5. Jiang LY, Liu XC, Pour Nejatian N, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023;619:357–62.
6. Zhou LQ, Wu XL, Huang SY, Wu GG, Ye HR, Wei Q, et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. Radiology. 2020;294:19–28.
7. Rim TH, Lee CJ, Tham YC, Cheung N, Yu M, Lee G, et al. Deep-learning-based cardiovascular risk stratification using coronary artery calcium scores predicted from retinal photographs. Lancet Digit Health. 2021;3:e306–e316.

8. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: Chances and challenges. J Dent Res. 2020;99:769–74.
9. Ahmed WM, Azhari AA, Fawaz KA, Ahmed HM, Alsadah ZM, Majumdar A, et al. Artificial intelligence in the detection and classification of dental caries. J Prosthet Dent. 2023;133:1326–32.
10. Li S, Liu J, Zhou Z, Zhou Z, Wu X, Li Y, et al. Artificial intelligence for caries and periapical periodontitis detection. J Dent. 2022;122:104107.
11. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80.
12. Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices. 2023;1:731–8.
13. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inform. 2024;151:104620.
14. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and other large language models are double-edged swords. Radiology. 2023;307:e230163.
15. Piwowar H, Priem J, Larivière V, Alperin JP, Matthias L, Norlander B, et al. The state of OA: A large-scale analysis of the prevalence and impact of open access articles. PeerJ. 2018;6:e4375.
16. McGrath SP, Kozel BA, Gracefo S, Sutherland N, Danford CJ, Walton N. A comparative evaluation of ChatGPT 3.5 and ChatGPT 4 in responses to selected genetics questions. J Am Med Inform Assoc. 2024;31:2271–83.
17. Nibali L, Zavattini A, Nagata K, Di Iorio A, Lin GH, Needleman I, et al. Tooth loss in molars with and without furcation involvement—a systematic review and meta-analysis. J Clin Periodontol. 2016;43:156–66.
18. Sanz M, Jepsen K, Eickholz P, Jepsen S. Clinical concepts for regenerative therapy in furcations. Periodontol 2000. 2015;68:308–32.
19. Al-Shammari KF, Kazor CE, Wang HL. Molar root anatomy and management of furcation defects. J Clin Periodontol. 2001;28:730–40.
20. Jepsen S, Deschner J, Braun A, Schwarz F, Eberhard J. Calculus removal and the prevention of its formation. Periodontol 2000. 2011;55:167–88.
21. Svärdström G, Wennström JL. Furcation topography of the maxillary and mandibular first molars. J Clin Periodontol. 1988;15:271–5.
22. Loos B, Nylund K, Claffey N, Egelberg J. Clinical effects of root debridement in molar and non-molar teeth: A 2-year follow-up. J Clin Periodontol. 1989;16:498–504.
23. Nordland P, Garrett S, Kiger R, Vanooteghem R, Hutchens LH, Egelberg J. The effect of plaque control and root debridement in molar teeth. J Clin Periodontol. 1987;14:231–6.
24. Graziani F, Gennai S, Karapetsa D, Rosini S, Filice N, Gabriele M, et al. Clinical performance of access flap in the treatment of class II furcation defects. J Clin Periodontol. 2015;42:169–81.
25. Jepsen S, Gennai S, Hirschfeld J, Kalemaj Z, Buti J, Graziani F. Regenerative surgical treatment of furcation defects: A systematic review and Bayesian network meta-analysis. J Clin Periodontol. 2020;47:352–74.
26. Dermata A, Arhakis MA, Makrygiannakis K, Giannakopoulos EG, Kaklamanos EG. Evaluating the evidence-based potential of six large language models in paediatric dentistry: A comparative study on generative artificial intelligence. Eur Arch Paediatr Dent. 2025;26:527–35.
27. Ahmed WM, Azhari AA, Alfaraj A, Alhamadani A, Zhang M, Lu CT. The quality of AI-generated dental caries multiple choice questions: A comparative analysis of ChatGPT and Google Bard language models. Heliyon. 2024;10:e28198.
28. Jeong H, Han SS, Yu Y, Kim S, Jeon KJ. How well do large language model-based chatbots perform in oral and maxillofacial radiology? Dentomaxillofac Radiol. 2024;53:390–5.
29. 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. 2024;cjae017.
30. Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, et al. Implications of ChatGPT in public health dentistry: A systematic review. Cureus. 2023;15:e40367.
31. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57:108–13.
32. Freire Y, Laorden AS, Pérez JO, Sánchez MG, García VDF, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131:659.e1–659.e6.

33. Albagieh H, Alzeer ZO, Alasmari ON, Alkadhi AA, Naitah AN, Almasaad KF, et al. Comparing artificial intelligence and senior residents in oral lesion diagnosis: A comparative study. Cureus. 2024;16:e51584.
34. Ozden I, Gokyar M, Ozden ME, SazakOvecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024;40:722–9.
35. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: Assessing their performance in clinically relevant questions. J Prosthet Dent. 2024;in press.
36. Koidou VP, Chatzopoulos GS, Tsalikis L, Kaklamanos EG. Large language models in peri-implant disease: How well do they perform? J Prosthet Dent. 2025;in press.
37. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J. 2023;57:305–14.
38. Sanz M, Herrera D, Kebschull M, Chapple I, Jepsen S, Beglundh T, et al. Treatment of stage I–III periodontitis—the EFP S3 level clinical practice guideline. J Clin Periodontol. 2020;47:4–60.
39. Chatzopoulos GS, Koidou VP, Tsalikis L. Local drug delivery in the treatment of furcation defects in periodontitis: A systematic review. Clin Oral Investig. 2023;27:955–70.
40. Das RK, Bharathwaj VV, Sindhu R, Prabu D, Rajmohan M, Dhamodhar D, et al. Comparative analysis of various forms of local drug delivery systems on a class 2 furcation: A systematic review. J Pharm Bioallied Sci. 2023;15:S742–S746.
41. Nibali L, Buti J, Barbato L, Cairo F, Graziani F, Jepsen S. Adjunctive effect of systemic antibiotics in regenerative/reconstructive periodontal surgery: A systematic review with meta-analysis. Antibiotics. 2021;11:8.
42. Chiou LL, Herron B, Lim G, Hamada Y. The effect of systemic antibiotics on periodontal regeneration: A systematic review and meta-analysis of randomized controlled trials. Quintessence Int. 2023;54:210–19.
43. Choi IGG, Cortes ARG, Arita ES, Georgetti MAP. Comparison of conventional imaging techniques and CBCT for periodontal evaluation: A systematic review. Imaging Sci Dent. 2018;48:79–86.
44. Walter C, Schmidt JC, Rinne CA, Mendes S, Dula K, Sculean A. Cone beam computed tomography (CBCT) for diagnosis and treatment planning in periodontology: Systematic review update. Clin Oral Investig. 2020;24:2943–58.
45. Assiri H, Dawasaz AA, Alahmari A, Asiri Z. Cone beam computed tomography (CBCT) in periodontal diseases: A systematic review based on the efficacy model. BMC Oral Health. 2020;20:191.
46. Jolivet G, Huck O, Petit C. Evaluation of furcation involvement with diagnostic imaging methods: A systematic review. Dentomaxillofac Radiol. 2022;51:20210529.
47. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
48. Hinkle DE, Wiersma W, Jurs SG. Applied statistics for the behavioral sciences. 5th ed. Boston, MA: Houghton Mifflin; 2003.
49. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023;124:101471.
50. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: A preliminary study on ChatGPT. J Am Dent Assoc. 2023;154:970–4.
51. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: A multicenter collaborative analysis. Otolaryngol Head Neck Surg. 2024;170:1492–503.
52. Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. J Med Internet Res. 2023;25:e51580.
53. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. J Dent. 2024;144:104938.