BMC Oral Health

RESEARCH Open Access

Dens invaginatus as a diagnostic challenge: evaluating large language models against expert endodontic reasoning

Damla Erkal[1*], Turgut Felek[2,3], Oana-Paula Butean[4] and Kürşat Er[5]

Abstract

Introduction This study hypothesized that large language models (LLMs) would underperform compared to expert clinicians in diagnosing and managing complex endodontic anomalies, such as dens invaginatus, when provided with periapical radiographs. Although LLMs have shown promise in dental education and basic diagnostics, their effectiveness in nuanced clinical reasoning has remained unclear.

Methods Nineteen anonymized periapical radiographs depicting challenging endodontic conditions were paired with clinical vignettes. Six advanced LLMs and one expert endodontist independently answered six structured clinical questions per case. Each response was scored against a reference key. Accuracy rates were compared using Kruskal-Wallis and Mann-Whitney U tests. Chi-square tests were used to evaluate model performance across question types.

Results The expert achieved 100% accuracy, while all LLMs performed significantly lower ($P < 0.05$). Copilot demonstrated the lowest scores across all questions. The most substantial performance drop was observed in anomaly classification tasks, particularly in identifying and categorizing dens invaginatus. No significant performance differences were found among the top-performing LLMs.

Conclusions While LLMs showed competence in basic diagnostic tasks, they failed to replicate expert-level decision-making in complex endodontic scenarios. Their current capabilities remain insufficient for unsupervised clinical use. This study is among the first to assess LLMs using real radiographic data in endodontics and highlights the need for further multimodal model development.

Keywords Dens invaginatus, Artificial intelligence, Large language models

*Correspondence:
Damla Erkal
damla.4716@gmail.com
[1]Department of Endodontics, Faculty of Dentistry, Burdur Mehmet Akif Ersoy University, Burdur, Türkiye
[2]Institute of Naturel and Applied Sciences, Akdeniz University, Antalya, Türkiye
[3]Faculty of Dentistry, Akdeniz University, Antalya, Türkiye
[4]Pharmacy, Science, and Technology of Târgu-Mureș, George Emil Palade University of Medicine, Târgu-Mureș, Romania
[5]Department of Endodontics, Faculty of Dentistry, Akdeniz University, Antalya, Türkiye

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Artificial intelligence (AI) systems hold great promise for transforming medical and dental practices by offering innovative approaches to complex clinical challenges and streamlining the decision-making processes faced by healthcare professionals [1]. One of the primary advantages of AI is its capacity to rapidly process and interpret extensive datasets. This feature is especially advantageous in dental radiographic analysis, where AI-based models can detect subtle patterns or abnormalities that may elude human observers [2]. The emergence of advanced large language models (LLMs), including ChatGPT 3.5 and 4 developed by OpenAI, has marked a significant milestone in the evolution and practical deployment of AI Technologies [3].

Dental radiographs frequently reveal complex endodontic conditions such as dens invaginatus (DI), internal resorption, canal obliteration, or broken instruments, each of which poses significant challenges for accurate diagnosis and treatment planning [4]. The accurate identification and management of such anomalies often rely on both advanced imaging interpretation skills and sound clinical reasoning.

Despite the expanding capabilities of LLMs in medical and dental applications, their potential to support or replicate expert-level performance in endodontic diagnosis and treatment planning has not been adequately explored. Current literature focuses predominantly on caries detection and basic radiographic assessments, leaving a significant gap in understanding how LLMs might perform in more intricate clinical situations [5]. This gap is particularly evident in cases involving developmental anomalies such as DI, a condition characterized by enamel and dentin invagination into the pulp space, which often leads to complex morphological presentations and diagnostic challenges on radiographs [6]. Accurate identification and classification of DI typically require advanced imaging interpretation skills and substantial clinical experience, making it an ideal test case for evaluating the reasoning capabilities of LLMs.

Misdiagnosis of DI may lead to delayed or inappropriate treatment, persistent infection, perforation risk during instrumentation, and compromised long-term prognosis. These consequences underscore the importance of accurate anomaly classification beyond detecting periapical pathology [4].

Therefore, this study aims to assess the diagnostic, therapeutic, and prognostic decision-making performance of advanced LLMs when presented with standardized clinical scenarios based on periapical radiographic images. By comparing their accuracy to that of an experienced endodontist, the study seeks to evaluate whether LLMs can serve as reliable tools in supporting expert-level decision-making in complex endodontic cases.

Materials and methods

Study design

This observational and comparative in vitro study was conducted to evaluate the diagnostic, treatment planning, and prognostic capabilities of LLMs in interpreting dental radiographic images. The periapical radiographs used in this study were obtained from DI cases referred to the Department of Endodontics, Akdeniz University, between 2015 and 2025.

Assessment form and question structure

A total of 19 dental periapical radiographs from past clinical cases were used in this study. The images were selected to represent clear endodontic pathologies, including internal resorption, DI, periapical radiolucency, canal obliteration, periapical cysts, and separated files. Each radiograph was accompanied by a case scenario structured to enhance clinical context, including patient age, gender, symptoms, and anamnesis.

A custom-designed questionnaire consisting of case-based diagnostic scenarios was developed to compare the clinical reasoning of LLMs with that of expert endodontists. The questionnaire was tailored to focus on dens invaginatus, a diagnostic challenge in endodontics, and has not been previously published. The English version of the questionnaire is available as Supplementary File 1. Each case was evaluated based on a structured questionnaire consisting of 6 clinical decision-making questions. The questions addressed: (1) presence of pathology (2), anomaly diagnosis (e.g., type of dens invaginatus) (3), need for treatment (4), proposed treatment type (5), treatment difficulty, and (6) prognostic evaluation (Fig. 1). Each question had a single correct answer and was scored in a binary fashion (1 = correct, 0 = incorrect).

Participant groups and model characteristics

Seven evaluators assessed the cases, including one expert endodontist and 6 LLMs. We evaluated the following models:

GPT-4o (OpenAI; released May 13, 2024), gpt-4o-mini (OpenAI; July 18, 2024), o1 (OpenAI; announced Sept 12, 2024), and o1-pro (OpenAI; introduced Dec 5, 2024). We also included Claude 3.5 Sonnet (Anthropic; June 20, 2024) and Microsoft Copilot (Microsoft; model routing service). For consistency, we refer to 'gpt-4o-mini' instead of 'GPT-o4-mini-high' and 'o1-pro' instead of 'GPT-o1 Pro' throughout the manuscript.

Two additional LLM systems, Gemini (Google) and DeepSeek (DeepSeek AI), were initially considered for inclusion. However, Gemini LLM was excluded as it did not support visual data input, and DeepSeek LLM was unable to process image uploads, rendering both models inapplicable for this visual diagnostic task.

1. On the periapical radiograph, what type of developmental malformation may be noticed?
2. What is the incidence of this developmental malformation reported in the literature?
3. Is there any classification for this developmental malformation?
4. If yes, which type of developmental malformation is present in this radiograph based on the classification?
5. Are there other names for this developmental malformation in the literature?
6. What additional anomalies could this malformation be associated with?
7. What treatment protocols can be applied?

Fig. 1 Structured questionnaire addressing diagnosis, classification, treatment need, options, difficulty, and prognosis of dens invaginatus cases

The expert evaluator was an endodontist with at least 5 years of clinical experience. All evaluators were presented with randomized and anonymized cases in a standardized order. Interactions with the AI models were conducted through text prompts, and their responses were manually transcribed into the evaluation forms by the researchers.

Scoring of responses

Each response provided by the models or the expert was scored based on the reference answer provided by the expert evaluator. Answers were classified as either correct or incorrect, resulting in a total of 114 responses per group (19 cases × 6 questions). Both total scores and question-level data were included in the analysis.

Before any model interaction, a reference answer key was developed a priori by the expert based on established diagnostic criteria, including Hülsmann's dens invaginatus classification, and evidence-based treatment principles outlined in Cohen's Pathways of the Pulp [12th edition]. The key defined a single correct answer for each of the six questions per case. Once finalized, the key was locked and not modified after data collection.

Statistical analysis

All statistical analyses were performed using SPSS version 26.0 (IBM Corp., Armonk, NY, USA). Normality of total scores for each evaluator was assessed using Kolmogorov-Smirnov and Shapiro-Wilk tests. As the data were not normally distributed, non-parametric statistical methods were employed.

Differences in total correct responses among the 7 groups were analyzed using the Kruskal-Wallis H test. When a significant difference was found, pairwise group comparisons were conducted using the Mann-Whitney U test as a post-hoc analysis. These comparisons included both 'expert vs LLM' and 'LLM vs LLM' evaluations. For 21 pairwise comparisons, Bonferroni-adjusted significance was $\alpha = 0.05/21 = 0.00238$.

Additionally, Chi-square tests were used to compare the distribution of response accuracy among different AI models and the expert clinician for each question. A significance level of $P < 0.05$ was adopted for all statistical analyses.

Table 1 Accuracy (%) with 95% confidence intervals (CI) for expert and AI models, Mann–Whitney U post-hoc pairwise comparisons with bonferroni correction, and *p*-values for comparisons against the expert

Groups 1	Accuracy (%)	95% CI	*p*-value (expert vs. model)
Expert	100.0	96.7–100.0	- [a]
Claude	90.4	83.5–94.5	0.000 [b]
Copilot	50.0	41.0–59.0	0.000 [c]
GPT-4o	74.6	65.9–81.7	0.001 [b]
GPT-o1	83.3	75.4–89.1	0.001 [b]
GPT-o1 Pro	93.9	87.9–97.0	0.004 [b]
GPT-4 Mini	87.7	80.4–92.5	0.002 [b]

Accuracy values are presented with corresponding 95% CIs. Different superscript letters indicate statistically significant differences between groups

Results

Total correct responses differed significantly across groups (Kruskal–Wallis, $P < 0.05$); post-hoc tests showed the expert outperformed all LLMs, with Copilot performing worst (Table 1).

To evaluate whether there were statistically significant differences in total correct response scores between the expert clinician and the various AI models, a Kruskal-Wallis H test was performed to compare the diagnostic performance among the groups. The results indicated a statistically significant difference between them ($P < 0.05$). To identify the specific

Table 2 Accuracy rates (%) of the expert and AI models for each clinical question (Q1–Q6) with Chi-square test results

	Expert	Claude	Copilot	GPT-4o	GPT-o1	GPT-o1 Pro	GPT-4 Mini	χ^2	*p*
Q1	100%	100%	57.9%	78.9%	89.5%	100%	78.9%	31.294	0.000
Q2	100%	100%	57.9%	78.9%	89.5%	100%	89.5%	31.294	0.000
Q3	100%	42.1%	10.5%	52.6%	52.6%	63.2%	57.9%	32.523	0,000
Q4	100%	100%	57.9%	78.9%	89.5%	100%	100%	31.294	0.000
Q5	100%	100%	57.9%	78.9%	89.5%	100%	100%	31.294	0,000
Q6	100%	100%	57.9%	78.9%	89.5%	100%	100%	31.294	0.000

Values indicate the percentage of correct responses for each clinical question. χ^2 and *p*-values correspond to Chi-square tests comparing accuracy distributions across all groups for each question

group differences responsible for this significance, post-hoc pairwise comparisons were carried out using the Mann-Whitney U test, as shown in Table 1. These analyses revealed that the expert clinician's performance differed significantly from that of all AI models. ($P < 0.05$). The expert clinician exhibited statistically significantly higher accuracy scores than all AI models ($P < 0.05$). Furthermore, Copilot performed significantly worse than Claude, GPT-4o, GPT-o1, GPT-o1 Pro, and GPT-4 Mini ($P < 0.05$), indicating that Copilot was the least accurate among all models. No statistically significant differences were observed among the other AI models (e.g., Claude vs. GPT-o1, GPT-o1 vs. GPT-4 Mini, etc.), suggesting broadly comparable performance across higher-performing AI systems.

Across all questions, the expert clinician consistently achieved 100% accuracy, demonstrating superior diagnostic reliability. Among AI models, performance varied considerably:

- GPT-o1 Pro exhibited high performance across most questions, but its accuracy decreased to 63.2% in Q3.
- Claude and GPT-4 Mini maintained perfect accuracy on all questions except Q3, where they dropped to 42.1% and 57.9%, respectively.
- GPT-4o and GPT-o1 showed moderate performance, typically scoring between 52.6% and 89.5%.
- Copilot consistently displayed the lowest performance, with a particularly sharp decline in Q3 (10.5% accuracy), suggesting substantial limitations in handling complex clinical reasoning.

The Chi-Square test results (χ^2 = 31.294–32.523, $P < 0.001$) for each question revealed statistically significant differences among the groups, as shown in Table 2. These findings indicate that the variation in accuracy scores across different models and the expert was unlikely due to chance. The most pronounced variability was observed in Q3, where model performance showed the widest spread.

Discussion

This study aimed to evaluate and compare the diagnostic, treatment planning, and prognostic decision-making accuracy of 6 advanced LLMs (GPT-4o, gpt-4o-mini, o1, o1-pro, Claude 3.5 Sonnet, and Microsoft Copilot). in interpreting endodontic conditions based on periapical radiographs, using a structured clinical question format. The results revealed that although all LLMs demonstrated a certain degree of agreement with the expert evaluator in some domains, statistically significant differences existed between the expert and each AI model in terms of overall accuracy. These findings suggest that current LLMs, despite their increasing sophistication, have not yet reached expert-level proficiency in managing complex endodontic scenarios.

The study employed a structured and clinically relevant protocol by using actual anonymized periapical radiographs accompanied by case scenarios. This approach allowed the LLMs to be tested in a context that mimics real-world diagnostic challenges. In particular, the study included different types of DI cases that required detailed interpretation of radiographic and clinical information. The questionnaire was structured to evaluate clinical reasoning by including questions on the type of malformation, the incidence of malformation, the classification type of malformation, other names of this malformation, as well as associated anomalies and treatment approaches. This format enabled a multidimensional evaluation of the diagnostic and decision-making performance of the LLMs based on radiographic case scenarios.

In contrast to previous studies [7, 8] that primarily relied on simulated cases or purely text-based prompts, the present study incorporated radiographic data, thereby offering one of the earliest assessments of LLM performance in diagnostic tasks grounded in visual clinical information. Recent advancements in multimodal AI models, such as GPT-4 V, have demonstrated potential in medical image analysis [9], yet applications in dentistry remain limited. Accordingly, the present study contributes to the early exploration of LLM capabilities in image-based diagnostic reasoning within endodontics.

Statistical analysis using the Kruskal-Wallis test revealed significant differences between the expert and

all LLMs in terms of total correct responses ($P < 0.05$). Mann-Whitney U post-hoc tests further confirmed that all LLMs performed significantly below expert-level accuracy. This aligns with previous findings in the literature, where LLMs have struggled with complex, context-driven decisions despite excelling at factual or guideline-based queries [10]. Interestingly, there were no statistically significant differences among the LLMs themselves. This suggests a shared ceiling of performance under the current design and perhaps limitations tied to the underlying GPT architecture. These models, though based on different versions of ChatGPT, may have similar language understanding capabilities when deprived of true visual integration [11, 12]. Chi square test results showed that the most significant discordance occurred in Question 3, which asked about the classification of the malformation. This finding suggests that LLMs may struggle to distinguish between nuanced diagnostic categories in developmental anomalies such as DI. Such classification requires not only recognizing the malformation but also understanding its subtype based on radiographic presentation, a task which remains challenging for current LLMs [11, 13].

The findings highlight that while LLMs have the potential to support clinical workflows, they are not yet suitable as autonomous diagnostic tools in endodontics. Their high performance in basic tasks (e.g., identifying pathology presence or estimating prognosis) shows promise, particularly for educational purposes or clinical decision support systems. However, their underperformance in scenarios involving anomaly classification (e.g., DI subtyping) limits their applicability in frontline care.

The findings of this study align with a growing body of literature demonstrating that, although LLMs may perform adequately on standard or factual tasks, they often fall short in complex clinical cases when compared with experienced clinicians [14]. The failure of LLMs to consistently recognize or appropriately manage such conditions underscores the importance of expert oversight. This reflects a broader trend observed in medical AI, where model hallucinations or shallow reasoning can lead to potentially harmful recommendations [15, 16].

These findings are consistent with prior research in dental AI applications that compared the performance of LLMs to human clinicians. For example, in a systematic review by Schwendicke et al. [5], it was reported that while AI systems, including deep learning models, performed well in caries and periapical lesion detection, expert clinicians consistently outperformed AI in complex diagnostic scenarios. Similarly, Kınıkoğlu et al. [17] compared ChatGPT and Gemini in the context of the Turkish Dental Specialty Exam (DUS) and found that while both models achieved high accuracy in knowledge-based questions, their performance notably declined in image-based tasks, and neither matched expert-level interpretation. Importantly, both studies, in line with our findings, noted that although different LLMs were tested, their diagnostic performance was statistically similar-highlighting a shared limitation in their interpretive reasoning capabilities, particularly when visual inputs are involved.

Although Claude and Gemini have performed well in other domains [18], Gemini was excluded here due to lack of image input support at the time. Future multimodal LLMs with direct image processing capabilities may close the gap observed in this study. Research into GPT-4 V and Gemini 1.5 Pro models has already demonstrated promising results in medical imaging interpretation [17].

A key limitation of this study is that LLMs relied solely on text prompts derived from radiographs rather than directly analyzing the images. This may have constrained their diagnostic depth. Future studies should evaluate multimodal LLMs capable of processing radiographic data directly. Another limitation of our study is that a single experienced endodontist served as the reference standard. Although appropriate for an initial benchmarking study, using one rater prevents estimation of inter-rater variability. Future work should involve multiple experts and report inter-rater agreement (e.g., ICC or κ) to quantify potential variability in the reference key. The small case set ($n = 19$) limits the generalizability of the findings, and heterogeneity in periapical radiograph quality (clarity, resolution, exposure) may have introduced input noise and biased decisions; future studies should standardize image acquisition or rate image quality, and include larger and more diverse case sets.

Furthermore, expanding the range of pathologies and including pediatric or traumatic cases may yield a more generalizable assessment of LLM competence. Evaluating native multimodal models and exploring fine-tuning or instruction-tuning on dental datasets are logical next steps.

Conclusion

These findings provide a realistic baseline for diagnostic support systems in endodontics, highlighting clear targets—robust visual integration and domain-specific training—for future development. While LLMs show potential in dental diagnostics, their current capabilities remain below expert-level accuracy in complex endodontic decision-making, suggesting they may be best suited as adjunctive tools, particularly in educational contexts or preliminary screening. However, continued advancement, especially in multimodal integration and domain-specific fine-tuning, will be critical before they can be reliably implemented in routine dental practice.

Abbreviations

AI	Artificial intelligence
LLM	Large language model
DI	Dens invaginatus

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-06987-z.

Supplementary Material 1.

Supplementary Material 2.

Acknowledgements
The authors would like to thank all participating clinicians for their valuable time and insights.

Authors' contributions
D. E. and K. E. conceptualized the study and designed the methodology. D. E. also wrote the original draft of the manuscript. T.F. was responsible for data curation and conducted the investigation. O-P. B. performed the software-related tasks and prepared the visualizations. K.E. supervised the project, validated the findings, and critically revised the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability
The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
This study was approved by the Clinical Research Ethics Committee of Akdeniz University, Antalya, Türkiye (Approval No: TBAEK 562, Date: 03 July 2025). All procedures were conducted in accordance with the ethical standards of the institutional and national research committees and with the 1964 Helsinki Declaration and its later amendments. Written informed consent was obtained from all participants prior to data collection and analysis.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 11 June 2025 / Accepted: 15 September 2025
Published online: 08 October 2025

References

1. Aminoshariae A, Kulild J, Nagendrababu V. Artificial intelligence in endodontics: current applications and future directions. J Endod. 2021;47(9):1352–7.
2. Mohammad-Rahimi H, Sohrabniya F, Ourang SA, Dianat O, Aminoshariae A, Nagendrababu V, et al. Artificial intelligence in endodontics: data preparation, clinical applications, ethical considerations, limitations, and future directions. Int Endod J. 2024;57(11):1566–95.
3. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare. 2025;13(6):603.
4. Patel S, Durack C, Abella F, Shemesh H, Roig M, Lemberg K. Cone beam computed tomography in endodontics – a review. Int Endod J. 2015;48(1):3–15.
5. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769–74.
6. Hülsmann M. Dens invaginatus: aetiology, classification, prevalence, diagnosis, and treatment considerations. Int Endod J. 1997;30(2):79–90.
7. Tosun B, Öztürk Z. Performance of five large Language models in managing acute dental pain: a comprehensive analysis. Turkish Endodontic J. 2025;10(1):39–49.
8. 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.
9. Li Y, Liu Y, Wang Z, Liang X, Wang L, Liu L, et al. A systematic evaluation of GPT-4V's multimodal capability for medical image analysis. arXiv. 2023;2310:20381.
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80.
11. Farhadi Nia M, Ahmadi M, Irankhah E. Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care. Front Dent Med. 2025;5:1456208.
12. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model, vol. 15. Springer Nature; 2023.
13. Siqueira JF, Rôças IN, Hernández SR, Brisson-Suárez K, Baasch AC, Pérez AR, et al. Dens invaginatus: clinical implications and antimicrobial endodontic treatment considerations. J Endod. 2022;48(2):161–70.
14. Sivari E, Senirkentli GB, Bostanci E, Guzel MS, Acici K, Asuroglu T. Deep Learning in Diagnosis of Dental Anomalies and Diseases: A Systematic Review. Diagnostics. Multidisciplinary Digital Publishing Institute (MDPI); 2023. p. 13.
15. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large language models in medicine: the potentials and pitfalls. Ann Intern Med. 2024;177(2):210–20.
16. Liu Y, Li Y, Wang Z, Liang X, Liu L, Wang L, et al. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. Meta-Radiology. 2024;2(4):100099.
17. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. Cureus. 2025.
18. Lim B, Seth I, Maxwell M, Cuomo R, Ross RJ, Rozen WM. Evaluating the efficacy of large language models in generating medical documentation: a comparative study of ChatGPT-4, ChatGPT-4o, and Claude. Aesthetic Plast Surg. 2025.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.