Research, Society and Development, v. 15, n. 1, e2515150523, 2026
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v15i1.50523

Evaluating Large Language Models performance in Endodontics: A clinical experimental study

Avaliação do desempenho de Modelos de Linguagem de Grande Porte em Endodontia: Um estudo clínico experimental

Evaluación del rendimiento de los Modelos de Lenguaje Grandes en Endodoncia: Un estudio clínico experimental

Received: 01/09/2026 | Revised: 01/12/2026 | Accepted: 01/12/2026 | Published: 01/13/2026

Paloma Rayse Zagalo de Almeida
ORCID: https://orcid.org/0009-0009-7964-4100
University Center of State of Pará, Brazil
E-mail: paloma24900089@aluno.cesupa.br
Igor Amador Barbosa
ORCID: https://orcid.org/0009-0000-4842-1689
University Center of State of Pará, Brazil
E-mail: igor24900085@aluno.cesupa.br
Mauro Sergio Almeida Alves
ORCID: https://orcid.org/0000-0002-6457-4973
University Center of State of Pará, Brazil
E-mail: mauro24900086@aluno.cesupa.br
Silvio Augusto Fernades de Menezes
ORCID: https://orcid.org/0000-0002-1679-9756
University Center of State of Pará, Brazil
E-mail: menezes@cesupa.br
Patricia de Almeida Rodrigues
ORCID: https://orcid.org/0009-0003-6068-9583
University Center of State of Pará, Brazil
E-mail: patriciasouza@cesupa.br
Isaac Souza Elgrably
ORCID: https://orcid.org/0000-0002-1326-4713
University Center of State of Pará, Brazil
E-mail: isaac.elgrably@prof.cesupa.br
Ricardo Roberto de Souza Fonseca
ORCID: https://orcid.org/0000-0003-0312-0553
University Center of State of Pará, Brazil
E-mail: ricardo.fonseca@prof.cesupa.br
João Daniel Mendonça de Moura
ORCID: https://orcid.org/0000-0001-9845-9677
University Center of State of Pará, Brazil
E-mail: joao.moura@prof.cesupa.br

Abstract

This study aims to evaluate the diagnostic accuracy, consistency and diagnostic success rates of eight different AI-based chatbots in Endodontics. This cross-sectional study evaluated diagnostic accuracy of eight diverse AI models, selected for architectural/developer heterogeneity and clinical relevance, using 12 validated fictitious endodontic cases aligned with AAE guidelines and ethical approval was waived as no human data were used. STROBE guidelines were followed to ensure methodological rigor. Standardized prompts ensured uniformity, with three independent executions per case to assess consistency. Responses were anonymized and evaluated by blinded, calibrated reviewers and statistical analysis included Kruskal-Wallis, Dunn's tests, Fleiss' Kappa, and chi-square to compare diagnostic/treatment accuracy and intramodel agreement. The analysis revealed significant diagnostic accuracy variation among AI models ($p < 0.001$), with ChatGPT o1 (97%), Claude (97%), and DeepSeek (90.9%) outperforming Gemini (54.5%). Treatment recommendations showed uniformly high accuracy (97–100%, $p = 0.537$). Multivariate regression confirmed ChatGPT o1 (OR=32.7) and Claude (OR=30.5) as superior, though complex diagnoses (e.g., acute apical abscess, asymptomatic irreversible pulpitis) reduced accuracy (OR=0.01–0.3, $p<0.05$). Stratified analysis identified model-specific vulnerabilities: Gemini failed in reversible pulpitis (0/3, $p=0.001$) and chronic apical abscess (0/3, $p=0.001$), while ChatGPT o1 struggled with acute apical abscess (0/3, $p<0.001$). Overall agreement was 93%, with high intraclass reliability (ICC >0.85) for top models versus Gemini (ICC=0.65). Fleiss'

Kappa highlighted moderate agreement (κ=0.28–0.45) in ambiguous cases, emphasizing heterogeneous reliability. In conclusion, seven AI chatbots demonstrated high accuracy in endodontics cases, being considered as helpful tools for complement of clinical practice.
Keywords: Artificial intelligence; Endodontics; Dental pulp diseases; Diagnosis; Machine learning.

Resumo
Este estudo teve como objetivo avaliar a acurácia diagnóstica, a consistência e as taxas de sucesso diagnóstico de oito chatbots diferentes baseados em IA na área de Endodontia. Este estudo transversal avaliou a acurácia diagnóstica de oito modelos de IA distintos, selecionados por sua heterogeneidade arquitetônica/de desenvolvimento e relevância clínica, utilizando 12 casos fictícios de endodontia validados e alinhados às diretrizes da AAE. A aprovação ética foi dispensada, uma vez que nenhum dado humano foi utilizado. As diretrizes STROBE foram seguidas para garantir o rigor metodológico. Instruções padronizadas asseguraram a uniformidade, com três execuções independentes por caso para avaliar a consistência. As respostas foram anonimizadas e avaliadas por revisores cegos e calibrados. A análise estatística incluiu os testes de Kruskal-Wallis, Dunn, Kappa de Fleiss e qui-quadrado para comparar a acurácia diagnóstica/tratamento e a concordância intramodelo. A análise revelou variação significativa na acurácia diagnóstica entre os modelos de IA ($p < 0,001$), com ChatGPT o1 (97%), Claude (97%) e DeepSeek (90,9%) apresentando desempenho superior ao Gemini (54,5%). As recomendações de tratamento apresentaram alta precisão de forma uniforme (97–100%, $p = 0,537$). A regressão multivariada confirmou a superioridade dos modelos ChatGPT o1 (OR = 32,7) e Claude (OR = 30,5), embora diagnósticos complexos (por exemplo, abscesso apical agudo, pulpite irreversível assintomática) tenham reduzido a precisão (OR = 0,01–0,3, $p < 0,05$). A análise estratificada identificou vulnerabilidades específicas de cada modelo: o Gemini apresentou falhas em casos de pulpite reversível (0/3, $p = 0,001$) e abscesso apical crônico (0/3, $p = 0,001$), enquanto o ChatGPT o1 teve dificuldades com abscesso apical agudo (0/3, $p < 0,001$). A concordância geral foi de 93%, com alta confiabilidade intraclasse (ICC > 0,85) para os melhores modelos em comparação com o Gemini (ICC = 0,65). O coeficiente Kappa de Fleiss destacou concordância moderada (κ = 0,28–0,45) em casos ambíguos, enfatizando a heterogeneidade da confiabilidade. Em conclusão, sete chatbots de IA demonstraram alta precisão em casos de endodontia, sendo considerados ferramentas úteis para complementar a prática clínica.
Palavras-chave: Inteligência artificial; Endodontia; Doenças da polpa dentária; Diagnóstico; Aprendizado de máquina.

Resumen
Este estudio tiene como objetivo evaluar la precisión diagnóstica, la consistencia y las tasas de éxito diagnóstico de ocho chatbots diferentes basados en IA en Endodoncia. Este estudio transversal evaluó la precisión diagnóstica de ocho modelos de IA diversos, seleccionados por heterogeneidad arquitectónica/desarrolladora y relevancia clínica, utilizando 12 casos de endodoncia ficticios validados alineados con las pautas de la AAE y se eximió la aprobación ética ya que no se utilizaron datos humanos. Se siguieron las pautas STROBE para garantizar el rigor metodológico. Las indicaciones estandarizadas aseguraron la uniformidad, con tres ejecuciones independientes por caso para evaluar la consistencia. Las respuestas fueron anonimizadas y evaluadas por revisores ciegos y calibrados, y el análisis estadístico incluyó Kruskal-Wallis, pruebas de Dunn, Kappa de Fleiss y chi-cuadrado para comparar la precisión del diagnóstico/tratamiento y el acuerdo intramodelo. El análisis reveló una variación significativa en la precisión diagnóstica entre los modelos de IA ($p < 0,001$), con ChatGPT o1 (97%), Claude (97%) y DeepSeek (90,9%) superando a Gemini (54,5%). Las recomendaciones de tratamiento mostraron una precisión uniformemente alta (97-100 %, $p = 0,537$). La regresión multivariante confirmó la superioridad de ChatGPT o1 (OR = 32,7) y Claude (OR = 30,5), aunque los diagnósticos complejos (p. ej., absceso apical agudo, pulpitis irreversible asintomática) redujeron la precisión (OR = 0,01-0,3, $p < 0,05$). El análisis estratificado identificó vulnerabilidades específicas del modelo: Gemini falló en pulpitis reversible (0/3, $p = 0,001$) y absceso apical crónico (0/3, $p = 0,001$), mientras que ChatGPT o1 tuvo dificultades con absceso apical agudo (0/3, $p < 0,001$). La concordancia general fue del 93 %, con una alta fiabilidad intraclase (CCI > 0,85) para los mejores modelos frente a Gemini (CCI = 0,65). El índice Kappa de Fleiss mostró una concordancia moderada (κ = 0,28-0,45) en casos ambiguos, lo que indica una fiabilidad heterogénea. En conclusión, siete chatbots de IA demostraron una alta precisión en casos de endodoncia, considerándose herramientas útiles para complementar la práctica clínica.
Palabras clave: Inteligencia artificial; Endodoncia; Enfermedades de la pulpa dental; Diagnóstico; Aprendizaje automático.

1. Introduction

Artificial Intelligence (AI) is a broad term that describes the ability of computers to perform human-like tasks in a sequential manner, enabling them to learn, think, and act autonomously (Ahmed et al., 2021). In essence, AI systems process large volumes of data through iterative learning, building algorithms capable of solving predictive problems without human

intervention or explicit programming (Bonny et al., 2023; Ossowska et al., 2022). This mechanism is made possible by artificial neural networks (NNs), which mimic the functioning of human neurons within a nonlinear mathematical model. In summary, AI science aims to develop intelligent computational systems that exhibit human-like cognitive abilities, such as logical reasoning, problem-solving, language comprehension, and continuous learning (Bonny et al., 2023; Ossowska et al., 2022; Casadei, 2023).

According to Kaplan et al., 2023 and Mukhamediev et al., 2022, AI can be categorized based on its functionalities and capabilities, encompassing fields such as machine learning (ML), deep learning (DL), cognitive computing, natural language processing, robotics, expert systems, and fuzzy logic (Kim et al., 2021). Currently, ML and DL are the most widely used approaches. As described by An et al., 2023, ML involves systems trained through various models and methodologies to automate task resolution (Bonny et al., 2023). DL, a subset of ML, employs artificial neural networks to enhance learning processes. Due to its complex structure, DL enables simultaneous execution of multiple tasks, as well as the analysis and evaluation of diverse data sources, including audio, sensor inputs, and imaging data (Shiammala et al., 2023; Torres et al., 2020).

Since its conceptualization by Turing and McCarthy in the 1950s, AI has continuously evolved, significantly impacting various aspects of human life and driving technological advancements (Ramoni et al., 2024). In medicine, AI-based technologies, including ML and DL, have revolutionized surgical procedures, improved disease diagnostics, and promoted the development of personalized and precision medicine. In dentistry, AI is an emerging field with applications ranging from administrative tasks, such as scheduling and coordinating appointments, to more complex functions, such as assisting in clinical diagnosis and treatment planning (Stanley, 2023; Nguyen et al., 2021). Currently, AI is widely employed in areas such as Radiology, Implantology, Restorative Dentistry, Orthodontics and Endodontics (Putra et al., 2022; Mangano et al., 2023; Revilla-León et al., 2022; Nordblom et al., 2024; Aminoshariae et al., 2021).

In Endodontics, one of the major challenges to successful treatment outcomes is achieving an accurate diagnosis of pulpal and periradicular pathologies. Limitations of conventional diagnostic methods, misinterpretation of radiographic examinations, complex root anatomy, and, most importantly, a clinician's level of experience can all compromise treatment efficacy (Karamifar et al., 2020). Given these challenges, Aminoshariae et al., 2021 and Karobari et al., 2023 have demonstrated that AI-based models, such as chatbots, can serve as viable alternatives for studying complex root canal anatomy, detecting periapical lesions and root fractures, predicting the success of retreatment procedures, facilitating access to obliterated canals, and interpreting clinical and radiographic data. These capabilities support clinical decision-making and enable more precise and individualized treatment planning (Decurcio et al., 2021; Ahmed et al., 2023; Setzer et al., 2024).

The number of studies investigating the use of AI chatbots in endodontic education and clinical practice has been steadily increasing, highlighting their numerous benefits (Aminoshariae et al., 2024). However, most available studies consist of narrative or systematic literature reviews, with a limited number of clinical studies evaluating key technical aspects such as accuracy, coherence, consistency, and the limitations of these AI-driven tools. According to Mendonça de Moura et al., 2024, significant challenges remain, particularly regarding data security, the potential misuse of sensitive patient information, risks of data breaches, privacy violations, and ethical concerns. Therefore, this study aims to evaluate the consistency and diagnostic success rates of eight AI-based chatbots responses in endodontics.

2. Methodology

This study was descriptive, observational, cross-sectional, in a qualitative and quantitative nature (Pereira et al., 2018) and with the use of simple descriptive statistics with absolute frequency values in numerical values and relative percentage

frequency (Shitsuka et al., 2018) and with statistical analysis (Vieira, 2021; Costa Neto & Bekman, 2009) and with the use of double-blind analysis conducted between February and March 2025. The study adhered to the guidelines set by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). Ethical approval was not required, as no human participants were involved; all clinical cases analyzed were entirely fictitious, created exclusively by the research team to avoid the use of real patient data.

To assess the diagnostic performance of AI chatbots, five categories of AI models were analyzed: GPT-4, GPT-3 mini, GPT-3 mini high, GPT-1 (OpenAI), Gemini (Google), Claude (Anthropic), Copilot (Microsoft), and Deepseek (Deepseek AI). The selection of models was based on three criteria: (1) architectural diversity to evaluate different AI frameworks, (2) scientific and clinical relevance with applications in health and research, and (3) developer heterogeneity to enable comparisons between companies and mitigate vendor-specific biases.

Twelve fictitious clinical cases were used, previously developed and validated in the study by Mendonça de Moura et al., 2024. These cases were constructed based on the terminology, guidelines, and diagnostic classification established by the American Association of Endodontists (AAE), with each case including a standardized set of clinical information, pulp sensitivity tests, and radiographic examinations, representing realistic and challenging diagnostic scenarios (American Association of Endodontists, 2003; Glickman, 2009; AAE Consensus Conference Recommended Diagnostic Terminology, 2009).

To ensure comparability among the AI models, a standardized prompt was used, adapted from the study by Mendonça de Moura et al., 2024. Prior to presenting the cases to the chatbots, Evaluator #1, who had been calibrated for this type of study, input the following command into the chatbots: "*As an endodontist, you must read the case report and suggest a diagnostic hypothesis and treatment. Use only the classifications provided by the American Association of Endodontics, which can be found in 'Glickman GN. American Association of Endodontists consensus conference on diagnostic terminology: background and perspectives. J Endod. 2009; 35(12):1619' and 'American Association of Endodontists Consensus Conference Recommended Diagnostic Terminology. J Endod. 2009; 35(12):1634.*"

Table 1 shows all and each case answered by the chatbots in three independent repetitions, with the browsing history and context reset between executions to prevent bias from prior learning of the AI. The prompt remained unchanged in all interactions, ensuring uniformity in the analysis of the different AI models. After receiving the chatbot responses, Evaluator #1 recorded the generated outputs in a Microsoft Excel spreadsheet without any subjective interpretation. To minimize bias, the identifiers of the AI models were removed, and the responses were anonymized (e.g., Chatbot 1, Chatbot 2, Chatbot 3). Then, Evaluator #2 and Evaluator #3, who were blinded to the chatbot identities and previously calibrated for AI studies, reviewed the responses and determined the accuracy of the diagnosis and treatment based on the established classification criteria.

The evaluation of diagnostic accuracy and treatment plans was carried out by the proportion of correct diagnoses and treatments compared to the reference classification. Response consistency was analyzed by considering the intramodel agreement rate, which represents the percentage of identical responses among the three executions of the same chatbot for a given case, along with the calculation of the variation coefficient (%) between responses. The creation and validation of these cases were carried out by four of the eight authors, all of whom had clinical and academic experience in endodontics. This ensured that the cases accurately reflected real diagnostic challenges, increasing their authenticity and applicability.

The data were statistically analyzed using the Kruskal-Wallis test, followed by Dunn's multiple comparisons adjusted using the Bonferroni method to identify statistically significant differences among the chatbots. The consistency of responses was evaluated using Fleiss' Kappa coefficient, which measures intra-rater reliability. The significance level was set at $p < 0.05$. For agreement analysis, Fleiss' Kappa for M Raters (exact value) was used.

Additionally, a stratified analysis was performed by diagnostic type and proposed treatment to investigate whether the

performance of each AI model varied according to the specific clinical condition. The chi-square test was applied to compare the frequency of correct ("Yes") and incorrect ("No") diagnoses and/or treatments independently, according to the AAE classification. This procedure allowed for the identification of potential accuracy differences in specific situations, highlighting whether a given chatbot had an easier or more difficult time with specific diagnoses or treatments, in addition to the overall result.

Table 1 - Fictional case reports presented to chatbots.

Diagnosis parameters **Pulpal diseases**	**Signs and symptoms text**	**Imaging information text**
Normal pulp	The patient referred for evaluation of tooth #46 reported no pain while chewing or drinking hot/cold beverages. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without mobility, no visible caries, and an adequate restoration on tooth #46. The patient positively responded to the pulp sensitivity test using cold spray; however, it immediately subsided when the thermal stimulus was removed. The patient reported no pain in vertical and horizontal percussion tests. Tomographic and radiographic images confirmed that the restoration had no infiltration and is quite distant from the pulp; in addition, periradicular tissues presented intact lamina dura and absence of lesions.	The patient referred for evaluation of tooth #46 reported no pain while chewing or drinking hot/cold beverages. Tomographic and radiographic images confirmed that the restoration had no infiltration and is quite distant from the pulp; in addition, periradicular tissues presented intact lamina dura and absence of lesions. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without mobility, no visible caries, and an adequate restoration on tooth #46. The patient positively responded to the pulp sensitivity test using cold spray; however, it immediately subsided when the thermal stimulus was removed. The patient reported no pain in vertical and horizontal percussion tests.
Reversible pulpitis	The patient attended the dental practice reporting a sharp, localized, and transient pain in tooth #12 only when drinking cold beverages. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and a small carious lesion on the mesial surface of tooth #12. The patient positively responded to the pulp sensitivity test using cold spray; however, it subsided 5 seconds after thermal stimulus removal. The patient reported no pain in vertical and horizontal percussion tests. Tomographic and radiographic images confirmed the carious lesion in the mesial surface of tooth #12, albeit the lesion had no direct contact with the pulp chamber; in addition, periradicular tissues presented intact lamina dura and absence of lesions.	The patient attended the dental practice reporting a sharp, localized, and transient pain in tooth #12 only when drinking cold beverages. Tomographic and radiographic images confirmed the carious lesion in the mesial surface of tooth #12, albeit the lesion had no direct contact with the pulp chamber; in addition, periradicular tissues presented intact lamina dura and absence of lesions. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and a small carious lesion on the mesial surface of tooth #12. The patient positively responded to the pulp sensitivity test using cold spray; however, it subsided 5 seconds after thermal stimulus removal. The patient reported no pain in vertical and horizontal percussion tests.
Symptomatic irreversible pulpitis	The patient reported a throbbing, excruciating, stabbing, continuous, and spontaneous pain in tooth #13, which persisted even after taking analgesics. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and an extensive carious lesion on the mesial surface of tooth #13. The patient positively responded to the pulp sensitivity test using cold spray and the intense pain lingered for approximately 50 seconds after thermal stimulus removal. The pain was exacerbated during vertical and horizontal percussion. Tomographic and radiographic images revealed that the extensive carious lesion directly contacted the pulp chamber. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions.	The patient reported a throbbing, excruciating, stabbing, continuous, and spontaneous pain in tooth #13, which persisted even after taking analgesics. Tomographic and radiographic images revealed that the extensive carious lesion directly contacted the pulp chamber. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and an extensive carious lesion on the mesial surface of tooth #13. The patient positively responded to the pulp sensitivity test using cold spray and the intense pain lingered for approximately 50 seconds after thermal stimulus removal. The pain was exacerbated during vertical and horizontal percussion.
Pulp necrosis	The patient was referred by another dentist reported that tooth #35 had been painful for some time but it spontaneously resolved. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and an extensive carious lesion on the occlusal surface of tooth #35. The patient reported almost no pain during the pulp sensitivity test using cold spray, which also did not get worse after vertical or horizontal percussion tests. Tomographic and radiographic images revealed that the extensive occlusal carious lesion directly contacted the pulp chamber. Moreover, periradicular tissues presented intact lamina dura and the absence	The patient was referred by another dentist reported that tooth #35 had been painful for some time but it spontaneously resolved. Tomographic and radiographic images revealed that the extensive occlusal carious lesion directly contacted the pulp chamber. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and an extensive carious lesion on the occlusal surface of tooth #35. The patient reported almost no pain during the pulp sensitivity test using cold spray, which also did

	of lesions.	not get worse after vertical or horizontal percussion tests.
Previously treated	The patient reported that tooth #24 was treated by another dentist, but did not recall the diagnosis. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, and normal probing depth without tooth mobility. The tooth #24 was restored with a ceramic crown and the patient negatively responded to the pulp sensitivity test using cold spray as well as reported no pain in vertical and horizontal percussion tests. Tomographic and radiographic images revealed an adequate filling of both buccal and palatal root canals with gutta-percha. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions.	The patient reported that tooth #24 was treated by another dentist, but did not recall the diagnosis. Tomographic and radiographic images revealed an adequate filling of both buccal and palatal root canals with gutta-percha. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, and normal probing depth without tooth mobility. The tooth #24 was restored with a ceramic crown and the patient negatively responded to the pulp sensitivity test using cold spray as well as reported no pain in vertical and horizontal percussion tests.
Previously initiated therapy	The patient reported previous pain in tooth #11, which was resolved by another dentist who accessed the root canal and placed a medication. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, and normal probing depth without tooth mobility. There was a provisional restoration on the occlusal surface of tooth #11 without a visible carious lesion. The patient negatively responded to the pulp sensitivity test using cold spray and reported no pain in vertical and horizontal percussion tests. Tomographic and radiographic images confirmed the access to the pulp chamber, which directly contacted the provisional restoration. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions.	The patient reported previous pain in tooth #11, which was resolved by another dentist who accessed the root canal and placed a medication. Tomographic and radiographic images confirmed the access to the pulp chamber, which directly contacted the provisional restoration. Moreover, periradicular tissues presented intact lamina dura and the absence of lesions. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, and normal probing depth without tooth mobility. There was a provisional restoration on the occlusal surface of tooth #11 without a visible carious lesion. The patient negatively responded to the pulp sensitivity test using cold spray and reported no pain in vertical and horizontal percussion tests.
Diagnosis parameters	**Signs and symptoms text**	**Imaging information text**
Periapical diseases		
Symptomatic apical periodontitis	The patient attended the dental practice reporting intense, spontaneous, and localized pain in tooth #36. Moreover, the tooth felt raised and painful upon chewing. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and an extensive carious lesion throughout the entire tooth #36. The patient negatively responded to the pulp sensitivity test using cold spray. The patient reported pain during the vertical percussion test but no symptoms after the horizontal percussion test. Tomographic and radiographic images revealed a thickened periodontal ligament space without visible bone resorption and periradicular lesion, as well as direct contact between the extensive carious lesion and the pulp chamber.	The patient attended the dental practice reporting intense, spontaneous, and localized pain in tooth #36. Moreover, the tooth felt raised and painful upon chewing. Tomographic and radiographic images revealed a thickened periodontal ligament space without visible bone resorption and periradicular lesion, as well as direct contact between the extensive carious lesion and the pulp chamber. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, normal probing depth without tooth mobility, and an extensive carious lesion throughout the entire tooth #36. The patient negatively responded to the pulp sensitivity test using cold spray. The patient reported pain during the vertical percussion test but no symptoms after the horizontal percussion test.
Asymptomatic apical periodontitis	The patient attended the dental practice for evaluation of the extensively restored tooth #41 and reported no pain. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, and normal probing depth without tooth mobility. The patient negatively responded to the pulp sensitivity test using cold spray as well as reported no pain in vertical and horizontal percussion tests. Tomographic and radiographic images revealed direct contact between the extensive restoration and the pulp chamber, as well as bone resorption/periradicular lesion in the periapex of tooth #41.	The patient attended the dental practice for evaluation of the extensively restored tooth #41 and reported no pain. Tomographic and radiographic images revealed direct contact between the extensive restoration and the pulp chamber, as well as bone resorption/periradicular lesion in the periapex of tooth #41. The patient exhibited facial symmetry, absence of fistula, healthy mucosa and periodontal tissues, and normal probing depth without tooth mobility. The patient negatively responded to the pulp sensitivity test using cold spray as well as reported no pain in vertical and horizontal percussion tests.
Chronic apical abscess	The patient attended the dental practice for evaluation of tooth #46, which was not painful during chewing or drinking hot/cold beverages. The patient exhibited facial symmetry, albeit a fistula and an extensive carious lesion were observed. The patient negatively responded to the pulp sensitivity test using cold spray as well as reported no pain in vertical and horizontal percussion tests. Tomographic and radiographic images revealed direct contact between the carious lesion and the pulp chamber as well as a periapical lesion and bone resorption.	The patient attended the dental practice for evaluation of tooth #46, which was not painful during chewing or drinking hot/cold beverages. Tomographic and radiographic images revealed direct contact between the carious lesion and the pulp chamber as well as a periapical lesion and bone resorption. The patient exhibited facial symmetry, albeit a fistula and an extensive carious lesion were observed. The patient negatively responded to the pulp sensitivity test using cold spray as well as reported no pain in vertical and horizontal percussion tests.
Acute apical abscess	The patient attended the dental practice and reported spontaneous, throbbing, stabbing, and localized pain in the extensively restored tooth #45, as well as intraoral swelling in the apical region. The	The patient attended the dental practice and reported spontaneous, throbbing, stabbing, and localized pain in the extensively restored tooth #45, as well as intraoral swelling in

	patient negatively responded to the pulp sensitivity test using cold spray but reported painful symptoms during vertical and horizontal percussion tests. Tomographic and radiographic images revealed a thickened periodontal ligament space and direct contact between the extensive restoration and the pulp chamber.	the apical region. Tomographic and radiographic images revealed a thickened periodontal ligament space and direct contact between the extensive restoration and the pulp chamber. The patient negatively responded to the pulp sensitivity test using cold spray but reported painful symptoms during vertical and horizontal percussion tests.

Source: Authors' archive (2025).

3. Results

Table 2 compares the accuracy rates of diagnosis and treatment recommendations among individual chatbots, presenting both absolute and relative frequencies of correct responses. Different letters indicate statistically significant differences ($p < 0.001$). Statistical analysis was performed using the Kruskal-Wallis test, followed by Dwass-Steel-Critchlow-Fligner pairwise comparisons.

The data presented highlight a statistically significant difference in the ability to provide correct diagnoses ($p < 0.001$) among the evaluated AI systems. ChatGPT o1 (97%), Claude (97%), and DeepSeek (90.9%) demonstrated the highest accuracy rates. In contrast, the Gemini system exhibited the lowest accuracy (54.5%). The remaining systems (ChatGPT o3 mini, ChatGPT o3 mini high, ChatGPT 4o, and Copilot) showed intermediate accuracy rates, ranging from 81.8% to 87.9%. According to the multiple comparison analysis, their performance was statistically comparable to both the highest and lowest accuracy groups.

Regarding treatment recommendations, nearly all systems achieved very high accuracy rates (97% or 100%), with no statistically significant differences among them ($p = 0.537$). Thus, the primary distinction lies in the ability to formulate correct diagnoses, as the performance in treatment recommendations was consistent across the chatbots analyzed.

Table 2 - Absolute and relative frequencies of correct responses regarding diagnostic accuracy and treatment recommendations.

Chatbots	**Correct Diagnoses**		**Correct Treatment Recommendation**	
	Absolute Frequency	Relative Frequency	Absolute Frequency	Relative Frequency
ChatGPT o1	32A	97.0%	33A	100%
ChatGPT o3 mini	29AB	87.9%	33A	100%
ChatGPT o3mini high	29AB	87.9%	33A	100%
ChatGPT 4o	27AB	81.8%	33A	100%
Gemini	18B	54.5%	32A	97%
Copilot	28AB	84.8%	32A	97%
DeepSeek	30A	90.9%	33A	100%
Claude	32A	97.0%	33A	100%
P-value	<0.001		P=0.537	

Source: Authors' archive (2025).

Multivariate logistic regression, adjusted for diagnosis type and AI model, revealed that diagnostic accuracy varied significantly among chatbots, even after controlling for clinical complexity. The ChatGPT o1 model (OR = 32.7; 95% CI: 8.1–132.1) and Claude (OR = 30.5; 95% CI: 7.5–124.3) had approximately 30 times higher odds of correct diagnosis compared to Gemini (reference), confirming their statistical superiority ($p < 0.001$). Notably, complex diagnoses, such as acute apical abscess (OR = 0.1; $p = 0.006$) and asymptomatic irreversible pulpitis (OR = 0.3; $p = 0.037$), significantly reduced diagnostic

accuracy, highlighting specific challenges even for high-performance models.

A critical interaction was identified: although ChatGPT o1 demonstrated overall exceptional performance, it exhibited pronounced vulnerability in diagnosing acute apical abscess (OR = 0.01; $p < 0.001$), failing in all attempts for this condition. These findings emphasize that accuracy is not uniform—beyond model quality, factors such as clinical ambiguity and the nature of the pathology significantly influence performance. These nuances reinforce the importance of stratified analyses to identify contextual limitations, aligning with the goal of assessing diagnostic robustness in heterogeneous endodontic scenarios (Table 3).

Table 3 - Multivariate analysis of chatbots and diagnosis.

Parameters **Chatbots comparative with Gemini AI**	Odds Ratio (OR)	CI 95%	P-value
ChatGPT o1	32.7	8.1 - 132.1	<0.001*
Claude	30.5	7.5 - 124.3	<0.001*
DeepSeek	12.4	3.1 - 50.1	0.002*
ChatGPT o3 mini	4.8	1.2 - 19.3	0.024*
Copilot	4.2	1.0 - 17.2	0.043*
Diagnosis comparative with normal pulp			
Asymptomatic Irreversible Pulpitis	0.3	0.1 - 0.9	0.037*
Acute Apical Abscess	0.1	0.02 - 0.5	0.006*

Source: Authors' archive (2025).

This analysis revealed heterogeneous results in some specific conditions, although most tests did not indicate statistically significant differences. For the diagnosis of normal pulp, for instance, nearly all AI models correctly identified all cases (3 correct/0 incorrect), except for Gemini, which had one error (2/1), without statistical relevance ($p = 0.39$). Similarly, all AIs correctly recommended the treatment for normal pulp (3/0), with no significant difference ($p = 1$).

For the diagnosis of reversible pulpitis, once again, all chatbots achieved 100% accuracy, except for Gemini, which failed in all cases (0/3), resulting in $p = 0.001$, indicating a statistically significant difference. However, regarding treatment recommendations for this condition, all AIs performed flawlessly (3/0), with no statistical distinction ($p = 1$). In symptomatic irreversible pulpitis, all models achieved perfect accuracy (3/0; $p = 1$). However, for the treatment of this condition, Gemini made one error (2/1), though statistical significance was not detected ($p = 0.39$).

For asymptomatic irreversible pulpitis, greater variability was observed among systems: some, such as Copilot, achieved perfect diagnostic accuracy (3/0), while others performed worse (e.g., ChatGPT o3 mini and ChatGPT o3 mini high, both scoring 0/3). Despite this, no statistically significant difference was found ($p = 0.08$). Regarding treatment, nearly all models maintained 100% accuracy (3/0), except for Copilot (2/1; $p = 0.39$). For necrotic pulp diagnosis, most AIs performed well (3/0), except for Copilot (1/2) and Claude (2/1), though without reaching statistical significance ($p = 0.10$). The treatment for this condition was correctly indicated by all systems (3/0; $p = 1$).

For the endodontic conditions “Previously treated” and “Symptomatic apical periodontitis,” unanimous accuracy was observed for both diagnosis and treatment (3/0; $p = 1$). In “Previously initiated therapy,” most models achieved 100% accuracy, although Copilot (1/2), Claude (2/1), and ChatGPT 4o (2/1) made errors, without statistical relevance ($p = 0.21$). Regarding “Asymptomatic apical periodontitis,” overall performance was also positive (3/0), except for Gemini (1/2), ChatGPT o3 mini (2/1), and ChatGPT o3 mini high (2/1), again without statistical significance ($p = 0.21$). For “Chronic apical abscess,” the only discrepancy was Gemini, which failed in all responses (0/3), yielding a statistically significant difference ($p = 0.001$). However, all AIs correctly recommended the treatment (3/0; $p = 1$).

Finally, in "Acute apical abscess," most models succeeded in diagnosis (3/0), particularly ChatGPT o1, ChatGPT o3 mini, ChatGPT o3 mini high, and Claude. However, ChatGPT o4, Gemini, and Copilot failed in all attempts (0/3), while DeepSeek had a partial success rate (2/1), producing statistical significance (p = 0.003). In contrast, all AIs demonstrated excellent performance in treatment recommendations for this condition (3/0; p = 1). In summary, these findings indicate that while the overall performance of AIs was largely satisfactory, specific vulnerabilities exist in certain diagnostic categories, reflecting varying degrees of robustness and consistency among the analyzed models.

A total of 176 responses were analyzed, with an overall agreement rate of 93%. When evaluating the different AI models individually, it was observed that ChatGPT o1, ChatGPT o3, ChatGPT o3 mini high, and Claude had the highest agreement rates, reaching 95%. Following them were ChatGPT 4.0, Copilot, and DeepSeek, all with 91%. In contrast, Gemini had the lowest agreement rate (86%).

The evaluation of response consistency revealed an overall agreement rate of 93% among the chatbots, indicating high consensus in the general responses. However, when stratified by AI model, notable disparities were observed: ChatGPT o1, Claude, and ChatGPT o3 mini high stood out with 95% agreement, while Gemini had the lowest rate (86%), suggesting intrinsic inconsistency in its responses. Fleiss' Kappa (κ) analysis corroborated these findings, showing almost perfect agreement (κ = 0.95) for low-complexity diagnoses, such as normal pulp, but moderate to weak values (κ = 0.28–0.45) for ambiguous conditions, such as asymptomatic irreversible pulpitis and acute apical abscess, where discordance between the models was pronounced (Table 4).

Table 4 - Quantitative assessment of inter-chatbot diagnostic agreement by Fleiss' Kappa.

Parameters Diagnosis	Fleiss' Kappa (κ)	Interpretation
Normal Pulp	0.95	Almost perfect agreement
Reversible Pulpitis	0.82	Substantial agreement
Acute Apical Abscess	0.45	Moderate agreement
Asymptomatic Irreversible Pulpitis	0.28	Weak agreement

Source: Authors' archive (2025).

Additionally, the Intraclass Correlation Coefficient (ICC) confirmed high intra-chatbot reliability for the higher-performing systems (ICC > 0.85), contrasting with Gemini (ICC = 0.65), which exhibited significant variability in repeated similar cases. Bland-Altman analysis between pairs of chatbots revealed systematic bias in models like Gemini, which underestimated correct responses by up to 43% compared to ChatGPT o1. These results emphasize that, while most chatbots exhibit high overall accuracy, reliability varies depending on clinical complexity and the model analyzed, with less robust systems displaying critical failures in challenging scenarios. This heterogeneity highlights the importance of validating not only accuracy but also the contextual consistency of these tools, aligning with the goal of identifying practical limitations for their safe use in Endodontics (Table 5).

Table 5 - ICC assessment of intra-chatbot consistency in responses across different cases.

Parameters Chatbots	ICC	CI 95%	Interpretation
ChatGPT o1	0.92	0.85 - 0.96	Almost perfect agreement
Claude	0.89	0.80 - 0.94	High agreement
Gemini	0.65	0.50 - 0.77	Moderate agreement
DeepSeek	0.85	0.75 - 0.91	High agreement
Copilot	0.78	0.65 – 0.87	High agreement

Source: Authors' archive (2025).

4. Discussion

To the best of the authors' knowledge, this is the first study in the literature to evaluate the performance variability of eight AI models in formulating endodontic diagnoses, with independent assessments conducted by two experienced endodontists regarding chatbot-generated responses. Based on the results of this study, we recognize that AI can serve as a valuable diagnostic aid for both endodontists and general dentists. However, it is crucial to emphasize that AI should neither replace the dentist nor undermine professional expertise.

Another key consideration is that the effective use of AI requires the continuous inclusion of data to refine machine learning (ML) and deep learning (DL) algorithms. According to Soori et al., 2023, the constant input of data into chatbots enhances their learning process and optimizes task performance. Sohrabniya et al., 2025 highlight the importance of promoting AI adoption in dentistry, noting that studies on deep learning in the field have increased significantly since 2020. Similar to this study, their review found that clinical diagnosis was the primary focus of 63.5% of studies, with the highest concentration in stomatology (21.5%), radiology (17.5%), and orthodontics (10.2%). However, they observed that 84.4% of studies utilized imaging data for clinical diagnoses. In contrast, this study employed highly detailed fictitious clinical cases, incorporating both clinical and radiographic information. This methodological distinction may explain the near-perfect and high agreement rates observed with ChatGPT, Claude, DeepSeek, and Copilot during the concordance and reliability analyses.

Setzer & Kratchman, 2022 emphasize the importance of integrating AI collaboratively, rather than as a replacement, for endodontic diagnosis and treatment planning. They argue that AI can enhance precision and efficiency. AI's objectivity in diagnosing endodontic pathologies is particularly relevant given the nuanced clinical, radiographic, and histological characteristics of periapical and pulpal lesions, which can introduce biases among less experienced endodontists or students. Such biases may lead to subjective diagnoses, erroneous prognoses, and increased failure rates in endodontic treatments. Setzer & Kratchman, 2022 and Schwendicke & Büttner, 2023 found that AI-assisted differential diagnosis, combined with clinical expertise, improves the ability to distinguish between apical granulomas and cysts, detect root fractures, and determine prognoses for retreatment cases.

Setzer et al., 2024; Aminoshariae et al., 2021; Uribe et al., 2024 advocate for incorporating chatbot technology into endodontic education and training. AI models possess continuous learning and adaptation capabilities based on new data and professional feedback. This facilitates theoretical learning enhancement and improvement in practical skills such as radiographic interpretation, differential diagnosis, treatment planning, risk-benefit assessment, and referral recommendations. Furthermore, AI can support personalized education by tracking individual student progress, ultimately improving endodontists' accuracy and efficiency.

A significant concern regarding AI in endodontics is the reliability and consistency of chatbots. Mendonça de Moura et al., 2024 found that chatbots with more sophisticated architectures and refined text-generation capabilities tend to exhibit greater consistency in repetitive tasks, such as those involving fictitious clinical cases in this study. While our findings indicate high concordance rates, subtle discrepancies in reliability and consistency among chatbots persist. The results generated by Gemini reinforce the need for critical interpretation of AI-generated responses by endodontists, especially in clinical scenarios where minor variations can impact decision-making.

Supporting the findings of this study, Mendonça de Moura et al., 2024 assessed diagnostic accuracy and treatment recommendation performance of four chatbots (ChatGPT 3.5, ChatGPT 4.0, Google Bard, and Bing). Their results showed that Bing and ChatGPT 4.0 had the highest diagnostic accuracy rates, at 86.4% and 85.3%, respectively. Notably, at the time of their study, Gemini was still referred to as Google Bard and demonstrated a low accuracy rate of only 28.6%. However, in our study, Gemini achieved a 54.5% accuracy rate, suggesting that this chatbot requires further refinement for endodontic applications. Regarding endodontic treatment recommendations, ChatGPT 4.0 (94.4%), Bing (93.2%), and ChatGPT 3.5

(86.3%) performed best, aligning with our results, while Google Bard exhibited a lower accuracy of 75%.

Contrary to the findings of this study and Mendonça de Moura et al., 2024, Mohammad-Rahimi et al., 2024 evaluated the validity and reliability of chatbot-generated responses (ChatGPT 3.5, Google Bard, and Bing) to frequently asked questions in endodontics. Their study reported a 95% accuracy rate for ChatGPT 3.5, while Google Bard performed better than Bing, achieving 85% and 75%, respectively. Künzle & Paris, 2024 conducted a similar study evaluating ChatGPT-4.0o, ChatGPT 4.0, ChatGPT 3.5, and Gemini in restorative dentistry and endodontics. They found that ChatGPT-4.0o had the highest accuracy, though its percentage was lower than reported in our study. The variability in results across the literature likely reflects differences in AI model architecture, natural language processing frameworks, and the reliance on generic datasets lacking specialized training in standardized terminology (Maltarollo et al., 2024). These factors can reduce diagnostic precision in complex cases requiring the integration of clinical and radiographic signs.

To mitigate methodological biases and ensure optimal chatbot performance, this study standardized the prompts used in AI interactions. The uniform command structure minimized bias, as variations in prompt phrasing can significantly influence AI-generated responses. Using identical standardized prompts across all models prevented discrepancies in diagnostic accuracy due to inconsistent case presentation. This approach ensured a controlled and comparable testing environment, allowing the study's findings to reflect the true diagnostic capabilities of each chatbot.

The results highlight ChatGPT 1o (97%), Claude (97%), and DeepSeek (90.9%) as the top-performing models, whereas Gemini (54.5%) exhibited intrinsic limitations, likely due to gaps in its dental training. However, stratified analysis revealed weaknesses in complex diagnoses, such as acute periapical abscesses, suggesting that less robust models struggle to integrate ambiguous clinical signs, even though treatment recommendations remained consistent. Nevertheless, this study has limitations, including the use of fictitious cases, which, while avoiding ethical biases, restrict generalizability to real clinical scenarios with inherent variability. The analysis was limited to 12 validated cases, potentially underestimating the complexity of atypical diagnoses. Standardized prompting may not fully capture dynamic interactions in practical contexts, and the short study duration (one month) does not account for continuous AI model updates. Despite the double-blind assessment, human evaluators' interpretation of chatbot responses introduces a risk of subjective bias.

5. Conclusion

Based on the results of this study, it can be concluded that, for endodontic clinical cases, among the eight chatbots used, seven demonstrated an accuracy ranging from 81.8% to 97% with high reliability, while only the Gemini chatbot showed an accuracy below 70% and weak reliability. These results suggest that the use of AI in endodontics will have a significant impact on everyday clinical practice, not as a substitute for human professionals but as a viable tool to enhance success rates in endodontic treatments. Therefore, the true challenge of integrating AI into endodontics lies not in replacing professionals with chatbots, but in the responsible, standardized, and critical integration of these technologies, while continually emphasizing the need for further research to assess their reliability, relevance, and cost.

Funding

No funding

Author’s Contributions

R.R.S.F. and J.D.M.M.: contributed to Conceptualization and Supervision. I.A.B., M.S.A.A., P.R.Z.A. and R.R.S.F.: contributed to Data Curation. I.S.E., I.A.B., M.S.A.A., P.R.Z.A. J.D.M.M.: contributed to Investigation. R.R.S.F. and

J.D.M.M.: contributed to Formal Analysis. I.S.E., I.A.B., M.S.A.A. and P.R.Z.A.: contributed to Methodology. P.A.R., S.A.F.M., I.A.B., R.R.S.F. and J.D.M.M.: contributed to Writing-Original Draft Preparation. P.A.R., S.A.F.M., R.R.S.F. and J.D.M.M.: contributed to Review and Editing of the manuscript. R.R.S.F. and J.D.M.M.: contributed to Project Administration.

Conflicts of interest

The authors declare no conflicts of interests.

References

AAE Consensus Conference Recommended Diagnostic Terminology. (2009). *Journal of Endodontics*, 35, 1634.

Ahmed, N., Abbasi, M. S., Zuberi, F., Qamar, W., Halim, M. S. B., Maqsood, A., & Alam, M. K. (2021). Artificial intelligence techniques: Analysis, application, and outcome in dentistry—A systematic review. *Biomed Research International*, 2021, 9751564.

Ahmed, Z. H., Almuharib, A. M., Abdulkarim, A. A., Alhassoon, A. H., Alanazi, A. F., Alhaqbani, M. A., Alshalawi, M. S., et al. (2023). Artificial intelligence and its application in endodontics: A review. *Journal of Contemporary Dental Practice*, 24, 912–917.

American Association of Endodontists. (2003). *Glossary of endodontic terms*. Chicago: American Association of Endodontists.

Aminoshariae, A., Kulild, J., & Nagendrababu, V. (2021). Artificial intelligence in endodontics: Current applications and future directions. *Journal of Endodontics*, 47, 1352–1357.

Aminoshariae, A., Nosrat, A., Nagendrababu, V., Dianat, O., Mohammad-Rahimi, H., O'Keefe, A. W., & Setzer, F. C. (2024). Artificial intelligence in endodontic education. *Journal of Endodontics*, 50, 562–578.

An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors (Basel)*, 23, 4178.

Bonny, T., Al Nassan, W., Obaideen, K., Al Mallahi, M. N., Mohammad, Y., & El-Damanhoury, H. M. (2023). Contemporary role and applications of artificial intelligence in dentistry. *F1000Research*, 12, 1179.

Casadei, R. (2023). Artificial collective intelligence engineering: A survey of concepts and perspectives. *Artificial Life*, 29, 433–467.

Costa Neto, P. L. O., & Bekman, O. R. (2009). *Statistical analysis of decision-making* (2nd ed.). São Paulo: Editora Blucher.

Decurcio, D. A., Bueno, M. R., Silva, J. A., Loureiro, M. A. Z., Sousa-Neto, M. D., & Estrela, C. (2021). Digital planning on guided endodontics technology. *Brazilian Dental Journal*, 32, 23–33.

de Moura, J. D. M., Fontana, C. E., da Silva Lima, V. H. R., de Souza Alves, I., de Melo Santos, P. A., & de Almeida Rodrigues, P. (2024). Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: A cross-sectional study. *Computers in Biology and Medicine*, 183, 109332.

Glickman, G. N. (2009). AAE consensus conference on diagnostic terminology: Background and perspectives. *Journal of Endodontics*, 35, 1619–1620.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, 65, 337–359.

Karobari, M. I., Adil, A. H., Basheer, S. N., Murugesan, S., Savadamoorthi, K. S., Mustafa, M., Abdulwahed, A., et al. (2023). Evaluation of the diagnostic and prognostic accuracy of artificial intelligence in endodontic dentistry: A comprehensive review of literature. *Computational and Mathematical Methods in Medicine*, 2023, 7049360.

Karamifar, K., Tondari, A., & Saghiri, M. A. (2020). Endodontic periapical lesion: An overview on the etiology, diagnosis and current treatment modalities. *European Endodontic Journal*, 5, 54–67.

Kim, D., Kim, S. H., Kim, T., Kang, B. B., Lee, M., Park, W., Ku, S., et al. (2021). Review of machine learning methods in soft robotics. *PLoS One*, 16, e0246102.

Künzle, P., & Paris, S. (2024). Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clinical Oral Investigations*, 28, 575.

Mangano, F. G., Admakin, O., Lerner, H., & Mangano, C. (2023). Artificial intelligence and augmented reality for guided implant surgery planning: A proof of concept. *Journal of Dentistry*, 133, 104485.

Maltarollo, T. F. H., Strazzi-Sahyon, H. B., Amaral, R. R., & Sivieri-Araújo, G. (2024). Is the field of endodontics prepared to utilise ChatGPT? *Australian Endodontic Journal*, 50, 176–177.

Mohammad-Rahimi, H., Ourang, S. A., Pourhoseingholi, M. A., Dianat, O., Dummer, P. M. H., & Nosrat, A. (2024). Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *International Endodontic Journal*, 57, 305–314.

Mukhamediev, R. I., Popova, Y., Kuchin, Y., Zaitseva, E., Kalimoldayev, A., Symagulov, A., Levashenko, V., et al. (2022). Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges. *Mathematics*, 10, 2552.

Nguyen, T. T., Larrivée, N., Lee, A., Bilaniuk, O., & Durand, R. (2021). Use of artificial intelligence in dentistry: Current clinical trends and research advances. *Journal of the Canadian Dental Association*, 87, l7.

Nordblom, N. F., Büttner, M., & Schwendicke, F. (2024). Artificial intelligence in orthodontics: Critical review. *Journal of Dental Research*, 103, 577–584.

Ossowska, A., Kusiak, A., & Świetlik, D. (2022). Artificial intelligence in dentistry—Narrative review. *International Journal of Environmental Research and Public Health*, 19, 3449.

Pereira, A. S., et al. (2018). *Scientific research methodology* [free e-book]. Santa Maria: Editora da UFSM.

Putra, R. H., Doi, C., Yoda, N., Astuti, E. R., & Sasaki, K. (2022). Current applications and development of artificial intelligence for digital dental radiography. *Dentomaxillofacial Radiology*, 51, 20210197.

Ramoni, D., Sgura, C., Liberale, L., Montecucco, F., Ioannidis, J. P. A., & Carbone, F. (2024). Artificial intelligence in scientific medical writing: Legitimate and deceptive uses and ethical concerns. *European Journal of Internal Medicine*, 127, 31–35.

Revilla-León, M., Gómez-Polo, M., Vyas, S., Barmak, A. B., Özcan, M., Att, W., & Krishnamurthy, V. R. (2022). Artificial intelligence applications in restorative dentistry: A systematic review. *Journal of Prosthetic Dentistry*, 128, 867–875.

Schwendicke, F., & Büttner, M. (2023). Artificial intelligence: Advances and pitfalls. *British Dental Journal*, 234, 749–750.

Setzer, F. C., & Kratchman, S. I. (2022). Present status and future directions: Surgical endodontics. *International Endodontic Journal*, 55, 1020–1058.

Setzer, F. C., Li, J., & Khan, A. A. (2024). The use of artificial intelligence in endodontics. *Journal of Dental Research*, 103, 853–862.

Shiammala, P. N., Duraimutharasan, N. K. B., Vaseeharan, B., Alothaim, A. S., Al-Malki, E. S., Snekaa, B., Safi, S. Z., et al. (2023). Exploring the artificial intelligence and machine learning models in the context of drug design difficulties and future potential for the pharmaceutical sectors. *Methods*, 219, 82–94.

Shitsuka, R., et al. (2014). *Fundamental mathematics for technology* (2nd ed.). São Paulo: Editora Érica

Sohrabniya, F., Hassanzadeh-Samani, S., Ourang, S. A., Jafari, B., Farzinnia, G., Gorjinejad, F., Ghalyanchi-Langeroudi, A., et al. (2025). Exploring a decade of deep learning in dentistry: A comprehensive mapping review. *Clinical Oral Investigations*, 29, 143.

Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics: A review. *Cognitive Robotics*, 3, 54–70.

Stanley, K. (2023). Artificial intelligence and the future of dentistry. *Compendium of Continuing Education in Dentistry*, 44, 250–253.

Torres, P. E. P., Torres, E. A., Hernández-Álvarez, M., & Yoo, S. G. (2020). EEG-based BCI emotion recognition: A survey. *Sensors (Basel)*, 20, 5083.

Uribe, S. E., Maldupa, I., Kavadella, A., El Tantawi, M., Chaurasia, A., Fontana, M., Marino, R., et al. (2024). Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. *European Journal of Dental Education*, 28, 865–876.

Vieira, S. (2021). *Introduction to biostatistics*. Rio de Janeiro: Editora GEN/Guanabara Koogan.