Journal of Dental Sciences and Education

Original Article

DOI: 10.51271/JDSE-0081

A multidisciplinary evaluation of AI-powered chatbots on apical root-end resection-assessing alignment with international endodontic guidelines: a comparative methodological study

Tuğçe Paksoy*[1], Selin Gaş[2], Seval Ceylan Şen[3], Özlem Saraç Atagün[3], Gülbahar Ustaoğlu[3], Şeyma Çardakçı Bahar[3], Savaş Özarslantürk[4]

[1]Department of Periodontology, Hamidiye Faculty of Dentistry, University of Health Sciences, İstanbul, Turkiye
[2]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, İstanbul Gelişim University, İstanbul, Turkiye
[3]Department of Periodontology, Gülhane Faculty of Dentistry, University of Health Sciences, Ankara, Turkiye
[4]Department of Oral and Maxillofacial Radiology, Gülhane Faculty of Dentistry, University of Health Sciences, Ankara, Turkiye

Cite this article as: Paksoy T, Gaş S, Ceylan Şen S, et al. A multidisciplinary evaluation of AI-powered chatbots on apical root-end resection-assessing alignment with international endodontic guidelines: a comparative methodological study. *J Dent Sci Educ.* 2026;4(3):76-83. doi:10.51271/JDSE-0081

Received: 22.05.2026 ◆ **Accepted**: 30.07.2026 ◆ **Published**: 20.08.2026

ABSTRACT

Aims: This study aimed to identify the clinically relevant patient questions about apical root-end resection based on international consensus guidelines, and to systematically evaluate the accuracy, quality, usefulness, and readability of generated responses by four Artificial Intelligence (AI)-powered conversational agents, ChatGPT-4, DeepSeek V3.1, Gemini 3, and Copilot.

Methods: A standardized set of 20 clinical questions was developed from the American Association of Endodontists guidelines and international consensus reports. A priori, an evidence-based answer key was established as the reference standard. Queries were submitted to ChatGPT-4, DeepSeek V3.1, Gemini 3, and Copilot, following a standardized interaction protocol designed to minimize personalization and carryover effects, such as incognito mode, newly created accounts, and no response regeneration. The outputs were independently assessed by three experts in periodontics and oral radiology, using validated tools, such as CLEAR criteria, modified Global Quality Score (mGQS), accuracy ratings, DISCERN, readability indices (Flesch reading ease, FRE and Flesch-Kincaid grade level, FKGL) and the patient education materials assessment tool (PEMAT).

Results: ChatGPT-4 and Copilot achieved significantly higher CLEAR scores than Gemini 3. Copilot achieved significantly higher mGQS and accuracy than DeepSeek V3.1 and Gemini 3 and received significantly lower, and therefore more favorable, usefulness scores. Gemini 3 scored higher on the FKGL test and lower on the understandability test compared to DeepSeek V3.1. Relationships between evaluation metrics were different between models.

Conclusion: The AI models evaluated demonstrated significant variability in clinical accuracy, quality, readability, and patient-centered usability. Copilot had the best overall mGQS and Accuracy scores and the best Usefulness scores, while the models varied in their strengths in readability, understandability, and actionability.

Keywords: Artificial Intelligence, chatbots, ChatGPT, patient education, readability

INTRODUCTION

Microbial infection is a key factor in the development of pulpal and periapical diseases.[1] The main goals of endodontic therapy are to eliminate or greatly reduce the number of microorganisms within the root canal system and to place a proper seal to prevent re-infection of the periapical tissues.[2] Conventional root canal therapy has an approximate 90% success rate, but failure may require nonsurgical retreatment. Where this is not possible, surgical endodontic therapy is indicated.[3] Apical root-end resection is mainly indicated in the surgical treatment of teeth with advanced endodontic-periodontal involvement among surgical approaches, but it is also indicated in some special cases of surgical endodontic management as described by the American Association of Endodontists (AAE) and Cohen's Pathways of the Pulp.[4,5]

Patients are increasingly supplementing the professional advice they receive from dentists with digital sources. Patients want to know the risks of procedures, how long recovery will take, and what the expected results are. The transition to Artificial Intelligence (AI)-powered platforms emphasizes the importance of delivering not just accurate but also patient-centered and accessible information.[6,7] The most complex part of the internal root canal anatomy is the apical root canal anatomy, in which irregularities such as isthmuses and accessory canals are often present.[8,9] These anatomical variations of the apical third may pose a challenge

***Corresponding Author:** Tuğçe Paksoy, tugce.paksoy@sbu.edu.tr

for endodontic procedures and are considered important risk factors for treatment failure.[10,11] When orthograde root canal therapy fails, apical root-end resection often is advocated to remove persistent microbial reservoirs in the complex apical anatomy. It has been reported that removal of the apical 3 mm of the root removes approximately 92% of such irregularities.[10] However, apical root-end resection may lead to thermal changes and procedural alterations in the root dentin, which may harm the periodontal ligament cells and adjacent alveolar bone.[12] Thus, careful control of temperature elevation and minimally invasive surgical techniques during apical root-end resection are essential to minimize the risk of damage to surrounding vital soft and hard tissues. Therefore, patients may seek information about indications, risks, success rates, and postoperative course in online resources such as AI chatbots.

Since its birth in 1956, AI, loosely defined as computer systems that mimic human thought processes, has become a revolutionary technology field.[13] While AI has the potential to revolutionize clinical dentistry, a major concern for patient safety is the risk of 'hallucinations', where models generate information that sounds convincing but is clinically wrong or out of date. Hence, a thorough assessment of AI-produced content is highly desirable to guarantee the correctness and understandability of information about complicated surgical interventions like apical root-end resection.[14,15] AI, which is the ability to accomplish tasks normally associated with human intelligence including learning, reasoning, problem-solving and decision-making, has made great strides in recent years due to improvements in computational power, algorithmic structures and big data analysis. These advances have paved the way for widespread use of AI in many areas, including the health sciences. The use of AI is gaining more acknowledgement in dentistry as a way to increase diagnostic accuracy, improve treatment planning, and advance education.[6,7,16]

While the performance of AI has been investigated in many dental fields including prosthetics, hypersensitivity and implantology, there are relatively fewer studies concerning the reliability of AI agents in endodontic surgery.[16-18] The present study provides a multidimensional evaluation of four prominent AI models to bridge this gap. A key component of this work is to assess the language difficulty of the AI outputs, as health information materials should be written at or below an eighth-grade reading level to optimize public understanding.[19,20] With the increasing use of AI chatbots by patients to seek oral health information, the objective of this study is to assess the clinical utility of AI chatbots as adjunctive educational tools in current endodontic practice using a comprehensive set of validated tools, namely, CLEAR, mGQS, DISCERN, and PEMAT.

METHODS

Ethics

No ethics committee approval was necessary, as the study did not involve human participants, identifiable patient data, or biological materials. All procedures were carried out in accordance with the ethical rules and the principles.

Study Design and Question Development

The standard question set was developed using authoritative clinical and anatomical references, including international guidelines, consensus-based references and relevant peer-reviewed studies on surgical endodontics. The core clinical themes and patient concerns were based on the AAE Glossary and Clinical Guidelines and Cohen's Pathways of the Pulp[4,5,10] and well-established anatomic success criteria, which demonstrate that a 3 mm apical root-end resection removes around 92% of the apical ramifications. Finally, a set of 20 questions was developed to cover clinical indications, success rates, procedural risks and postoperative protocols. Two authors (TP and SCŞ, periodontology specialists) reviewed the items for linguistic and clinical accuracy and they eliminated duplicates and semantically equivalent questions and refined the items to ensure grammatical clarity.

AI Interaction and Data Collection

A primary account was created for each of the four platforms (ChatGPT-4, Gemini 3, DeepSeek V3.1 and Copilot) with unique credentials and accessed in incognito mode to remove any interference from previous user data or search history. Each query was posed as a 'new chat' by a member of the research team (SG, specialist in oral and maxillofacial surgery) to ensure independence between responses. Only the first output was collected and the regeneration feature was avoided strictly for the sake of consistency of the data. All interactions occurred on 20th November, 2025 between 09:00 and 16:00.

Reference Standard and Outcome Assessment

To facilitate an objective evaluation of the AI-generated responses, a standardized "answer key" (reference standard) was created prior to the evaluation phase grounded on clinical consensus statements.[4,5,10] The answers were evaluated by three independent specialists, who were unaware of the question formulation team: OSA and GUS (periodontology specialists) and SÖ (specialist in oral and maxillofacial radiology). The multidisciplinary panel graded the AI agents using the evidence-based reference key and the CLEAR criteria, modified Global Quality Score (mGQS), accuracy ratings, the DISCERN instrument, standard readability indices (Flesch Reading Ease, FRE and Flesch-Kincaid Grade Level, FKGL), and the patient education materials assessment tool (PEMAT). To reduce subjective inter-observer variability and guarantee a thorough comparison with existing clinical evidence, the validation and data curation process was overseen by another periodontology expert (ŞÇB).

To examine whether the repeated values of the same variables are similar, intra-class correlation (two-way mixed and consistency approach) was used. The analyses showed that the inter-rater reliability level was above the minimum value of 0.700 and the inter-rater reliabilities were statistically significant and highly consistent ($p<0.05$).

The qualitative integrity of the AI outputs was assessed with the CLEAR tool, which divides reliability into five core domains: content completeness, no misinformation, evidence-based grounding, appropriate formatting, and topical relevance. The evaluators scored each category (1-5) and a total quality index was calculated; higher scores indicated better reliability of the information.[21]

The mGQS was used to assess overall contextual depth and accuracy of the answers. This 5-point scale ranges from 'strongly disagree' (indicating entirely flawed content) to 'strongly agree' (signifying a comprehensive and error-free response).[22]

Accuracy was measured using a five-point Likert scale. The scale is as follows; Score 1: responses are completely wrong; Score 2: responses contain more wrong than right information; Score 3: responses contain an equal amount of wrong and right information; Score 4: responses contain more right than wrong information; and Score 5: responses are completely right.[23]

The usefulness of the AI-generated responses was assessed using a structured 4-point rating scale. This scale characterizes the responses as follows: (score 1) very useful: the responses are complete and correct, like what a domain expert would provide. (Score 2) useful; the responses are mostly correct but may miss some important details. (Score 3) Partially useful: the responses contain both correct and incorrect/misleading information. (Score 4) Not useful: the responses contain only incorrect/misleading information, with no reliable content.[24] In this study, usefulness was defined as the extent to which a response provided practical information that was needed by the patient, and provided relevant, understandable and clinically applicable guidance. Lower scores indicated more usefulness. A negative correlation with positively oriented measures like CLEAR, mGQS and Accuracy was thus due to an opposite direction of scoring scales, and not to an opposite conceptual relationship.

The linguistic accessibility was quantified using the FRE and FKGL metrics. The FRE is a numerical representation of the difficulty of a text (100 being the most easily understood) and the FKGL transforms this into the American educational grade level required for complete comprehension.

To systematically evaluate these two key aspects of patient education materials, a standardized and validated tool called the PEMAT was employed to assess the comprehensibility and actionability of the responses generated by AI. The generated advice was tested for understandability by lay people and actionability in a clinical setting using PEMAT.[19,20]

Individuals with different cultural and educational backgrounds and varying levels of health literacy must be able to understand and correctly interpret the main messages of instructional materials for them to be considered intelligible. Understandability refers to the extent to which individuals with varying levels of health literacy can comprehend the principal messages of educational material. Actionability refers to the degree to which users are able to identify specific actions that can be taken based on the information. The understandability domain of the PEMAT contains 17 items that are rated 1 (agree), 0 (disagree), or NA (not applicable). The Actionability domain has 7 elements and they are all scored the same way. Total scores for both areas range from 0% to 100%, with higher percentages indicating information that is more understandable and useful for guiding patient behavior.[25]

Statistical Analysis

Descriptive statistics were presented as means, standard deviations, medians and interquartile ranges. Data distribution was evaluated using the Shapiro-Wilk test. Overall differences among AI models were assessed using the Kruskal-Wallis H test. When significant, pairwise post hoc comparisons were performed with Bonferroni correction because the assumptions of normality were not met. The Spearman's rank correlation coefficient was used to assess the associations between variables. Statistical analyses were performed using IBM SPSS Statistics version 27.

RESULTS

Kruskal-Wallis tests were used to compare differences in response characteristics by type of AI. There were significant differences between AI types for CLEAR, mGQS, accuracy, usefulness, FKGL and understandability% scores ($p<0.05$). Bonferroni post hoc tests showed that both Copilot and ChatGPT scored significantly higher on CLEAR than Gemini 3 ($p=0.001$ and $p<0.001$). For mGQS and Accuracy, Copilot was much higher than DeepSeek V3.1 and Gemini 3 ($p<0.001$). Copilot scored significantly lower and therefore more favorable Usefulness scores than DeepSeek V3.1 and Gemini 3 ($p<0.001$). Gemini 3 had a significantly higher FKGL score than DeepSeek V3.1 ($p=0.027$) and DeepSeek V3.1 was significantly better than Gemini 3 for understandability% ($p=0.040$) (**Table 1**).

Then, Spearman correlation analyses were performed separately for each AI model. In the case of ChatGPT-4, the CLEAR score was highly and positively correlated with mGQS and Accuracy ($r=0.856$, $p<0.05$) but highly and negatively correlated with Usefulness ($r=-0.856$, $p<0.05$). mGQS was found to be perfectly and positively correlated with accuracy

Table 1. Distribution and comparison of features of responses according to AI types

	ChatGPT-4		Copilot		DeepSeek V3.1		Gemini 3			
	Mean±SD	Median (25%-75% q)	Mean±SD	Median (25%-75% q)	Mean±SD	Median (25%-75% q)	Mean±SD	Median (25%-75% q)	Test statistics	p
CLEAR	24±0.86	24 (24-24.5)	24.2±0.89	24 (24-25)	23.25±1.77	24 (24-24)	23.05±0.51	23 (23-23)	24.299	<0.001*
mGQS	4.25±0.44	4 (4-4.5)	4.5±0.51	4.5 (4-5)	4±0	4 (4-4)	4±0	4 (4-4)	22.282	<0.001*
Accuracy	4.25±0.44	4 (4-4.5)	4.5±0.51	4.5 (4-5)	4±0	4 (4-4)	4±0	4 (4-4)	22.282	<0.001*
Usefulness	1.75±0.44	2 (1.5-2)	1.5±0.51	1.5 (1-2)	2±0	2 (2-2)	2±0	2 (2-2)	22.282	<0.001*
FRE	53.95±17.68	57.4 (44.55-66.4)	47.47±12.46	46.7 (37.75-59.1)	56.11±20.89	55.3 (39.75-62.85)	46.48±13.54	47.8 (37.25-56.35)	4.311	0.230
FKGL	8.48±2.8	8.35 (7.25-9.5)	9.81±2.24	10.25 (8.05-11.25)	7.32±3.43	7.4 (6.6-9.35)	10.82±2.23	10.7 (9.15-12.1)	16.868	<0.001*
DISCERN	27.7±0.98	28 (28-28)	28±2.15	28 (27.5-28)	27.5±1.54	28 (27.5-28)	27.35±1.87	28 (28-28)	0.404	0.939
U%	75.05±18.76	83 (67-87)	72.15±7.21	75 (67-79)	78.75±10.19	82.5 (76-85.5)	70.35±9.94	70 (66-77)	11.322	0.010*
A%	57±17.5	60 (40-70)	61±17.74	60 (60-70)	54.9±21.59	58.5 (40-73.5)	54±9.4	60 (40-60)	1.905	0.592

*p<0.05, AI: Artificial Intelligence, SD: Standard deviation, q: Quartile, mGQS: Modified Global Quality Score, FRE: Flesch reading ease, FKGL: Flesch-Kincaid grade level, U: Understandability, A: Actionability

(r=1.000, p<0.05) and both were found to be perfectly and negatively correlated with usefulness (r=-1.000, p<0.05). FRE and FKGL were very strongly correlated (r=-0.967, p<0.05). There was a moderate negative correlation between FKGL and DISCERN (r=-0.456, p<0.05). Correlation between understandability% and actionability% was moderate (r=0.587, p<0.05) and DISCERN with understandability% and actionability% was moderate (r=0.585 and r=0.537, p<0.05) (**Table 2**).

Regarding Copilot, CLEAR score was strongly and positively correlated with mGQS and accuracy (r=0.877, p<0.05), and strongly and negatively correlated with usefulness (r=-0.877, p<0.05). It also indicated a moderate positive correlation with understandability% (r=0.485, p<0.05). mGQS and accuracy were perfectly correlated (r=1.000, p<0.05) and both were perfectly negatively correlated with usefulness (r=-1.000, p<0.05). FRE was very strongly and negatively correlated to FKGL (r=-0.908, p<0.05). FKGL was moderately negatively correlated to DISCERN (r=-0.579, p<0.05). DISCERN had moderate positive correlations with understandability% and actionability% (r=0.600 and r=0.544, p<0.05), and understandability% and actionability% were also moderately positively correlated (r=0.455, p<0.05) (**Table 3**).

In DeepSeek V3.1, CLEAR score showed a moderate positive correlation with the actionability% (r=0.589, p<0.05). mGQS had perfect positive correlations with accuracy and usefulness (r=1.000, p<0.05). Accuracy and usefulness were perfectly correlated (r=1.000). FRE and FKGL were highly negatively correlated (r=-0.971, p<0.05). There was a moderate positive correlation between understandability% and DISCERN (r=0.499, p<0.05) and understandability% and actionability% (r=0.640, p<0.05) (**Table 4**).

For Gemini 3, mGQS was perfectly positively correlated with both accuracy and usefulness (r=1.000, p<0.05) and accuracy was also perfectly positively correlated with usefulness (r=1.000,

Table 2. Relationships between metrics for ChatGPT-4

		mGQS	Accuracy	Usefulness	FRE	FKGL	DISCERN	U%	A%
CLEAR	r	0.856	0.856	-0.856	-0.161	0.248	0.094	0.248	0.360
	p	<0.001*	<0.001*	<0.001*	0.497	0.292	0.694	0.291	0.119
mGQS	r		1.000	-1.000	-0.250	0.351	-0.040	0.293	0.159
	p		<0.001*	<0.001*	0.287	0.130	0.868	0.210	0.504
Accuracy	r			-1.000	-0.250	0.351	-0.040	0.293	0.159
	p			<0.001*	0.287	0.130	0.868	0.210	0.504
Usefulness	r				0.250	-0.351	0.040	-0.293	-0.159
	p				0.287	0.130	0.868	0.210	0.504
FRE	r				1.000	-0.967	0.420	0.225	-0.138
	p					<0.001*	0.065	0.341	0.561
FKGL	r						-0.456	-0.218	0.099
	p						0.043*	0.357	0.677
DISCERN	r							0.585	0.537
	p							0.007*	0.015*
Understandability%	r								0.587
	p								0.007*

*p<0.05 and r: Correlation coefficient, mGQS: Modified Global Quality Score, FRE: Flesch reading ease, FKGL: Flesch-Kincaid grade level, U: Understandability, A: Actionability

Table 3. Relationships between metrics for Copilot

		mGQS	Accuracy	Usefulness	FRE	FKGL	DISCERN	U%	A%
CLEAR	r	0.877	0.877	-0.877	-0.309	0.051	0.322	0.485	0.268
	p	<0.001*	<0.001*	<0.001*	0.185	0.831	0.166	0.030*	0.253
mGQS	r		1.000	-1.000	-0.434	0.200	0.059	0.315	0.096
	p		<0.001*	<0.001*	0.056	0.399	0.805	0.176	0.689
Accuracy	r			-1.000	-0.434	0.200	0.059	0.315	0.096
	p			<0.001*	0.056	0.399	0.805	0.176	0.689
Usefulness	r				0.434	-0.200	-0.059	-0.315	-0.096
	p				0.056	0.399	0.805	0.176	0.689
FRE	r				1.000	-0.908	0.357	0.085	0.121
	p					<0.001*	0.122	0.722	0.612
FKGL	r						-0.579	-0.310	-0.345
	p						0.007*	0.183	0.137
DISCERN	r							0.600	0.544
	p							0.005*	0.013*
Understandability%	r								0.455
	p								0.044*

*p<0.05 and r: Correlation coefficient, mGQS: Modified Global Quality Score, FRE: Flesch reading ease, FKGL: Flesch-Kincaid grade level, U: Understandability, A: Actionability

p<0.05). FRE was strongly and negatively correlated with FKGL (r=-0.964, p<0.05), and was moderately and positively correlated with understandability% (r=0.530, p<0.05). FKGL was moderately correlated with understandability% (r=-0.539, p<0.05). Finally, understandability% and actionability% were strongly positively correlated (r=0.778, p<0.05) (**Table 5**).

DISCUSSION

This study is among the first to systematically identify the clinically relevant patient questions about apical root-end resection and to evaluate the accuracy, content quality, readability, and patient-centered usability of responses generated by four AI-powered conversational agents (ChatGPT-4, DeepSeek V3.1, Copilot, and Gemini 3). Our findings revealed significant differences in response quality, accuracy, and readability across the evaluated platforms. Models differed overall in terms of readability. Some responses were suitable for patient education and some were beyond the recommended grade level. There were also limitations in the accuracy and depth of content. The results are consistent with the previous literature emphasizing the potential and limits of AI-based tools in dentistry. Compared to prior research that mainly assesses readability metrics,[16,18,23,26] the application of the PEMAT and DISCERN tools in our study provides a more complex assessment of whether content produced by AI is both readable and actionable and of high quality from a patient education standpoint.

Bahadir et al.[27] found that patients have a positive attitude towards AI-assisted dental diagnostics and trust their dentists. Similarly, Kosan et al.[28] showed that patients tend

Table 4. Relationships between metrics for DeepSeek V3.1

		mGQS	Accuracy	Usefulness	FRE	FKGL	DISCERN	U%	A%
CLEAR	r	0.427	0.427	0.427	0.228	-0.274	0.065	0.437	0.589
	p	0.061	0.061	0.061	0.334	0.242	0.785	0.054	0.006*
mGQS	r		1.000	1.000	0.139	-0.338	0.442	0.342	0.183
	p		<0.001*	<0.001*	0.558	0.145	0.051	0.139	0.440
Accuracy	r			1.000	0.139	-0.338	0.442	0.342	0.183
	p			<0.001*	0.558	0.145	0.051	0.139	0.440
Usefulness	r				0.139	-0.338	0.442	0.342	0.183
	p				0.558	0.145	0.051	0.139	0.440
FRE	r					-0.971	0.161	0.042	0.243
	p					<0.001*	0.499	0.861	0.301
FKGL	r						-0.242	-0.075	-0.183
	p						0.303	0.752	0.440
DISCERN	r							0.499	0.227
	p							0.025*	0.335
U%	r								0.640
	p								0.002*

*p<0.05 and r: Correlation coefficient, mGQS: Modified Global Quality Score, FRE: Flesch reading ease, FKGL: Flesch-Kincaid grade level, U: Understandability, A: Actionability

Table 5. Relationships between metrics for Gemini 3

		mGQS	Accuracy	Usefulness	FRE	FKGL	DISCERN	U%	A%
CLEAR	r	0.026	0.026	0.026	-0.211	0.298	-0.018	-0.404	-0.374
	p	0.913	0.913	0.913	0.371	0.202	0.939	0.077	0.104
mGQS	r		1.000	1.000	-0.060	-0.020	0.392	0.083	-0.150
	p		<0.001*	<0.001*	0.803	0.934	0.087	0.729	0.527
Accuracy	r			1.000	-0.060	-0.020	0.392	0.083	-0.150
	p			<0.001*	0.803	0.934	0.087	0.729	0.527
Usefulness	r				-0.060	-0.020	0.392	0.083	-0.150
	p				0.803	0.934	0.087	0.729	0.527
FRE	r					-0.964	0.225	0.530	0.435
	p					0.001*	0.339	0.016*	0.055
FKGL	r						-0.233	-0.539	-0.388
	p						0.322	0.014*	0.091
DISCERN	r							0.309	0.298
	p							0.186	0.201
U%	r								0.778
	p								<0.001*

*p<0.05 and r: Correlation coefficient, mGQS: Modified Global Quality Score, FRE: Flesch reading ease, FKGL: Flesch-Kincaid grade level, U: Understandability, A: Actionability

to have a positive attitude towards AI in dentistry and that AI-assisted diagnostics could enhance patients' caries detection on radiographs without significantly affecting their trust in dental professionals. Previous studies[27,28] have reported generally positive patient attitudes towards AI in dentistry, although they stress the continued importance of trust in dental professionals. The findings underscore the importance of AI-generated information being accurate and understandable, and used as a supplement, not a substitute, to professional consultation. Thus, AI-powered conversational agents seem to have increasing potential in both clinical education and patient information. In this context, we sought to evaluate the accuracy of AI responses on apical root-end resection, a procedure where patients, even with detailed in-office explanations, frequently retain residual uncertainty.

Previous studies indicate that AI systems can achieve remarkable accuracy in areas such as dental implant prognosis, peri-implant bone loss detection, and apical surgery. For instance, Alqutaibi et al.[29] described that AI algorithms predicting implant prognosis from radiographic data achieved up to 99.8% accuracy, while Mugri et al.[30] reported that AI models detecting peri-implant bone loss demonstrated 61-94.7% accuracy with high specificity. While the high performance of AI in image-based dental diagnostic tasks has been demonstrated, such findings are not directly comparable with the evaluation of generative chatbot responses. Generative models are evaluated separately, as they could generate linguistically plausible but clinically incorrect information. In endodontics, Durmazpınar et al.[31] showed that the diagnostic accuracy of ChatGPT-4 was higher than that of 3rd- and 5th-year dental students, suggesting the possibility of AI as a clinical education tool. Ekmekci et al.[32] also reported that ChatGPT-4 with a PDF plugin achieved higher accuracy than ChatGPT-4o and Gemini in answering questions about regenerative endodontic treatment. The studies support the proposition of AI generating clinically relevant and accurate responses, consistent with our finding that all four chatbots scored high on the CLEAR, mGQS, and Accuracy scales. Our query set was created from established clinical consensus statements and the AAE and Cohen's Pathways of the Pulp guidelines, thus establishing an evidence-based reference standard based on established clinical guidelines and consensus sources in evaluating the AI models, rather than arbitrary patient inquiries.[4,5,10] We find more subtle differences across platforms as well. Esmailpour et al.[14] found that Gemini is better in factual correctness and ChatGPT-4 and Copilot showed more uniform overall performance in a broader set of questions. There was no difference in DISCERN scores between the four models. However, Copilot scored the highest for mGQS and Accuracy. These findings partly align with previous work[33] that indicates AI-generated answers may offer high-quality patient information overall but are still not deep enough for clinical decision-making. Copilot scored the highest mean CLEAR score followed by ChatGPT-4, both significantly higher than Gemini 3, suggesting that readability and understandability may not fully align with factual accuracy, which is also noted in the evaluation of AI responses to emergency care questions by Yau et al.[34] Gemini 3 had a higher FKGL and lower understandability than DeepSeek V3.1, suggesting that increased linguistic complexity may limit the accessibility of otherwise informative responses. The information in Gemini 3 was generally accurate, but the higher reading grade level appeared to compromise practical understandability. Importantly, the negative correlations between Usefulness and CLEAR, mGQS, or Accuracy should not be interpreted as indicating that more accurate or higher-quality responses were less useful. This pattern primarily resulted from the opposite scoring directions: lower Usefulness scores denoted greater usefulness, whereas higher CLEAR, mGQS, and Accuracy scores indicated better performance. Because the Usefulness scale was reverse-coded, with lower scores indicating greater usefulness, negative correlations with positively oriented measures such as CLEAR, mGQS, and Accuracy represented concordant rather than contradictory performance.

A multidisciplinary panel, which included periodontists and an oral and maxillofacial radiologist, broadly assessed the AI-generated responses. This multidisciplinary approach enabled evaluation of the responses from the perspectives of the periodontist/surgeon and the radiologist relevant to apical root-end resection and follow-up.

In our study, the analysis of readability indicated that the responses from Gemini 3 and Copilot had higher FKGL scores, showing that generally, a higher educational reading level was required for responses. DeepSeek V3.1 had the lowest mean FKGL score. Büker and Mercan[26] reported that none of the AI chatbots met the recommended level of readability for patient education materials, but Gemini was superior to the other models in terms of readability, accuracy and quality of response. Dursun and Geçer[35] found that all AI chatbots provided generally accurate and moderate-to-good quality information on clear aligners, but the responses from Gemini were significantly more readable than ChatGPT-3.5, ChatGPT-4 and Copilot. Özcivelek and Özcan[17] found that the readability of AI chatbots was different when patients asked questions about dental and maxillofacial prostheses. The best readability was DeepSeek-R1 and the worst readability was Dental GPT. Esmailpour et al.[14] evaluated the performance of AI chatbots in responding to patients' questions regarding dental prosthesis. They observed that Google Gemini scored the highest quality scores, whereas ChatGPT offered more complex and less readable responses. Variability in readability across studies may be attributed to differences in model architecture, training data, prompt design, domain-specific fine-tuning, and question complexity. The choice of the AI architecture, the training datasets, the prompt design and the domain-specific fine-tuning can impact the complexity of the sentence, the choice of words and the overall structure of the text. Also, the type of question asked, ranging from a technical clinical procedure to a general patient concern, will probably affect the language complexity required to give an accurate, yet comprehensible, answer. Together, these factors help to explain why some AI outputs have greater readability and others are above the recommended levels for patient education materials.

Limitations

This study has a few limitations. First, only four AI-powered chatbots were assessed (ChatGPT-4, DeepSeek V3.1, Copilot, and Gemini 3), which limits the generalizability of the findings to other AI models or future versions. Second, the responses were collected on one day and only the initial responses were considered. As AI models are constantly updating and improving their responses, the results might be

different if the same questions were asked at a different time or with different prompts. Third, although we systematically assessed the readability by established indices, the patient-centered understandability and actionability were assessed by the PEMAT tool and not by testing on patients, which may not fully reflect the real-world comprehension. Finally, the complexity and wording of questions could have affected the quality and clarity of the AI-generated responses. Together, these limitations highlight the need for careful interpretation and the need for further studies to assess AI chatbots in other languages, clinical domains, patient populations, and longitudinal settings.

CONCLUSION

The tested models presented considerable potential for the delivery of generally accurate and clear information regarding apical root-end resection, demonstrating different performance profiles. Copilot had the highest overall mGQS and Accuracy scores, the most favorable usefulness scores, and the highest mean CLEAR score, with ChatGPT-4 close behind, both significantly outperforming Gemini 3 on CLEAR. There was no significant difference in the DISCERN scores between the four models, indicating comparable performance for this measure. Gemini 3 scored higher on FKGL and lower on understandability compared with DeepSeek V3.1, suggesting that clinically informative content may still be difficult to understand for patients when presented at a higher reading level. Because the usefulness scale was reverse-coded, negative correlations with positively oriented measures such as CLEAR, mGQS and Accuracy reflected concordant rather than contradictory performance. Overall, the findings suggest that technical accuracy, readability, understandability and actionability are related but distinct dimensions of patient information generated by AI. AI chatbots can be useful additions to answering patient questions, but clinician oversight is advised to make sure that the information provided is clinically accurate, understandable, and practically actionable. Future developments should therefore take into account the trade-off between technical accuracy and patient-centered communication as well as existing health literacy standards.

ETHICAL DECLARATIONS

Ethics Committee Approval

This study did not require ethical approval as it did not involve any human subjects or animal experiments.

Informed Consent

Because the study did not involve human participants, written informed consent was not required.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: TP, SG, SCŞ, ÖSA, GU, ŞÇB, SÖ; Design: TP, SG, SCŞ, ÖSA, GU, ŞÇB, SÖ; Control: TP, SG, SCŞ, ÖSA, GU, ŞÇB, SÖ; Resources: TP, SG; Materials: TP, SG; Data Collection and/or Processing: TP, SG, SCŞ, ÖSA, GU, ŞÇB, SÖ; Analysis and/or Interpretation: TP, SG, SCŞ, ÖSA, GU; Literature Review: SCŞ, ÖSA, GU, ŞÇB, SÖ; Writing the Article: TP, SG, SCŞ, ÖSA, GU, ŞÇB, SÖ; Critical Review: TP, SCŞ, ÖSA, GU.

REFERENCES

1. Möller AJ, Fabricius L, Dahlén G, et al. Influence on periapical tissues of indigenous oral bacteria and necrotic pulp tissue in monkeys. *Scand J Dent Res.* 1981;89(6):475-484. doi:10.1111/j.1600-0722.1981.tb01711.x
2. Song M, Nam T, Shin SJ, Kim E. Comparison of clinical outcomes of endodontic microsurgery: 1 year versus long-term follow-up. *J Endod.* 2014;40(4):490-494. doi:10.1016/j.joen.2013.10.034
3. Setzer FC, Shah SB, Kohli MR, Karabucak B, Kim S. Outcome of endodontic surgery: a meta-analysis of the literature-part 1: comparison of traditional root-end surgery and endodontic microsurgery. *J Endod.* 2010;36(11):1757-1765. doi:10.1016/j.joen.2010.08.007
4. American Association of Endodontists. Glossary of Endodontic Terms. 10th ed. American Association of Endodontists; 2020.
5. Berman LH, Hargreaves KM. Cohen's Pathways of the Pulp. 12th ed. Elsevier; 2020.
6. Jiang F, Jiang Y, Zhi H, et al. Artificial Intelligence in healthcare: past, present and future. *Stroke Vasc Neurol.* 2017;2(4):230-243. doi:10.1136/svn-2017-000101
7. Schwendicke F, Samek W, Krois J. Artificial Intelligence in dentistry: chances and challenges. *J Dent Res.* 2020;99(7):769-774. doi:10.1177/0022034520915714
8. Keleş A, Keskin C. A micro-computed tomographic study of band-shaped root canal isthmuses, having their floor in the apical third of mesial roots of mandibular first molars. *Int Endod J.* 2018;51(2):240-246. doi:10.1111/iej.12842
9. Xu T, Fan W, Tay FR, et al. Micro-computed tomographic evaluation of the prevalence, distribution, and morphologic features of accessory canals in Chinese permanent teeth. *J Endod.* 2019;45(8):994-999. doi:10.1016/j.joen.2019.04.001
10. Ordinola-Zapata R, Martins JNR, Niemczyk S, Bramante CM. Apical root canal anatomy in the mesiobuccal root of maxillary first molars: influence of root apical shape and prevalence of apical foramina-a micro-CT study. *Int Endod J.* 2019;52(8):1218-1227. doi:10.1111/iej.13109
11. Xu T, Tay FR, Gutmann JL, et al. Micro-computed tomography assessment of apical accessory canal morphologies. *J Endod.* 2016;42(5):798-802. doi:10.1016/j.joen.2016.02.006
12. Stropko JJ, Doyon GE, Gutmann JL. Root-end management: resection, cavity preparation, and material placement. *Endod Topics.* 2005;11(1):131-151. doi:10.1111/j.1601-1546.2005.00158.x
13. Thurzo A, Urbanová W, Novák B, et al. Where is the Artificial Intelligence applied in dentistry? Systematic review and literature analysis. *Healthcare (Basel).* 2022;10(7):1269. doi:10.3390/healthcare10071269
14. Esmailpour H, Rasaie V, Babaee Hemmati Y, Falahchai M. Performance of Artificial Intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health.* 2025;25(1):574. doi:10.1186/s12903-025-05965-9
15. Jacobs T, Shaari A, Gazonas CB, Ziccardi VB. Is ChatGPT an accurate and readable patient aid for third molar extractions? *J Oral Maxillofac Surg.* 2024;82(10):1239-1245. doi:10.1016/j.joms.2024.06.177
16. Ceylan Şen S, Çardakci Bahar Ş, Saraç Atagün Ö, et al. Quality and reliability of AI information on dental implant failure: a comparative multi-model analysis. *J Craniofac Surg.* 2026;37(3-4):675-680. doi:10.1097/SCS.0000000000012415
17. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. *BMC Oral Health.* 2025;25(1):871. doi:10.1186/s12903-025-06267-w
18. Ceylan Şen S, Saraç Atagün Ö, Ustaoğlu G, et al. Do AI chatbots tell the truth about dentin hypersensitivity? A comparative evaluation of quality, accuracy, and readability. *Cumhuriyet Dent J.* 2026;29(1):138-147. doi:10.7126/cumudj.1848545
19. Helvacioglu-Yigit D, Demirturk H, Ali K, et al. Evaluating Artificial Intelligence chatbots for patient education in oral and maxillofacial radiology. *Oral Surg Oral Med Oral Pathol Oral Radiol.* 2025;139(6):750-759. doi:10.1016/j.oooo.2025.01.001

20. National Library of Medicine. How to write easy-to-read health materials. Updated November 2020. Accessed December 12, 2022. https://medlineplus.gov/all_easytoread.html
21. Sallam M, Barakat M, Sallam M. Pilot testing of a tool to standardize the assessment of the quality of health information generated by Artificial Intelligence-based models. *Cureus*. 2023;15(11):e49373. doi:10.7759/cureus.49373
22. Bernard A, Langille M, Hughes S, et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am J Gastroenterol*. 2007;102(9):2070-2077. doi:10.1111/j.1572-0241.2007.01325.x
23. Hatia A, Doldo T, Parrini S, et al. Accuracy and completeness of ChatGPT-generated information on interceptive orthodontics: a multicenter collaborative study. *J Clin Med*. 2024;13(3):735. doi:10.3390/jcm13030735
24. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol*. 2023;29(3):721-732. doi:10.3350/cmh.2023.0089
25. Alnsour MM, Alenezi R, Barakat M, et al. Assessing ChatGPT's suitability in responding to the public's inquires on the effects of smoking on oral health. *BMC Oral Health*. 2025;25(1):1207. doi:10.1186/s12903-025-06377-5
26. Büker M, Mercan G. Readability, accuracy and appropriateness and quality of AI chatbot responses as a patient information source on root canal retreatment: a comparative assessment. *Int J Med Inform*. 2025;201:105948. doi:10.1016/j.ijmedinf.2025.105948
27. Bahadir HS, Keskin NB, Çakmak EŞK, et al. Patients' attitudes toward Artificial Intelligence in dentistry and their trust in dentists. *Oral Radiol*. 2025;41(1):52-59. doi:10.1007/s11282-024-00775-1
28. Kosan E, Krois J, Wingenfeld K, et al. Patients' perspectives on Artificial Intelligence in dentistry: a controlled study. *J Clin Med*. 2022;11(8):2143. doi:10.3390/jcm11082143
29. Alqutaibi AY, Algabri RS, Alamri AS, et al. Advancements of Artificial Intelligence algorithms in predicting dental implant prognosis from radiographic images: a systematic review. *J Prosthet Dent*. 2025;134(6):2177-2188. doi:10.1016/j.prosdent.2024.10.036
30. Mugri MH. Accuracy of Artificial Intelligence models in detecting peri-implant bone loss: a systematic review. *Diagnostics (Basel)*. 2025;15(6):655. doi:10.3390/diagnostics15060655
31. Durmazpinar PM, Ekmekci E. Comparing diagnostic skills in endodontic cases: dental students versus ChatGPT-4o. *BMC Oral Health*. 2025;25(1):457. doi:10.1186/s12903-025-05857-y
32. Ekmekci E, Durmazpinar PM. Evaluation of different Artificial Intelligence applications in responding to regenerative endodontic procedures. *BMC Oral Health*. 2025;25(1):53. doi:10.1186/s12903-025-05424-5
33. Terzi M, Yavuz MC, Bicer T, Buyuk SK. Evaluation of Artificial Intelligence robot's knowledge and reliability on dental implants and peri-implant phenotype. *Sci Rep*. 2025;15(1):9519. doi:10.1038/s41598-025-94576-z
34. Yau JYS, Saadat S, Hsu E, et al. Accuracy of prospective assessments of 4 large language model chatbot responses to patient questions about emergency care: experimental comparative study. *J Med Internet Res*. 2024;26:e60291. doi:10.2196/60291
35. Dursun D, Bilici Geçer R. Can Artificial Intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak*. 2024;24(1):211. doi:10.1186/s12911-024-02619-8