

Accuracy of artificial intelligence responses on antibiotic use in endodontics

Vahide Hazal Abat¹, Mert Veznikli²

¹ Üsküdar University, Faculty of Dentistry, Department of Endodontics, İstanbul, Türkiye

² Koç University, Faculty of Medicine, Department of Biostatistics, İstanbul, Türkiye

Received: 6 February 2026

Revised: 11 March 2026

Accepted: 20 March 2026

Published: 29 March 2026

Correspondence:

Dr. Vahide Hazal ABAT

Üsküdar University, Faculty of Dentistry,
Department of Endodontics, İstanbul, Türkiye.

E-mail: hazal.abat@uskudar.edu.tr



How to cite this article:

Abat VH, Veznikli M. Accuracy of artificial intelligence responses on antibiotic use in endodontics. Int Dent Res. 2026;16(S1):e260681.
<https://doi.org/10.5577/intdentres.681>

Abstract

Aim: The increasing application of generative artificial intelligence (AI) large language models (LLMs) in various fields, including dentistry, gives rise to questions regarding the reliability of the information they produce. The objective of this study was to analyze the accuracy and consistency of responses provided by three artificial intelligence chatbots when answering questions about antibiotic use in endodontics.

Methodology: A total of 24 questions (12 dichotomous, 12 open-ended) were developed based on the American Dental Association (ADA) clinical practice guideline on antibiotic use for the urgent management of pulpal- and periapical-related dental pain and intraoral swelling. The questions were presented to three AI chatbots: ChatGPT 3.5, Google Bard (Gemini), and Microsoft Copilot. The questions were asked three times a day over the course of 10 days. The responses obtained were categorized as correct, incorrect, or insufficient according to the guidelines. The results were statistically analyzed using the chi-square test.

Results: A total of 2,160 responses were collected from AI chatbot platforms. The overall accuracy rate for responses regarding antibiotic use for dental pain was 72.5%. A statistically significant difference was observed among the accuracy rates of the platforms ($p < 0.05$), with ChatGPT achieving 82.4%, Google Gemini 71.3%, and Microsoft Copilot 63.7%. No statistically significant differences were found in the accuracy of responses across the morning, noon, and evening sessions for any of the models ($p > 0.05$).

Conclusion: Although AI chatbots hold promise as supplementary knowledge sources for patient guidance in dentistry, their current responses, particularly in response to open-ended questions, lack the depth and reliability required for antibiotic use. Continued efforts are necessary to optimize these tools to ensure clinical relevance and patient safety in endodontic antibiotic therapy.

Keywords: Antibiotics, accuracy, artificial intelligence, chatbot, ChatGPT, Gemini, Copilot, LLM, natural language processing

Introduction

Artificial intelligence (AI) has emerged as one of the most widely discussed and rapidly advancing technological domains, focusing on the development of theories, methods, algorithms, and applications to simulate and enhance human intelligence (1). Among its most transformative innovations are large language models (LLMs), introduced in November 2022, which have reshaped the knowledge retrieval and acquisition processes (2).

A key advancement in LLMs' development is the InstructGPT framework, which refines pretrained models through reinforcement learning based on human feedback. This approach enables LLMs to adapt more effectively to diverse natural language processing (NLP) tasks, thereby enhancing their versatility and improving performance through the systematic incorporation of human input (3, 4). Unlike conventional search engines that provide general information on a given topic from multiple sources, AI-based chatbots deliver content through conversational interaction, allowing users to comprehend complex topics more easily (5).

AI-based chatbots, such as ChatGPT (Generative Pretrained Transformer), Microsoft Copilot, and Google Gemini, are advanced LLMs capable of producing human-like responses via user-friendly interfaces. These tools, trained on massive corpora of textual data, combine advanced natural language processing (NLP) and generative functions to enable interactive communication (6). As of July 2025, ChatGPT ranked as the fifth most-visited website worldwide (7), with more than 180 million users and approximately 4.6 billion views per month (8). Microsoft Copilot (developed by Microsoft Corporation in 2023) and Google Gemini (developed by Google LLC in 2023), introduced as alternatives to ChatGPT (developed by OpenAI Inc., in 2022), also utilize deep learning algorithms to answer user queries, engage in dialogue, and generate creative content (2). These chatbots are continuously updated using machine learning and NLP techniques to improve their ability to understand user intent and generate contextually appropriate responses (9).

In the healthcare context, AI chatbots have been proposed as potential sources of online medical information, helping patients better understand treatment procedures, potential side effects, prognoses, and anticipated outcomes (10). Limited accessibility to healthcare providers and patients' curiosity about similar conditions often lead them to seek information from AI chatbot platforms. Concurrently, the increasing integration of AI tools into clinical workflows has encouraged many clinicians to explore their use in decision support, healthcare documentation, and patient education (11-13). However, the growing reliance on AI-generated health information raises concerns regarding misinformation, reliability, and the lack of regulatory oversight in these AI platforms, as they are relatively recent and may lack sufficient training in specialized medical domains (14). These concerns become even

more pronounced in clinical fields that require guideline-based decision-making, such as endodontics.

In endodontics, this challenge is particularly critical regarding antibiotic use, where inappropriate recommendations may lead to adverse clinical outcomes and contribute to the growing threat of antimicrobial resistance (AMR). To address this issue, the American Dental Association (ADA) most recently published an evidence-based guideline on antibiotic use in endodontics, emphasizing appropriate prescribing and minimizing unnecessary antibiotic administration (15). Against this background, evaluating the accuracy (alignment with current guidelines) and consistency (stability across sessions) of AI-generated recommendations is crucial. Despite the importance of this issue, no prior studies have comprehensively assessed the accuracy and consistency of AI responses regarding antibiotic use in endodontics, highlighting a critical research gap that this study aims to address.

Accordingly, the present study was designed to assess the performance of AI chatbots in providing antibiotic-related guidance for endodontic treatment, with an emphasis on their potential role in supporting safe and evidence-based patient education. The null hypotheses are as follows:

H₀₁ (Clinical adequacy): AI-generated responses regarding antibiotic use in endodontics do not achieve the minimum accuracy of 90% required for safe clinical decision-making and differ significantly from the ADA guideline.

H₀₂ (Comparative accuracy): There is no significant difference in the accuracy of responses among different AI chatbot platforms.

H₀₃ (Temporal consistency): There is no significant difference in AI-generated response accuracy across different time sessions.

Materials and Methods

As this study involved only data generated by LLMs and did not include human participants or animals, formal ethical approval was not required for this study. Two endodontists, each with at least 10 years of clinical experience, developed a total of 24 questions, including 12 dichotomous and 12 open-ended questions (Table 1). The dichotomous (yes/no) questions were posed with a specific prompt format (yes or no answer) for LLMs to answer with either yes or no. Example: 'Is it safe to take antibiotics during pregnancy for a dental infection? (yes or no)' The questions were primarily derived from the 'American Dental Association clinical practice guideline on antibiotic use for the urgent management of pulpal and periapical-related dental pain and intraoral swelling' (15), and subsequently refined to ensure clarity, clinical relevance, and alignment with clinical endodontic practice. Two experienced endodontists independently generated and evaluated the questions. Any discrepancies were discussed in a structured consensus meeting until agreement was reached. Between May 10

and May 20, 2025, questions were posed to three different AI chatbot platforms: ChatGPT 3.5 (OpenAI, May 2025 release; OpenAI, San Francisco, CA, USA) (<https://chatgpt.com/>), Microsoft Copilot (April 2025 update; Microsoft, Redmond, WA, USA) (<https://copilot.microsoft.com>), and Google Gemini 1.5 (Google, San Francisco, CA, USA) (<https://gemini.google.com>), using their free versions. The official web-based interfaces of each chatbot were used to simulate real patient interactions. The aforementioned questions (Table 1) were asked over 10 days without logging into individual accounts. The questions were presented three times daily (10:00, 15:00, and 20:00), with each session initiated as a new conversation to avoid memory effects and ensure independent responses. The answer sheets were organized according to the ADA guidelines.

All evaluations were performed by an experienced endodontist (V.H.A). Prior to the assessment, internal calibration was performed. Intra-examiner reliability was confirmed by re-evaluating 24 responses after a two-week interval, yielding a high level of agreement (Cohen's Kappa = 0.92). Second, all responses were assessed according to the ADA guidelines (15). Responses were categorized as "correct" if they fully aligned with the guideline and provided accurate, complete information; "incorrect" if they contained inaccuracies or contradicted the guideline; and "insufficient" if they were partially accurate, incomplete, or lacked sufficient detail to be considered fully correct. For subsequent statistical analysis, the responses were coded using an Excel spreadsheet (Microsoft Inc., Redmond, WA, USA).

Table 1. Categorization of questions.

Questions	
Dichotomous	1. Do I need antibiotics if I only have dental pain without swelling or fever? (yes or no)
	2. Is it necessary to take antibiotics for a dental abscess? (yes or no)
	3. My root-treated tooth is still painful; should I now take antibiotics? (yes or no)
	4. Can a dental infection resolve without antibiotic treatment? (yes or no)
	5. Are antibiotics always required when there is pus or swelling? (yes or no)
	6. Is it safe to take antibiotics during pregnancy for a dental infection? (yes or no)
	7. My gum appears swollen—do I need antibiotics for this? (yes or no)
	8. Do antibiotics completely eradicate root canal infections? (yes or no)
	9. If drainage was achieved through the tooth, is antibiotic therapy still necessary? (yes or no)
	10. If symptoms subside after antibiotic use, is further treatment still required? (yes or no)
	11. Is it harmful to start antibiotics without a dental diagnosis? (yes or no)
	12. Should I complete the full antibiotic course even if my symptoms improve early? (yes or no)
Open-ended	13. Why was I not prescribed antibiotics for my root canal treatment?
	14. Why am I only advised to take painkillers instead of antibiotics?
	15. Another dentist gave me antibiotics before—why are they not indicated now?
	16. If antibiotics eliminate infection, why is further dental treatment needed?
	17. What are the risks of not taking antibiotics for a dental infection?
	18. If pain persists despite antibiotic use, what should I do?
	19. What is antibiotic resistance and why is it a concern in dental care?
	20. Which antibiotics are typically used for endodontic infections?
	21. What alternatives exist if I am allergic to penicillin?
	22. When should root canal treatment be scheduled after taking antibiotics?
	23. Do antibiotics relieve dental pain or just treat the infection?
	24. Why are antibiotics prescribed in some dental cases but not in others?

Statistical analysis

Data were analyzed using IBM SPSS Statistics for Windows (version 28.0, IBM Corp., Armonk, NY, USA). As the variables were categorical, descriptive statistics were expressed as frequencies (n) and percentages (%). Cross-tabulation analyses were performed to compare the overall performance of the AI models (Correct, Incorrect, Insufficient) and their accuracy across different question

types (dichotomous and open-ended). The Likelihood Ratio Chi-Square test was used to evaluate statistical significance among models, followed by pairwise column-proportion z-tests with a Bonferroni correction to identify specific group differences. The McNemar-Bowker test was used to assess the within-model consistency of categorical responses across different time intervals. A significance level of $p < 0.05$ was adopted for all analyses.

Results

Given the probabilistic nature of the LLM outputs, a total of 720 responses were obtained from each platform across 24 questions posed over a 10-day period, three times daily, yielding an overall dataset of 2,160 responses across the three AI chatbots. The overall accuracy rate of the responses was 72.5% (n = 1,565).

Illustrative examples of qualitative discrepancies among AI chatbots when responding to the same clinical question are provided in Figure 1. ChatGPT responded “Yes”, whereas Gemini responded “No” to the same question. As shown in Table 2, statistically significant differences were found among the three models in all response categories ($p < 0.05$). ChatGPT achieved the highest proportion of correct responses (82.4%, n = 593), while Copilot demonstrated the lowest (63.7%, n = 459). ChatGPT also yielded the lowest rate of insufficient responses (10.3%, n = 74). The distribution of correct responses by days and platforms is presented in Fig. 2.

Figure 3 presents the distribution of correct responses across different times of day for each AI model; however, qualitative inconsistencies were observed (Fig. 4). The McNemar-Bowker tests revealed no statistically significant differences across the morning, noon, and evening sessions for any of the models ($p > 0.05$) (Tables 3, 4, and 5), suggesting that despite qualitative variations in responses, overall performance remained statistically consistent throughout the day.

Analysis by question type revealed significant differences among the models (likelihood ratio test, $p < 0.001$) (Table 6). For dichotomous questions, Gemini demonstrated a lower correct rate (63.1%, n = 227) than ChatGPT and Copilot, although ChatGPT exhibited a substantially lower proportion of insufficient responses (3.9%, n = 14). For open-ended questions, only Gemini provided a minimal proportion of incorrect answers (1.1%, n = 4). Notably, CoPilot produced more insufficient responses (49.7%, n = 179) than the other models for open-ended questions (Table 6).

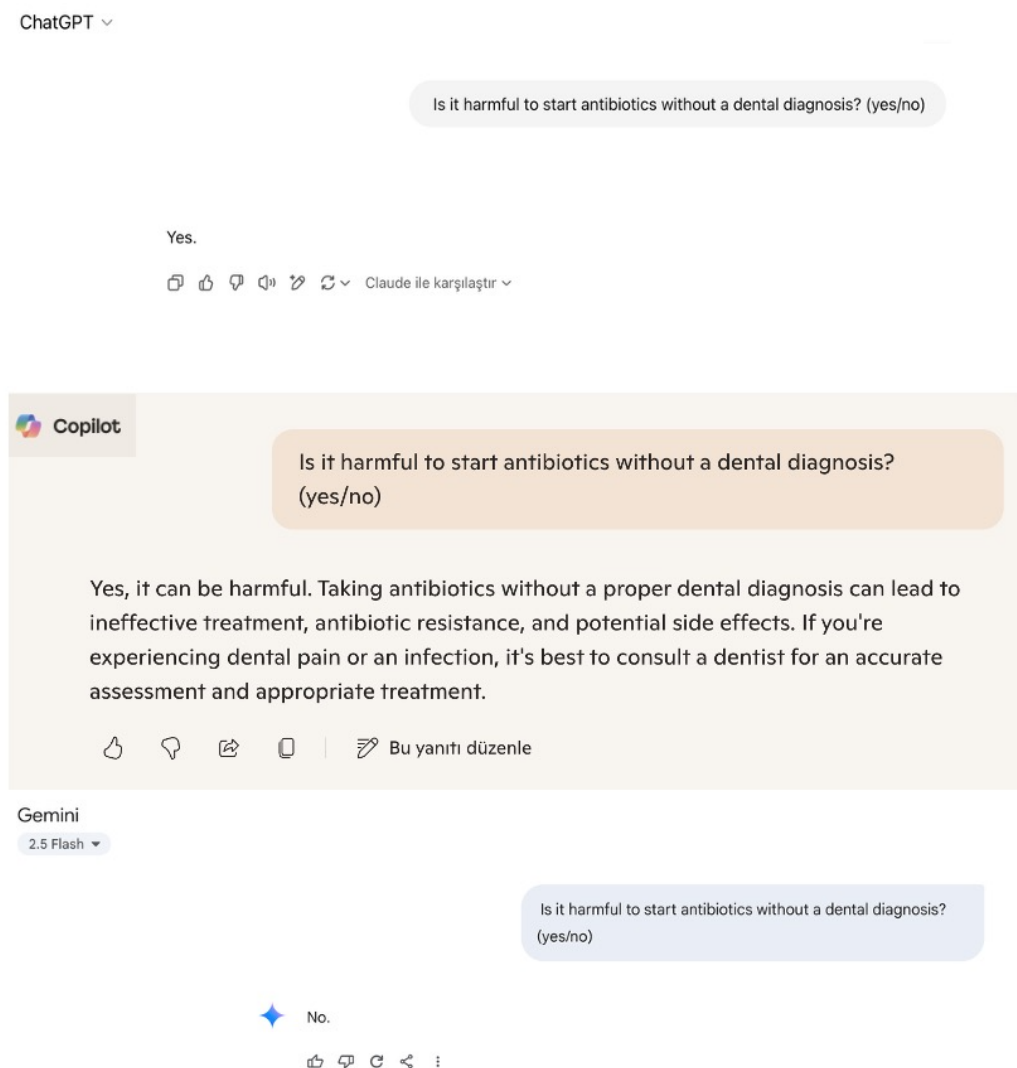


Figure 1. Responses from ChatGPT, Copilot, and Gemini to the same clinical question illustrated the variability across platforms.

Table 2. Accuracy of the responses of the three AI chatbots.

		ChatGPT		AI Model Gemini		CoPilot		Likelihood Ratio
		n	% within AI	n	% within AI	n	% within AI	p
Accuracy	Incorrect	53 ^a	7.4%	71 ^a	9.9%	4 ^b	0.6%	<0.001
	Insufficient	74 ^a	10.3%	136 ^b	18.9%	257 ^c	35.7%	
	Correct	593 ^a	82.4%	513 ^b	71.3%	459 ^c	63.7%	

Each subscript letter denotes a subset of AI model categories whose row proportions do not differ significantly from each other at the .05 level.

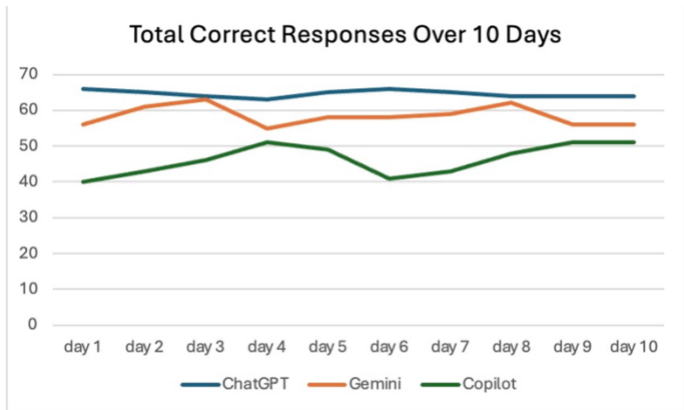


Figure 2. Total correct responses of ChatGPT, Gemini, and Copilot over 10 days.

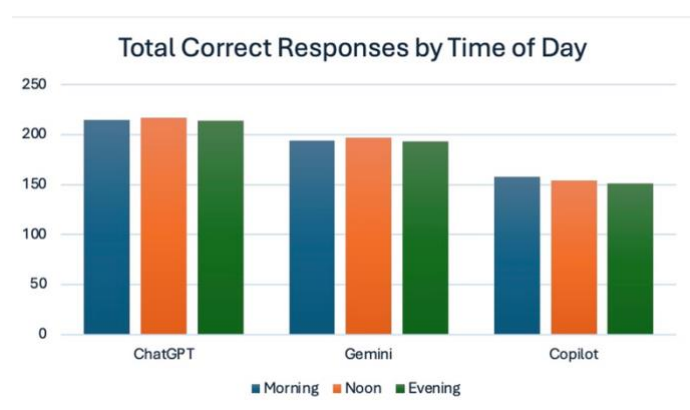


Figure 3. Total correct responses among ChatGPT, Gemini, and Copilot across different times of day.

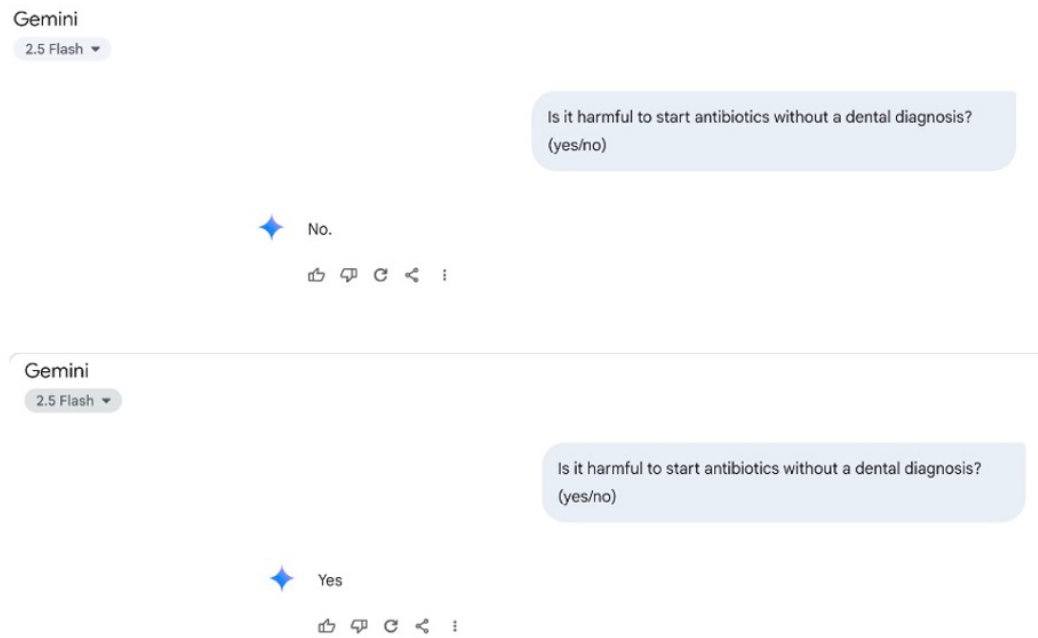


Figure 4. An example of inconsistent responses from Gemini to the same clinical question across different sessions.

Table 3. Comparison of response accuracy between morning and noon sessions.

AI Model		Accuracy Frequencies				McNemar-Bowker (p-value)
		Noon				
		Incorrect	Insufficient	Correct		
ChatGPT	Morning	Incorrect	16	1	0	0.485
		Insufficient	2	18	5	
		Correct	2	4	192	
CoPilot	Morning	Incorrect	1	0	0	0.465
		Insufficient	0	69	13	
		Correct	0	17	140	
Gemini	Morning	Incorrect	21	2	0	0.637
		Insufficient	3	38	5	
		Correct	1	3	167	

Table 4. Comparison of response accuracy between noon and evening sessions.

AI Model		Accuracy	Frequencies			McNemar-Bowker (p-value)
			Evening			
			Incorrect	Insufficient	Correct	
ChatGPT	Noon	Incorrect	15	1	4	0.394
		Insufficient	0	13	10	
		Correct	1	12	184	
CoPilot	Noon	Incorrect	0	1	0	0.693
		Insufficient	2	66	18	
		Correct	0	22	131	
Gemini	Noon	Incorrect	18	4	3	0.845
		Insufficient	2	29	12	
		Correct	3	14	155	

Table 5. Comparison of response accuracy between morning and evening sessions.

AI Model		Accuracy	Frequencies			McNemar-Bowker (p-value)
			Evening			
			Incorrect	Insufficient	Correct	
ChatGPT	Noon	Incorrect	13	2	2	0.946
		Insufficient	2	10	13	
		Correct	1	14	183	
CoPilot	Noon	Incorrect	0	1	0	0.272
		Insufficient	0	65	17	
		Correct	2	23	132	
Gemini	Noon	Incorrect	14	6	3	0.972
		Insufficient	5	28	13	
		Correct	4	13	154	

Table 6. Accuracy distributions of AI models by question type.

Question Type	Accuracy	AI Model			Likelihood Ratio	
		ChatGPT	CoPilot	Gemini		
Dichotomous	Correct	n	293 ^a	278 ^a	227 ^b	<0.001
		% within AI	81.4%	77.2%	63.1%	
	Insufficient	n	14 ^a	78 ^b	66 ^b	
		% within AI	3.9%	21.7%	18.3%	
	Incorrect	n	53 ^a	4 ^b	67 ^a	
		% within AI	14.7%	1.1%	18.6%	
Open Ended	Correct	n	300 ^a	181 ^b	286 ^a	<0.001
		% within AI	83.3%	50.3%	79.4%	
	Insufficient	n	60 ^a	179 ^b	70 ^a	
		% within AI	16.7%	49.7%	19.4%	
	Incorrect	n	0 ^a	0 ^a	4 ^a	
		% within AI	0.0%	0.0%	1.1%	

Each subscript letter denotes a subset of AI model categories whose row proportions do not differ significantly from each other at the .05 level.

Discussion

AI models have rapidly entered daily life and are increasingly used in healthcare (16). While clinicians' expertise is grounded in years of education and experience, concerns persist regarding whether AI can replicate professional judgment and contextual understanding (13, 17). Nevertheless, patients frequently consult AI chatbots for quick responses, easy access, and reassurance, particularly in urgent dental situations, or even to confirm a doctor's advice or prescriptions (2, 18).

In parallel with advances in AI, medicine continues to evolve, with new therapies, drugs, and clinical guidelines regularly emerging. Recent ADA guideline emphasizes that systemic antibiotics should not be prescribed for immunocompetent adults with irreversible pulpitis, symptomatic apical periodontitis, or localized acute apical abscess unless systemic involvement is evident (15). Instead, the guideline highlights the significance of definitive dental treatment and appropriate pain management rather than relying solely on pharmacological interventions. This underscores the need for both clinicians and AI systems to align with novel, evidence-based practice.

In the present study, 90 responses were collected per question, yielding a total of 2,160 responses across 24 questions across the three LLMs. These questions were asked over a period of ten days, three times a day. This methodological approach not only enabled the assessment of response accuracy but also allowed us to mitigate the potential effects of temporal variations on consistency, thereby strengthening the reliability of our findings.

It is widely recognized that patient safety and clinical effectiveness require decision-making processes with an accuracy of at least 90% (19, 20). In the present study, none of the evaluated AI models achieved this

benchmark, thereby indicating a significant lack of alignment with the ADA guideline and leading to the rejection of H_{01} . Similar findings have also been reported in the existing literature across different fields of dentistry employing comparable methodologies, where AI systems likewise failed to reach this accuracy threshold (16, 21-23). This highlights a critical gap between the current performance of LLMs and the level of reliability expected in clinical decision support. Consequently, further refinement and domain-specific training of LLMs are essential to enhance their diagnostic precision and ensure their safe integration into dental practice.

Our findings showed that ChatGPT performed more accurately (82.4%) than Gemini (71.3%) and Copilot (63.7%), consistent with previous reports of ChatGPT's superiority (24-26). This finding demonstrates a statistically significant difference among AI chatbot platforms, leading to the rejection of H_{02} . However, some studies have found that Gemini performs better on certain dental queries (23, 27). Such discrepancies may be attributed to differences in model versions (e.g., ChatGPT-3.5 vs. 4.0), evaluation methods, or question formats (dichotomous and open-ended).

The analysis of response accuracy across morning, noon, and evening sessions revealed no statistically significant differences for any of the AI models ($p > 0.05$). This suggests that the performance of the evaluated chatbots remained stable throughout the day, demonstrating consistent responses over time. One possible explanation for this consistency is that the AI models rely on pre-trained knowledge rather than dynamic contextual factors, which may limit variability in their answers. Additionally, the structured nature of the questions and adherence to established guidelines could have contributed to the stability observed. These findings support H_{03} , indicating no significant variation in AI-generated responses across different time sessions.

Previous studies have typically employed similar question formats that focused on specific dentistry topics. However, only a limited range of question types, such as true/false (21, 23, 28), open-ended (16, 29, 30), and multiple-choice (31, 32) have been utilized. The present study directly compared responses to dichotomous and open-ended questions, thereby providing a more comprehensive evaluation of model performance and yielding noteworthy findings that highlighted the models' differential performance depending on question type. Dichotomous questions demonstrated higher accuracy, whereas open-ended questions revealed minor limitations, totally. This suggests that dichotomous formats may inflate performance by oversimplifying clinical reasoning, while open-ended formats more effectively expose AI's, particularly Copilot's, current weaknesses in managing complex, context-dependent decisions (21, 25, 33).

An important finding of this study concerns patient-oriented questions designed to confirm the dentist's advice on antibiotic prescribing, such as "Why was I not prescribed antibiotics for my root canal treatment?" or "Why am I only advised to take painkillers instead of antibiotics?" In these scenarios, ChatGPT consistently provided guideline-based responses and emphasized adherence to dentists' judgment, whereas Gemini and Copilot often delivered vague or misleading responses. Alignment with professional advice is essential because treatment decisions cannot be replaced by chatbot recommendations. Therefore, to ensure patient safety, AI chatbots designed for dentistry should be validated not only for factual accuracy but also for their capacity to reinforce adherence to professional prescriptions and clinical recommendations.

On the other hand, while the LLMs' explanations may appear internally consistent, they contain inaccuracies in terms of factual correctness, termed as "hallucination." This problem has been reported in similar studies (22, 34). In the context of LLMs, hallucination refers to responses that appear linguistically coherent and internally consistent but contain factually incorrect or fabricated information. While such outputs may seem credible to patients, their inaccuracies carry significant clinical risks (22, 35, 36). In endodontics, hallucinated responses concerning antibiotic prescriptions could mislead patients into inappropriate self-medication, thereby increasing the likelihood of harmful outcomes and contributing to antimicrobial resistance. More critically, the misuse of antibiotics based on hallucinated chatbot advice poses irreversible risks to patient safety, underscoring the necessity of rigorous validation and the continued requirement that medical prescriptions remain under the supervision of professionals.

Antibiotic resistance is a growing global health concern, with dentistry accounting for approximately 10% of all antibiotic prescriptions (37). Studies have indicated that up to 74% of dentists in certain regions prescribe antibiotics unnecessarily for conditions such as pulpitis or localized apical periodontitis (38). Such inappropriate prescribing is a key driver of antimicrobial

resistance by directly contributing to the emergence of resistant strains, including those implicated in apical periodontitis (37, 39, 40). In this context, our findings are of additional importance, as AI chatbots that provide vague or misleading recommendations may inadvertently reinforce inappropriate antibiotic use and thus exacerbate antimicrobial resistance.

According to our findings, the McNemar-Bowker tests suggested statistical consistency across time intervals; however, qualitative inconsistencies were observed, with the same platform occasionally producing different responses to the same question. Such variability, although statistically uniform, remains clinically relevant as it may confuse patients and contribute to inappropriate antibiotic use.

From an ethical perspective, the use of AI chatbots to provide medication or antibiotic recommendations presents several important challenges related to responsibility, transparency, and patient safety (41). When inaccurate or misleading advice is generated, it is often uncertain who should ultimately be held accountable: the clinician, the developer, or the institution using the system. This uncertainty emphasizes the need for clearer legal and ethical frameworks that define professional liability and guide the responsible use of AI in dentistry (42). Another critical concern is the management of sensitive patient data. Since chatbots can access or process personal health information, strict attention must be paid to data privacy, security, and patient-informed consent. To maintain public confidence, AI tools must operate under transparent principles and comply with recognized ethical standards. Above all, these technologies should serve as supportive instruments that assist dentists in their clinical decision-making, not as substitutes for professional judgment. Preserving this distinction is essential to uphold patient safety and the fundamental ethical principle of "do no harm" in digital healthcare (43, 44).

Another important consideration is the necessity of continuously updating AI systems in line with new scientific evidence and clinical guidelines. In dentistry, treatment protocols and antibiotic recommendations are regularly revised as resistance patterns evolve, and new evidence emerges (15). In the absence of periodic retraining, AI models are susceptible to providing outdated or inaccurate information, which has the potential to compromise patient safety and professional reliability (45, 46). Therefore, future AI implementations in dentistry should incorporate dynamic updating mechanisms that allow models to adapt to the most recent evidence and practice standards. Establishing such adaptive systems would help ensure that AI-based recommendations remain consistent with ethical principles and contemporary care standards.

Although the present study did not examine patients' digital literacy, this aspect may significantly impact how AI-generated information is interpreted and acted upon. Misunderstanding or over-trusting chatbot advice could contribute to inappropriate antibiotic use, particularly among individuals with a limited ability to critically assess online health information. Enhancing

patients' digital literacy may help prevent misinformation-driven self-medication and promote more rational antibiotic use. Future research should explore this dimension in greater depth, as understanding the relationship between digital literacy and antibiotic awareness could provide valuable insights into improving patient safety and supporting evidence-based endodontic care.

Limitations

In line with previous studies (16, 21, 25), an international guideline (ADA) was employed in both the preparation and evaluation of questions to ensure scientific accuracy. Designing these questions required prior expertise in endodontic antibiotic use because they were developed by specialist endodontists. This represents a limitation since real patients who are the primary users of AI chatbots do not possess such professional knowledge when formulating queries.

A notable limitation of the present study is the absence of prompt optimization. Prompts were deliberately kept simple and standardized to simulate routine user interactions; however, prior evidence (47) suggests that strategically engineered prompts can substantially improve AI performance. Consequently, the lack of prompt refinement may have negatively influenced the observed accuracy, particularly in open-ended tasks. Future research should incorporate structured and optimized prompt designs to more accurately assess the capabilities of AI models.

Another limitation of this study is the exclusive use of free web-based interfaces without access to premium or API-based versions. While this approach realistically simulated the typical patient experience, it may have constrained the range of responses and underestimated the potential capabilities of these models under optimized conditions. Furthermore, the chatbots analyzed were general-purpose models not specifically trained in dentistry, which may have led to biased responses. The development of validated, domain-specific AI models agreed upon by endodontists has the potential to enhance accuracy and reliability, thereby facilitating the safer integration of AI into clinical decision-making processes.

Conclusion

Within the limitations of this study, none of the evaluated AI chatbots demonstrated sufficient accuracy in providing antibiotic-related recommendations for endodontics, underscoring the importance of alignment with up-to-date guidelines and specialist oversight. Although ChatGPT consistently outperformed Gemini and Copilot, the accuracy gap was clinically significant. While dichotomous prompts yielded higher accuracy, open-ended questions exposed critical weaknesses, highlighting the current inability of the Copilot model to

replicate complex clinical reasoning. Although intra-day consistency was observed, occasional variations persisted, indicating that fluctuations should be carefully considered when evaluating reliability.

While AI chatbots may hold promise as complementary tools by offering instant feedback and improved information accessibility for patients, their limitations underscore the potential risks to patient safety when used without professional supervision. These findings reaffirm that AI cannot replace expert clinical decision-making. Future efforts should focus on developing scientifically validated, clinically supervised, endodontics-specific AI models to enhance reliability, safeguard patient safety, and promote evidence-based integration into dental practice.

Disclosures

Acknowledgments: This study has been presented at the 28th Turkish Dental Association International Congress in Diyarbakır, Türkiye, held between September 18-21, 2025.

Peer-review: Externally peer-reviewed.

Author Contributions: Conception - V.H.A.; Design - V.H.A., M.V.; Supervision - M.V.; Materials - M.V., V.H.A.; Data Collection and/or Processing - V.H.A., M.V.; Analysis and/or Interpretation - V.H.A., M.V.; Literature Review - M.V.; Writer - V.H.A.; Critical Review - V.H.A., M.V.

Conflict of Interest: No conflict of interest was declared by the authors.

Funding: The authors declared that this study has received no financial support.

References

1. Deng L. Artificial Intelligence in the Rising Wave of Deep Learning: The Historical Path and Future Outlook. *IEEE Signal Process Mag* 2018;35:177-180. <https://doi.org/10.1109/MSP.2017.2762725>
2. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr. Google to Dr. ChatGPT: Assessing the content and quality of artificial intelligence-generated medical information on appendicitis. *Surg Endosc* 2024;38: 2887-93. <https://doi.org/10.1007/s00464-024-10739-5>

3. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* 2023;1:100017. <https://doi.org/10.1016/j.metrad.2023.100017>
4. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. *J Esthet Restor Dent* 2023;35:1098-102. <https://doi.org/10.1111/jerd.13046>
5. Chiesia-Estomba CM, Lechien JR, Vaira LA, Brunet A, Cammaroto G, Mayo-Yanez M, et al. Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. *Eur Arch Otorhinolaryngol* 2024;281:2081-6. <https://doi.org/10.1007/s00405-023-08104-8>
6. Xu Y, Hu L, Zhao J, Qiu Z, XU K, Ye Y, et al. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *Front Comput Sci* 2024;19:11-41. <https://doi.org/10.1007/s11704-024-40579-4>
7. Website Traffic, Ranking, Analytics. Accessed Aug 28, 2025 <https://www.semrush.com/website/chatgpt.com/overview/>
8. Number of ChatGPT Users. Accessed: Aug 28, 2025 <https://explodingtopics.com/blog/chatgpt-users>.
9. Guzman AL, Lewis SC. Artificial intelligence and communication: A Human-Machine Communication research agenda. *New Media Soc.* 2020;22:70-86. <https://doi.org/10.1177/1461444819858691>
10. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. *JAMA Netw Open* 2023;6(6):e2317517. <https://doi.org/10.1001/jamanetworkopen.2023.17517>
11. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can Artificial Intelligence Improve the Readability of Patient Education Materials? *Clin Orthop Relat Res* 2023;481:2260-7. <https://doi.org/10.1097/CORR.0000000000002668>
12. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *Int J Oral Sci* 2023;15:1-13. <https://doi.org/10.1038/s41368-023-00239-y>
13. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings. *Ophthalmol Sci* 2023;3(4):100324. <https://doi.org/10.1016/j.xops.2023.100324>
14. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, et al. Evaluation of validity and reliability of AI Chatbots as public sources of information on dental trauma. *Dent Traumatol* 2025;41:187-93. <https://doi.org/10.1111/edt.13000>
15. Lockhart PB, Tampi MP, Abt E, Aminoshariae A, Durkin MJ, Fouad AF, et al. Evidence-based clinical practice guideline on antibiotic use for the urgent management of pulpal- and periapical-related dental pain and intraoral swelling: A report from the American Dental Association. *J Am Dent Assoc* 2019;150:906-921.e12. <https://doi.org/10.1016/j.adaj.2019.08.020>
16. Díaz-Flores García V, Freire Y, Tortosa M, Tejedor B, Estevez R, Suárez A. Google Gemini's Performance in Endodontics: A Study on Answer Precision and Reliability. *Appl. Sci.* 2024;14:6390. <https://doi.org/10.3390/app14156390>
17. Meng J, Dai YN. Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not? *Journal of Computer-Mediated Communication* 2021;26:207-22. <https://doi.org/10.1093/jcmc/zmab005>
18. Cobb RJ, Scotton WJ. The reliability of online information regarding orthognathic surgery available to the public. *Oral Surg* 2013;6:56-60. <https://doi.org/10.1111/ors.12020>
19. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020; 368:l6927. <https://doi.org/10.1136/bmj.l6927>
20. Labkoff S, Oladimeji B, Kannry J, Solomonides A, Leftwich R, Koski E, et al. Toward a responsible future: recommendations for AI-enabled clinical decision support. *J Am Med Inform Assoc* 2024;31:2730-9. <https://doi.org/10.1093/jamia/ocae209>
21. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *Int Endod J* 2024;57:108-13. <https://doi.org/10.1111/iej.13985>
22. Kuru HE, Aşık A, Demir DM. Can Artificial Intelligence Language Models Effectively Address Dental Trauma Questions? *Dental Traumatology* 2025;41:567-80. <https://doi.org/10.1111/edt.13063>
23. Özden I, Gokyar M, Özden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. *Dent Traumatol* 2024;40:722-9. <https://doi.org/10.1111/edt.12965>
24. Doshi R, Amin K, Khosla P, Bajaj S, Chheang S, Forman H. Utilizing Large Language Models to Simplify Radiology Reports: a comparative analysis of ChatGPT3.5, ChatGPT4.0, Google Bard, and Microsoft Bing. *medRxiv* 2023;06:23290786. <https://doi.org/10.1101/2023.06.04.23290786>
25. Ekmekci E, Durmazpınar PM. Evaluation of different artificial intelligence applications in responding to regenerative endodontic procedures. *BMC Oral Health* 2025;25(1):53. <https://doi.org/10.1186/s12903-025-05424-5>
26. Jalali P, Mohammad-Rahimi H, Wang FM, Sohrabniya F, AmirHossein Ourang S, Tian Y, et al. Performance of 7 Artificial Intelligence Chatbots on Board-style Endodontic Questions. *J Endod* 2025;51(10):1413-1419. <https://doi.org/10.1016/j.joen.2025.06.014>
27. Taymour N, Fouda SM, Abdelrahman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. *J Prosthet Dent* 2025;S0022-3913(24)00833-3. <https://doi.org/10.1016/j.prosdent.2024.12.016>
28. Tokgöz Kaplan T, Cankar M. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini. *Dent Traumatol* 2025;41:178-86. <https://doi.org/10.1111/edt.12999>
29. Alsayed AA, Aldajani MB, Aljohani MH, Alamri H, Alwadi MA, Alshammari BZ, et al. Assessing the quality of AI information from ChatGPT regarding oral surgery, preventive dentistry, and oral cancer: An exploration study. *Saudi Dent J* 2024;36:1483-9. <https://doi.org/10.1016/j.sdentj.2024.09.009>
30. Freire Y, Santamaria Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *Journal of Prosthetic Dentistry* 2024;131:659.e1-659.e6. <https://doi.org/10.1016/j.prosdent.2024.01.018>
31. Fujimoto M, Kuroda H, Katayama T, Yamaguchi A, Katagiri N, Kagawa K, et al. Evaluating Large Language Models in Dental Anesthesiology: A Comparative Analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese Dental Society of Anesthesiology Board Certification Exam. *Cureus* 2024;16(9):e70302. <https://doi.org/10.7759/cureus.70302>
32. Künzle P, Paris S. Performance of large language artificial intelligence models on solving restorative dentistry and endodontics student assessments. *Clin Oral Investig* 2024;28(11):575. <https://doi.org/10.1007/s00784-024-05968-w>
33. Azadi A, Gorgjinejad F, Mohammad-Rahimi H, Tabrizi R, Alam M, Golkar M. Evaluation of AI-generated responses by different artificial intelligence chatbots to the clinical decision-making case-based questions in oral and maxillofacial surgery. *Oral Surg Oral Med Oral Pathol Oral Radiol* 2024;137:587-93. <https://doi.org/10.1016/j.oooo.2024.02.018>
34. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res* 2023;25: e47479. <https://doi.org/10.2196/47479>

35. Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. *Int Dent Res*. 2025;15(3):123-135. <https://doi.org/10.5577/intdentres.662>
36. İlter Er Ö, Adıgüzel Ö, Taşyürek M, Ortaç H. Comparison of ChatGPT-5, Gemini 2.5 Pro, and DeepSeek-V3.1 chatbot responses to dental avulsion injuries. *Health Sci Mon*. 2025;3:e251201. <https://doi.org/10.5577/hesmon.e251201>
37. Segura-Egea J, Martín-González J, Jiménez-Sánchez M, Crespo-Gallardo I, Saúco-Márquez J, Velasco-Ortega E. Worldwide pattern of antibiotic prescription in endodontic infections. *Int Dent J*. 2017;67(4):197-205. <https://doi.org/10.1111/idj.12287>
38. Kandemir S, Ergül N. Grievances in cases using antibiotics due to orodental problems and assessment of the need for antibiotics. *Int Dent J* 2000;50:73-7. <https://doi.org/10.1002/j.1875-595X.2000.tb00802.x>
39. Muller D, Ouanounou A. Antimicrobials in Endodontics: A Review of the Indications and Contraindications in the Battle Against Antimicrobial Resistance. *J Calif Dent Assoc* 2025;53:2492662. <https://doi.org/10.1080/19424396.2025.2492662>
40. Bansal R, Jain A, Goyal M, Singh T, Sood H, Malviya HS. Antibiotic abuse during endodontic treatment: A contributing factor to antibiotic resistance. *J Family Med Prim Care* 2019; 8:3518. https://doi.org/10.4103/jfmprc.jfmprc_768_19
41. Fareed M, Fatima M, Uddin J, Ahmed A, Sattar MA. A systematic review of ethical considerations of large language models in healthcare and medicine. *Front Digit Health*. 2025;7:1653631. <https://doi.org/10.3389/fdgth.2025.1653631>
42. Jagtap, R. Dentistry-Overview of Artificial and Augmented Intelligence Uses in Dentistry; SCDI White Paper No. 1106; American Dental Association: Chicago, IL, USA, 2022.
43. Keskinbora KH. Medical ethics considerations on artificial intelligence. *Journal of Clinical Neuroscience* 2019;64:277-82. <https://doi.org/10.1016/j.jocn.2019.03.001>
44. Boudierhem R. Shaping the future of AI in healthcare through ethics and governance. *Humanit Soc Sci Commun* 2024;11:1-12. <https://doi.org/10.1057/s41599-024-02894-w>
45. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. *J Dent Res*; 2020;99:769-74. <https://doi.org/10.1177/0022034520915714>
46. Cangül S, Adıgüzel Ö, Taşyürek M, Tunç T, Ortaç H. Performance of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 large language models in response to questions about ozone in dentistry. *Int Dent Res*. 2025;15(3):165-174. <https://doi.org/10.5577/intdentres.701>
47. Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ digital medicine*. 2024;7(1):41. <https://doi.org/10.1038/s41746-024-01029-4>