European Annals of Dental Sciences
Official Publication of Ankara University, Faculty of Dentistry

EADS, 2025, 52 (1), 10–16

e-ISSN : 2757-6744
Article Received/Accepted : November, 4 2024 / March, 18 2025
doi : 10.52037/eads.2025.0002

ORIGINAL RESEARCH ARTICLE

Evaluation of the Competency of Large Language Models GPT-4o and Claude 3.5 Sonnet in Endodontic Emergencies

Merve Sarı [1ROR], * and Pelin Tüfenkçi [1ROR]

[1]Department of Endodontics, Faculty of Dentistry, Hatay Mustafa Kemal University, Hatay, Türkiye

*Corresponding Author; mervess94.ms@gmail.com

Abstract

Purpose: This study aimed to evaluate the accuracy and comprehensiveness of the responses generated by GPT-4o and Claude-3.5 Sonnet to the most frequently asked questions about endodontic emergencies.
Materials and Methods: The most frequently asked questions about nine different topics (inferior alveolar nerve block, sodium hypochlorite accidents, aspiration of dental materials, separated instruments, perforation, transportation, $Ca(OH)_2$ extrusion, root filling, and flare-up) in endodontics were generated by GPT 3.5. Each question was asked to both GPT-4o and Claude 3.5 Sonnet. Two authors independently scored the responses. Accuracy and comprehensiveness were assessed for each question using Likert scales. The data were statistically analyzed using the Mann–Whitney U test and the Kruskal–Wallis test. The significance level was set at 0.05.
Results: Responses generated by both GPT-4o and Claude 3.5 Sonnet to a total of 81 open-ended questions were evaluated. The two models yielded similar results in terms of accuracy and comprehensiveness ($p > 0.05$). The topics of root filling, perforation, and flare-up have the lowest accuracy scores, and root filling and separated instruments have the lowest comprehensiveness scores for GPT-4o ($p < 0.05$). The accuracy of Claude 3.5's responses did not show significant differences between the topics ($p > 0.05$); however, separated instruments had the lowest comprehensiveness scores ($p < 0.05$).
Conclusions: The accuracy and comprehensiveness scores of GPT-4 and Claude 3.5 Sonnet are statistically similar. Despite the high levels of accuracy and comprehensiveness shown by GPT-4o and Claude 3.5 Sonnet, they do not yet have the effect of replacing the operator in endodontic procedures.

Keywords: Artificial intelligence; Claude 3.5; Endodontic emergencies; GPT-4o

Introduction

Artificial intelligence (AI) can build systems to perform tasks that require human intelligence, such as language understanding, reasoning, learning, and problem solving. In a short period of time, its potential applications in various fields, including dentistry, have been investigated and are the subject of numerous articles. [1] The applications of AI in endodontics have also been investigated, and studies have reported varying levels of effectiveness in detecting radiographic features such as root/canal morphology, [2] periapical lesions, [3] minor apical foramen, [4] and vertical root fractures. [5]

Natural language processing (NLP) is a subfield of AI that aims to enable computers to understand and interpret human language in a way that is both meaningful and useful. Large language models (LLMs) are neural network models trained on vast amounts of text data that demonstrate high performance in NLP-related tasks. [6,7] GPT-3 is the first large-scale language model developed by OpenAI, followed by GPT-3.5, GPT-4, GPT-4 Turbo, and an improved version, GPT-4o, released on May 13, 2024. [8] Claude-3.5 Sonnet, developed by Anthropic and released on June 20, 2024, is a large language model that competes with OpenAI's GPT-4o. According to Anthropic's model description, it was trained on data up to April 2024. [9] However, the effectiveness of LLMs in the field of endodontics has been evaluated in a few studies, and to the best of our knowledge, GPT-4o and Claude-3.5 Sonnet have not yet been studied.

Endodontic emergencies can be challenging for both the clinician and the patient. Clinicians need to know how various complications may arise, how to manage them predictably when they do occur, and what they need to do to prevent such complications. [10] Understanding how a particular complication may affect treatment prognosis is important, as the same complication can lead to dif-

How to cite: Sari M, Tufenkci P. Evaluation of the Competency of Large Language Models GPT-4o and Claude 3.5 Sonnet in Endodontic Emergencies. EADS. 2025;52(1): 10-16

ferent outcomes in different cases. During the treatment process, it is necessary to evaluate the clinical factors that determine the prognosis and to identify appropriate treatment strategies for the patient.[11] In these clinical situations where quick decision-making and stress management are crucial for the success of treatment, an AI-based model that can generate accurate and comprehensive responses can be beneficial for practitioners. Therefore, this study aims to evaluate the accuracy and comprehensiveness of the responses generated by GPT-4o and Claude-3.5 Sonnet to the most frequently asked questions about endodontic emergencies. The null hypothesis of the study is that there will be no difference in the evaluated parameters between the two different LLMs.

Material and Methods

Ethical approval was not provided for this study as no individual or patient details were included.

The authors identified nine different topics related to endodontic emergencies. The GPT 3.5 program was used to generate the ten most frequently asked questions about each of the nine topics. These topics include inferior alveolar nerve block, sodium hypochlorite accidents, aspiration of dental materials, separated instruments, perforation, transportation, $Ca(OH)_2$ extrusion, root filling, and flare-up. Overlapping and irrelevant questions were removed (e.g., What are the potential reasons for inferior alveolar nerve block failure in endodontics? How can the success rate of inferior alveolar nerve blocks be improved to prevent failure in endodontics?) All the questions are available in Table 1.

Each question was directed to two LLMs (GPT-4o and Claude 3.5 Sonnet). The main application programming interface of each LLM was used (GPT-4o available at: https://chatgpt.com/; Claude 3.5 Sonnet available at: https://claude.ai/). It has been reported that LLMs may provide different responses when the same question is asked again or at different time points.[12] Therefore, all the questions were asked only once. To ensure consistency and standardization, each question was directed to both GPT 4o and Claude 3.5 at the same time, respectively. Additionally, the response times of LLMs were recorded. Between July 8, 2024, and July 16, 2024, a different topic was practiced each day. A new chat was created for each question to prevent any influence from previous responses. The questions were asked by an independent researcher using the same computer throughout the study. All the responses were saved as a Word file (Microsoft, Redmond, Washington, USA).

After the LLMs generated responses, two authors independently scored the responses. Accuracy and comprehensiveness were assessed for each question using Likert scales. The authors were blinded to which bot generated the evaluated response. The questions were randomly evaluated. Randomization was performed using an online software (www.randomizer.org) by an independent researcher who asked the questions.

The Accuracy Likert Scale: 1) Completely incorrect; 2) More incorrect than correct; 3) Approximately equal correct and incorrect; 4) More correct than incorrect; 5) Completely correct

The Comprehensiveness Likert Scale:1) Not adequate; 2) Somewhat adequate; 3) Adequate; 4) Very adequate; 5) Extremely adequate

Recent literature and guidelines covering widely accepted treatment approaches were used as references in the evaluation of the responses. When inconsistencies arose in the evaluations, a third endodontic specialist was consulted to reach a consensus. Consequently, a single scoring chart was prepared for each LLM's responses for statistical analysis.

Statistical Analysis

IBM SPSS Statistics (SPSS Inc., Chicago, IL, USA; Version 22.0) software was used to analyze the collected data. The data were nonnormally distributed according to the Shapiro–Wilk test. Therefore, nonparametric tests were performed. The accuracy and comprehensiveness scores of LLMs and their response times, were compared using the Mann–Whitney U test.

The accuracy and comprehensiveness scores of each LLM's responses to different topics were evaluated using the Kruskal–Wallis test. The statistical significance level was set at 0.05.

Results

Responses generated by both GPT-4o and Claude 3.5 Sonnet to a total of 81 open-ended questions on nine different topics related to endodontic emergencies were evaluated in terms of accuracy, comprehensiveness, and response generation time. The two models yielded similar results in terms of correct information transfer and comprehensiveness ($p > 0.05$), and high scores were obtained. The time required to generate responses was shorter for Claude 3.5 Sonnet than for GPT-4o. ($p < 0.05$) (Table 2). In the intragroup evaluation of the LLMs' responses, a statistically significant difference was observed in the accuracy and comprehensiveness scores for GPT-4o ($p < 0.05$). The accuracy of responses to the topic of root filling was statistically lower than that for the topics of inferior alveolar nerve block, sodium hypochlorite accidents, aspiration, and transportation. The accuracy of responses to the topic of perforation was statistically lower than that for the topics of inferior alveolar nerve block and transportation. The accuracy of responses to the topic of flare-up was statistically lower than that of the topic of transportation ($p < 0.05$). The comprehensiveness of the responses given to the root filling and separated instruments topics is statistically lower than that given to the transportation topic ($p < 0.05$) (Table 3). The accuracy of the responses given by Claude 3.5 Sonnet did not show a statistically significant difference between topics ($p > 0.05$). The comprehensiveness of the responses given to the separated instruments topic is statistically lower than that given to the transportation and flare-up topics ($p < 0.05$) (Table 4).

Discussion

Whether LLMs can be used as auxiliary tools for establishing diagnosis[13] and treatment protocols[14] or for education[15] in healthcare applications has been the subject of many recent studies. However, incorrect or inadequate answers[15] and the citing of nonexistent or erroneous sources[16] are concerning. Ramezanzade et al.[17] reported that although the reported accuracy metrics of AI-based models for the detection of radiographic features in endodontic treatments seem promising, most of the articles present methodological biases. Currently, there is a consensus that AI has the potential to facilitate healthcare practice and education, but more extensive research is needed to overcome its limitations.[18,19] Therefore, this study aimed to evaluate the accuracy, comprehensiveness, and response generation time of two different LLMs in the subject of endodontic emergencies whose clinical management requires knowledge, attention, and quickness.

In this study, the accuracy and comprehensiveness scores of the responses did not show a significant difference between the two LLMs. In the intragroup evaluation of the LLMs, the accuracy and comprehensiveness of GPT-4o responses showed significant differences between different topics. The topics of root filling, perforation, and flare-up have the lowest accuracy scores. The topics of root filling and separated instruments have the lowest comprehensiveness scores. The accuracy of Claude 3.5's responses did not show significant differences between different topics; however, the

Table 1. Questions

1) Inferior alveolar nerve block	
1	What are the potential reasons for inferior alveolar nerve block failure in endodontics?
2	What are the signs and symptoms of inferior alveolar nerve block failure for patients?
3	How is inferior alveolar nerve block failure managed in endodontics?
4	Are there any alternative anesthesia techniques that can be used if an inferior alveolar nerve block fails?
5	How to prevent trismus after inferior alveolar nerve block?
6	How do you treat trismus after an inferior alveolar nerve block?
7	How long does facial palsy typically last after an inferior alveolar nerve block?
8	Are there any long-term consequences of facial palsy after an inferior alveolar nerve block?
9	What treatment options are available for facial palsy after an inferior alveolar nerve block?
10	How can I prevent facial palsy after an inferior alveolar nerve block in the future?
2) Sodium hypochlorite accidents	
1	What are the potential risks and complications associated with the injection of sodium hypochlorite in endodontics?
2	Are there specific safety measures that should be followed when using sodium hypochlorite in endodontics?
3	What causes sodium hypochlorite extrusion during endodontic procedures?
4	How is sodium hypochlorite extrusion managed in endodontics?
5	What are the signs and symptoms of sodium hypochlorite extrusion in patients?
6	How can sodium hypochlorite extrusion be prevented during endodontic procedures?
7	Are there any alternative irrigation solutions that can be used to minimize the risk of sodium hypochlorite extrusion in endodontics?
8	Can sodium hypochlorite accidents lead to legal issues for dentists?
9	What concentration of sodium hypochlorite is considered safe for endodontic use?
3) Aspiration	
1	What are the potential risks and complications associated with aspiration during endodontic procedures?
2	How common is aspiration during endodontic treatments?
3	What are the signs and symptoms of aspiration for patients in endodontics?
4	How can aspiration be prevented in endodontics?
5	What should be done in the event of aspiration occurring during an endodontic procedure?
6	Can certain patient conditions increase the risk of aspiration during dental procedures?
7	What are the legal and ethical considerations if a patient aspirates an object during an endodontic procedure?
8	How can dental professionals ensure they are trained and prepared to handle aspiration incidents during endodontic procedures?
4) Separated instruments	
1	What causes instrument separation during root canal treatment?
2	How can instrument separation be prevented during endodontic procedures?
3	What are the potential consequences of instrument separation in root canal treatment?
4	How is instrument separation managed or treated when it occurs during endodontic treatment?
5	Are certain instruments more prone to separation than others during endodontic procedures?
6	How common is instrument separation in endodontics?
7	What are the consequences of leaving a separated instrument in the root canal?
8	What are the different techniques for retrieving separated instruments from the root canal?
9	How long does it take to retrieve a separated instrument from the root canal?
5) Perforation	
1	What is a perforation in endodontics?
2	What are the potential causes of perforation during endodontic procedures?
3	Can a perforated tooth be saved?
4	What are the potential consequences of perforation on the success of endodontic treatment?
5	How is perforation diagnosed and confirmed during an endodontic procedure?
6	What are the treatment options for managing perforation in endodontics?
7	What are the best practices for preventing perforation during endodontic treatments?
8	Are there any specific tools or materials that can help repair perforations in endodontics?
9	What are the challenges associated with treating perforations in different areas of the tooth?
6) Transportation	
1	What is root canal transportation and how does it occur during endodontic procedures?
2	What is the most important reason for canal transportation in endodontics?
3	What are the signs and symptoms that may indicate canal transportation has occurred?
4	What are the treatment options for managing and correcting canal transportation during endodontic procedures?
5	How can dental professionals improve their techniques and skills to reduce the likelihood of canal transportation?
6	What are the potential long-term consequences of untreated canal transportation in endodontic cases?
7	What are the key steps that should be followed to manage apical transportation in endodontics?
8	What is the success rate of endodontic treatment of vital and nonvital teeth with apical transportation?
7) $Ca(OH)_2$ extrusion	
1	What is calcium hydroxide and why is it used in endodontics?
2	What are the potential consequences of calcium hydroxide extrusion in endodontics?
3	How common is calcium hydroxide extrusion during root canal treatments?
4	What are the signs and symptoms of calcium hydroxide extrusion for patients?
5	How is calcium hydroxide extrusion diagnosed and managed by dental professionals?
6	Are there any preventive measures that can be taken to minimize the risk of calcium hydroxide extrusion?
7	What are the challenges associated with removing extruded calcium hydroxide from the periapical region?
8	How can dental professionals improve their techniques and skills to reduce the likelihood of calcium hydroxide extrusion in endodontic procedures?
9	What are the legal and ethical considerations if calcium hydroxide extrusion occurs?

8) Root filling	
1	What is underfilling and overfilling in the context of endodontics?
2	How can I prevent extrusion of filling material beyond the apex?
3	What are the best techniques to avoid voids in the obturation?
4	How do I manage a situation where the root filling material has been overextended?
5	How can I ensure complete obturation in curved canals?
6	How can I minimize the risk of root fracture during lateral condensation?
7	What are the pros and cons of different obturation techniques in endodontics?
8	How do I address inadequate apical seal after obturation?
9	What should I do if I notice air bubbles in the obturation on the post-operative radiograph?
10	How can I ensure proper adaptation of the master cone to reduce the risk of overfilling or underfilling?
9) Flare-up	
1	What is an endodontic flare-up and what are the common signs and symptoms?
2	What are the potential causes of endodontic flare-ups during or after treatment?
3	How common are endodontic flare-ups in endodontic procedures?
4	What are the risk factors that may contribute to a higher likelihood of experiencing a flare-up in endodontics?
5	How can endodontic flare-ups be prevented in patients undergoing root canal treatment?
6	How is an endodontic flare-up managed and treated by dental professionals?
7	Are there any specific pain management strategies that can help relieve discomfort associated with endodontic flare-ups?
8	How can dental professionals communicate with patients about the possibility of a flare-up and ensure they are prepared for any potential complications post-treatment?
9	Is an endodontic flare-up a sign of treatment failure?

Table 2. Accuracy, comprehensiveness, and response time, generated with the GPT-4o and Claude 3.5 Sonnet models.

	GPT-4o	Claude 3.5 Sonnet	Z	95% CI	P value
	Mean ± SD	Mean ± SD			
Accuracy	4.64 ± 0.53	4.61±0.6	-.115	.89-.90	0.908*
Comprehensiveness	3.9 ± 0.7	3.95±0.81	-.596	.56-.57	0.551*
Response time	25.08 ± 11.22	11.01±3.77	-8.748	0	P<0.001*

The data represented as Mean±SD. CI: Confidence Interval, SD: standard deviation, *Mann-Whitney U test. $p < 0.05$ indicates statistical significance.

Table 3. Accuracy and comprehensiveness scores for GPT-4o generated answers to questions on key topics in endodontics.

		Accuracy	Comprehensiveness
Inferior alveolar nerve block	Median (min-max)	5 (4–5) ^{CD}	4 (3–5) ^{AB}
	Mean±SD	4.9 ± 0.32	4.2 ± 0.63
Sodium hypochlorite accidents	Median (min-max)	5 (4–5) ^{BCD}	4 (4–5) ^{AB}
	Mean±SD	4.78 ± 0.44	4.22 ± 0.44
Aspiration	Median (min-max)	5 (4–5) ^{BCD}	4 (4–5) ^{AB}
	Mean±SD	4.88 ± 0.35	4.38 ± 0.52
Separated instruments	Median (min-max)	5 (4–5) ^{ABCD}	3 (3–4) ^{A}
	Mean±SD	4.67 ± 0.5	3.44 ± 0.53
Perforation	Median (min-max)	4 (4–5) ^{AB}	4 (2–4) ^{AB}
	Mean±SD	4.44 ± 0.53	3.67 ± 0.71
Transportation	Median (min-max)	5 (5–5) ^{D}	4.5 (4–5) ^{B}
	Mean±SD	5 ± 0	4.5 ± 0.54
$Ca(OH)_2$ extrusion	Median (min-max)	5 (4–5) ^{ABCD}	4 (3–5) ^{AB}
	Mean±SD	4.56 ± 0.53	3.89 ± 0.6
Root filling	Median (min-max)	4 (3–5) ^{A}	3 (3–4) ^{A}
	Mean±SD	4.2 ± 0.63	3.4 ± 0.52
Flare-up	Median (min-max)	5 (3–5) ^{ABC}	4 (2–5) ^{AB}
	Mean±SD	4.44 ± 0.73	3.56 ± 0.88
	P value	0.023*	<0.001*

The data represented as Median (min-max) and Mean±SD. SD: standard deviation. *Kruskal Wallis test. $p < 0.05$ indicates statistical significance. Different superscript capital letters indicate statistically significant differences in the same column.

Table 4. Accuracy and comprehensiveness scores for Claude 3.5 Sonnet generated answers to questions on key topics in endodontics.

		Accuracy	Comprehensiveness
Inferior alveolar nerve block	Median (min-max)	5 (4–5)	4 (3–5) ^{AB}
	Mean±SD	4.78 ± 0.44	4.3 ± 0.68
Sodium hypochlorite accidents	Median (min-max)	5 (3–5)	4 (4–5) ^{AB}
	Mean±SD	4.56 ± 0.73	4.11 ± 0.33
Aspiration	Median (min-max)	5 (4–5)	4 (3–5)^{AB}
	Mean±SD	4.88 ± 0.35	4.13 ± 0.64
Separated instruments	Median (min-max)	5 (4–5)	3 (2–4) ^{A}
	Mean±SD	4.56 ± 0.53	3 ± 0.87
Perforation	Median (min-max)	4 (4–5)	4 (3–4) ^{AB}
	Mean±SD	4.44 ± 0.53	3.56 ± 0.53
Transportation	Median (min-max)	5 (3–5)	5 (3–5) ^{B}
	Mean±SD	4.75 ± 0.71	4.5 ± 0.76
$Ca(OH)_2$ extrusion	Median (min-max)	4 (4–5)	4 (3–5) ^{AB}
	Mean±SD	4.33 ± 0.5	4.11 ± 0.6
Root filling	Median (min-max)	5 (2–5)	3.5 (2–5) ^{AB}
	Mean±SD	4.4 ± 0.97	3.5 ± 0.85
Flare-up	Median (min-max)	5 (4–5)	5 (3–5) ^{B}
	Mean±SD	4.89 ± 0.33	4.44 ± 0.73
	P value	0.171*	<0.001*

The data represented as Median (min-max) and Mean±SD. SD: standard deviation. *Kruskal Wallis test. $p < 0.05$ indicates statistical significance. Different superscript capital letters indicate statistically significant differences in the same column.

topic of separated instruments had the lowest comprehensiveness scores. Separated instruments, root fillings, and flare-ups are topics where clinical scenarios can show high variability, which may be a reason for receiving less accurate and inadequate responses. This indicates the limitations of LLMs in risk assessment and treatment planning for certain situations. It should be noted that despite the high accuracy scores for both LLMs, they can provide irrelevant responses and commit crucial errors (Table 5).

Research on LLMs is not yet widespread in the endodontic literature. Suarez et al. [20] reported that 57.33% of GPT-4 responses to clinical endodontic questions were correct. In that study, questions were asked in a dichotomous format (only yes or no) to objectively measure the accuracy of the answers. Similarly, Ozden et al. [21] received 51% and 64% correct answers from GPT-3.5 and Google Bard, respectively, to dichotomous questions asked in the field of dental traumatology. LLMs are not specifically trained for healthcare applications. They use data from publicly available research articles, books, texts, and sources containing health information on the internet, [22] which may explain the different levels of effectiveness shown by the systems. Another disadvantage of LLMs is their current inability to filter out the latest developments effectively. Health sciences have a dynamic nature, and when LLMs fail to keep up with the most recent clinical guidelines, they can provide misleading guidance. [23] For example, GPT-4o reported in Question 2.7 that hydrogen peroxide could be used as an alternative to sodium hypochlorite (Table 5). Hydrogen peroxide has no place in current endodontic irrigation protocols and has not been routinely recommended for a long period of time.

Although there does not seem to be a significant difference in scores between the two LLMs, Claude 3.5 shows a more consistent performance across different topics and has a significantly shorter response generation time. There are no studies evaluating the effectiveness of these two LLMs in health sciences with which we can directly compare our results. However, a study evaluating vulnerability detection with three different prompts for GPT-4o and Claude 3.5 Sonnet reported that Claude 3.5 Sonnet outperformed GPT-4o. Similar to our study, Claude 3.5 Sonnet consistently showed high performance across all prompt types, while GPT-4o's effectiveness is reported to be variable. [24] In LiveBench, a platform for evaluating and comparing the performance of LLMs, Claude 3.5 Sonnet was rated as the best-performing model. [9] Differences in LLM responses can be attributed to the models' architectural design parameters, datasets used for training, optimization techniques and algorithms, response generation strategies (fine-tuning), and response evaluation techniques. [25] However, in this study, the two LLMs were compared on a text–based task. A study evaluating the phonological skills of LLMs reported that no single model consistently outperformed others in all tested tasks. [26] This indicates that in the future, performance in the required field should be considered when selecting LLMs.

In this study, the questions most frequently asked about endodontic emergencies were created using GPT-3.5. GPT-3.5 was thought to be more inclusive due to its free access, but questions created by students, specialty students, and general or specialist dentists could have been different and more detailed. Although the questions were not interfered with to limit the models' responses, and the given responses were evaluated for accuracy and comprehensiveness for two different models, another limitation of this study is that the consistency of the responses at different time points was not evaluated. Additionally, the accuracy and comprehensiveness of the responses were scored by only two endodontics specialists (with 5 and 15 years of experience); evaluation by a broader team could have been more valid. Although there are not yet sufficient articles on the use of AI technologies in endodontics, there is a widespread expectation that LLMs will benefit quickly and effectively. Considering the increasing use and ease of access to these models, it is important to continue studies on scientifically proven datasets with consensus to prevent potential factual errors, the spread of misinformation, and legal problems in medical applications, and to understand how to best utilize AI.

Conclusion

Despite the level of accuracy and comprehensiveness shown in this study, neither GPT-4o nor Claude 3.5 Sonnet yet has the effect of replacing the operator in endodontic procedures; clinicians should be aware of the limitations of LLMs and use them judiciously. However, integrating these technologies into the practice of endodontics and using human intelligence and AI in collaboration will certainly be beneficial, but it is clear that numerous studies are needed to use them as reliable sources.

Table 5. Inaccurate/irrelevant answers

2.1. (GPT-4o)	Ineffective Irrigation: If not used correctly, sodium hypochlorite may not effectively irrigate the entire root canal system, potentially leaving behind bacteria or necrotic tissue.
	Interaction with Other Materials: Sodium hypochlorite can react with other substances used in endodontics, such as chlorhexidine, forming a precipitate that can clog the root canal system or interfere with sealing.
	Effect on Dentin: Prolonged exposure to sodium hypochlorite can weaken the dentin, potentially leading to fractures.
2.1. (Claude 3.5 Sonnet)	Chlorine gas formation: If sodium hypochlorite is mixed with certain other chemicals, it can produce toxic chlorine gas.
2.7. (GPT-4o)	Hydrogen Peroxide (H_2O_2): Hydrogen peroxide has been used as an endodontic irrigant for its effervescent effect, which helps in dislodging debris and providing some antimicrobial action. It is often used in conjunction with other irrigants.
4.5. (Claude 3.5 Sonnet)	Smaller diameter files: Instruments with smaller diameters (e.g., #10, #15) are more susceptible to separation due to their reduced mass and flexibility.
4.6. (GPT-4o)	Rotary nickel-titanium (NiTi) instruments are more prone to separation than stainless steel hand files due to their flexibility and tendency to undergo cyclic fatigue.
5.2. (GPT-4o)	Over-instrumentation: Extending the instruments or filling materials beyond the confines of the root canal system can cause perforation at the apex or in the furcation area.
6.2. (Claude 3.5 Sonnet)	Avoiding canal transportation is crucial because: 1. It helps maintain the structural integrity of the tooth root.
	2. It reduces the risk of weakening the root structure or creating perforations.
	3. It allows for more predictable and efficient cleaning and shaping of the canal.
	4. It facilitates better obturation (filling) of the canal system.
	5. It minimizes the risk of leaving untreated areas of the canal, which could harbor bacteria and lead to treatment failure.
7.3. (Claude 3.5 Sonnet)	Detection methods: The reported frequency can depend on the method of detection. Extrusion visible on radiographs is less common than microscopic extrusion that might only be detectable in histological studies.
7.5. (Claude 3.5 Sonnet)	Diagnosis: Pulp vitality tests: To assess the status of the pulp and surrounding tissues
8.1. (GPT-4o)	Underfilling: Definition: Underfilling occurs when the root canal filling does not extend to the entire length of the prepared root canal. The filling material falls short of the root apex (the tip of the root).
8.2. (GPT-4o)	Selection of Filling Material: Sealers: Choose biocompatible sealers that have good flow properties but do not tend to extrude beyond the apex easily.
	Filling Technique: Vertical Condensation: This technique, such as warm vertical condensation or the continuous wave technique, can help achieve a dense fill without overextension.
8.2. (Claude 3.5 Sonnet)	Appropriate irrigation techniques: - Use side-vented needles - Keep the needle 1-2 mm short of working length - Apply gentle pressure during irrigation
	Employ warm vertical compaction: This technique allows for better control of filling material.
8.4. (Claude 3.5 Sonnet)	When dealing with overextended filling material in dentistry, here are some steps to manage the situation: 1. Remove excess: Carefully remove the overextended material using appropriate dental instruments like finishing burs, polishing discs, or interproximal strips.
	2. Reshape and contour: Adjust the filling to match the natural tooth shape and ensure proper occlusion.
	3. Check contacts: Verify that interproximal contacts are correct and not too tight or loose.
	4. Polish: Smooth the adjusted areas to prevent plaque accumulation and ensure patient comfort.
	5. Evaluate occlusion: Check the bite to ensure the filling doesn't interfere with normal function.
8.6. (GPT-4o)	Moisture Control: Ensure the canal is dry before lateral condensation to prevent the weakening of dentinal walls.
	Employ warm vertical compaction: This technique allows for better control of filling material.
8.10. (Claude 3.5 Sonnet)	Use of sealers: Apply a thin layer of sealer to improve adhesion and fill small gaps. Avoid excess sealer, which can lead to overfilling.

Ethical Approval

Ethical approval was not required for this study.

Acknowledgements

Not applicable.

Financial Support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions

Conceptualization : All Authors
Methodology : All Authors
Investigation : All Authors
Formal Analysis : M.S.
Writing – Original Draft : M.S.
Writing – Review and Editing : All Authors

Conflict of Interest

The authors deny any conflicts of interest related to this study.

Authors' ORCID(s)

M.S. 0000-0002-9432-3809
P.T. 0000-0001-9881-5395

References

1. Li ZQ, Wang XF, Liu JP. Publication Trends and Hot Spots of ChatGPT's Application in the Medicine. J Med Syst. 2024;48(1):52. doi:10.1007/s10916-024-02074-y.
2. Jeon SJ, Yun JP, Yeom HG, Shin WS, Lee JH, Jeong SH, et al. Deep-learning for predicting C-shaped canals in mandibular second molars on panoramic radiographs. Dentomaxillofac Radiol. 2021;50(5):20200513. doi:10.1259/dmfr.20200513.
3. Brignardello-Petersen R. Artificial intelligence system seems to be able to detect a high proportion of periapical lesions in cone-beam computed tomographic images. J Am Dent Assoc. 2020;151(9):e83. doi:10.1016/j.adaj.2020.04.006.
4. Saghiri MA, Garcia-Godoy F, Gutmann JL, Lotfi M, Asgar K. The reliability of artificial neural network in locating minor apical foramen: a cadaver study. J Endod. 2012;38(8):1130–4. doi:10.1016/j.joen.2012.05.004.
5. Fukuda M, Inamoto K, Shibata N, Ariji Y, Yanashita Y, Kutsuna S, et al. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. Oral Radiol. 2020;36(4):337–343. doi:10.1007/s11282-019-00409-x.
6. 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(5):2887–2893. doi:10.1007/s00464-024-10739-5.
7. Kanthavel R, Anathajothi K, Balamurugan S, Ganesh RK. Artificial Intelligent Techniques for Wireless Communication and Networking. John Wiley & Sons; 2022.
8. OpenAI. Hello GPT-4o [Web Page]; 2024. Available from: https://openai.com/index/hello-gpt-4o/.
9. LiveBench [Web Page]; 2024. Available from: :https://livebench.ai/.
10. Nouroloyouni A, Nazi Y, Mikaieli Xiavi H, Noorolouny S, Kuzekanani M, Plotino G, et al. Cone-Beam Computed Tomography Assessment of Prevalence of Procedural Errors in Maxillary Posterior Teeth. Biomed Res Int. 2023;2023:4439890. doi:10.1155/2023/4439890.
11. Johnsen I, Bårdsen A, Haug SR. Impact of Case Difficulty, Endodontic Mishaps, and Instrumentation Method on Endodontic Treatment Outcome and Quality of Life: A Four-Year Follow-up Study. J Endod. 2023;49(4):382–389. doi:10.1016/j.joen.2023.01.005.
12. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes Metab Syndr. 2023;17(4):102744. doi:10.1016/j.dsx.2023.102744.
13. Hirosawa T, Harada Y, Mizuta K, Sakamoto T, Tokumasu K, Shimizu T. Evaluating ChatGPT-4's Accuracy in Identifying Final Diagnoses Within Differential Diagnoses Compared With Those of Physicians: Experimental Study for Diagnostic Cases. JMIR Form Res. 2024;8:e59267. doi:10.2196/59267.
14. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. 2023;23(4):405–406. doi:10.1016/s1473-3099(23)00113-5.
15. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6). doi:10.3390/healthcare11060887.
16. Manohar N, Prasad SS. Use of ChatGPT in Academic Publishing: A Rare Case of Seronegative Systemic Lupus Erythematosus in a Patient With HIV Infection. Cureus. 2023;15(2):e34616. doi:10.7759/cureus.34616.
17. Ramezanzade S, Laurentiu T, Bakhshandah A, Ibragimov B, Kvist T, Bjørndal L. The efficiency of artificial intelligence methods for finding radiographic features in different endodontic treatments - a systematic review. Acta Odontol Scand. 2023;81(6):422–435. doi:10.1080/00016357.2022.2158929.
18. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Netw Open. 2023;6(11):e2343689. doi:10.1001/jamanetworkopen.2023.43689.
19. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023;25:e51580. doi:10.2196/51580.
20. 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(1):108–113. doi:10.1111/iej.13985.
21. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024. doi:10.1111/edt.12965.
22. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med. 2023;388(13):1233–1239. doi:10.1056/NEJMsr2214184.
23. Betzler BK, Chen H, Cheng CY, Lee CS, Ning G, Song SJ, et al. Large language models and their impact in ophthalmology. Lancet Digit Health. 2023;5(12):e917–e924. doi:10.1016/s2589-7500(23)00201-7.
24. Bae J, Kwon S, Myeong SJE. Enhancing Software Code Vulnerability Detection Using GPT-4o and Claude-3.5 Sonnet: A Study on Prompt Engineering Techniques. Electronics. 2024;13(13):2657. doi:10.3390/electronics13132657.
25. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. Front Comput Sci. 2024;18(6):186345. doi:10.1007/s11704-024-40231-1.
26. Suvarna A, Khandelwal H, Peng N. PhonologyBench: Evaluating Phonological Skills of Large Language Models. ACL. 2024:1–14. doi:10.18653/v1/2024.knowllm-1.1.