

Performance of five large language models in managing acute dental pain: A comprehensive analysis

Büşra Tosun,¹ Zeynep Öztürk²

¹Department of Prosthodontics, Bolu Abant İzzet Baysal University, Bolu, Türkiye

²Department of Pediatric Dentistry, Bolu Abant İzzet Baysal University, Bolu, Türkiye

Purpose: This study aimed to evaluate and compare five large language models (LLMs) used in the pharmacologic management of acute dental pain based on the following parameters: Comprehensiveness, scientific accuracy, clarity, relevance, and similarity of information they provide.

Methods: For this study, 20 open-ended questions were asked from five LLMs, namely ChatGPT-4.0, Gemini Advanced, Claude, Copilot, and Perplexity, and their responses were evaluated by two experts based on American Dental Association guidelines. Their scores ranged from 0 to 10, and the iThenticate program was used to assess the similarity indices. Statistical analyses included the Friedman and Dunn tests, with significance level set at $p < 0.05$.

Results: Claude and ChatGPT-4o scored the highest in terms of comprehensiveness, scientific accuracy, clarity, and relevance, while Copilot and Perplexity scored the lowest. Claude had the lowest similarity index ($3 \pm 5\%$), and ChatGPT-4o had the highest ($7 \pm 8\%$). Statistical analysis showed significant differences among the five LLMs ($p < 0.001$). While Claude, ChatGPT-4o, and Gemini Advanced performed in similar ways, they significantly outperformed Copilot and Perplexity.

Conclusion: According to the findings of this study, Claude and ChatGPT-4o provided the most accurate and comprehensive responses; however, LLMs cannot replace clinical guidelines. These findings highlight the potential of LLMs in supporting clinicians and underscore the scope for further improvement.

Keywords: Artificial intelligence; ChatGPT; Claude; dental pain; large language models.

Introduction

Large language models (LLMs) utilize artificial intelligence (AI) algorithms to produce natural and engaging content similar to that produced by humans. AI is a system used to perform complex tasks and processes and solve complex problems with the help of computers trained through the transfer of human thoughts to machines (1).

While LLMs can benefit patient care, research, and education in healthcare sector, they also pose risks such as misinformation dissemination, privacy breaches, and data bias (2,3). In recent years, AI has gained significant traction in dentistry, especially in answering questions in several areas such as dental trauma, maxillofacial surgery, and pediatric dentistry (4,5). LLMs have emerged as an effective tool to facilitate easier access to best practices in dentistry by pro-

Cite this article as: Tosun B, Öztürk Z. Performance of five large language models in managing acute dental pain:

A comprehensive analysis. Turk Endod J 2025;10:39-49.

Correspondence: Büşra Tosun. Department of Prosthodontics, Faculty of Dentistry, University of Bolu Abant İzzet Baysal, Bolu, Türkiye

Tel: +90 505 – 614 28 48 e-mail: dtbusra86@hotmail.com

Submitted: January 11, 2025 **Revised:** January 28, 2025 **Accepted:** February 04, 2025 **Published:** April 29, 2025

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence



viding dentists with vast information rapidly and ensuring minimal scope of human error (6). AI falls into the following two groups: Knowledge-based AI, which models human expertise, and data-driven AI, which learns from large datasets. In the field of dentistry, general interest in data-driven AI has increased due to the availability of extensive electronic data resources (7,8).

ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), and Copilot (Microsoft) are examples of popular large language models that have the potential to improve the online information environment. ChatGPT is a model developed by OpenAI and released in 2022. It is part of the generative pretraining transformer (GPT) series of language models and has been trained on large datasets (9). ChatGPT is a tool that is equipped with advanced technology and AI techniques, and, tailored to users' needs, its topic-relevant responses are quick and effective (10). It is widely used in academics as a highly effective tool for various tasks such as language translation, article summarization, and outlining (11).

Google's Gemini model was introduced as a step forward in AI. This multimodal large language model—capable of processing diverse data formats, such as texts, images, audios, and videos—demonstrates its versatility, making its use suitable in a wide range of fields (12). A study of Gemini's capabilities shows its competitive abilities against existing models such as ChatGPT-4 in terms of visual comprehension and reasoning (13).

Claude is a large language model developed by Anthropic. With its advanced capabilities in natural language processing, Claude can be used in a variety of tasks, such as text generation, data analysis, and code writing. A study assessed the radiological diagnostic capabilities of Claude 3 Opus and Claude 3.5 Sonnet models and confirmed their effectiveness in recognizing differential diagnoses (14).

Copilot, an AI-based assistive tool developed by Microsoft, has the potential to facilitate the execution of many tasks—from analyzing medical records to building health applications—that are relevant in the development of health software (15).

Like ChatGPT, Perplexity AI is built on OpenAI's GPT model. It is positioned as a tool that focuses on research and answering questions. The aforementioned is evident in the simpler structure of its answers and the less conversational nature of its responses than ChatGPT (16).

Typically, pain is categorized as either acute or chronic. Acute pain, often linked to tissue damage, is characterized by its short duration and temporary nature (17). In contrast, chronic pain may have no clear diagnosis of the cause and may persist long after the injury has healed. In

general, dental pain is acute and associated with procedures such as tooth extraction, root canal treatment, and periodontal cleaning. However, in some cases, pain may take longer to subside than that caused after more invasive procedures, such as bone surgery or surgical removal of impacted teeth. In addition, patients with chronic orofacial pain may require long-term pain management (18).

In general, in nonsurgical dental treatments, anesthesia is not required once restorative preparation is complete. Effective pain management remains crucial in surgical interventions and the care of patients following traumatic injuries. However, since these procedures are more invasive than conservative treatments, pain management is required not only during the procedure but also in the postoperative period (19). Education plays a crucial role in ensuring the effective management of pain management. Thus, evaluating how LLMs can contribute to providing dentists with up-to-date information on dental pain management remains crucial. For such an evaluation, the collection of existing guidelines can be used as a reference point to measure the performance of LLMs in this area (20,21).

Although studies on AI and health applications have increased in recent years, such studies in dentistry remain limited. While AI is viewed as a promising tool in dentistry, the potential impact of inaccurate or potentially harmful responses should not be ignored. Examining whether LLMs can provide accurate and reliable access to content related to dental pain management remains paramount. The aim of this study was to evaluate and compare the information provided by five LLMs (ChatGPT-4o, Gemini, Claude, Copilot, Perplexity) in terms of comprehensiveness, scientific accuracy, clarity, relevance, and similarity in the information they provide in the management of acute dental pain. For this study, the null hypothesis is that the answers produced by these models do not differ in terms of comprehensiveness, scientific accuracy, clarity, relevance, and similarity of information among LLMs, and the collection of guidelines is used as a standard against these LLMs.

Materials and Methods

For this study, a collaborative panel comprising experts from the American Dental Association's Institute for Science and Research, the University of Pittsburgh, and the University of Pennsylvania conducted systematic reviews and meta-analyses to offer evidence-based recommendations. These recommendations addressed the management of acute pain following simple and surgical tooth extractions and the management of temporary dental pain related to pulp and periapical diseases. This effort led to the de-

velopment of the following two guidelines: The Evidence-Based Clinical Practice Guideline for the Pharmacologic Management of Acute Dental Pain in Adolescents, Adults, and Older Adults (20) and the Evidence-Based Clinical Practice Guideline for the Pharmacologic Management of Acute Dental Pain in Children (21). These guidelines were used as the gold standard. Referring to the guidelines, 20 open-ended questions were created using appropriate terminology and required text-based responses.

These 20 questions are provided in Table 1. The questions were asked only once from each LLM by one of the authors (B.T.) in December 2024, and they were presented without further explanation or reformulation. The LLMs were the following chatbots that are very popular among AI programs: ChatGPT-4o, Gemini, Claude, Copilot, and Perplexity AI. All responses were recorded in a spreadsheet (Excel; Microsoft Corp.).

By modeling real-life scenarios in which dentists need

Table 1. Questions included in the study

Question number	Questions
1	Is an analgesic prescription necessary after all tooth extractions? In what situations is an analgesic prescribed after extractions?
2	In adolescents, adults, and elderly adults, which analgesic and in what dosage should be used as the first-line analgesic treatment for dental pain or acute tooth pain following tooth extraction?
3	In adolescents, adults, and elderly adults, which analgesic and in what dosage should be used when NSAIDs alone are insufficient for managing dental pain or acute tooth pain following tooth extraction?
4	In adolescents, adults, and elderly adults, which analgesic and in what dosage should be used when the combination of NSAIDs and acetaminophen is insufficient for managing dental pain or acute tooth pain following tooth extraction?
5	If NSAIDs are contraindicated, which analgesic and in what dosage should be used for acute pain following a simple tooth extraction in adolescents, adults, and elderly adults?
6	If NSAIDs are contraindicated, which analgesic and in what dosage should be used for managing dental pain or acute tooth pain after surgical tooth extraction in adolescents, adults, and elderly adults?
7	Should long-acting or short-acting local anesthetics be recommended for managing postsurgical dental pain in adolescents, adults, and elderly adults?
8	Should long-acting or short-acting local anesthetics be recommended for managing acute tooth pain in adolescents, adults, and elderly adults?
9	After simple and surgical extractions, which anesthetic and in what dosage should be administered via additional local anesthesia postprocedure, and by which injection technique?
10	In cases in which definitive dental treatment is not immediately possible, which local anesthetic and in what dosage should be used for the temporary management of tooth pain (symptomatic pulpitis [i.e., reversible or symptomatic irreversible pulpitis, with or without symptomatic apical periodontitis] or symptomatic apical periodontitis or pulp necrosis with acute apical abscess)?
11	What are the reasons opioids are not preferred for managing dental pain or acute tooth pain following tooth extractions in adolescents, adults, and elderly adults?
12	Which opioids, in what dosages, and for how long should be used for managing dental pain or acute tooth pain following tooth extractions in adolescents, adults, and elderly adults?
13	Why is the use of corticosteroid not recommended in dental pain management?
14	Which topical anesthetic and in what concentration should be used for managing acute dental pain?
15	What are the maximum daily doses of the preferred medications (Ibuprofen, Naproxen Sodium, and Acetaminophen) for temporary management of postoperative pain and toothache following simple and surgical extractions in adolescents, adults, and elderly adults?
16	At what ages are codeine and tramadol suitable for managing dental pain or acute tooth pain following tooth extractions, and what are their side effects?
17	Which analgesics and in what dosages should be used for managing dental pain or acute tooth pain in children following tooth extractions?
18	If NSAIDs are contraindicated, which analgesic should be recommended for postoperative pain management in children after simple or surgical extractions?
19	Why is the use of long-acting local anesthetics not recommended for pain management in children?
20	What should be considered when prescribing topical anesthetics for dental pain management in children?

quick information with a single question, this study aimed to examine the capacity of LLMs to provide accurate information and offer clarity. Limiting the interactions to a single question allowed for a more focused LLMs' assessment to provide direct, clear, and relevant answers to complex questions, without additional prompting. Each LLM's response was independently reviewed by a specialist (B.T., Z.Ö.) with 15 years of experience. The scores of the responses were calculated on a scale of 0–10 according to predetermined criteria for comprehensiveness, scientific accuracy, clarity, and relevance (22). The similarity index in the responses of the LLMs was also evaluated using the widely used iThenticate program, and similarity rates were recorded as percentages. To prevent any bias during the evaluation, each LLM was assigned a letter so that the evaluators did not know which response belonged to which LLM. The correct answer, considered the “gold standard” in terms of comprehensiveness, scientific accuracy, clarity, and relevance, was provided to the evaluators as a reference, and this answer was assigned a full score of 10/10.

The similarity index was employed to assess the resemblance of chatbot responses to source materials. Each response was individually analyzed using the iThenticate program (<http://www.ithenticate.com>). Similarity rates were expressed as percentages and categorized into the following four levels: 0–10% denoted high originality, 10–20% reflected acceptable similarity, 20–40% represented high similarity, and over 40% signified very high similarity (23).

The data were analyzed using version 4.4.1 of the R programming language. Compliance with normal distribution was examined using the Shapiro–Wilk test. The Friedman test was used to compare dependent and nonnormally distributed variables measured according to AI tools, and multiple comparisons were analyzed with Dunn's test. Interevaluator agreement was analyzed using the intraclass correlation coefficient. The median (minimum–maximum) was used to represent the quantitative data, and the significance level was set as $p < 0.05$.

Results

Using the five LLMs ChatGPT-4o, Gemini, Claude, Copilot, and Perplexity AI, responses to the 20 dental pain management–related questions were evaluated on a scale of 0–10 in terms of comprehensiveness, scientific accuracy, clarity, relevance, and similarity of information provided. Each response was evaluated by two evaluators, and each evaluator evaluated 100 responses. The correlation between the scores given by the two evaluators to the individual responses provided by ChatGPT-4o, Gemini, Claude, Copilot, and Perplexity AI was examined using the intraclass correlation coefficient (Table 2), and most correlations were generally good or excellent. These results indicated uniformity in the way the responses of the five LLMs were evaluated.

The pair of scores provided by the two evaluators for each response of an LLM was evaluated separately. Descriptive statistics for the scores given by the two evaluators to the responses provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity are presented in Table 3. Claude and ChatGPT-4o received the highest scores in terms of comprehensiveness, scientific accuracy, clarity, and relevance, while Copilot received the lowest scores. The lowest similarity index percentage was observed in Claude AI ($3 \pm 5\%$) and the highest in ChatGPT-4o ($7 \pm 8\%$). The similarity index percentages of the responses provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity are shown in Figure 1.

Since there were good or excellent correlations between the responses of the evaluators, an average score was calculated for the answers provided by the LLMs by averaging the scores given by the two evaluators. When the performance of each LLM was evaluated, the Friedman test found significant differences between the scores of the answers produced by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity and the average scores for comprehensiveness, scientific accuracy, clarity, relevance, similarity index, and overall score (Table 4).

Table 2. Interclass Correlation Coefficient (ICC) for scores given by 2 evaluators to answers provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity

	Comprehensiveness		Scientific accuracy		Clarity		Relevance	
	ICC (95% CI)	P	ICC (95% CI)	p	ICC (95% CI)	p	ICC (95% CI)	p
ChatGPT-4o	0.906 (0.778–0.962)	< 0.001	0.914 (0.797–0.965)	< 0.001	0.922 (0.814–0.968)	< 0.001	0.935 (0.843–0.974)	< 0.001
Claude	0.956 (0.892–0.982)	< 0.001	0.967 (0.92–0.987)	< 0.001	0.966 (0.915–0.986)	< 0.001	0.969 (0.923–0.987)	< 0.001
Copilot	0.944 (0.864–0.977)	< 0.001	0.923 (0.816–0.969)	< 0.001	0.887 (0.738–0.954)	< 0.001	0.915 (0.797–0.965)	< 0.001
Gemini Advanced	0.888 (0.74–0.954)	< 0.001	0.869 (0.699–0.946)	< 0.001	0.9 (0.766–0.959)	< 0.001	0.876 (0.713–0.949)	< 0.001
Perplexity	0.931 (0.835–0.972)	< 0.001	0.936 (0.845–0.974)	< 0.001	0.923 (0.817–0.969)	< 0.001	0.941 (0.857–0.976)	< 0.001

ICC (95% CI): Intra-class correlation coefficient (95% confidence interval).

Table 3. Descriptive statistics for scores given by 2 evaluators to answers provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity

	Evaluator 1					Evaluator 2				
	ChatGPT-4o	Claude	Copilot	Gemini Advanced	Perplexity	ChatGPT-4o	Claude	Copilot Advanced	Gemini	Perplexity
Comprehensiveness										
Mean	5.70	6.25	4.25	5.40	4.35	5.80	6.30	4.05	5.75	4.50
SD	1.56	1.86	1.68	1.31	1.76	2.04	2.20	1.99	1.80	2.24
Median	5.50	7.00	4.00	5.00	4.00	6.00	7.00	3.00	5.50	3.50
Min	3.00	3.00	2.00	3.00	2.00	2.00	2.00	1.00	2.00	2.00
Max	8.00	9.00	7.00	8.00	8.00	9.00	9.00	8.00	8.00	9.00
SEM	0.35	0.42	0.38	0.29	0.39	0.46	0.49	0.44	0.40	0.50
CoV	2.43	3.46	2.83	1.73	3.08	4.17	4.85	3.95	3.25	5.00
Scientific accuracy										
Mean	5.65	6.05	4.20	5.45	4.35	5.55	6.05	4.00	5.60	4.35
SD	1.69	2.11	1.58	1.36	1.76	2.09	2.28	1.95	1.79	2.25
Median	5.50	7.00	4.00	5.50	4.00	5.50	7.00	3.00	5.00	3.50
Min	3.00	2.00	2.00	3.00	2.00	2.00	1.00	1.00	2.00	1.00
Max	8.00	9.00	7.00	8.00	8.00	9.00	9.00	8.00	8.00	9.00
SEM	0.38	0.47	0.35	0.30	0.39	0.47	0.51	0.44	0.40	0.50
CoV	2.87	4.47	2.48	1.84	3.08	4.37	5.21	3.79	3.20	5.08
Clarity										
Mean	5.50	6.20	4.25	5.65	4.45	5.65	6.20	4.35	5.75	4.60
SD	1.79	2.04	1.52	1.42	1.70	1.98	2.24	2.03	1.77	2.09
Median	5.00	7.00	4.00	6.00	4.00	5.50	7.00	3.50	5.00	4.00
Min	3.00	2.00	2.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00
Max	8.00	9.00	7.00	8.00	8.00	9.00	9.00	8.00	8.00	9.00
SEM	0.40	0.46	0.34	0.32	0.38	0.44	0.50	0.45	0.40	0.47
CoV	3.21	4.17	2.30	2.03	2.89	3.92	5.01	4.13	3.15	4.36
Relevance										
Mean	5.70	6.30	4.30	5.70	4.40	5.90	6.80	4.80	5.90	4.75
SD	1.56	1.98	1.56	1.38	1.76	1.83	2.12	1.77	1.68	2.12
Median	6.00	7.00	4.00	6.00	4.00	6.00	8.00	4.00	6.00	4.50
Min	3.00	3.00	2.00	3.00	2.00	3.00	3.00	2.00	2.00	2.00
Max	8.00	9.00	7.00	8.00	8.00	9.00	9.00	8.00	8.00	9.00
SEM	0.35	0.44	0.35	0.31	0.39	0.41	0.47	0.39	0.38	0.48
CoV	2.43	3.91	2.43	1.91	3.10	3.36	4.48	3.12	2.83	4.51
Similarity index										
Mean	7	3	6	5	7	7	3	6	5	7
SD	8	5	7	6	7	8	5	7	6	7
Median	5	2	4	4	4	5	2	4	4	4
Min	0	0	0	0	0	0	0	0	0	0
Maxi	25	18	23	22	22	25	18	23	22	22
SEM	2	1	2	1	1	2	1	2	1	1
CoV	1	0	1	0	0	1	0	1	0	0

ICC (95% CI): Intra-class correlation coefficient (95% confidence interval).

A statistical difference was found among the LLMs in terms of comprehensiveness ($p < 0.001$). The highest scores were obtained in Claude, ChatGPT-4o, and Gemini Advanced, and no statistically significant difference was observed among the three AI programs. A significant dif-

ference between Copilot and Perplexity was also observed in terms of comprehensiveness, with the lowest scores between Perplexity and Copilot, and no statistically significant difference was observed between these two LLMs. In terms of scientific accuracy values, a statistically significant

Table 4. Descriptive statistics for average scores to answers provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity based on examined criteria including comprehensiveness, scientific accuracy, clarity, relevance, similarity index and overall score

	ChatGPT-4o	Claude	Copilot	Gemini Advanced	Perplexity	Test Statistics	p
Comprehensiveness	5.75 (2.5-8.5) ^a	7 (2.5-9) ^a	3.5 (1.5-7.5) ^b	5.25 (2.5-8) ^b	3.75 (2-8.5) ^b	28.444	< 0.001 ^x
Scientific accuracy	5.5 (2.5-8.5) ^a	6.75 (1.5-9) ^a	3.5 (1.5-7.5) ^b	5.5 (2.5-8) ^{ab}	3.75 (1.5-8.5) ^b	21.323	< 0.001 ^x
Clarity	5.25 (2.5-8.5) ^{ab}	6.75 (2-9) ^a	3.5 (2-7.5) ^b	5.5 (2.5-8) ^{ab}	3.75 (2.5-8.5) ^b	16.654	0.002 ^x
Relevance	6 (3-8.5) ^{ab}	7.5 (3-9) ^a	3.75 (2-7.5) ^{bc}	5.75 (2.5-8) ^{abc}	4 (2-8.5) ^c	23.409	< 0.001 ^x
Similarity index	5 (0-25)	2 (0-18)	4 (0-23)	4 (0-22)	4 (0-22)	4.973	0.290 ^x
Averall score	5.69 (2.63-8.5) ^a	7.13 (2.25-9) ^a	3.57 (1.75-7.5) ^b	5.51 (2.5-8) ^{ab}	3.88 (2-8.5) ^b	21.959	< 0.001 ^x

*Friedman test; ^{a-c} There is no difference between groups that have the same letter.

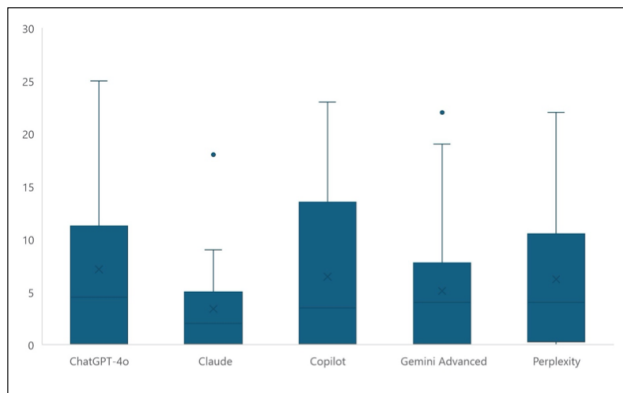


Fig. 1. Box plot of the similarity index percentages of the responses provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity.

difference was observed between five LLMs ($p < 0.001$). The highest scores were observed for Claude, ChatGPT-4o, and Gemini Advanced, while no statistically significant differences were observed among these three AI programs. The lowest scientific accuracy scores were obtained between Perplexity and Copilot. A statistically significant difference was observed among LLMs in terms values for clarity ($p < 0.001$). The highest scores were observed in Claude, ChatGPT-4o, and Gemini Advanced, and no statistically significant difference was observed among the three AI programs. The lowest openness scores were obtained between Perplexity and Copilot. The only statistically significant difference among the LLMs was observed among Claude, Perplexity, and Copilot. A statistically significant difference among LLMs was found in terms of relevance ($p < 0.001$). The highest scores were observed in Claude, ChatGPT-4o, and Gemini Advanced, and no statistically significant difference was observed among the three AI programs. While a significant difference was observed between ChatGPT-4o and Perplexity, a statistically significant difference was also observed between Claude and Perplexity and Copilot. In terms of similarity index, although the highest similarity index percentage was ob-

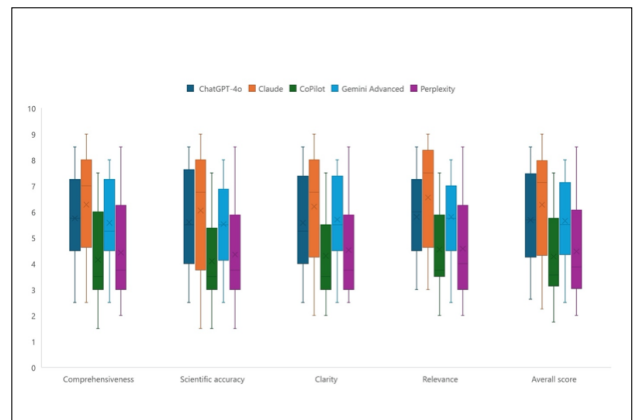


Fig. 2. Box plot of average scores for responses provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity.

tained in ChatGPT-4o and the lowest in the Claude AI model, no statistically significant difference was observed between five LLMs ($p > 0.05$). The similarity index scores ranged from 2 (0-18) to 5 (0-25), and as the similarity index values were below 10% for all LLMs, all models demonstrated high originality. When the overall scores of the LLMs were analyzed, the highest scores were observed for Claude 7.13 (2.25-9), ChatGPT-4o 5.69 (2.63-8.5), Gemini Advanced 5.51 (2.5-8), Perplexity 3.88 (2-8.5), and Copilot 3.57 (1.75-7.5). No statistically significant difference was observed among ChatGPT-4o, Claude AI, and Gemini Advanced. A statistically significant difference was observed between Claude and ChatGPT-4o and between Perplexity and Copilot (Table 4). The box plot of the average scores for responses provided by ChatGPT-4o, Claude, Copilot, Gemini Advanced, and Perplexity is given in Figure 2. As an example, the responses provided by the LLMs to question 10, "In cases where definitive dental treatment is not immediately possible, which local anesthetics and in what dosage should be used for the temporary management of tooth pain (symptomatic pulpitis [i.e., reversible or symptomatic irreversible pulpitis, with or without symptomatic apical periodontitis]

Table 5. The responses provided by large language models to a sample question (Question 10)

Large language models	Answers
ChatGPT-4o	1. Benzocaine, formulation: Available in 10% to 20% concentrations. Dosage: Apply a small amount directly to the painful area up to four times a day. 2. Lidocaine, formulation: Commonly used as a 2% solution with or without epinephrine. Dosage: For adults: Maximum recommended dose is 4.4 mg/kg, not exceeding 500 mg per appointment. For children: Generally, the maximum dose is 7 mg/kg. 3. Articaine, formulation: Typically used as a 4% solution with epinephrine (1:100,000 or 1:200,000). Dosage: For adults: Up to 7 mg/kg is generally acceptable. 4. Bupivacaine, formulation: Often used as a 0.5% solution with epinephrine (1:200,000). Dosage: Maximum of 1.3 mg/kg, typically not exceeding 90 mg per appointment.
Claude	1. Lidocaine, formulation: Typically used as a 2% solution with epinephrine (1:100,000), dosage: Adults: The maximum recommended dose is 4.4 mg/kg, not exceeding 500 mg per appointment, children: The maximum dose is usually 7 mg/kg. 2. Articaine, formulation: Typically used as a 4% solution with epinephrine (1:100,000), dosage: Adults: Generally acceptable up to 7 mg/kg Important warnings: Temporary solution until definitive treatment is possible, the patient's general health status and possible allergies must be evaluated, professional dental treatment must be planned as soon as possible, in cases of infection where antibiotics may be required, a physician must be consulted, dosage and application must be evaluated individually by the dentist.
Copilot	Lidocaine: 2-4 ml, bupivacaine: 0.5-1 ml, mepivacaine: 3-5 ml. These dosages can be adjusted according to the patient's pain level and treatment needs. The effects of local anaesthetics usually last 30-60 minutes.
Gemini Advanced	1. Lidocaine 2%: One of the most commonly used local anesthetics. It has a rapid onset of action and lasts approximately 1–2 hours. The maximum dose for adults is 4.4 mg/kg (approximately 300 mg), while the dose for children is calculated based on their weight. 2. Articaine 4%: A stronger and longer-acting anesthetic compared to lidocaine. Its effect lasts 2–4 hours. The maximum dose for adults is 7 mg/kg (approximately 500 mg), and for children, the dose is calculated based on their weight. 3. Bupivacaine 0.5%: A long-acting anesthetic. Its effect lasts 4–8 hours. The maximum dose for adults is 2 mg/kg (approximately 140 mg), and for children, the dose is calculated based on their weight.
Perplexity	1. Benzocaine, formulation: Available in concentrations of 10% to 20%, dosage: A small amount can be applied directly to the painful area up to four times a day. 2. Bupivacaine, formulation: Typically used as a 0.5% solution with epinephrine (1:200,000), dosage: The maximum dose is 1.3 mg/kg, usually not exceeding 90 mg per appointment
True answer	1. %2 lidokain + 1:100.000 epinefrin veya 2. %4 artikain + 1:100.000 epinefrin

or symptomatic apical periodontitis or pulp necrosis with acute apical abscess)?” directed to the chatbots, are presented in Table 5.

Discussion

This study aimed to examine and compare the accuracy of the responses of five major language models (Claude, ChatGPT-4o, Gemini Advanced, Copilot, and Perplexity) for questions related to acute dental pain management. The ideal responses were evaluated with reference to guidelines established by the American Dental Association Institute for Science and Research, based on systematic reviews and meta-analyses. The results revealed that Claude

received the highest mean score among the examined LLMs; however, no statistically significant difference was observed among Claude, ChatGPT-4o, and Gemini Advanced models. In contrast, statistically significant differences were observed among these three LLMs along with Copilot and Perplexity in terms of comprehensiveness, scientific accuracy, clarity, and relevance. The similarity index values of the LLMs exhibited high originality. Therefore, the null hypothesis stating that the models' responses are consistent with each other and with the standard guidelines was rejected.

AI, especially LLM chatbot applications, is of great interest in the medical field, thanks to its fast information access

and algorithmic decision-making capabilities (24). Studies in the literature have shown that these models have the potential to improve patient education in pre- and postoperative processes and provide recommendations consistent with guidelines (25-27). Dental pain management is central in clinical dentistry and plays a critical role in increasing patients' adherence to the treatment process. However, the lack of expertise in dental treatment in emergency services and the inadequacy of an effective referral system have led to a lack of definitive treatment and resulted in temporary solutions based on analgesics (28,29), further leading to negative outcomes, especially due to the prescription of opioids (30). LLMs have the potential to support clinical decision-making processes in this area by providing rapid recommendations based on guideline information.

Evidence-based guidelines developed by professional organizations, such as the American Dental Association, provide a fundamental reference point and aid clinical decision-making processes. However, continuously updating the guidelines and increasing their accessibility remain necessary. In this context, LLMs offer significant innovation in providing quick access to information and supporting the need for updating guides (2,31). In this study, Claude, ChatGPT-4o, and Gemini Advanced models demonstrated this potential by scoring high in terms of comprehensiveness, scientific accuracy, clarity, and relevance. However, the lower performance of Copilot and Perplexity on these parameters highlights the heterogeneity among the models and the impact of the training datasets on performance.

A study by Wu et al. (32) in the field of nephrology has shown that GPT-4 performs significantly better than Claude 2 (GPT-4: 73.3%; Claude 2: 54.4%). However, in the current study, Claude (version 3.5) and ChatGPT-4o exhibited similar performances. This situation may be because Claude 3.5 has a more advanced language processing capacity and an optimized model structure than Claude 2. In addition, such performance differences among the different versions of the models highlight the importance of context-specific model evaluations.

ChatGPT-4o, one of the most established and popular chatbots, has been exposed to more training data, which may have positively impacted its performance (33). However, the strong features of Claude and Gemini suggest that these models can also exhibit similar performances. For example, Claude's capacity to remember long conversations and Gemini's features of real-time internet browsing and visual addition indicate that these models can be effective in context-specific usage (34). In a study comparing the performances of Gemini and ChatGPT4,

Rana et al. (35) highlighted the promising potential of Gemini in providing accurate and comprehensive information but stated that these AI tools need further development to increase their accuracy and reliability, especially in academic environments that require expertise. In addition, in the current study, the performance of the responses provided by Copilot and Perplexity LLMs was the lowest. In their study, Huo et al. (36) compared chatbots such as ChatGPT-4, Copilot, Google Bard, and Perplexity AI to offer recommendations for the surgical treatment of gastroesophageal reflux disease. In line with the results of this study, they found that Copilot and Perplexity AI had lower accuracy rates compared to other models. Similarly, in their study on abdominoplasty, Lim et al. (37) evaluated the performance of LLM-connected chatbots, such as ChatGPT-3.5, Claude, Gemini, and Copilot, and Claude LLM emerged as the most reliable model. In the current study, the highest score among LLMs was that of Claude AI.

In terms of the similarity index, Claude's low rate ($3 \pm 5\%$) reflected the model's capacity to provide context-specific and creative responses. In contrast, ChatGPT-4o's higher similarity index ($7 \pm 8\%$) indicated a tendency to closely follow guiding information. This situation highlighted that models should be preferred according to their intended use; for example, ChatGPT-4o may be more suitable for conveying guided information, while Claude might be a better option in situations requiring creative solutions. Some studies show that language and structure imitation capabilities of LLMs have reached a level that makes it difficult for even experts to distinguish between fake and real research papers, especially in their summarization abilities (38,39). This situation emphasizes that the authors of scientific papers should carefully evaluate the potential risks and the benefits of the LLMs they choose to use. Clearly, LLMs still have significant limitations, particularly in their capacity to provide evidence-based information. These limitations need to be addressed.

In this study, the potential of LLMs to provide rapid access to information in the management of acute dental pain has been demonstrated; however, the limitations of these models have also been clearly stated. LLMs are quite sensitive to the content and details of a question. In some cases, when questions are not formulated clearly or in detail, it can be difficult for these models to understand the questions correctly and generate appropriate responses (40). In addition, in the field of medical and dental AI, the lack of scope of datasets used in training different LLMs can negatively affect the accuracy and consistency of the generated responses (41). In particular, the following two findings are emphasized: The responses of the models are

largely dependent on the quality of the datasets used in the training processes, and these datasets are generally sourced from the internet; however, they do not always provide evidence-based recommendations (42). This situation indicates that LLMs can only be used as complementary tools and cannot replace guides.

This study has a few limitations. Only five language models were evaluated in this study, and the performance of the other existing models was not examined. In addition, the accuracy and comprehensiveness of the responses are related to diversity and quality of the models' training datasets. Models supported by broader and higher-quality datasets are expected to be more effective, particularly in clinical decision support systems. The high performance exhibited by Claude, ChatGPT-4o, and Gemini Advanced highlights the potential of these models to be used as complementary tools in clinical applications, while the low performance of models such as Copilot and Perplexity draws attention to the heterogeneity among the models. This study sheds light on the potential role of these technologies in the healthcare sector by providing relevant data on the capabilities and limitations of LLMs.

Conclusion

This research has shown that LLMs can be effective tools for providing quick access to relevant information sought in dental pain management. Claude, ChatGPT-4o, and Gemini Advanced have demonstrated superior performance compared to other models in terms of parameters such as comprehensiveness and scientific accuracy. However, these models are not yet enough to replace guides. Hence, healthcare professionals must use these tools only complementarily, along with their critical thinking skills.

Future studies should focus on strategies such as dataset expansion and model optimization to enhance the accuracy of LLMs. In addition, improvements can be made in the integration of these models into clinical applications through user feedback and analyses based on a broader range of questions. These findings further clarify the role of this technology in the healthcare sector by highlighting the potential of LLMs to support clinical processes and improve patient care.

Authorship Contributions: Concept: B.T.; Design: B.T, Z.O.; Supervision: B.T.; Fundings: B.T, Z.O.; Materials: B.T, Z.O.; Data Collection and/or Processing: B.T., Z.O.; Analysis and/or Interpretation: B.T., Z.O.; Literature Review: B.T., Z.O.; Writer: B.T., Z.O.; Critical review: B.T.

Use of AI for Writing Assistance: Artificial intelligence (AI)–assisted technologies, such as Large Language Models (LLMs), chatbots, or image creators, were not used in the

preparation of the submitted work. The responses of following chatbots: ChatGPT-4o, Gemini, Claude, Copilot, and Perplexity AI; were recorded in a spreadsheet (Excel; Microsoft Corp.).

Source of Funding: None declared.

Conflict of Interest: None declared.

Informed consent: N/A

References

1. Tamkin A, Brundage M, Clark J, et al. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv* 2021; arXiv: 210202503.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023; 620: 172–80. [CrossRef]
3. Henderson P, Krass M, Zheng L, et al. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Adv Neural Inf Process Syst* 2022; 35: 29217–34.
4. Ozden I, Gokyar M, Ozden ME, et al. Assessment of artificial intelligence applications in responding to dental trauma. *Dent Traumatol* 2024; 40: 722–9. [CrossRef]
5. Rokhshad R, Zhang P, Mohammad-Rahimi H, et al. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. *J Dent* 2024; 144: 104938. [CrossRef]
6. Rewthamrongris P, Burapachee P, Trachoo V, et al. Accuracy of large language models for infective endocarditis prophylaxis in dental procedures. *Int Dent J* 2025; 75: 206–12. [CrossRef]
7. Steels L, López de Mantaras R. The Barcelona declaration for the proper development and usage of artificial intelligence in Europe. *AI Commun* 2018; 31: 485–94. [CrossRef]
8. Shan T, Tay F, Gu L. Application of artificial intelligence in dentistry. *J Dent Res* 2021; 100: 232–44. [CrossRef]
9. Liu H, Peng H, Song X, et al. Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interv* 2022; 27: 100495. [CrossRef]
10. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care* 2023; 27: 75. [CrossRef]
11. Hutson M. Could AI help you to write your next paper? *Nat* 2022; 611: 192–3. [CrossRef]
12. Pal A, Sankarasubbu M. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. *arXiv* 2024; arXiv: 240207023. [CrossRef]
13. Fu C, Zhang R, Wang Z, et al. A challenger to GPT-4V? Early explorations of gemini in visual expertise. *arXiv*

- 2023; arXiv: 231212436.
14. Kurokawa R, Ohizumi Y, Kanzawa J, et al. Diagnostic performances of Claude 3 Opus and Claude 3.5 Sonnet from patient history and key images in Radiology's "Diagnosis Please" cases. *Jpn J Radiol* 2025; 2025: 1–4. [CrossRef]
 15. Mitsea E, Drigas A, Skianis C. Digitally assisted mindfulness in training self-regulation skills for sustainable mental health: a systematic review. *Behav Sci* 2023; 13: 1008. [CrossRef]
 16. Deike M. Evaluating the performance of ChatGPT and Perplexity AI in Business Reference. *J Bus Finance Librarianship* 2024; 29: 125–54. [CrossRef]
 17. Merskey HE. Classification of chronic pain: Descriptions of chronic pain syndromes and definitions of pain terms. *Pain* 1986; 3: S1–226.
 18. Mehlisch DR. The efficacy of combination analgesic therapy in relieving dental pain. *J Am Dent Assoc* 2002; 133: 861–71. [CrossRef]
 19. Malamed SF. Pain management following dental trauma and surgical procedures. *Dent Traumatol* 2023; 39: 295–303. [CrossRef]
 20. Carrasco-Labra A, Polk DE, Urquhart O, et al. Evidence-based clinical practice guideline for the pharmacologic management of acute dental pain in adolescents, adults, and older adults: A report from the American Dental Association Science and Research Institute, the University of Pittsburgh, and the University of Pennsylvania. *J Am Dent Assoc* 2024; 155: 102–17.e109. [CrossRef]
 21. Carrasco-Labra A, Polk DE, Urquhart O, et al. Evidence-based clinical practice guideline for the pharmacologic management of acute dental pain in children: A report from the American Dental Association Science and Research Institute, the University of Pittsburgh School of Dental Medicine, and the Center for Integrative Global Oral Health at the University of Pennsylvania. *J Am Dent Assoc* 2023; 154: 814–25.e812. [CrossRef]
 22. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: A comparative study of ChatGPT, Google Bard, and Microsoft Bing. *Eur J Orthod* 2024; 2024: cjae017. [CrossRef]
 23. Yurdakurban E, Topsakal KG, Duran GS. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. *J Stomatol Oral Maxillofac Surg* 2024; 125: 101757. [CrossRef]
 24. Allauzen C, Mohri M, Roark B. Generalized algorithms for constructing statistical language models. *Proc Meet Assoc Comput Linguist* 2003; 1: 40–7. [CrossRef]
 25. Durairaj KK, Baker O, Bertossi D, et al. Artificial intelligence versus expert plastic surgeon: Comparative study shows ChatGPT "Wins" rhinoplasty consultations: Should we be worried? *Facial Plast Surg Aesthet Med* 2024; 26: 270–5. [CrossRef]
 26. Capelleras M, Soto-Galindo GA, Cruellas M, et al. ChatGPT and rhinoplasty recovery: An exploration of AI's role in postoperative guidance. *Facial Plast Surg* 2024; 40: 628–31. [CrossRef]
 27. Soto-Galindo GA, Capelleras M, Cruellas M, et al. Effectiveness of chatgpt in identifying and accurately guiding patients in rhinoplasty complications. *Facial Plast Surg* 2024; 40: 623–7. [CrossRef]
 28. Naavaal S, Kelekar U, Shah S. Opioid and nonopioid analgesic prescriptions for dental visits in the emergency department, 2015-2017 national hospital ambulatory medical care survey. *Prev Chronic Dis* 2021; 18: E58. [CrossRef]
 29. Roberts RM, Bohm MK, Bartoces MG, et al. Antibiotic and opioid prescribing for dental-related conditions in emergency departments: United States, 2012 through 2014. *J Am Dent Assoc* 2020; 151: 174–81.e171. [CrossRef]
 30. Schroeder AR, Dehghan M, Newman TB, et al. Association of opioid prescriptions from dental clinicians for US adolescents and young adults with subsequent opioid use and abuse. *JAMA Intern Med* 2019; 179: 145–52. [CrossRef]
 31. Omiye JA, Gui H, Rezaei SJ, et al. Large language models in medicine: The potentials and pitfalls: A narrative review. *Ann Intern Med* 2024; 177: 210–20. [CrossRef]
 32. Wu S, Koo M, Blum L, et al. A comparative study of open-source large language models, GPT-4 and Claude 2: Multiple-choice test taking in nephrology. *arXiv* 2023; arXiv: 230804709. [CrossRef]
 33. Sarkar S. AI Industry Analysis: 50 Most Visited AI Tools and Their 24B+ Traffic Behavior. *Writerbuddy*. <https://bit.ly/3TUVtBK>, 2023.
 34. Rebelo M. The best AI chatbots in 2024. Available at: <https://zapier.com/blog/best-ai-chatbot/>. Accessed Apr 17, 2025.
 35. Rana N, Katoch N. AI for biophysical phenomena: A comparative study of ChatGPT and Gemini in explaining liquid–liquid phase separation. *Appl Sci* 2024; 14: 5065. [CrossRef]
 36. Huo B, Calabrese E, Sylla P, et al. The performance of artificial intelligence large language model-linked chatbots in surgical decision-making for gastroesophageal reflux disease. *Surg Endosc* 2024; 38: 2320–30. [CrossRef]
 37. Lim B, Seth I, Cuomo R, et al. Can AI answer my questions? Utilizing artificial intelligence in the perioperative assessment for abdominoplasty patients. *Aesthetic Plast Surg* 2024; 48: 4712–24. [CrossRef]
 38. Májovský M, Černý M, Kasal M, et al. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res* 2023; 25: e46924. [CrossRef]
 39. Kim HJ, Yang JH, Chang D-G, et al. Assessing the repro-

- ducibility of the structured abstracts generated by ChatGPT and Bard compared to human-written abstracts in the field of spine surgery: Comparative analysis. *J Med Internet Res* 2024; 26: e52001. [CrossRef]
40. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. *Cureus* 2023;15: e42133. [CrossRef]
41. Roganović J, Radenković M, Miličić B. Responsible use of artificial intelligence in dentistry: Survey on dentists' and final-year undergraduates' perspectives. *Healthcare* 2023; 11: 1480. [CrossRef]
42. Li J, Dada A, Puladi B, et al. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed* 2024; 245: 108013. [CrossRef]