



Performance of large language models in providing patient- oriented information on gingival recession: a comparative study

Sultan Albeshri¹

Received: 11 August 2026 / Accepted: 20 September 2026
© The Author(s) 2026

Abstract

Patients are increasingly turning to large language models (LLMs) for health information, but the quality of patient-oriented information remains uncertain. This article evaluated the performance of three major LLMs in answering clinically relevant patient's questions related to gingival recession. A descriptive comparative study was carried out utilizing twenty standardized patient-centered questions related to gingival recession. Each question was submitted once to ChatGPT 5.5, Gemini 3.5, and Claude Sonnet 5, generating 60 responses. Five board-certified periodontists independently evaluated the responses using the validated Accuracy of Information (AOI) Index, a completeness scale, and expert preference rankings. Word count and response generation time were also recorded. Non-parametric tests and generalized estimating equations (GEE) were applied for statistical analysis, with statistical significance set at $p \leq 0.05$. No significant differences were found among the three LLMs in accuracy ($p=0.993$) or completeness ($p=0.972$). ChatGPT generated responses significantly faster (median: 5.82 s) and with fewer words (median: 232) than Gemini (8.46 s; 438 words) and Claude (14.61 s; 333.5 words) (both $p < 0.001$). Expert preference rankings differed significantly ($p=0.002$): Gemini and Claude were each ranked first in 38% of evaluations, compared with 24% for ChatGPT. For all three models, the highest accuracy was observed in Post-treatment Care. The three LLMs offered broadly comparable levels of accuracy and completeness when answering patient-oriented questions with respect to gingival recession. ChatGPT was faster and more concise, while Gemini and Claude were more often preferred by the expert evaluators. These LLMs may be useful as adjuncts to patient education, but they should complement rather than replace professional periodontal consultation.

Keywords Large language models · Chatbots · Artificial intelligence · Gingival recession · Dentistry

1 Introduction

Artificial intelligence (AI) improved the patient's ability to access healthcare information. (Hirani et al. 2024) Large language models (LLMs), including ChatGPT (OpenAI), Deep seek, Gemini (Google), and Claude (Anthropic), are globally available and can produce responses related to healthcare questions. (Huang et al. 2023; Wang and Zhang 2024) Distinct from traditional search engines, that require users to access information from different sources, LLMs provide direct answers in natural language. This feature has

made them increasingly attractive to patients seeking for quick medical and dental information. (Salehin et al. 2025; Sivaramakrishnan et al. 2025; Wang and Zhang 2024)

The increased use of AI chatbots facilitate the gathering of patient's information with respect to diseases etiology, treatment options, and follow-up care prior they see a healthcare professional. (Huang et al. 2023; Sivaramakrishnan et al. 2025; Tiwari et al. 2023) Although the LLMs has promising potential for patient education, concerns remain about the accuracy, completeness, consistency, and reliability of its responses, particularly when patients may use that information to guide clinical decisions. (Khan et al. 2023; Schwendicke et al. 2020; Zhang et al. 2025a, b)

In dentistry, LLMs application has grown among specialties, including, endodontics, orthodontics, oral maxillofacial surgery, prosthodontics, and periodontology. (Daraqel et al. 2024; Umer et al. 2024; Zhang et al. 2025a, b) Contemporary reviews have reported potential applications in

✉ Sultan Albeshri
salbeshri@ksu.edu.sa

¹ Department of Periodontics and Community Dentistry,
College of Dentistry, King Saud University, Riyadh
12372, Saudi Arabia

patient education, clinical decision-making, and the interpretation of clinical guidance. At the same time, these reviews emphasize that AI-generated information should not replace professional judgment because responses may contain factual errors, outdated information, or incomplete recommendations. (Umer et al. 2024; Zhang et al. 2025a, b) Performance can also vary among LLMs depending on the topic and clinical relevance, making repeated assessment important as these systems are frequently updated. (Sivaramakrishnan et al. 2025; Zhang et al. 2025a, b)

Gingival recession (GR) is a frequently observed clinical condition characterized by apical shift of the gingival margin pass the cemento-enamel junction, resulting in root surface exposure. (Cortellini and Bissada 2018) GR can increase dentin hypersensitivity, root caries, cervical abrasion, plaque retention, esthetic issues, and affect oral health-related quality of life. (Merijohn 2016) Prior periodontal consultation, patients frequently seek for information related to etiology, progression, treatment choices, expected outcomes, postoperative care, and long-term prognosis. As AI chatbots become another source of this information, the quality of their responses becomes relevant to patient understanding and treatment expectations.

Numerous studies have investigated the quality and accuracy of LLM responses to questions and inquiries asked by patients. (Daraqel et al. 2024; Sivaramakrishnan et al. 2025; Zhang et al. 2025a, b) Systematic reviews typically demonstrate acceptable level of clarity and precision, conclusions differ according to the clinical domain and assessment tools employed. (Albalawi et al. 2025; Ma 2025; Zhang et al. 2025a, b) In periodontology, a study by Babayiğit et al. evaluated how effectively ChatGPT addressed frequently asked questions. While the responses generally exhibited high accuracy and completeness, inquiries regarding gingival recession scored significantly lower in accuracy than those related to other periodontal diseases. (Babayiğit 2023) More recent comparisons of several LLMs have found broadly similar objective accuracy, while differences remain in presentation, comprehensiveness, readability, and user preference. (Sivaramakrishnan et al. 2025; Zhang et al. 2025a, b)

Despite this growing body of work, a comprehensive comparison of contemporary LLMs focused specifically on patient-oriented information about gingival recession is still lacking. Previous studies have also seldom considered several outcomes together, including expert preference, response generation time, response length, and performance across specific clinical domains. Therefore, this study compared ChatGPT, Gemini, and Claude with respect to accuracy, completeness, response generation time, response length, expert preference, and domain-specific performance

when answering standardized patient-centered questions about gingival recession.

2 Methods

2.1 Study design

A descriptive comparative design was used to assess responses generated by three artificial intelligence (AI) models: ChatGPT 5.5 (OpenAI), Gemini 3.5 (Google), and Claude Sonnet 5 (Anthropic). The study included 20 patient-oriented questions addressing gingival recession (GR). (Supplement 1) The questions were developed using patient education materials from the American Academy of Periodontology (AAP), including its patient FAQ database, as well as clinically relevant topics identified in the 2018 World Workshop consensus on mucogingival conditions (Cortellini and Bissada 2018; Tonetti et al. 2025). Questions addressing issues commonly discussed during periodontal consultations were also included, particularly those related to the causes, prevention, prognosis, and treatment of gingival recession.

The initial set of questions was reviewed by two board-certified periodontists, who discussed and modified the questions until a consensus was reached. The final questions were grouped into five clinical domains: (1) Assessment and Diagnosis (Q1–Q4); (2) Efficacy and Outcomes (Q5–Q8); (3) Complications and Management (Q9–Q12); (4) Post-treatment Care (Q13–Q16); and (5) Precautions and Clinical Considerations (Q17–Q20). Each question was entered separately into all three AI models, producing a total of 60 responses. To minimize the confounding effects of prompt variability, the analysis was strictly restricted to the initial response generated by each model.

All questions were submitted using the same computer with a stable high-speed internet connection. The AI models were accessed through free access mode using standardized browser setup (Google Chrome version 147) during the data-collection period of (May 2026). Each question was submitted independently in a new conversation, with previous conversation history and browser cookies cleared before each submission to minimize potential carryover effects from earlier interactions. For each response, the generation time and total word count were recorded.

Five independent expert evaluators were recruited using predefined eligibility criteria. All evaluators were required to be full-time Assistant or Associate Professors in Periodontics, Consultants in Periodontics, American Board-certified in Periodontics, and have at minimum 3 years of experience in the specialty. All five evaluators met these requirements. Each response was independently assessed

by all five evaluators, resulting in 300 evaluations for each primary outcome measure.

2.2 Outcome measures

Once the responses had been collected, the three answers to each question were provided to the expert evaluators for assessment. The evaluators were blinded to the identity of the AI model that generated each response. In addition, the order in which the responses were presented was randomized for each question to reduce the possibility of bias related to response sequence or knowledge of the AI model.

Each evaluator assessed the responses using two primary measures: (1) the Accuracy of Information (AOI) Index (Daraqel et al. 2024) and (2) a completeness score recorded on a visual analog scale ranging from 0 to 10. To assess each response, the AOI Index was utilized, an instrument previously validated for measuring the accuracy of generated information (Daraqel et al. 2024). The index consists of five domains: factual accuracy, corroboration with scientific evidence and guidelines, consistency, clarity and specificity, and relevance. Each domain was assigned a score from 0 to 2, giving a possible total score of 0–10. Higher scores indicated greater accuracy.

In addition to the AOI and completeness metrics, the evaluators performed a comparative review of the three responses for each item, assigning ranks from highest to lowest quality according to their aggregate preference and quality. This ranking was performed separately for each question.

2.3 Statistical analysis

Statistical analysis was completed using the Statistical Package for the Social Sciences (SPSS) software version 30 (Chicago, IL, USA). Descriptive statistics were summarized

as mean, standard deviation, median, interquartile range (IQR), frequency, and percentage. As the Shapiro–Wilk test indicated that the quantitative data were not normally distributed, non-parametric tests were used. The Kruskal–Wallis test was applied to compare quantitative variables across more than two groups, and the Mann–Whitney U test was used for pairwise (two-group) comparisons. The Chi-square test was used to compare frequencies between groups. Univariable Generalized Estimating Equations (GEE) were used to evaluate the association of the study variables with accuracy and completeness. A *p* value of ≤ 0.05 was considered statistically significant.

3 Results

The median accuracy of Gemini (8.40) was highest followed by ChatGPT (8.30) and Claude (8.00). ChatGPT showed highest median completeness (7.60) compared with Gemini (7.30) and Claude (7.20). The generation time for ChatGPT was least with a median generation time of 5.82 s, while Claude showed highest generation time of 14.61 s. The median word count for ChatGPT, Gemini, and Claude was 232, 438, and 333.5, respectively. (Table 1)

The response to each question was ranked by five expert evaluators. The response was graded with ranking range from 1 to 3. The proportion of 1st rank for ChatGPT, Gemini, and Claude was 24.0%, 38.0%, and 38.0% respectively. On the other hand, the proportion for 3rd rank was greatest in ChatGPT (49%) followed by Claude (27%) and Gemini (24%). (Table 2)

With respect to study variables, statistically significant difference was observed between the three AI models for only two variables, that is, generation time and number of words. ChatGPT took least time of 5.82 s followed by Gemini (8.46 s) and Claude (14.61 s). The ChatGPT generated

Table 1 Descriptive table of study variables across three AI models

AI Tool		Min.	Max.	Mean	SD	Median	IQR
ChatGPT	Accuracy	4	10	8.10	1.72	8.30	6.55–9.85
	Completeness	1	9	7.13	1.75	7.60	6.65–8.15
	Ranking	1	3	1.75	0.82	2.00	1.00–2.00
	Generation time	3.50	7.85	5.81	1.19	5.82	4.80–6.69
	No. of words	153	412	249.10	68.53	232.00	198.25–298.75
Gemini	Accuracy	5	10	8.21	1.45	8.40	7.10–9.15
	Completeness	1	8.80	7.07	1.61	7.30	7.00–8.00
	Ranking	1	3	2.14	0.78	2.00	2.00–3.00
	Generation time	6.92	9.88	8.54	0.87	8.46	7.82–9.36
	No. of words	285	581	440.45	74.10	438.00	407.00–486.00
Claude	Accuracy	7	10	8.21	1.14	8.00	7.25–9.60
	Completeness	3	9	7.21	1.40	7.20	6.25–8.00
	Ranking	1	3	2.11	0.80	2.00	1.00–3.00
	Generation time	11.91	17.50	14.57	1.50	14.61	13.23–15.64
	No. of words	266	422	329.90	39.25	333.50	303.00–349.25

Table 2 Distribution of ranking among three AI models

	Ranking Categories	N	%
ChatGPT	1st (Best)	24	24.0
	2nd	27	27.0
	3rd	49	49.0
Gemini	1st (Best)	38	38.0
	2nd	38	38.0
	3rd	24	24.0
Claude	1st (Best)	38	38.0
	2nd	35	35.0
	3rd	27	27.0

least words (232) and Gemini produced the most words (438). (Table 3)

Post-hoc analysis for variables generation time and number of words revealed statistically significant difference for all pairwise comparisons. ($p=0.001$). (Table 4) The median accuracy scores for all three AI models were highest for “Post-treatment Care” domain. The least median accuracy score for ChatGPT (7.3) and Gemini (7.7) was observed for “Assessment and Diagnosis” domain. However, for Claude, the least accuracy was recorded for “Efficacy and Outcomes” domain (7.3). Regarding completeness, ChatGPT showed highest median score for “Post-treatment Care” domain with 8.0 score. Gemini showed 7.5 median completeness score for all domains except “Assessment and Diagnosis” domain which showed lowest completeness score of 7.1. Claude showed highest completeness score for “Precautions and Clinical Considerations” domain (8.5) (Table 5; Figs. 1 and 2).

The univariable GEE analysis of accuracy revealed an OR of 1.08 for both Gemini and Claude in reference to ChatGPT. This OR is close to 1, indicating no meaningful association between the AI models and accuracy. Similarly, for completeness, the OR for Gemini was 0.97 and for Claude was 1.04. These ORs are near 1, which shows little difference. In addition, the association of the AI models with both accuracy and completeness was statistically non-significant. Regarding the domains, none showed a statistically significant association with completeness. However, only the “Post-treatment Care” domain showed a significant association with accuracy, with a high OR of 4.04. (Table 6)

The Chi-square test reported statistically significant difference in the distribution of ranking categories across the three tools ($p=0.002$). Both Gemini and Claude secured

Table 4 Post-hoc pairwise comparisons of response time and word count between AI models

	Pairwise comparison		
	ChatGPT Vs Gemini	ChatGPT Vs Claude	Gemini Vs Claude
Generation time	0.001*	0.001*	0.001*
No. of words	0.001*	0.001*	0.001*

Mann Whitney U test, *statistical significance at $p \leq 0.05$

first rank for 38% whereas ChatGPT was ranked first for only 24%. The ChatGPT was ranked third for 49.0% and while the Gemini and Claude were ranked third for 24% and 27% respectively (Table 7; Fig. 3).

Rankings were transformed into weighted scores which a 1st rank was assigned 3 points, a 2nd rank assigned 2 points, and a 3rd rank assigned 1 point. This scoring suggests that a higher median indicates a more favorable preference. Although all three tools shared a median score of 2.00, the Kruskal–Wallis test revealed a statistically significant overall difference ($p < 0.001$). Post-hoc Mann–Whitney U comparisons reported that ChatGPT rank score is statistically significant different with both Gemini and Claude. In contrast, there was no statistically significant difference noted between Gemini and Claude ($p=0.81$). (Table 8; Fig. 4).

4 Discussion

The present study compared ChatGPT, Google Gemini, and Claude in their responses to patient-oriented questions about gingival recession. The three models performed similarly in terms of accuracy and completeness, and none of the differences in these measures reached statistical significance. In contrast, the models differed clearly in expert preference, response generation time, and response length. Thus, the main differences were not in the overall factual quality of the information, but in how the information was presented.

A key finding was the lack of a significant difference in accuracy or completeness among the three models. Median accuracy scores were above 8/10 for all three systems, and completeness scores were also relatively high. These findings suggest that the models demonstrated relatively high expert-rated accuracy for common patient questions about gingival recession. This is in line with recent dental studies

Table 3 Comparison of accuracy, completeness, response time, and response length across AI models

AI tools	ChatGPT		Gemini		Claude		P value
	Median	IQR	Median	IQR	Median	IQR	
Accuracy	8.30	6.55–9.85	8.40	7.10–9.15	8.00	7.25–9.60	0.993
Completeness	7.60	6.65–8.15	7.30	7.00–8.00	7.20	6.25–8.00	0.972
Generation time	5.82	4.80–6.69	8.46	7.82–9.36	14.61	13.23–15.64	0.001*
No. of words	232.00	198.25–298.75	438.00	407.00–486.00	333.50	303.00–349.25	0.001*

Kruskal Wallis test, *statistical significance at $p \leq 0.05$

Table 5 Domain-specific accuracy and completeness scores across evaluated AI models

	Accuracy			Completeness		
	ChatGPT		Gemini	Claude		Gemini
	Median	IQR	Median	Median	IQR	Median
Assessment and Diagnosis	7.30	6.25–8.35	7.70	7.80	7.40–8.35	7.10
Efficacy and Outcomes	8.30	6.65–9.20	8.40	7.30	7.05–8.00	7.50
Complications and Management	7.90	4.60–9.85	8.60	8.00	7.25–9.50	7.50
Post-treatment Care	9.50	8.25–10.00	9.50	9.00	8.00–10.00	7.50
Precautions and Clinical Considerations	8.90	6.45–10.00	8.00	8.55	7.03–10.00	7.50

reporting generally acceptable factual accuracy among leading LLMs, despite differences in response style. (Sivaramakrishnan et al. 2025) Zhang et al. likewise reported high comprehensiveness across several LLMs when answering patient questions about gingival and endodontic conditions (Zhang et al. 2025a). Taken together, these findings suggest that contemporary LLMs may have a useful role in preliminary patient education, provided that professional oversight remains available.

The expert rankings, however, did not follow the same pattern as the accuracy and completeness scores. Gemini and Claude were each ranked first in 38% of evaluations, compared with 24% for ChatGPT ($p < 0.05$). ChatGPT also had a significantly less favorable weighted ranking despite having similar AOI-based accuracy and completeness. One possible explanation is that the AOI measures factual quality, whereas expert preference can also be influenced by organization, readability, flow, and the balance between technical detail and patient-friendly language. Similar observations have been reported in other studies in which clinicians rated chatbot responses favorably because of their structured presentation even when factual content was comparable. (Maruska et al. 2025) Objective accuracy measures may therefore not capture every feature that affects clinicians' overall preference.

A range of methods has been used to assess AI-generated healthcare responses. Previous studies have commonly relied on expert ratings to examine accuracy, completeness, and overall quality. (He et al. 2024; Singhal et al. 2023) Other tools include DISCERN-AI, PEMAT-AI, Global Quality Scores, and modified DISCERN, which assess areas such as understandability, actionability, safety, and educational value. (Garcia et al. 2024; Ghanem et al. 2024) Readability has also been examined using measures such as the Flesch-Kincaid Grade Level, Flesch Reading Ease, Gunning Fog Index, and SMOG Index. (Garcia et al. 2024; Ghanem et al. 2024; Grabeel et al. 2018; Tam et al. 2024) In the present study, the AOI index was selected because its domains closely matched the clinical focus of the questions. It assesses factual accuracy, corroboration with evidence, consistency, clarity and specificity, and relevance. (Daraqel et al. 2024) These areas are particularly relevant when responses include treatment instructions, contraindications, expectations, or safety information. The inclusion of evidence corroboration is also useful given concerns about hallucinations and unwarranted confidence in LLM-generated content.

Response generation time and length differed substantially between the models. ChatGPT was consistently faster and produced shorter responses than Gemini and Claude. Gemini generated the longest responses, whereas Claude had the longest processing time. These differences may

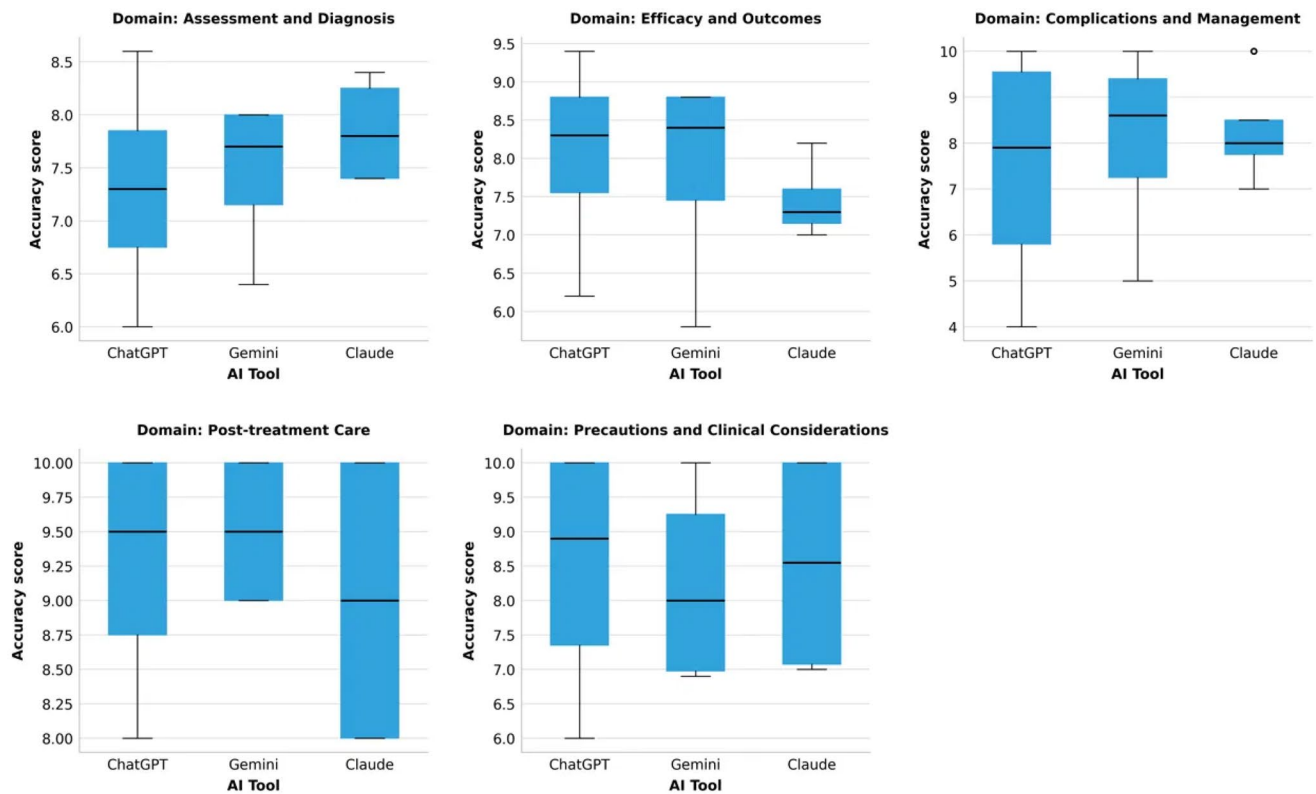


Fig. 1 Accuracy of the AI models among each domain

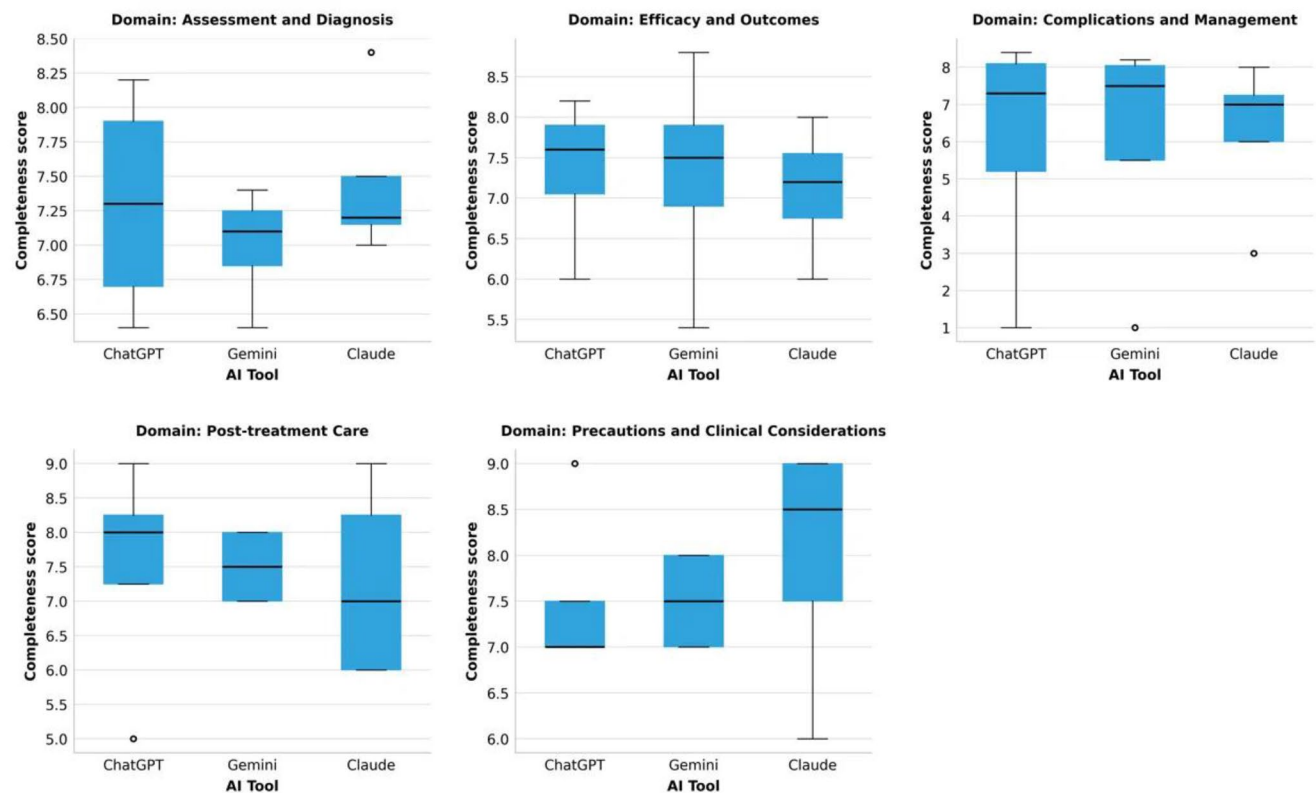


Fig. 2 Completeness of the AI models among each domain

Table 6 Univariable GEE analysis of factors associated with accuracy and completeness

	Accuracy OR [95% CI]	P value	Completeness OR [95% CI]	P value
AI Tool				
ChatGPT	1.00 (reference)		1.00 (reference)	
Gemini	1.08 (0.73–1.60)	0.70	0.97 (0.73–1.28)	0.84
Claude	1.08 (0.74–1.58)	0.69	1.04 (0.82–1.33)	0.75
Domain				
Assessment and Diagnosis	1.00 (reference)		1.00 (reference)	
Efficacy and Outcomes	1.15 (0.83–1.59)	0.40	1.00 (0.84–1.19)	1.00
Complications and Management	1.24 (0.48–3.23)	0.65	0.59 (0.21–1.72)	0.34
Post-treatment Care	4.04 (1.60–10.20)	0.01*	1.09 (0.68–1.74)	0.72
Precautions and Clinical Considerations	1.72 (0.58–5.06)	0.33	1.25 (0.92–1.69)	0.16

Univariable Generalized Estimating Equations (GEE) test, *statistical significance at $p \leq 0.05$

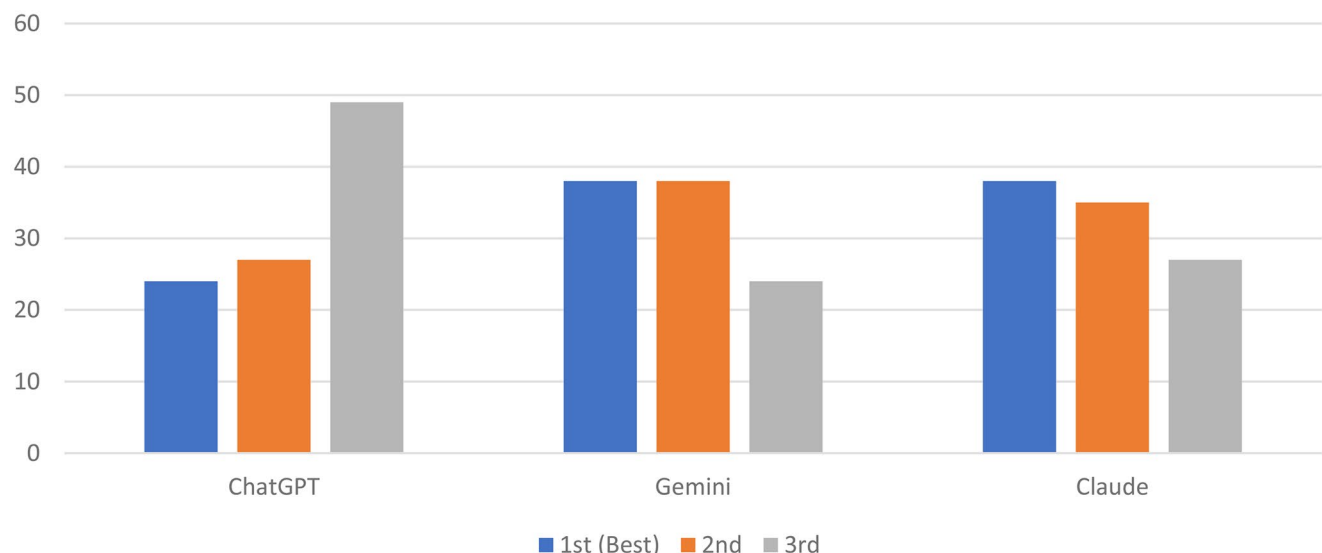
Table 7 Distribution and comparison of expert ranking categories across AI models

	Ranking categories			P value
	1st N (%)	2nd N (%)	3rd N (%)	
ChatGPT	24 (24.0)	27 (27.0)	49 (49.0)	0.002*
Gemini	38 (38.0)	38 (38.0)	24 (24.0)	
Claude	38 (38.0)	35 (35.0)	27 (27.0)	

Chi-square test, *statistical significance at $p \leq 0.05$

reflect variation in model architecture, inference strategies, and optimization rather than differences in clinical knowledge. Importantly, a shorter response is not necessarily less useful, and a longer response is not automatically more

informative. For instance, Sezer et al. found no significant relationship between response time and accuracy when comparing several AI models on dental specialty examination questions. (Sezer and Aydoğdu 2025) Very short answers may leave out clinically relevant information, while overly long responses may be difficult for patients to follow, particularly when health literacy is limited. In addition, Arslan et al. similarly reported differences in response detail between ChatGPT and Gemini despite comparable accuracy in patient questions about refractive surgery. (Arslan and Usta 2025) The useful response, therefore, is not necessarily the longest one; it is the one that delivers sufficient information in a way that patients can comprehend. (Ayers et al. 2023).

**Fig. 3** Comparison of AI models ranking**Table 8** Ranking scores comparison across three AI models

		Median	IQR	P value	Post hoc comparison	P value
AI Tool	ChatGPT	2.00	1.00–2.00	0.001*	ChatGPT vs. Gemini	0.001*
	Gemini	2.00	2.00–3.00		ChatGPT vs. Claude	0.002*
	Claude	2.00	1.00–3.00		Gemini vs. Claude	0.81

Kruskal Wallis test, Post hoc comparison with Mann Whitney U test*statistical significance at $p \leq 0.05$

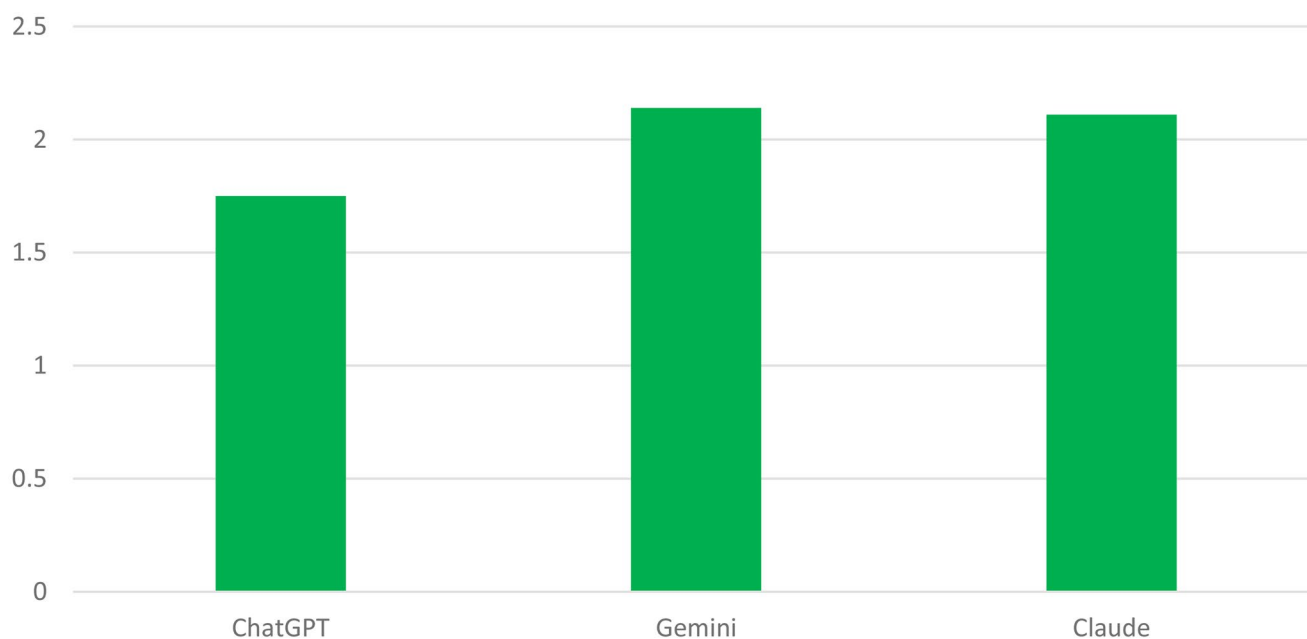


Fig. 4 Comparison of AI models ranking (weighted score)

The domain-specific findings provide another important point. All three LLMs performed best in the Post-treatment Care domain, and GEE analysis identified this domain as the only category significantly associated with higher accuracy. Batool et al. reported that LLMs could generate clinically appropriate postoperative dental care instructions and suggested that they may help reinforce patient education after dental procedures, although clinician oversight remains necessary. (Batool et al. 2024)

The stronger performance on post-treatment care may be related to the standardized nature of this type of information. In contrast, assessment and diagnosis require consideration of individual factors such as recession classification, periodontal phenotype, esthetic demands, smoking status, and systemic health. These factors are difficult for an LLM to assess without a clinical examination and patient-specific information. (Kaplan and Akyol 2026) The results therefore support the potential use of LLMs for educational information more strongly than for individualized treatment recommendations.

From a clinical perspective, these findings are relevant because patients increasingly use AI chatbots before and after dental appointments to learn about disease progression, procedures, and postoperative care. Clear and accurate responses may improve understanding and reinforce instructions. At the same time, using AI without professional interpretation could contribute to unrealistic expectations or inappropriate self-diagnosis which could potentially have negative impact for patients. (Hager et al. 2024) AI-generated information should consequently be considered an educational aid rather than a replacement for clinical

examination or professional judgment. This view is consistent with the broader dental AI literature, which supports the use of LLMs to assist rather than replace clinician–patient communication. (Rajan et al. 2026; Rokhshad et al. 2024; Thorat et al. 2024)

Several limitations should be considered. First, only the initial single-turn response to each standardized prompt was assessed. In practice, patients may have multi-turn conversations with an LLM, and subsequent interaction could change the accuracy or consistency of the information. Second, although all five evaluators met strict clinical and board-certification criteria, readability from the patient's perspective was not formally measured thus, may limit the generalizability of the findings. Third, all questions were written in English and were assessed by periodontists practicing within a single institutional setting, which may limit generalizability to other languages and populations. Future investigations would benefit from exploring longitudinal patient–LLM interactions and evaluating informational accessibility across populations with varying socio-educational backgrounds.

5 Conclusion

The three LLMs offered broadly comparable levels of accuracy and completeness when answering patient-oriented questions with respect to gingival recession. ChatGPT was faster and more concise, while Gemini and Claude were more often preferred by the expert evaluators. These LLMs may be useful as adjuncts to patient education, but they

should complement rather than replace professional periodontal consultation.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s44445-026-00247-8>.

Acknowledgements The author acknowledges and appreciate the support of this study by the Ongoing Research Funding Program (ORF-2026-1718), King Saud University, Riyadh, Saudi Arabia. The author gratefully acknowledges R.A. (Raed Alrowis) and R.A. (Razan Alaqueely) for their valuable assistance in validating the patient-centered questions used in this study.

Author contributions The author was (SA) responsible for the conception and design of the study, data collection, data analysis and interpretation, manuscript drafting, critical revision, and approval of the final manuscript.

Funding The author acknowledges and appreciate the support of this study by the Ongoing Research Funding Program (ORF-2026-1718), King Saud University, Riyadh, Saudi Arabia.

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate The study and protocols were approved by King Saud University-Institutional Review Board, Riyadh, Saudi Arabia (IRB #: R26-IRB-432). This work was conducted in accordance with all relevant institutional regulations, ethical guidelines, and applicable research protocols.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Albalawi F, Khanagar SB, Iyer K, Alhazmi N, Alayyash A, Alhazmi AS, Awawdeh M, Singh OG (2025) Evaluating the performance of artificial intelligence-based large language models in orthodontics—a systematic review and meta-analysis. *Appl Sci* 15:893
- Arslan S, Usta Küçükbezirci G (2025) Evaluating the Accuracy, Completeness, and Readability of Chatbot Responses to Refractive Surgery-Related Patient Questions: A Comparative Analysis of ChatGPT and Google Gemini. *Cureus* 17:e88980
- Ayers JW, Poliak A, Dredze M et al (2023) Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med* 183:589–596
- Babayigit O, Tastan Eroglu Z, Ozkan Sen D, Ucan Yarkac F (2023) Potential Use of ChatGPT for Patient Information in Periodontology: A Descriptive Pilot Study. *Cureus* 15:e48518
- Batool I, Naved N, Kazmi SMR, Umer F (2024) Leveraging Large Language Models in the Delivery of Post-Operative Dental Care: A Comparison between an Embedded GPT Model and ChatGPT. *BDJ Open* 10:48
- Cortellini P, Bissada NF (2018) Mucogingival Conditions in the Natural Dentition: Narrative Review, Case Definitions, and Diagnostic Considerations. *J Periodontol* 89:S204–S213
- Daraqel B, Wafaie K, Mohammed H, Cao L, Mheissen S, Liu Y et al (2024) The Performance of Artificial Intelligence Models in Generating Responses to General Orthodontic Questions: ChatGPT vs Google Bard. *Am J Orthod Dentofac Orthop* 165:652–662
- Garcia P, Ma SP, Shah S, Smith M, Jeong Y, Devon-Sand A, Tai-Seale M, Takazawa K, Clutter D, Vogt K et al (2024) Artificial Intelligence-Generated Draft Replies to Patient Inbox Messages. *JAMA Netw Open* 7:e243201
- Ghanem Y, Rouhi A, Al-Houssan A, Saleh Z, Moccia M, Joshi H et al (2024) Dr. Google to Dr. ChatGPT: Assessing the Content and Quality of Artificial Intelligence-Generated Medical Information on Appendicitis. *Surg Endosc* 38:2887–2893
- Grabeel K, Russomanno J, Oelschlegel S, Tester E, Heidel R (2018) Computerized Versus Hand-Scored Health Literacy Tools: A Comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in Printed Patient Education Materials. *J Med Libr Assoc* 106:38–45
- Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, Rueckert D et al (2024) Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making. *Nat Med* 30:2613–2622
- He Z, Bhasuran B, Jin Q, Tian S, Hanna K, Shavor C, Garcia Arguello L, Murray P, Lu Z (2024) Quality of Answers of Generative Large Language Models vs Peer Patients for Interpreting Lab Test Results for Lay Patients: Evaluation Study. *J Med Internet Res* 26:e56655
- Hirani R, Noruzi K, Khuram H, Hussaini AS, Aifuwa EI, Ely KE, Lewis JM, Gabr AE, Smiley A, Tiwari RK et al (2024) Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. *Life* 14:557
- Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, Yin H, Xu C, Yang R, Zheng Q et al (2023) ChatGPT for Shaping the Future of Dentistry: The Potential of Multi-Modal Large Language Model. *Int J Oral Sci* 15:29
- Kaplan TT, Akyol GE (2026) The Role of Large Language Models in Dental Diagnosis, Decision-Making, and Communication: A Systematic Review. *Jpn Dent Sci Rev* 62:57–67
- Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, Abdullah S (2023) Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. *Biomed Mater Devices* 1:731–738
- Ma X, Pan W, Yu XN (2025) Evaluating AI-generated examination papers in periodontology: a comparative study with human-designed counterparts. *BMC Med Educ* 25:1099
- Maruska EE, Elgreatly A, Madaio W, Razoky K, Bay C, Mahrous A (2025) Comparing Dentist and Chatbot Answers to Dental Questions for Quality and Empathy. *JADA Found Sci* 4:100044
- Merijohn GK (2016) Management and Prevention of Gingival Recession. *Periodontol* 2000 71:228–242

- Rajan J, Rathan ACL, Subramanian A, Kesavan H (2026) Influence of Large Language Models on Clinical Decision-Making in Dentistry: A Narrative Review. *SRM J Res Dent Sci* 17:98–104
- Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F (2024) Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. *J Dent* 144:104938
- Salehin I, Ahmed Sajib MT, Huda Badhon N, Sakibul Hassan Rifat M, Amin N, Nessa Moon N (2025) Systematic literature review of LLM-large language model in medical: digital health, technology and applications. *Eng Rep* e70365
- Schwendicke F, Samek W, Krois J (2020) Artificial Intelligence in Dentistry: Chances and Challenges. *J Dent Res* 99:769–774
- Sezer B, Aydoğdu T (2025) Performance of Advanced Artificial Intelligence Models in Traumatic Dental Injuries in Primary Dentition: A Comparative Evaluation of ChatGPT-4 Omni, DeepSeek, Gemini Advanced, and Claude 3.7 in Terms of Accuracy, Completeness, Response Time, and Readability. *Appl Sci* 15:7778
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S et al (2023) Large Language Models Encode Clinical Knowledge. *Nature* 620:172–180
- Sivaramakrishnan G, Almuqahwi M, Ansari S, Lubbad M, Alagamaw E, Sridharan K (2025) Assessing the Power of AI: A Comparative Evaluation of Large Language Models in Generating Patient Education Materials in Dentistry. *BDJ Open* 11:59
- Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, Osterhoudt H, Wu X, Visweswaran S, Fu S et al (2024) A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review. *NPJ Digit Med* 7:258
- Thorat V, Rao P, Joshi N, Talreja P, Shetty AR, Rao P (2024) Role of Artificial Intelligence (AI) in Patient Education and Communication in Dentistry. *Cureus* 16:e59799
- Tiwari A, Kumar A, Jain S, Dhull KS, Sajjanar A, Puthenkandathil R, Paiwal K, Singh R (2023) Implications of ChatGPT in public health dentistry: A systematic review. *Cureus* 15:e40367
- Tonetti M, Sanz M, Cairo F, Nart J, Chapple I, Aimetti M, Aroca S, Avila-Ortiz G, Blanco J, Bujaldón A, Cavalcanti R (2025) Aesthetics and patient-reported outcomes in periodontology and implant dentistry: consensus report. *J Clin Periodontol* 52:1222–1244
- Umer F, Batool I, Naved N (2024) Innovation and Application of Large Language Models (LLMs) in Dentistry—A Scoping Review. *BDJ Open* 10:90
- Wang D, Zhang S (2024) Large Language Models in Medical and Healthcare Fields: Applications, Advances, and Challenges. *Artif Intell Rev* 57:299
- Zhang Q, Wu Z, Song J, Luo S, Chai Z (2025a) Comprehensiveness of Large Language Models in Patient Queries on Gingival and Endodontic Health. *Int Dent J* 75:151–157
- Zhang R, Pan Y, Liu Y, Deng Y, Pow EHN (2025b) Leveraging large language models for patient instructions in dentistry—a systematic review and meta-analysis. *J Prosthodont* 1–10

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.