

# Comparative analysis of artificial intelligence chatbots in orthodontic emergency scenarios: ChatGPT-3.5, ChatGPT-4.0, Copilot, and Gemini

Buket Erdem<sup>a</sup>; Mustafa Özcan<sup>a</sup>; Çağla Şar<sup>b</sup>

## ABSTRACT

**Objectives:** To evaluate and compare the accuracy of four AI chatbots, ChatGPT-3.5, ChatGPT-4.0, Copilot, and Gemini, in response to orthodontic emergency scenarios.

**Materials and Methods:** Forty frequently asked questions related to orthodontic emergencies were posed to the chatbots. These questions were categorized as fixed orthodontic treatment, clear aligner treatment, eating and oral hygiene, pain and discomfort, general concerns, retention, and sports and travel. The responses were evaluated by three orthodontic experts using a five-point Likert scale, and statistical analysis was conducted to assess variations in accuracy across chatbots.

**Results:** Statistical analysis revealed significant differences among the chatbots. Gemini and ChatGPT-4.0 demonstrated the highest accuracy in response to orthodontic emergencies, followed by Copilot, whereas ChatGPT-3.5 had the lowest accuracy scores. Additionally, the “Fixed Orthodontic Treatment” category showed a statistically significant difference ( $P = .043$ ), with Gemini outperforming the other chatbots in this category. However, no statistically significant differences were found in other categories.

**Conclusions:** AI chatbots show potential in providing immediate assistance for orthodontic emergencies, but their accuracy varies across different models and question categories. (*Angle Orthod.* 2026;96:100–105.)

**KEY WORDS:** Artificial intelligence; Chatbot; Orthodontic emergency

## INTRODUCTION

Artificial intelligence (AI) technologies have witnessed rapid advancement in recent years, transforming various areas, including healthcare. Within dentistry, AI-driven innovations have demonstrated remarkable potential to enhance diagnostic accuracy, optimize treatment planning, and improve patient management practices.<sup>1</sup> These developments in AI technology are progressing in a way that will not only make the work of doctors

easier, but will also be beneficial for dental patients. Specific AI chatbot models, which are the computer programs that have been developed for conversing with human users, offer 24/7 availability, cost-effectiveness, and consistency in responding to user inquiries.<sup>2</sup>

The application of AI-powered chatbots in orthodontics shows great promise, with improved patient satisfaction rates and faster response times. In the context of orthodontic treatment, patients are more likely to seek assistance from chatbots during moments of uncertainty, such as broken appliances or cracked aligners. This behavior can be attributed to the immediacy, accessibility, and user-friendly nature of such platforms, which provide instant responses to questions without the need for in-person consultation. Patients often prefer this approach in situations that they perceive as urgent or when they seek clarification outside regular clinical hours.

Chatbots like ChatGPT, Google Gemini (formerly Bard), and Microsoft Copilot offer potential solutions by providing patients with reliable, accessible information in real time. However, evaluating their performance based on metrics like response accuracy is crucial to determine their suitability for applications. It should be considered

<sup>a</sup> Assistant Professor, Department of Orthodontics, Faculty of Dentistry, Istanbul Health and Technology University, Istanbul, Türkiye.

<sup>b</sup> Professor and Chair, Department of Orthodontics, Faculty of Dentistry, Istanbul Health and Technology University, Istanbul, Türkiye.

Corresponding author: Dr Buket Erdem, Department of Orthodontics, Faculty of Dentistry, Istanbul Health and Technology University, Sütlüce Mah. İmrahor Cad. No: 82 Beyoğlu, Istanbul, Türkiye  
(e-mail: buket.erdem@istun.edu.tr)

Accepted: September 1, 2025. Submitted: February 18, 2025.

Published Online: September 26, 2025

© 2026 by The EH Angle Education and Research Foundation, Inc.

that medical chatbots may use biased algorithms, resulting in potentially harmful, incorrect responses.<sup>3</sup>

There are no studies in the literature specifically addressing orthodontic emergencies; however, there are a limited number of chatbot studies related to general questions that may arise in the orthodontic field. A recent study indicated that responses generated by ChatGPT 3.5 and Google Bard (Gemini) were rated highly for accuracy and completeness in addressing general orthodontic inquiries from patients.<sup>4</sup> Tanaka et al.<sup>5</sup> evaluated the reliability of ChatGPT 3.5 in responding to orthodontic inquiries concerning mini-implants, clear aligners, and digital imaging. Naureen et al.<sup>6</sup> conducted a comparative study evaluating the performance of ChatGPT-4 and Google Bard (Gemini) in orthodontic knowledge. Their findings highlighted significant limitations in both models, including the provision of incorrect and fabricated references across various categories. They concluded that such observations aligned with the growing concern about the effectiveness of AI-based tools in clinical orthodontics, emphasizing the need for cautious interpretation of their output. In another recent study, Vassis et al.<sup>7</sup> evaluated the reliability and effectiveness of AI-generated information regarding orthodontic treatment side effects. They concluded that, although patients found the AI-generated content more comprehensible and user-friendly compared to traditional methods, it lacked the depth and quality necessary for comprehensive education. These findings highlight the potential of AI tools in patient education while underscoring their current limitations and the need for cautious application in clinical settings.

This study aimed to compare the response accuracy of ChatGPT-3.5, ChatGPT-4.0, Copilot, and Google Gemini in addressing orthodontic emergency scenarios. The null hypothesis was that all AI chatbots could guide patients in any emergency situation during their orthodontic treatment with the same level of accuracy.

## MATERIALS AND METHODS

This study was conducted at the Istanbul Health and Technology University, Faculty of Dentistry, Department of Orthodontics, Istanbul, Türkiye. Ethical approval was not required as no participants were involved in the study.

### Identifying Frequently Asked Questions (FAQs)

In the current study, ChatGPT 3.5 was asked 50 FAQs related to “orthodontic emergencies” and the responses were recorded. From the initial set of 50 questions, 40 that were considered to be the most frequently encountered in clinical practice were selected by two experienced orthodontists (Table 1). The first answer to each prompt was considered the final answer.

**Table 1.** Orthodontic Emergency Questions by Categories

General Concerns	
1.	What is considered as an orthodontic emergency?
2.	Can orthodontic emergencies cause permanent damage?
3.	Can an orthodontic emergency delay my treatment?
4.	How soon should I see my orthodontist after an emergency?
5.	What are the risks of attempting to fix an orthodontic issue on my own?
Fixed Orthodontic Treatment	
1.	What should I do if a bracket comes off?
2.	How do I handle a wire that is poking my cheek or gums?
3.	What can I use to cover a sharp wire until I see my orthodontist?
4.	Is it safe to cut a long wire that is causing discomfort?
5.	What if multiple brackets or wires break at once?
6.	What happens if I run out of rubber bands?
7.	I lost my rubber bands. Is it okay to go without rubber bands for a few days?
8.	What happens if I use rubber bands incorrectly (from a different tooth)?
9.	What if I swallow a part of my braces?
Clear Aligner Treatment	
1.	What should I do if my aligner cracks?
2.	Can I continue using a cracked aligner until I get a replacement?
3.	What if my aligner doesn't fit properly anymore?
4.	How do I clean a damaged aligner?
5.	My aligner was cracked. Is it okay to skip wearing my aligner until I get a new one?
6.	My aligner attachment came off. What should I do?
Pain and Discomfort	
1.	How can I relieve pain caused by braces?
2.	How can I relieve pain caused by aligners?
3.	What should I do if I experience severe pain after an orthodontic adjustment?
4.	Is it normal to have mouth sores from braces?
5.	How do I reduce swelling caused by orthodontic appliances?
6.	What over-the-counter pain relief is safe to use with braces?
Eating and Oral Hygiene	
1.	What should I do if food gets stuck in my braces and I can't remove it?
2.	Are there foods I should avoid to prevent orthodontic emergencies?
3.	How can I brush and floss effectively with broken braces?
4.	Is it okay to eat with a loose bracket or wire?
Sports and Travel	
1.	How can I protect my braces during sports?
2.	What should I do if I get hit in the mouth while wearing braces?
3.	Is it safe to wear a mouthguard with braces?
4.	What should I do if my braces are damaged during a game?
5.	How can I find an orthodontist if I'm away from home?
6.	What should I do if my braces break on a holiday or weekend?
Retention	
1.	What should I do if my retainer breaks?
2.	How long can I go without wearing my retainer if it's broken?
3.	Is it normal for my retainer to feel tight?
4.	Is it normal for my retainer to feel loose?

An example of a question along with responses from all four chatbots is provided in Figure 1.

The questions were categorized into ten distinct groups as “General Concerns” (12.5%), “Fixed Orthodontic

<b>Question 3: What should I do if a bracket comes off?</b>
<b>ChatGPT-3.5:</b>
Contact your orthodontist as soon as possible to schedule a repair. In the meantime, try to keep the area clean and avoid eating sticky or hard foods.
<b>ChatGPT-4.0:</b>
If a bracket comes off, try to save it and contact your orthodontist as soon as possible. Use orthodontic wax to cover any sharp edges to prevent irritation until you can get it repaired.
<b>Copilot:</b>
If a bracket comes off, contact your orthodontist as soon as possible. In the meantime, you can use orthodontic wax to cover any sharp edges to prevent discomfort.
<b>Gemini:</b>
If a bracket comes off: <ul style="list-style-type: none"> <li>• Don't panic. While it's inconvenient, it's a common occurrence.</li> <li>• Contact your orthodontist: Schedule an appointment to have the bracket reattached.</li> <li>• Save the bracket: If possible, clean and save the bracket to bring to your appointment.</li> <li>• Manage discomfort: If the wire is poking you, you can use orthodontic wax to cover the sharp end.</li> </ul> It's important to get the bracket reattached as soon as possible to keep your treatment on track.

**Figure 1.** Example of a question with answers from all four AI chatbots.

Treatment” (22.5%), “Clear Aligner Treatment” (15%), “Pain and Discomfort” (15%), “Eating and Oral Hygiene” (10%), “Sports and Travel” (15%), and “Retention” (10%).

A single investigator submitted each question to all AI models: ChatGPT (GPT-3.5), ChatGPT (GPT-4.0), Copilot, and Gemini. Since AI models can learn from previous conversations, a new account was created for each chatbot to ensure that no prior interactions influenced the responses. Subsequently, all questions were asked, and the responses were recorded by the investigator.

The investigator then completed the form by randomly arranging the four separate answers under each question without indicating which chatbot provided each response. Subsequently, experts were asked to rate the responses. A five-point Likert scale was utilized for accuracy.<sup>8</sup> The comprehensive scoring methodology was as follows:

- Completely incorrect: It was entirely incorrect.
- Not quite correct: The response contained significant inaccuracies and did not correspond with the answer.
- Slightly correct: The response exhibited several inaccuracies that did not align with the true answer.
- Moderately correct: The response was largely precise, exhibiting only slight discrepancies from the actual information.
- Completely correct. The response corresponded with the main point comprehensively.

**Table 2.** Inter-rater Agreement

N coders	3
N cases	160
N decisions	480
<b>Krippendorff's alpha (ordinal)</b>	<b>0.871</b>

## Expert Evaluation

Three orthodontists (with 8, 12, and 20 years of orthodontic experience) reviewed the responses to each of the 40 FAQs. They evaluated the anonymized responses independently, were blinded to the identity of the chatbot that generated each answer, and were unaware of each other's evaluations.

All collected data were entered in Microsoft Excel (Microsoft Corp., Redmond, Wash.) for further analysis.

## Statistical Analysis

Numerical data were summarized using means and standard deviations. Comparisons of numerical data were performed using one-way analysis of variance (one-way ANOVA) since the assumption of variance homogeneity was met, and effect sizes (eta-squared) were also reported. When the overall test result was found to be significant, pairwise comparisons were conducted using the least significant difference (LSD) test. Inter-rater reliability was assessed using Krippendorff's alpha, which is appropriate for ordinal data. The Krippendorff analysis was performed using the tool available at <http://dfreelon.org/recal/recal-oir.php>. All other statistical analyses were conducted using IBM SPSS Statistics 27.0 (IBM Corp., Armonk, New York, USA). A significance level of 5% was considered statistically significant.

## RESULTS

Findings related to inter-rater reliability are presented in Table 2. The inter-rater agreement (Krippendorff's alpha = 0.871) was found to be nearly perfect.

The accuracy of chatbot responses was calculated as the average rating given by the three evaluators. The calculated accuracy scores were compared for a total of 40 questions and seven categories. Comparisons of chatbot accuracy scores are presented in Table 3.

For the total score, the mean score of Gemini ( $4.29 \pm 0.82$ ) was significantly higher than those of ChatGPT 3.5 ( $3.57 \pm 0.79$ ) and Copilot ( $3.8 \pm 0.78$ ). However, the mean score of ChatGPT 4.0 ( $3.84 \pm 0.74$ ) did not show statistically significant differences compared to the other chatbot means.

A statistically significant difference was found among chatbot mean scores in the “Fixed Orthodontic

**Table 3.** Comparison of Chatbots' Response Based on Averages<sup>a,b</sup>

	ChatGPT-3.5	ChatGPT-4.0	Copilot	Gemini	F	P	Eta-squared
	$\bar{X} \pm SD$	$\bar{X} \pm SD$	$\bar{X} \pm SD$	$\bar{X} \pm SD$			
General concerns	3.67 ± 0.47	3.8 ± 0.65	3.93 ± 0.6	3.87 ± 0.61	0.19	.902	0.034
Fixed orthodontic treatment	3.67 ± 0.71 <sup>a</sup>	3.78 ± 0.6 <sup>a</sup>	3.67 ± 0.88 <sup>a</sup>	4.52 ± 0.6 <sup>b</sup>	3.032	<b>.043</b>	0.221
Clear aligner treatment	3.44 ± 1	4.22 ± 0.78	4.11 ± 0.69	4.06 ± 1.36	0.746	.537	0.101
Pain and discomfort	3 ± 1.1	3.28 ± 1.18	3.28 ± 1.18	3.72 ± 0.83	0.458	.715	0.064
Eating and oral hygiene	3.5 ± 0.58	3.75 ± 0.5	3.58 ± 0.5	4.58 ± 0.83	2.588	.101	0.393
Sports and travel	4.17 ± 0.41	4.28 ± 0.44	4.28 ± 0.44	4.67 ± 0.42	1.566	.229	0.19
Retention	3.42 ± 0.79	3.75 ± 0.5	3.75 ± 0.5	4.67 ± 0.47	3.441	.052	0.462
Total score	3.57 ± 0.79 <sup>a</sup>	3.84 ± 0.74 <sup>a,b</sup>	3.8 ± 0.78 <sup>a</sup>	4.29 ± 0.82 <sup>b</sup>	6.049	<b>.001</b>	0.104

<sup>a</sup>  $\bar{X}$  indicates arithmetic mean, SD, standard deviation.

<sup>b</sup> In pairwise comparisons, groups that do not show a statistically significant difference at the 5% level are indexed with the same letter.

$P < .05$  was considered statistically significant.

Treatment" category ( $P = .043$ ) and for the total score ( $P = .001$ ). The effect sizes for the observed differences were small-to-moderate in the Fixed Orthodontic Treatment category ( $\eta^2 = 0.043$ ) and moderate-to-high for the total score ( $\eta^2 = 0.104$ ).

Pairwise comparisons to identify the source of the differences showed that the mean score of Gemini ( $4.52 \pm 0.6$ ) in the "Fixed Orthodontic Treatment" category was significantly higher than the means of the other chatbots (ChatGPT 3.5 =  $3.67 \pm 0.71$ , ChatGPT 4.0 =  $3.78 \pm 0.6$ , Copilot =  $3.67 \pm 0.88$ ).

## DISCUSSION

The rise of AI technology, particularly large language models (LLMs), has transformed how information is accessed and utilized across various domains, including healthcare. These systems leverage advanced architectures, like transformer models and extensive training datasets, to give answers to a wide range of questions that make sense and are relevant to the situation. However, differences in model architectures, training methodologies, and capabilities result in varying performance outcomes across specific applications.<sup>9</sup>

This study evaluated the comparative performance of ChatGPT-3.5, ChatGPT-4.0, Copilot, and Gemini in addressing orthodontic treatment emergencies. The findings indicated notable variations in the accuracy of responses, with Gemini achieving the highest mean score (4.29) and ChatGPT-3.5 the lowest (3.57). Gemini's superior performance can be attributed to its advanced model architecture, which integrates multi-modal capabilities, and its dual reliance on extensive pre-trained datasets and real-time internet access.<sup>10</sup> These features enable Gemini to provide both comprehensive and up-to-date responses, setting it apart from models like ChatGPT that rely solely on pre-existing data. However, pairwise comparisons indicated no statistically significant difference between ChatGPT-4.0 and Gemini,

as reflected in the groupings assigned to each model. This suggests that, while Gemini achieved the highest overall score, ChatGPT-4.0 performed at a comparable level in terms of response accuracy.

During question categorization, it was observed that while other categories contained 4–6 questions, the "Fixed Orthodontic Treatment" category included a greater number of questions. This is likely because patients encounter more complications during fixed orthodontic treatment, leading to a greater volume of inquiries in this area.

The "Fixed Orthodontic Treatment" category ( $P = 0.043$ ) also showed statistically significant variations, with Gemini outperforming the other chatbots. However, no statistically significant differences were found in other categories ( $P > 0.05$ ). Most likely, questions related to "Fixed Orthodontic Treatment" are among the most commonly asked topics in general conversational AI interactions, potentially providing the AI models with a disproportionately higher volume of training data in this area. Despite this, Gemini emerged as the most accurate chatbot in this category. Fixed orthodontic treatment has been a well-established method for many years, and the necessary protocols for handling emergencies have remained largely unchanged over time. Therefore, access to more recent information would not provide a significant advantage in this category. Consequently, Gemini's ability to retrieve real-time data is unlikely to be the primary factor contributing to its superior performance. Instead, given that all chatbots likely possess a comprehensive dataset in this well-documented area, Gemini's higher accuracy in responding to these queries suggests a more effective utilization of its training data. Additionally, the longstanding presence of fixed orthodontic treatment in clinical literature may have provided a richer foundation of training data for the chatbots, further enhancing their performance in this category.

Clear aligners have been used in a significant percentage of orthodontic treatments in recent years. Due to the advantage of remote treatment management, patients who will be staying abroad for extended periods may prefer clear aligner therapy. It can also be assumed that these patients are more likely to consult a chatbot directly in case of emergencies. According to the results of this study, there was no statistically significant difference among the AI chatbots for consultation by clear aligner patients. The lowest-rated chatbot response belonged to ChatGPT 3.5, with a score of 3.44. Therefore, if aligner patients are unable to reach their doctors, they can safely consult any of the AI chatbots for assistance.

In the present study, ChatGPT 4.0 showed better results than ChatGPT 3.5 in all categories. The performance levels of ChatGPT 3.5 and ChatGPT 4.0 were influenced by their respective knowledge cutoff dates. ChatGPT 3.5, with a cutoff date of September 2021, did not include developments or data beyond that point, which limited its ability to address recent advancements. In contrast, ChatGPT 4.0 extended its knowledge base to April 2023, providing more up-to-date information and a broader scope for generating contextually relevant responses.<sup>11</sup>

In a study by Dursun et al.<sup>12</sup>, the same four chatbots were evaluated for questions related to clear aligners and, unlike the results of the current study, ChatGPT-4.0 was found to provide the most reliable and comprehensive responses. This discrepancy highlights how chatbot performance can vary depending on the context. None of the questions in that study were related to emergency situations with clear aligners, so a question-based comparison was, unfortunately, not possible.

In a study evaluating the same AI chatbots with previous versions of some, Microsoft Bing (Copilot), ChatGPT 3.5, ChatGPT 4.0, and Google Bard (Gemini), in addressing orthodontic clinical questions, only one question could be related to orthodontic emergencies: "Do painkillers, taken before or after orthodontic treatment, help relieve pain? If so, which painkillers work best?"<sup>13</sup> Gemini got the lowest score for this question. In contrast to that previous study, Gemini was the best-performing chatbot in the "Pain and Discomfort" category in the present study. It would be insufficient to compare the results of a chatbot in a single question with the broader evaluation in the current study, as this category comprised six questions.

The timing of the questions asked to chatbots is crucial when considering the influence of real-time data. Since some chatbots retrieve real-time data, the outcomes could be affected by the information available at the time of the query. Future studies should focus

on evaluating questions asked to chatbots at different time intervals to explore how their responses change over time.

It is also important to note that the knowledge base of AI chatbots is not always derived from peer-reviewed or verified sources. Some information may be outdated or inaccurate, which can compromise the reliability of the responses. Therefore, chatbot responses should be interpreted cautiously, especially in clinical contexts.

## CONCLUSIONS

- In summary, the application of AI chatbots in orthodontic emergency situations has shown promise.
- Among the evaluated models, Gemini and ChatGPT-4.0 demonstrated the highest reliability and accuracy.
- In addition to their accuracy, AI chatbots offer practical advantages, since patients are more inclined to use digital tools in urgent situations, making them a convenient alternative to printed guides.

## REFERENCES

1. Agrawal P, Nikhade P. Artificial intelligence in dentistry: past, present, and future. *Cureus*. 2022;14(7):e27405.
2. Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee S-S. Overview of chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. *Front Artif Intell*. 2023;6:1237704.
3. Sharma D, Kaushal S, Kumar H, Gainer S. Chatbots in healthcare: challenges, technologies and applications. In: 2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST). IEEE; 2022:1–6.
4. Daraqel B, Wafaie K, Mohammed H, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. *Am J Orthod Dentofacial Orthop*. 2024; 165(6):652–662.
5. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs, and digital imaging. *Dent Press J Orthod*. 2023;28(5):e2323183.
6. Naureen S, Kiani HG. Assessing the accuracy of AI models in orthodontic knowledge: a comparative study between ChatGPT-4 and Google Bard. *J Coll Physicians Surg Pak*. 2024;34(7):761–766.
7. Vassis S, Powell H, Petersen E, et al. Large-language models in orthodontics: assessing reliability and validity of ChatGPT in pretreatment patient education. *Cureus*. 2024; 16(8):e68085.
8. Likert R. A technique for the measurement of attitudes. *Arch Psychol*. 1932;(140):1–55.
9. Bhattacharya P, Prasad VK, Verma A, et al. Demystifying ChatGPT: an in-depth survey of OpenAI's robust large language models. *Arch Comput Methods Eng*. 2024;31: 4557–4600.

10. Rahaman MS, Ahsan MMT, Anjum N, Rahman MM, Rahman MN. The AI race is on! Google's Bard and OpenAI's ChatGPT head to head: an opinion article. *SSRN J*. 2023. [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=4351785](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4351785). Accessed September 8, 2025.
11. OpenAI. GPT-4 technical report. Published March 2023. <https://cdn.openai.com/papers/gpt-4.pdf>. Accessed February 15, 2025.
12. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? *BMC Med Inform Decis Mak*. 2024;24:211.
13. 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;cjae017.