IDR International Dental Research

Int Dent Res 2025; 15(3): 165-174
https://doi.org/10.5577/intdentres.701

Original Article Open Access

Performance of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 large language models in response to questions about ozone in dentistry

Suzan Cangül[1], Özkan Adıgüzel[2,3], Makbule Taşyürek[2], Tuba Tunç[1], Hatice Ortaç[4]

[1] Dicle University, Faculty of Dentistry, Department of Restorative Dentistry, Diyarbakır, Türkiye
[2] Dicle University, Faculty of Dentistry, Department of Endodontics, Diyarbakır, Türkiye
[3] Nelcotech R&D Center, Department of Dentistry, Sheridan, WY, USA
[4] Dicle University, Faculty of Medicine, Department of Biostatistics, Diyarbakır, Türkiye

Received: 31 October 2025
Accepted: 10 December 2025

Correspondence:

Dr. Suzan CANGÜL
Dicle University, Faculty of Dentistry, Department of Restorative Dentistry, Diyarbakır, Türkiye.
E-mail: suzanbali@outlook.com

How to cite this article:

Cangül S, Adıgüzel Ö, Taşyürek M, Tunç T, Ortaç H. Performance of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 large language models in response to questions about ozone in dentistry. Int Dent Res. 2025;15(3):165-174. https://doi.org/10.5577/intdentres.701

Abstract

Aim: The aim of this study was to compare and evaluate the accuracy and completeness of responses to open-ended questions about the use of ozone in dentistry that were addressed to three recent large language models (LLMs): ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2.

Methodology: A total of 25 open-ended questions on the subject of ozone in dentistry were asked to three different LLMs (ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2). The chat history was deleted after each response, and a new session was started for each question. All questions were asked only once using the same computer IP address, on the same day, on the same fixed fiber Internet network. The questions were asked to the chatbots in English, and English responses were obtained and recorded. The responses were evaluated independently by two restorative dentistry specialists using two separate Likert scales for accuracy and completeness. In the statistical analyses, the Intraclass Correlation Coefficient (ICC) was calculated, and the Shapiro-Wilk, Kruskal-Wallis, Dunn-Bonferroni, and Spearman tests were applied.

Results: The agreement between the evaluators was found to be high in respect of the accuracy points and varied according to the model for the completeness points. Significant differences were determined between the LLMs in the comparisons of the accuracy scores ($p < 0.001$). The median points of the ChatGPT-5 and DeepSeek-V3.2 models were determined to be statistically significantly higher than those of Gemini 2.5 Flash ($p = 0.002$ and $p < 0.001$, respectively). Significant differences were observed between the LLMs in terms of completeness scores ($p = 0.004$). In the subgroup analyses, the mean points of the ChatGPT-5 and DeepSeek-V3.2 models were determined to be statistically significantly higher than those of Gemini 2.5 Flash ($p = 0.006$ and $p = 0.021$, respectively).

Conclusion: The responses to questions about the use of ozone in dentistry showed differences among the LLMs in terms of accuracy and completeness. ChatGPT-5 and DeepSeek-V3.2 demonstrated better performance than Gemini 2.5 Flash. Despite the high accuracy rates, it must not be forgotten that incorrect healthcare information can lead to serious outcomes. Therefore, it should be recommended that LLMs be used under human supervision in clinical applications and that future research focus on increasing the diagnostic reliability of artificial intelligence.

Keywords: Artificial intelligence, large language models, ChatGPT, Gemini, DeepSeek, ozone, dentistry

Introduction

Artificial intelligence (AI) emerged with the development of the first program by Christopher Strachey in 1951, and John McCarthy initiated the modern age with the term "artificial intelligence" in 1956. In the early years, progress was slow because research focused on rule-based systems with limited data and difficulties in calculations. With the development of machine learning and neural networks in the 1980s and 1990s, AI transformed to systems that could increase performance by learning from data. A symbol of the success of this period was the 1997 chess match defeat of Garry Kasparov by the IBM Deep Blue computer. Since the 2000s, great advances have been recorded in AI in the fields of natural language processing and computer vision (1).

Modern natural language processing (NLP) approaches are based on large language models (LLMs) and are generally AI applications trained on extensive text data of hundreds of terabytes. These are widely used in various scientific disciplines. A subfield of AI is machine learning, which encompasses algorithmic approaches, enabling certain difficulties of computers to be dealt with and insights to be obtained from the data that has been input. Deep learning has become the predominant methodology in this field. AI systems have been examined in the medical sciences in the literature. Different forms of AI are used in dentistry, such as for the detection of oral lesions, the identification of significant points in orthodontics, scheduling, the classification of implants, the determination of root anatomy in endodontics, and the diagnosis of periodontitis (2-4).

The vast majority of studies conducted in the field of dentistry have examined the performance of LLMs, such as ChatGPT, Gemini, and DeepSeek, by evaluating the responses given to questions taken from the national dentistry, dental hygienist, and dental technician qualification examinations. The findings obtained have shown that these models have significant potential in respect of understanding the questions and producing appropriate responses. Thus, it has been suggested that LLMs can serve as assistive educational tools for dental students. However, it has also been emphasized that the performance of the existing models has not yet reached the desired level of excellence (5-9).

ChatGPT (OpenAI, San Francisco, CA, USA) is an NLP model using the latest technology and deep learning techniques to produce responses similar to those of a human. It offers superior capability in accessing and analyzing important information, creating a text with high consistency and accuracy, differentiating categories, and summarizing. As ChatGPT is user-friendly, it has become the preferred tool for tasks such as academic writing and examinations, attracting interest in this field. However, in a study that examined ChatGPT responses, there was seen to be incomplete information in the responses given by ChatGPT on the subject of diagnosis. In another study, it was seen that ChatGPT gave different responses when the same question was asked more than once with the same method (10, 11).

ChatGPT version 4 was launched on the market in March 2024, followed by a further developed version, ChatGPT-4o, in May 2024. The developers stated that these new versions offered the user a more rapid, productive, and multifaceted experience. ChatGPT-4o is distinguished particularly by its ability to both process and produce different types of data, such as text, audio, and visual. Moreover, compared to previous models, this version has shown a significantly faster processing speed, shorter delays, and noteworthy improvements in code processing of text in languages other than English. Finally, while the database of ChatGPT-4 is limited to data up to 2021, the database of ChatGPT-4o has been updated up to 2023 (12).

Gemini is an LLM developed by Google DeepMind (Mountain View, CA, USA), which entered the market on 19 December 2024. By combining advanced language understanding ability with strong reasoning skills, this model was designed to be effective in tasks such as responding to questions, explaining concepts, and solving problems in different areas (13). Introduced to the market on March 26, 2025, Gemini 2.5 was described by its developers as an advanced "thinking model" designed to solve increasingly complex problems. This model family first passes responses through a cognitive reasoning filter rather than producing the response directly from the information. As a result of this approach, Gemini 2.5 is not limited to presenting progression solely in terms of processing power; it also significantly improves the accuracy and reliability of the responses produced (14).

The DeepSeek AI system is an open-access AI platform, developed by DeepSeek AI (Hangzhou DeepSeek AI Basic Tec. Res. Co., Ltd.; Beijing, China), which became accessible on January 5, 2024. It was optimized for academic and professional use by focusing on science, technology, engineering, and mathematics (STEM) fields. The system comprises two specific models. DeepSeek-V3 is suitable for comprehensive NLP tasks such as AI speech, translation, and content production in many languages, and shows superior performance in multi-language text processing and creating lengthy content (13). DeepSeek-V3.1 is an advanced model that symbolizes the start of the "agent age" in AI. Flexible use is provided by combining the "Think" and "Non-Think" modes with a hybrid output architecture. Compared to previous versions, this model is more rapid and productive and has shown significant advances in the skills of tool use and fulfilling multi-step tasks. According to the developers, DeepSeek-V3.1 has created a new standard, not only in terms of increased performance but also in AI research with advanced agent capabilities (14). A new experimental model, DeepSeek-V3.2-Exp, was introduced, and the developers stated that this version was built on V3.1-Terminus and worked more quickly and productively in long contexts owing to DeepSeek Sparse Attention (DSA) technology (15).

Ozone is a three-atom molecule that is used in various fields because of its oxidizing properties and strong antimicrobial effects. Ozone has retained an important place in dentistry applications, as clinicians are continuously seeking innovative solutions to improve oral health. When it is considered that dental diseases are increasing overall worldwide, there is a need for the use of assistive therapies, such as ozone, for more effective treatments (16, 17).

Ozone is used in dentistry in three forms; ozonated water, ozonated olive oil, and ozone gas. These forms can be used alone or in combination to treat the oral tissues. When ozone is applied for just a few minutes, it eliminates the bacteria causing decay. By penetrating fissures and dentin tubules, ozone also has a relaxing effect on the nerve endings of the sensory organs and stimulates blood flow. Therefore, ozone is an effective painkiller. At high concentrations and in different forms, ozone has antimicrobial effects against bacteria, fungi, and viruses. Unlike many other known antiseptics, ozone does not destroy human skin. Moreover, teeth are remineralized and become hard tissue again in the weeks after treatment with ozone therapy. The use of ozone with remineralizing agents facilitates the complete removal of decayed tissue (18, 19).

LLMs do not actually have knowledge like humans do. For example, to the question, “Who first introduced the use of phosphoric acid abrasion in dental bonding?”, LLMs may give the correct answer of “Dr Michael Bunocore”, but this response is not due to the model actually knowing the subject but because this is the most likely answer as the most frequently written in many texts produced by humans. In short, LLMs produce a response based on statistical patterns in data; they cannot understand information and only make probability calculations. Nevertheless, LLMs are extremely effective, as they can produce fluent and consistent texts, respond to questions, make translations, and successfully fulfil tasks based on different languages (20).

The aim of this study was to evaluate the responses of three recently developed LLMs (ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2) to open-ended questions about the use of ozone in dentistry by comparing the responses in terms of accuracy and completeness. The first null hypothesis was that there would be no difference between ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 in terms of the accuracy of the responses related to ozone knowledge. The second null hypothesis was that there would be no difference between ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 in terms of the completeness of the responses related to ozone knowledge.

Materials and Methods

As this cross-sectional study did not include any patient information or data, and the aim was to analyze freely accessible online information sources, it was exempted from ethical approval by the institutional review boards of Dicle University Dentistry Faculty.

A dentist with 15 years of experience and a specialist in restorative dental treatment and endodontics developed a question pool formed from 25 open-ended questions, which conformed to the literature related to the use of ozone in dentistry (16, 21-24). Before being addressed to the LLMs, the questions were carefully examined for grammatical accuracy, appropriate structure, and readability, and, if necessary, corrected by two of the authors (S.C. and Ö.A.). The 25 questions started with the chemical properties of ozone and then encompassed its applications in dentistry, effect mechanism, material interactions, clinical outcomes, contraindications, and comparative evaluations. The prepared questions were recorded in a Microsoft Word document (Microsoft Corp., Version 16.0, Redmond, WA, USA) (Table 1).

To achieve consistency, all procedures were performed on the same day using a single laptop computer with a cable Internet connection on the same software and network. This was done to minimize the effect of network fluctuations on the results. All data clusters from the personal Internet browser were deleted before the research began.

On October 14, 2025, all questions were copied and pasted manually to newly created accounts on the following platforms: ChatGPT-5 Auto mode (OpenAI, https://chatgpt.com/), Gemini 2.5 Flash Deep Research mode (Google, https://gemini.google.com/), and DeepSeek-V3.2 DeepThink mode (Hangzhou DeepSeek, https://chat.deepseek.com/).

All questions were addressed to the chatbots on October 14, 2025, by a volunteer who was independent of the research team. To minimize learning bias, be able to determine accuracy more clearly, and eliminate the potential effects of previous interactions, the following precautions were taken: (a) a new account was created and used for access to each chatbot, (b) no positive or negative feedback was given for any of the responses given, (c) before asking each new question, the previous chats were deleted, and a new chat session was not started, thus allowing the models to focus only on the current input, (d) no pre-test, instant improvement, or additional intervention was performed, (e) no production parameters (including heat or sampling settings) were changed, (f) each model had the right to only one response, and it was not possible to change the responses, reproduce them, or make an additional request. Accordingly, the “recreate the response” option was disabled in the chat environment. Thus, the first responses obtained from the models were recorded objectively.

The interactions in the scope of the study were performed only with the predefined English questions, and the English language was used at both the entry and exit stages.

The responses obtained from the three AI models were anonymized and presented to a restorative dental treatment specialist and an endodontics specialist, each with 15 years of experience. To provide neutrality of the

evaluation process, the researchers were blinded to which responses belonged to which model. Before being presented to the researchers for evaluation, the responses of each chatbot were recorded on a separate Microsoft Word document (Microsoft Corp., Redmond, WA, USA) and labelled Chatbot 1, Chatbot 2, and Chatbot 3.

In the evaluation process, the literature from which the questions were prepared was used as the reference standard (16, 21-24). These reference texts were presented to the specialist evaluators along with anonymized responses. Each evaluator used a standard scoring table, prepared in Microsoft Excel (Microsoft, Redmond, WA, USA), to ensure a structured and consistent scoring process.

In the evaluations made by the specialists, the responses were analyzed in the two main dimensions of content accuracy and completeness.

The accuracy of the responses given by the LLMs was evaluated using a 5-point Likert scale, and the completeness of the responses was evaluated using a 3-point Likert scale (14, 25, 26) (Table 2). To minimize the risk of bias, the evaluation process was conducted using a double-blind design. The detailed workflow of this study is shown in Fig. 1.

Statistical analysis

The data obtained in the study were analyzed statistically using SPSS 25.0 software (IBM, Armonk, NY, USA).

The Intraclass Correlation Coefficient (ICC) was calculated to evaluate the agreement in scoring between the two evaluators. The conformity of continuous variables to a normal distribution was assessed using the Shapiro-Wilk test. Descriptive statistics were stated as mean ± standard deviation (SD) values for variables with normal distribution and as median (minimum-maximum) values for data not showing normal distribution.

The Kruskal-Wallis test was applied to compare more than two groups when the distribution was not normal. When overall significance was determined after the Kruskal-Wallis test, subgroup analyses were performed using the Dunn-Bonferroni test.

The Spearman correlation coefficient was calculated to examine the relationship between accuracy and completeness. Type 1 error was accepted at 5%, and the level of statistical significance was set at $p < 0.05$.

Table 1. The questions asked of the chatbots.

Questions	
1.	What is ozone therapy?
2.	What are the applications of ozone in dentistry?
3.	In what form does ozone occur naturally?
4.	How is ozone obtained?
5.	What are ozone gas production systems?
6.	What are the application methods of ozone therapy?
7.	What are the industrial applications of ozone?
8.	What is the mechanism of action of ozone?
9.	Why is anesthesia not required during ozone application in dentistry?
10.	What is the effect of ozone on biofilms compared to other cavity disinfectants?
11.	At what concentrations is ozone used in dentistry?
12.	How does the application of ozone to the cavity affect the bonding of restorative materials?
13.	How long does the antiseptic effect of ozone last in dentistry?
14.	What are the contraindications for ozone use in dentistry?
15.	How does the application of ozone affect the surface properties of restorative materials?
16.	Does the use of ozone as a cavity disinfectant affect microleakage?
17.	What are the differences between classic caries removal and cavity preparation methods using ozone?
18.	How does the treatment time with ozone compare to traditional caries removal methods?
19.	Which is more successful in teeth with initial enamel caries: fluoride, laser, or ozone applications?
20.	Is ozone effective on its own in cavitated caries lesions?
21.	How does the effectiveness of ozonated oils and gels compare to traditional desensitizing agents such as potassium nitrate?
22.	What are the current limitations on the application of ozone therapy in dental practice?
23.	How does the duration of ozone contact with tissues affect the rate of killing of pathogenic micro-organisms?
24.	What are the side-effects of ozone?
25.	How is ozone intoxication treated?
Total 25 questions	

Table 2. Evaluation dimensions and Likert scales.

Scale dimension	Score	Definition
Accuracy	1= very poor	The responses are not consistent with the facts, with serious inconsistencies and missing information; poor flow and impaired integrity of meaning.
	2= poor	Some correct elements are present but there is missing or incomplete information, weak and fragmented expression.
	3= moderate	Generally understandable and partially correct; some details are missing or unclear, incomplete integration
	4= good	Mostly correct and consistent; fluent with only small deficiencies or insignificant errors.
	5= excellent	Completely consistent with the facts, comprehensive, fluent, and well-organized: fully integrated information with no important errors.
Completeness of information	1= incomplete	The response only addresses some aspects of the question; basic elements or important information is missing.
	2= sufficient	The response encompasses all the important components of the question and includes the necessary information but does not provide additional context.
	3= comprehensive	The response completely meets all the elements of the question and exceeds expectations, presenting additional context or related information.

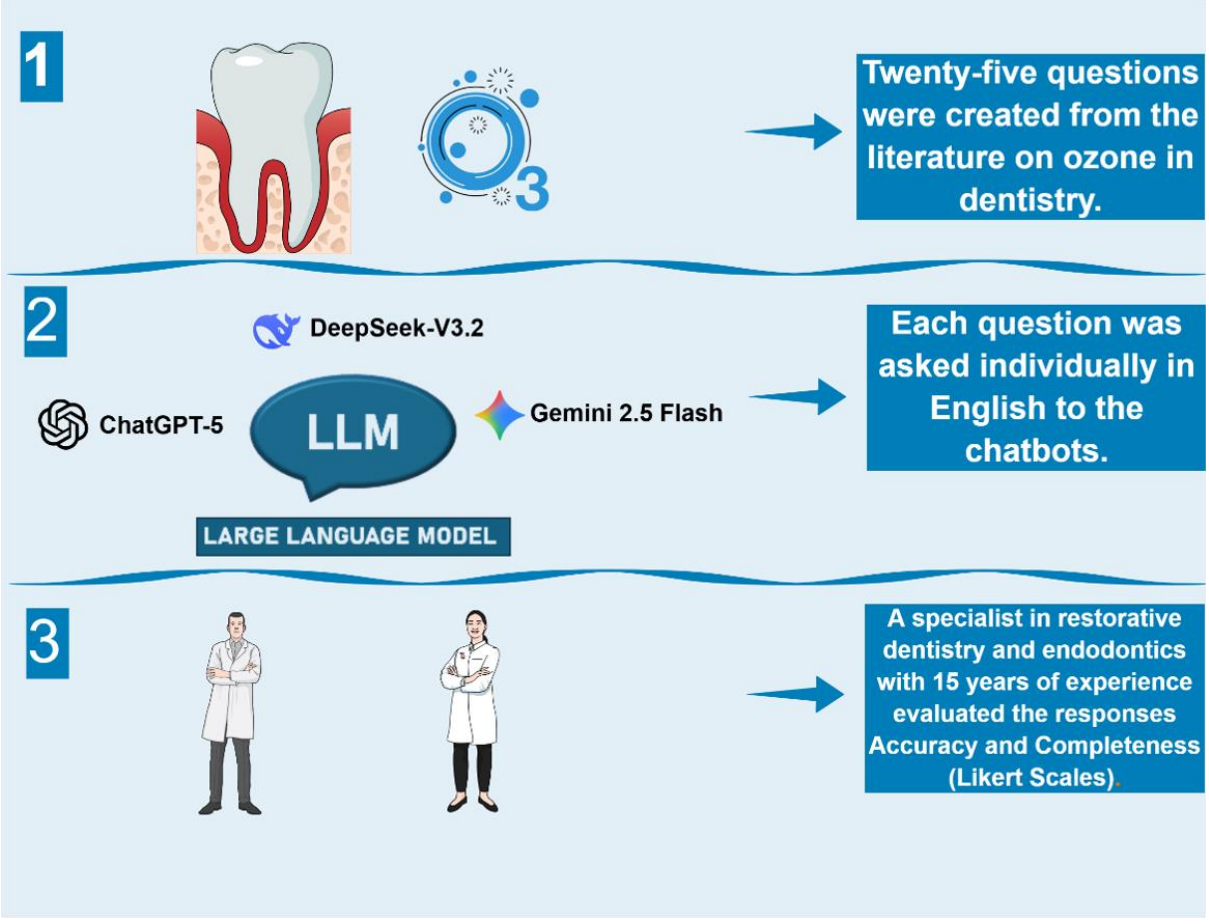

Figure 1. Study design and an overview of the workflow of the evaluation process.

Results

The inter-evaluator agreement of the accuracy points was evaluated using the two-way consistency type single measurement Intraclass Correlation Coefficient (ICC 2,1). The levels of inter-evaluator agreement for all the models were found to be statistically significant ($p < 0.001$). The ICC was calculated to be 0.786 for ChatGPT-5 (95% CI: 0.572–0.900, $p < 0.001$), 0.872 for Gemini 2.5 Flash (95% CI: 0.732–0.942, $p < 0.001$), and 0.938 for DeepSeek V3.2 (95% CI: 0.866–0.972, $p < 0.001$). These

results demonstrated a high level of reliability between the evaluators for all the models (Table 3).

The inter-evaluator agreement of the completeness points was evaluated using the two-way consistency type single measurement Intraclass Correlation Coefficient (ICC 2,1). The ICC value was calculated as 1.000 for ChatGPT-5, indicating an excellent level of agreement between the evaluators. The ICC was found to be 0.780 for Gemini 2.5 Flash (95% CI: 0.562–0.897, $p < 0.001$), and 0.657 for DeepSeek V3.2 (95% CI: 0.361–0.833, $p < 0.001$). These results demonstrated an excellent level of reliability between the evaluators in respect of the completeness points for ChatGPT-5, a high level for Gemini 2.5 Flash, and a moderate level for DeepSeek-V3.2 (Table 4).

The accuracy scores of the responses given by the three AI models to the open-ended questions related to the use of ozone in dentistry were compared, and significant differences were observed ($p < 0.001$). In the subgroup analyses, the median points of Gemini 2.5 Flash were determined to be 4 (min-max: 2-5), which was statistically significantly lower than the points of 5 (min-max: 3-5) for ChatGPT-5, and 5 (min-max:3-5) for DeepSeek-V3.2 ($p = 0.002$, $p > 0.001$, respectively). No significant difference was observed in the other paired group comparisons ($p > 0.05$) (Table 5) (Fig. 2).

The completeness scores of the responses given by the three AI models to the open-ended questions related to the use of ozone in dentistry were compared, and significant differences were observed ($p = 0.004$). In the subgroup analyses, the mean points of Gemini 2.5 Flash were determined to be 2.6 ± 0.58, which was statistically significantly lower than the mean points of 2.96 ± 0.2 for ChatGPT-5, and 2.92 ± 0.28 for DeepSeek-V3.2 ($p = 0.006$, $p = 0.021$, respectively). No significant differences were observed in the other paired group comparisons ($p > 0.05$) (Table 5) (Fig. 3).

A statistically significant positive correlation was found between the accuracy and completeness scores ($rs = 0.72$; $p < 0.001$). These findings demonstrated that an increase in the accuracy score was related to an increase in the completeness score (Table 6).

Table 3. Inter-evaluator agreement of the accuracy points.

	ICC(2,1)	95%CI	p-value
Model			
ChatGPT-5	0.786	0.572-0.900	<0.001
Gemini 2.5 Flash	0.872	0.732-0.942	<0.001
DeepSeek-V3.2	0.938	0.866-0.972	<0.001

Table 4. Inter-evaluator agreement of the completeness points.

	ICC (2,1)	95%CI	p-value
Model			
ChatGPT-5	1.000	1.000-1.000	<0.001
Gemini 2.5 Flash	0.780	0.562-0.897	<0.001
DeepSeek-V3.2	0.657	0.361-0.833	<0.001

Table 5. Comparison of accuracy and completeness scores for ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2.

		Chatbot			
		ChatGPT-5	Gemini 2.5 Flash	DeepSeek-V3.2	p-value
Accuracy	Mean ± SD	4.64±0.57	3.84±0.85	4.72±0.61	<0.001[a]
	Median (min.-max.)	5(3-5)	4(2-5)	5(3-5)	
Completeness	Mean ± SD	2.96±0.2	2.6±0.58	2.92±0.28	0.004[a]
	Median (min.-max.)	3(2-3)	3(1-3)	3(2-3)	

Data are expressed as mean ± standard deviation and median (minimum: maximum) values.
a: Kruskal-Wallis test

Table 6. Spearman correlation analysis between accuracy and completeness scores for ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1.

	Accuracy
Completeness	
rs	0.72
p value	<0.001

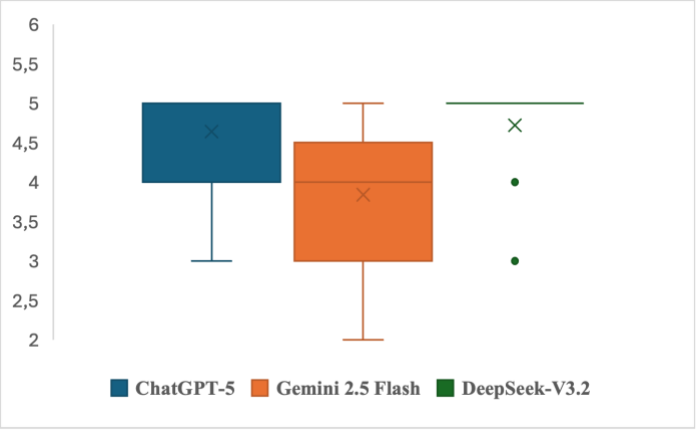

Figure 2. Mean accuracy scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 in responding to questions related to the use of ozone in dentistry.

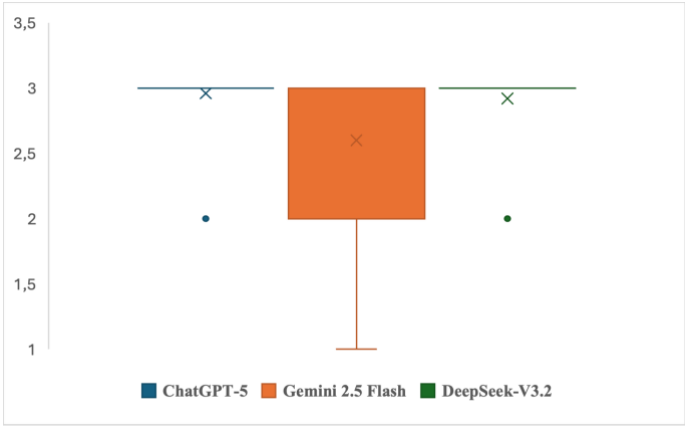

Figure 3. Mean completeness scores of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 in responding to questions related to the use of ozone in dentistry.

Discussion

The results obtained in this study demonstrate significant differences among the ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2 AI chatbots in terms of the accuracy and completeness of responses produced to questions related to knowledge of ozone in dentistry, and thus, both the first and second null hypotheses of the study were rejected. In terms of accuracy and completeness, the median scores of Gemini 2.5 Flash were significantly lower than those of ChatGPT-5 and DeepSeek-V3.2. It was also seen that an increase in the accuracy score was related to an increase in the completeness score.

The development of AI-based LLMs, such as ChatGPT, Google Gemini, and the recently prominent DeepSeek, has created an important transformation in the field of education, primarily medical education. Students of medicine, dentistry, and other health sciences benefit from these systems in the learning processes, preparation of assignments, examination studies, and various other academic activities. This is due to the capacity of these models to immediately present personalized information that is research-focused and consistent with the lesson content. As these LLMs are trained on extensive datasets, they have high levels of ability to analyze questions, understand, and produce current and consistent responses (27, 28).

The current study, which evaluated the information accuracy and completeness of responses of AI-based LLMs in the field of dentistry, is specific research to fill a gap in the literature. Previous studies have compared the performances of models such as ChatGPT, Gemini, and DeepSeek on various dentistry subjects, including fixed prostheses, dental implantology, traumatic dental injuries, saliva gland cancer, and pulp treatment in immature permanent teeth, and significant findings have been obtained related to the quality of the responses of these models in these fields (29-33). However, there is no study in the literature that has examined the levels of accuracy and completeness of these models in respect of the use of ozone in dentistry.

Thus, the current study is the first to have analyzed the responses to questions about ozone in dentistry to evaluate the performance of three advanced-level LLMs, which are the most up-to-date and most frequently encountered in the literature, namely ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2. Therefore, the results of this study make a significant scientific and practical contribution to both increasing the accuracy of clinical information and to the development of LLM-based educational and decision-making support systems.

In a study of pulp treatment in immature permanent teeth, it was reported that in terms of accuracy and completeness, the ChatGPT-4o and DeepSeek models showed a better performance than the Gemini Advanced model (31). Although the subject and the versions of the models used were different (ie, the focus was on ozone application and new-generation models were used), the results of the current study showed that ChatGPT-5 and DeepSeek-V3.2 demonstrated better performance than Gemini 2.5 Flash in terms of both accuracy and completeness. These findings emphasize the critical importance of model selection of LLMs to obtain information on subjects that include multidisciplinary approaches, such as ozone therapy, as in this study, and other dentistry subjects.

In a previous study that compared the performance of five LLMs, ChatGPT 4.5, DeepSeek R1, Gemini 2.5 Pro, Claude 3.7 Sonnet, and Microsoft Copilot, in the treatment planning of restoration of teeth that had undergone endodontic treatment, the highest performance was obtained from Gemini 2.5 Pro in all sessions, far exceeding both DeepSeek R1 and Microsoft Copilot to a significant degree (34). In contrast to previous findings, the accuracy and completeness scores of DeepSeek-V3.2 in the current study were found to be significantly higher than those of Gemini 2.5 Flash. Although different areas of dentistry were examined using LLMs, these findings show that not only the model

of AI chatbots but also different versions could have a determinant effect on the results obtained. It is thought that updates in the knowledge base of the model versions, algorithmic improvements, and developments in the language processing capacity could create differences in terms of accuracy and completeness. Therefore, in scientific studies, including AI-based analyses, the knowledge of the model version should be evaluated with care, as much as the model selection. It is of great importance that in similar studies in the future, both the model selected and differences in the version used are evaluated as a methodological variable in respect of the validity and comparability of the results.

Marcaccini et al. compared the ChatGPT, DeepSeek, and Gemini platforms in terms of diagnosis and treatment strategies for hand fractures. The highest accuracy rate was obtained by ChatGPT-4o (98.28%), followed by DeepSeek-V3 (63.79%), and the lowest accuracy rate was shown by Gemini 1.5 (18.97%). The researchers emphasized that these findings indicated significant limitations, especially in terms of clinical decision support (35). Similarly, in the current study, Gemini 2.5 Flash showed the lowest performance.

The latest generation of LLMs has shown noteworthy advances in the field of medicine. Comprehensive comparative analyses of the USMLE and other international medical qualification examinations have shown that the GPT-4 and GPT-4o models are superior to previous versions, such as GPT-3.5, and to alternative LLMs that have demonstrated a significantly lower performance (36). This trend shows that LLMs have become increasingly more reliable in medical specialization examinations and have reached a certain level of success in evaluating clinical information.

In a study by Latkowska et al., the performance of ChatGPT-5 was evaluated in the Polish endocrinology specialization examination, and it was reported to have passed the threshold by providing correct answers to 91 of 119 questions (76.47%) and incorrect answers to 28 (23.53%). The researchers stated that this result indicated a notable improvement in AI technology, but it was also emphasized that for this finding to be confirmed, there was a need for further studies to be conducted in different medical specialism areas and with various test materials (37). The high accuracy rate of the ChatGPT-5 model in the current study is consistent with the findings of Latkowska et al., and supports that new-generation LLMs exhibit a more consistent and reliable performance in the evaluation of clinical information. However, whether these successes directly reflect clinical decision-making processes and are effective at a similar level in real clinical scenarios should be evaluated more comprehensively in future studies.

In the evaluation made by Barnes et al., published in the study entitled, “Can medical students use artificial intelligence to learn transfusion? Evaluating ChatGPT responses to the American Society of Hematology medical student transfusion learning objectives”, the authors emphasized that the results were based on data obtained from an old model, and this led to incorrect interpretations and a potentially invalid educational tool. It was also stated that there was a lack of comparison with other LLMs (eg, Claude and Gemini), which was a significant methodological deficiency. In more recent evaluations conducted to overcome this limitation, analyses of Claude Opus 4 and Gemini 2.5 Pro models were included, in addition to ChatGPT-5. The results showed that the highest performance was obtained from ChatGPT-5 with a mean of 3.52 points, followed by Gemini 2.5 with 3.23 points, and Claude with 3.03 points (38). Although the current study dealt with a different subject the use of ozone in dentistry and used different question types, the performance of ChatGPT-5 in responding to these questions was determined to be higher than that of Gemini 2.5 Flash. These findings demonstrate that ChatGPT-5 can exhibit a superior performance to that of Gemini 2.5 Flash, especially in medicine and dentistry education and in clinical decision support processes. However, there are insufficient studies in the current literature that have comprehensively compared ChatGPT-5 with other LLMs. Therefore, further large-scale, detailed analyses encompassing different subjects are needed.

Limitations

There were some limitations of this study to be considered. First, the study was limited to three LLMs (ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.2), and other current models were not included in the evaluation. A second limitation was that the evaluation process was only conducted on responses in the English language, and the quality of responses in other languages may show variability. The fact that the questions were asked in a single session and a single response was evaluated can also be considered a limitation, as this may not reflect differences in learning or consistency, which could be seen in different interaction scenarios of the models.

Conclusion

In this study, the levels of accuracy and completeness of responses to questions about the use of ozone in dentistry were examined through comparisons of three LLMs. Statistically significantly higher accuracy and completeness points were obtained from ChatGPT-5 and DeepSeek-V3.2 than from Gemini 2.5 Flash. Despite the high accuracy scores obtained in this study, it must not be forgotten that AI models can sometimes produce incorrect or incomplete information. Therefore, it is recommended that AI-supported information systems are not used directly in clinical decision-making processes but as complementary tools under the supervision of a specialist. There remains a need for further studies to focus on increasing the diagnostic reliability and contextual understanding capacity of AI models in clinical areas such as dentistry.

Disclosures

Peer-review: Externally peer-reviewed.

Author Contributions: Conception — S.C.; Design — S.C., M.T.; Supervision — Ö.A.; Materials — M.T., T.T.; Data Collection and/or Processing — T.T., H.O.; Analysis and/or Interpretation — H.O.; Literature Review — S.C., Ö.A.; Writer — S.C.; Critical Review — Ö.A.

Conflict of Interest: No potential conflict of interest relevant to this article was reported.

Funding: None to declare.

References

1. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689. https://doi.org/10.1186/s12909-023-04698-z
2. Ozdemir ZM, Yapici E. Evaluating the accuracy, reliability, consistency, and readability of different large language models in restorative dentistry. J Esthet Restor Dent. 2025. https://doi.org/10.1111/jerd.13447
3. Mohammad-Rahimi H, Rokhshad R, Bencharit S, Krois J, Schwendicke F. Deep learning: a primer for dentists and dental researchers. J Dent. 2023;130:104430. https://doi.org/10.1016/j.jdent.2023.104430
4. Çelik B, Çelik ME. Automated detection of dental restorations using deep learning on panoramic radiographs. Dentomaxillofac Radiol. 2022;51(8):20220244. https://doi.org/10.1259/dmfr.20220244
5. Lin CCC, Sun JS, Chang CH, Chang YH, Chang JZC. Performance of artificial intelligence chatbots in National dental licensing examination. J Dent Sci. 2025;20(4):2307-2314. https://doi.org/10.1016/j.jds.2025.05.012
6. Wu YH, Tso KY, Chiang CP. Performance of ChatGPT in answering the oral pathology questions of various types or subjects from Taiwan National Dental Licensing Examinations. J Dent Sci. 2025;20(3):1709-1715. https://doi.org/10.1016/j.jds.2025.03.030
7. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. Cureus. 2025;17(1):e77292. https://doi.org/10.7759/cureus.77292
8. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. J Prosthet Dent. 2025;134(6):2427-2434. https://doi.org/10.1016/j.prosdent.2024.12.016
9. Fukuda H, Morishita M, Muraoka K, et al. Evaluating the image recognition capabilities of GPT-4V and Gemini Pro in the Japanese national dental examination. J Dent Sci. 2025;20(1):368-372. https://doi.org/10.1016/j.jds.2024.06.015
10. Kim HW, Shin DH, Kim J, Lee GH, Cho JW. Assessing the performance of ChatGPT's responses to questions related to epilepsy: a cross-sectional study on natural language processing and medical information retrieval. Seizure. 2024;114:1-8. https://doi.org/10.1016/j.seizure.2023.11.013
11. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. J Med Internet Res. 2023;25:e48009. https://doi.org/10.2196/48009
12. Arılı Öztürk E, Turan Gökduman C, Çanakçi BC. Evaluation of the performance of ChatGPT-4 and ChatGPT-4o as a learning tool in endodontics. Int Endod J. 2025. https://doi.org/10.1111/iej.14217
13. Huang CY, Lee YP, Sun A, Chiang CP. Performance of ChatGPT-4, Gemini, and DeepSeek-V3 on answering the multiple choice questions from Taiwan national dental technician licensing examinations and their self-learning abilities over a three-week period. J Dent Sci. 2025;20(3):2154-2162. https://doi.org/10.1016/j.jds.2025.07.011
14. Taşyürek M, Adıgüzel Ö, Gündoğar M, Goncharuk-Khomyn M, Ortaç H. Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. Int Dent Res. 2025;15(3):123-135.https://doi.org/10.5577/intdentres.662
15. DeepSeek. Introducing DeepSeek-V3.2-Exp. Accessed March 6, 2026. https://api-docs.deepseek.com/news/news250929
16. Rezaeianjam M, Khabazian A, Khabazian T, et al. Efficacy of ozone therapy in dentistry with approach of healing, pain management, and therapeutic outcomes: a systematic review of clinical trials. BMC Oral Health. 2025;25(1):433. https://doi.org/10.1186/s12903-025-05790-0
17. Tricarico G, Rodrigues Orlandin J, Rocchetti V, Ambrosio C, Travagli V. A critical evaluation of the use of ozone and its derivatives in dentistry. Eur Rev Med Pharmacol Sci. 2020;24(17):9071-9093. https://doi.org/10.26355/eurrev_202009_22854
18. Balcı Yüce H, Uçan Yarkaç F, Tulu F. Clinical Parameters and CRP Levels in Generalized Aggressive Periodontitis Patient: Case Report. Int Dent Res 2018;8(1):39-44. https://doi.org/10.5577/intdentres.2018.vol8.no1.7
19. Smith NL, Wilson AL, Gandhi J, Vatsia S, Khan SA. Ozone therapy: an overview of pharmacodynamics, current research, and clinical utility. Med Gas Res. 2017;7(3):212-219. https://doi.org/10.4103/2045-9912.215752
20. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098-1102. https://doi.org/10.1111/jerd.13046
21. Liaqat S, Tariq S, Hayat I, et al. Therapeutic effects and uses of ozone in dentistry: a systematic review. Ozone Sci Eng. 2023;45(4):387-397. https://doi.org/10.1080/01919512.2022.2125363
22. El Meligy OA, Elemam NM, Talaat IM. Ozone therapy in medicine and dentistry: a review of the literature. Dent J (Basel). 2023;11(8):187. https://doi.org/10.3390/dj11080187
23. Veneri F, Lanteri V, Consolo U, Checchi V, Generali L. Ozone in dentistry: an updated overview of current applications and future perspectives. Curr Oral Health Rep. 2025;12(1):3. https://doi.org/10.1007/s40496-024-00395-y
24. Dinc Ata G, Mujdeci A. Effect of ozone on bond strength of different restorative materials to enamel and dentin. Ozone Sci Eng. 2023;45(1):89-101. https://doi.org/10.1080/01919512.2022.2115976
25. De Vito A, Colpani A, Moi G, et al. Assessing ChatGPT's potential in HIV prevention communication: a comprehensive evaluation of accuracy, completeness, and inclusivity. AIDS Behav. 2024;28(8):2746-2754. https://doi.org/10.1007/s10461-024-04391-2
26. Çîtîr M. ChatGPT and oral cancer: a study on informational reliability. BMC Oral Health. 2025;25(1):86. https://doi.org/10.1186/s12903-025-05479-4
27. Erdem Hepşenoğlu Y, Değirmencioğlu D, Topbaş C. Analyzing biomimetic dentistry YouTube videos. Int Dent Res. 2023;13(S1):44-49. https://doi.org/10.5577/idr.2023.vol13.s1.7
28. Singal A, Goyal S. Comparative evaluation of AI platforms "Google Gemini 2.5 Flash, Google Gemini 2.0 Flash, DeepSeek V3 and ChatGPT 4o" in solving multiple-choice questions from

different subtopics of anatomy. Surg Radiol Anat. 2025; 47(1):1-8. https://doi.org/10.1007/s00276-025-03707-8
29. Wu X, Cai G, Guo B, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. BMC Oral Health. 2025;25(1):1272. https://doi.org/10.1186/s12903-025-06619-6
30. Sezer B, Aydoğdu T. 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. 2025;15(14):7778. https://doi.org/10.3390/app15147778
31. Sezer B, Aydoğdu T. Performance of advanced artificial intelligence models in pulp therapy for immature permanent teeth: a comparison of ChatGPT-4 Omni, DeepSeek, and Gemini Advanced in accuracy, completeness, response time, and readability. J Endod. 2025;51(11):1675-1684. https://doi.org/10.1016/j.joen.2025.08.011
32. Bashah A, Salem A, Al-Waqeerah A, et al. Evaluation of DeepSeek, Gemini, ChatGPT-4o, and Perplexity in responding to salivary gland cancer. BMC Oral Health. 2025;25(1):1-15. https://doi.org/10.1186/s12903-025-06726-4
33. Shirani M. Comparing the performance of ChatGPT 4o, DeepSeek R1, and Gemini 2 Pro in answering fixed prosthodontics questions over time. J Prosthet Dent. 2025. https://doi.org/10.1016/j.prosdent.2025.04.038
34. Shirani M, Emami M. Performance comparison of large language models in treatment planning for the restoration of endodontically treated teeth over time. J Dent. 2025:105998. https://doi.org/10.1016/j.jdent.2025.105998
35. Marcaccini G, Seth I, Xie Y, et al. Breaking bones, breaking barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. J Clin Med. 2025;14(6):1983. https://doi.org/10.3390/jcm14061983
36. Chen Y, Huang X, Yang F, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ. 2024;24(1):1372. https://doi.org/10.1186/s12909-024-06309-x
37. Latkowska A, Sawina P, Dolata T, et al. Assessment of the ability of the ChatGPT-5 model to pass the endocrinology specialization exam. Cureus. 2025;17(9): e93479. https://doi.org/10.7759/cureus.93479
38. Troy-Barnes E, Shen L, Alimam S. Comment on McBride et al.'s 'Can medical students use artificial intelligence to learn transfusion? Evaluating ChatGPT responses to the American Society of Hematology medical student transfusion learning objectives'. Vox Sang. 2025;121(1):92-93. https://doi.org/10.7759/cureus.93479