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

*EADS*, 2025, 52 (2), 71–78

e-ISSN : 2757-6744
Article Received/Accepted : April, 18 2025 / June, 17 2025
doi : 10.52037/eads.2025.0011

ORIGINAL RESEARCH ARTICLE

# ChatGPT vs Google Gemini: Assessment of Performance Regarding the Accuracy and Repeatability of Responses to Questions in Implant-Supported Prostheses

Deniz Yılmaz [1] and Emine Dilara Çolpak [1], *

[1]Department of Prosthodontics, Faculty of Dentistry, Alanya Alaaddin Keykubat University, Antalya, Turkiye

*Corresponding Author; dilara.colpak@alanya.edu.tr

## Abstract

**Purpose:** This study aimed to assess the accuracy and repeatability of the responses of different large language models (LLMs) to questions regarding implant-supported prostheses and assess the impact of pre-prompting and the time of day.
**Materials and Methods:** A total of 12 open-ended questions related to implant-supported prostheses were generated. The content validity of questions was verified by a specialist. Following that, questions were posed to two different LLMs: ChatGPT-4.0 and Google Gemini (morning, afternoon, and evening; with and without pre-prompt). The responses were evaluated by two expert prosthodontists with a holistic rubric. The concordance between the graders' responses and repeated responses by ChatGPT-4.0 and Gemini was calculated using the Brennan and Prediger coefficient, Cohen's kappa coefficient, Fleiss's kappa, and Krippendorff's alpha coefficients. Kruskal-Wallis, Mann-Whitney U, and independent t-test, as well as ANOVA analyses, were used to compare the responses obtained in the implementations.
**Results:** The results displayed that the accuracies of ChatGPT and Google Gemini were 34.7% and 17.4%, respectively. The implementation of pre-prompt significantly increased accuracy in Gemini ($p = 0.026$). No significant difference was found according to the time of day (morning, afternoon, or evening) or inter-week implementations. In addition, inter-rater reliability and repeatability displayed high levels of consistency.
**Conclusions:** The use of pre-prompt positively affected accuracy and repeatability in both ChatGPT and Google Gemini. However, LLMs can still produce hallucinations. Therefore, LLMs may assist clinicians, but they should be aware of these limitations.

**Keywords**: Chatbot; ChatGPT; Prostheses; Implant

## Introduction

Large language models (LLMs) are enhanced artificial intelligence (AI) systems replicating human language processing skills by training on large datasets. LLMs are based on natural language processing (NLP) and machine learning, an aspect of AI that aims to enable computers to understand natural language input. [1]

Healthcare providers frequently use LLMs to address questions related to patient care management due to their access to numerous articles, textbooks, and guidelines, along with their rapid information retrieval capabilities and 24/7 availability. [2,3]A systematic review of the literature evaluating the performance of LLMs in answering medical questions revealed that the overall accuracy was 56%. [4] The effectiveness of LLMs in responding to medical questions has been assessed variably across different medical specialties. Concerns about using these models in healthcare contexts include inaccurate and unreliable responses, the risk of bias, and ethical and legal considerations. [5] In dentistry, AI has the potential to assist dental professionals with diagnosis, treatment planning, image analysis, outcome prediction, record keeping, and workflow efficiency. [6,7] In prosthodontics, AI provides a wide range of applications, including implant-supported and maxillofacial prostheses, computer-aided design, and computer-aided manufacturing (CAD-CAM), as well as fixed and removable prostheses. [8–10] In a systematic review, Revilla-León et al. observed that AI models were potential instruments for implant type recognition, implant success prediction, and implant design optimization. [11] However, the performance of AI-assisted LLMs in generating accurate responses to questions regarding implant-supported prostheses has been insufficiently assessed. [12] Furthermore, LLMs were not specifically created to offer medical advice and may generate misinformation or disinformation responses for clinical decisions that appear coherent but lack significant meaning. These responses are reportedly called hallucinations. [13–15] Therefore, the assessment of LLM per-

formance becomes crucial considering the increasing awareness of their capacity to answer dental questions. [8]

Various organizations and companies, including ChatGPT (Chat Generative Pre-trained Transformer; OpenAI, San Francisco, California, USA), Gemini (Google, Mountain View, California, USA), Copilot (Microsoft, Redmond, Washington, USA), Meta AI (Facebook, Menlo Park, California, USA), and DeepSeek (DeepSeek Artificial Intelligence Co., Beijing, China), have developed LLMs. [6,15–17] Among LLM tools, chatbot platforms like ChatGPT and Gemini have received the most attention and show the potential to comprehend clinical expertise and deliver relevant information. [18,19] The effectiveness and accuracy of these LLMs often depend on their training, expertise, model updates, and question complexity. [20] ChatGPT, an AI language model developed by OpenAI, is based on generative pre-trained transformer (GPT) architecture. ChatGPT-4.0 was launched in February 2023. [21–23] It is also a retrained transformative AI model that can incorporate human feedback into the training process. On the other hand, Gemini is an advanced AI model, introduced in December 2023 by Google DeepMind, that uses a transformer-based architecture. Gemini Advanced was launched in February 2024. It is Google's next-generation AI model with a one million token context window and improvements in logical reasoning, coding, and creative collaboration over Gemini. [22]

This study aimed to assess the performance of two different LLMs [ChatGPT-4.0 (C) and Google Gemini Pro Advanced 1.5 (G)] in terms of accuracy and repeatability regarding implant-supported prostheses in Turkish-language responses. The null hypotheses were that there would be no difference in the (1) accuracy or (2) repeatability of responses regarding implant-supported prostheses information between C and G.

## Material and Methods

Twelve specific questions related to implant-supported prostheses were generated to evaluate the accuracy and repeatability of the C and G software programs' responses. This study did not require ethical approval because no human participants were involved, and no personal data were collected.

Two experienced prosthodontists (D.Y. and E.D.C.) generated a total of 15 open-ended questions using the Fixed Bridges and Dental Implants Guidelines [24] published by The British Society for Restorative Dentistry and then translated them into Turkish. A measurement and evaluation specialist (D.K.) collaborated to verify the content validity of the questions. The questions were edited based on the feedback provided (Table 1). Twelve questions were selected in accordance with expert opinions, and necessary arrangements were made. A small pilot study was conducted to determine the comprehensibility of the questions. These questions were asked to both chatbot software in January 2025 from two different computers connected to the same internet network at the same time, according to the parameters presented in Table 2. Twelve questions were asked to both C and G software programs at three different times on the same day (in the morning, afternoon, and evening) twice at one-week intervals. Another condition considered in the study was to compare the responses of AI chatbots with and without pre-prompting. For these reasons, both conditions were manipulated, using the following pre-prompt to provide the chatbots with a perspective: "I would like you to answer my questions as a prosthodontist." In the non-pre-prompted case, the chatbots were asked the questions directly without any explanation or perspective. A total of 24 randomized order questions were answered using the "new chat" option in each implementation to minimize memory retention bias and reset the memory. These questions were administered 12 times for the C software program and 12 times for the G software program. The responses were graded by two prosthodontists affiliated with the Department of Prosthodontics, Faculty of Dentistry, Alanya Alaaddin Keykubat University, using the holistic rubric in Table 3, and mean values were calculated. The holistic rubric was used to evaluate the answers to the questions due to the higher grading reliability (objective scoring) and internal validity of the rubrics compared to Likert-type scales in assessing cognitive characteristics such as knowledge and ability. [25]

The performance of C and G software programs was evaluated by calculating accuracy and repeatability. For the purposes of this study, accuracy was defined as the ratio of accurate responses to total repetitions within the set of questions created by C and G software programs, and its 95% confidence interval was calculated using the Wald binomial technique. [26,27] The assessment of repeatability required performing concordance analyses that were weighted for ordinal categories, incorporating multiple repetitions of the gradings provided by the experts. This included evaluating the Brennan and Prediger coefficient, Cohen's kappa coefficient, Fleiss's kappa, and Krippendorff's alpha coefficients, along with their respective 95% confidence intervals. The estimated coefficients were categorized based on the benchmark scale suggested by Gwet: <0.0 Poor, 0.0–0.2 Slight, 0.2–0.4 Fair, 0.4–0.6 Moderate, 0.6–0.8 Substantial, 0.8–1.0 Almost Perfect. [28]

In determining the technique for hypothesis testing, the normality assumption was first examined with Kolmogorov-Smirnov and Shapiro-Wilk tests. Based on the normality test results, parametric or nonparametric techniques were used. Accordingly, the Mann-Whitney U test was used to determine whether there was a significant difference between the means of the responses obtained in the implementations (implementations 1-24) and the accuracy of the C and G, because the data were not normally distributed. In the comparison of the grades obtained from C and G software programs with and without pre-prompting, as well as in the comparison of the responses received initially and one week later, each answer was analyzed with an independent t-test due to the normal distribution of the data. A one-way analysis of variance was used to compare the responses obtained during morning, afternoon, and evening implementations. The concordance between the graders' responses and repeated responses by C and G software programs was calculated with the Brennan and Prediger coefficient, Cohen's kappa coefficient, Fleiss's kappa, and Krippendorff's alpha coefficients. The data were analyzed with a statistical software program (IBM SPSS Statistics for Windows, v25.0; IBM Corp., Armonk, New York, USA) for accuracy assessments, while another statistical software program (R Foundation for Statistical Computing, R Core Team, Vienna, Austria) was used for repeatability analysis ($p<.05$).

## Results

Using C and G software programs, a total of 288 responses were generated for each of the 144 questions asked before and after one week at different hours of the day. Figure 1 shows the distribution of responses generated by the chatbots. Figure 2 shows the distribution of mean responses generated from the implementations. Figure 3 shows the distribution of the number of questions answered with completely correct responses from the implementations.

### Comparison of Implementations Across All Conditions

Statistically significant p-values were found by comparing the implementations with the Kruskal-Wallis test (Table 4). There were no significant differences between the other implementation groups ($p > 0.050$).

### Effect of Pre-prompting Within Weeks

When analyzing the effect of pre-prompting on the accuracy of responses generated by chatbots within the same weeks, there were no statistically significant differences between non-pre-prompted

**Table 1.** Questions used in the study

| Question Number | Question Text* |
|---|---|
| 1 | What factors should be considered when selecting implant systems for implant-supported fixed crown and bridge restorations? |
| 2 | What factors should be considered during the intraoral examination for implant-supported fixed crown and bridge restorations? |
| 3 | What factors should be considered when evaluating edentulous spaces in dental implant planning? |
| 4 | What factors affect the number and position of implants in dental implant planning? |
| 5 | How many implants are required in the maxilla and mandible for implant-supported fixed crown and bridge restorations and overdentures in completely edentulous patients? |
| 6 | What are the differences between immediate implant placement and immediate loading protocols? |
| 7 | What factors influence the decision between cement-retained and screw-retained abutments in implant-supported fixed crown and bridge restorations? |
| 8 | What systematic procedure should be followed for the delivery of screw-retained, implant-supported fixed crown and bridge restorations? |
| 9 | What systematic procedure should be followed for the delivery of cement-retained implant-supported fixed crown and bridge restorations? |
| 10 | What radiological criteria should be considered before the fabrication of implant-supported prostheses? |
| 11 | What are the clinical and radiographic success criteria for implant-supported prostheses? |
| 12 | What criteria should be considered in the cantilever design of implant-supported fixed crown and bridge restorations? |

* The questions presented were translated into Turkish.

**Table 2.** Definitions regarding question implementations

| Time | Preprompt | Chatbot | Initial | After One week |
|---|---|---|---|---|
| Morning | Pre-prompt | Gemini Pro Advanced 1.5 | Implementation 1 (I1) | Implementation 13 (I13) |
| | | ChatGPT 4.0 | Implementation 2 (I2) | Implementation 14 (I14) |
| | Non-pre-prompt | Gemini Pro Advanced 1.5 | Implementation 3 (I3) | Implementation 15 (I15) |
| | | ChatGPT 4.0 | Implementation 4 (I4) | Implementation 16 (I16) |
| Afternoon | Pre-prompt | Gemini Pro Advanced 1.5 | Implementation 5 (I5) | Implementation 17 (I17) |
| | | ChatGPT 4.0 | Implementation 6 (I6) | Implementation 18 (I18) |
| | Non-pre-prompt | Gemini Pro Advanced 1.5 | Implementation 7 (I7) | Implementation 19 (I19) |
| | | ChatGPT 4.0 | Implementation 8 (I8) | Implementation 20 (I20) |
| Evening | Pre-prompt | Gemini Pro Advanced 1.5 | Implementation 9 (I9) | Implementation 21 (I21) |
| | | ChatGPT 4.0 | Implementation 10 (I10) | Implementation 22 (I22) |
| | Non-pre-prompt | Gemini Pro Advanced 1.5 | Implementation 11 (I11) | Implementation 23 (I23) |
| | | ChatGPT 4.0 | Implementation 12 (I12) | Implementation 24 (I24) |

I, Implementation

**Table 3.** Holistic Rubric for answers

| Holistic Rubric Criteria | |
|---|---|
| **Completely Incorrect Response (1 Point)** | Content: The response is entirely unrelated to the main topic of the question or contains incorrect information.<br>Scientific Basis: The provided information is scientifically inaccurate or inconsistent with reliable sources.<br>Language and Terminology: Incorrect or inappropriate terminology is used.<br>Clarity: The response is unclear and difficult to understand, failing to convey meaning to the reader. |
| **Correct Response Containing Errors (2 Points)** | Content: The response is partially relevant to the main topic but includes incorrect information.<br>Scientific Basis: While containing some correct information, it also includes erroneous details that contradict the accurate ones.<br>Language and Terminology: The terminology is generally correct, but some inaccuracies are present.<br>Clarity: The response is comprehensible; however, the presence of incorrect information diminishes the overall accuracy. |
| **Partially Correct Response (3 Points)** | Content: The response is largely relevant to the main topic but lacks certain key details or is insufficiently developed.<br>Scientific Basis: The provided information is accurate but does not comprehensively address the topic.<br>Language and Terminology: The terminology is correctly used, but explanations need further elaboration.<br>Clarity: The response is clear and understandable; however, the missing information prevents a thorough evaluation. |
| **Completely Correct Response (4 Points)** | Content: The response comprehensively and thoroughly addresses the main topic of the question.<br>Scientific Basis: The information is accurate, up-to-date, and aligned with reliable sources.<br>Language and Terminology: The terminology is used correctly and in a professional manner.<br>Clarity: The response is clear, well-structured, and easy to understand, presenting information in a logical sequence. |

**Table 4.** Statistically significant p-values by comparisons of implementations using the Kruskal-Wallis Test and the accuracy percentages of implementations

| Implementations | | | | | Accuracy Percentages (%) | I1 | I4 | I5 | I6 |
|---|---|---|---|---|---|---|---|---|---|
| Initial | nonpreprompt | Morning (INM) | G | I1 | 66.5 | | | | |
| | | | C | I2 | 79 | 0.068 | | | |
| | | Afternoon (INA) | G | I3 | 70.75 | 0.597 | | | |
| | | | C | I4 | 79 | 0.038* | | | |
| | | Evening (INE) | G | I5 | 66.5 | 1.000 | 0.038* | | |
| | | | C | I6 | 85.25 | 0.007* | 0.284 | 0.007* | |
| | preprompt | Morning (IPM) | G | I7 | 75 | 0.164 | 0.488 | 0.164 | 0.107 |
| | | | C | I8 | 81.25 | 0.041* | 0.675 | 0.041* | 0.590 |
| | | Afternoon (IPA) | G | I9 | 79 | 0.038* | 1.000 | 0.038* | 0.284 |
| | | | C | I10 | 83.25 | 0.013* | 0.468 | 0.013* | 0.724 |
| | | Evening (IPE) | G | I11 | 79 | 0.038* | 1.000 | 0.038* | 0.284 |
| | | | C | I12 | 85.25 | 0.038* | 1.000 | 0.038* | 0.284 |
| After 1 week | nonpreprompt | Morning (ANM) | G | I13 | 70.75 | 0.484 | 0.166 | 0.484 | 0.032* |
| | | | C | I14 | 81.25 | 0.023* | 0.710 | 0.023* | 0.479 |
| | | Afternoon (ANA) | G | I15 | 81.25 | 0.023* | 0.710 | 0.023* | 0.479 |
| | | | C | I16 | 77 | 0.159 | 0.823 | 0.159 | 0.288 |
| | | Evening (ANE) | G | I17 | 75 | 0.164 | 0.488 | 0.164 | 0.107 |
| | | | C | I18 | 75 | 0.238 | 0.557 | 0.238 | 0.158 |
| | preprompt | Morning (APM) | G | I19 | 77 | 0.107 | 0.763 | 0.107 | 0.214 |
| | | | C | I20 | 77 | 0.159 | 0.823 | 0.159 | 0.288 |
| | | Afternoon (APA) | G | I21 | 77 | 0.107 | 0.763 | 0.107 | 0.214 |
| | | | C | I22 | 81.25 | 0.041* | 0.675 | 0.041* | 0.590 |
| | | Evening (APE) | G | I23 | 77 | 0.060 | 0.691 | 0.060 | 0.146 |
| | | | C | I24 | 81.25 | 0.023* | 0.710 | 0.023* | 0.479 |

*Statistical significance was set at p<.05. I, Implementation; G, Google Gemini Pro Advanced 1.5; C, Chat GPT 4.0; INM, Initial Non-pre-prompt Morning; INA, Initial Non-pre-prompt Afternoon; INE, Initial Non-pre-prompt Evening; IPM, Initial Pre-prompt Morning; INA, Initial Pre-prompt Afternoon; INE, Initial Pre-prompt Evening; ANM, After One Week Non-pre-prompt Morning; ANA, After One Week Non-pre-prompt Afternoon; ANE, After One Week Non-pre-prompt Evening; APM, After One Week Pre-prompt Morning; ANA, After One Week Pre-prompt Afternoon; ANE, After One Week Pre-prompt Evening

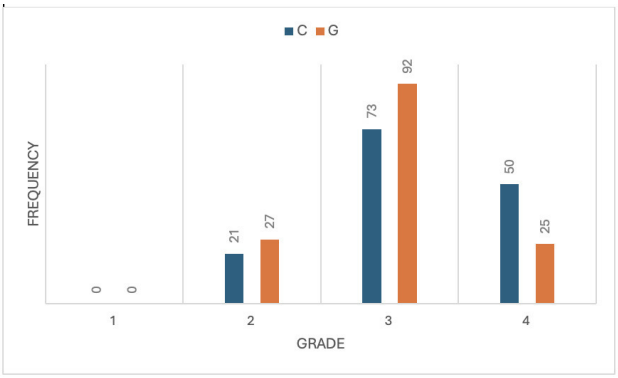


**Figure 1.** Distribution of responses generated by chatbots

G, Google Gemini Pro Advanced 1.5; C, Chat GPT 4.0

C and pre-prompted C (p = 0.460). However, a statistically significant difference was found between non-pre-prompted G and pre-prompted G (p = 0.026).

## Effect of Pre-prompting Throughout the Day

When analyzing the effect of pre-prompting on the accuracy of chatbot responses throughout the day, the analysis showed no statistically significant differences for non-pre-prompted C (p = 0.889), pre-prompted C (p = 0.676), non-pre-prompted G (p = 0.229), and pre-prompted G (p = 0.854).

## Effect of Time of Day on Accuracy (Without Pre-prompt)

The analysis showed no statistically significant difference in the accuracy of answers provided by both C (p = 0.822) and G (p = 0.314) when evaluating responses from the chatbots throughout the day (morning, afternoon, and evening), without the effect of pre-prompt variables.

## Week-to-Week Comparison Without Pre-prompt

The analysis of chatbot accuracy across different weeks, without the effect of pre-prompting, also revealed no statistically significant differences for ChatGPT between the initial week (p = 0.889) and one week later (p = 0.676). Similarly, Gemini's performance showed no significant difference between the initial implementation (p = 0.229) and one week later (p = 0.854).

## General Effect of Pre-prompting

Overall, the effect of pre-prompting on chatbot accuracy, when analyzed irrespective of time or repetition, was not statistically significant for either ChatGPT (p = 0.217) or Gemini (p = 0.217).

## Week-Based Analysis of Pre-prompting

When the effect of pre-prompting was examined within weeks, there were no significant differences for non-pre-prompted C (p = 0.583), pre-prompted C (p = 0.620), or pre-prompted G (p = 0.699). However, a statistically significant difference was found in the non-pre-prompted G (p = 0.005), indicating variation in response accuracy over time.

## Inter-rater Reliability and Repeatability

Table 5 presents the inter-rater reliability and repeatability coefficients based on expert evaluations. The results demonstrate a high level of agreement among the raters across all statistical indices. Specifically, the Brennan and Prediger ($\kappa$ = 0.87, SE = 0.01, 95% CI:

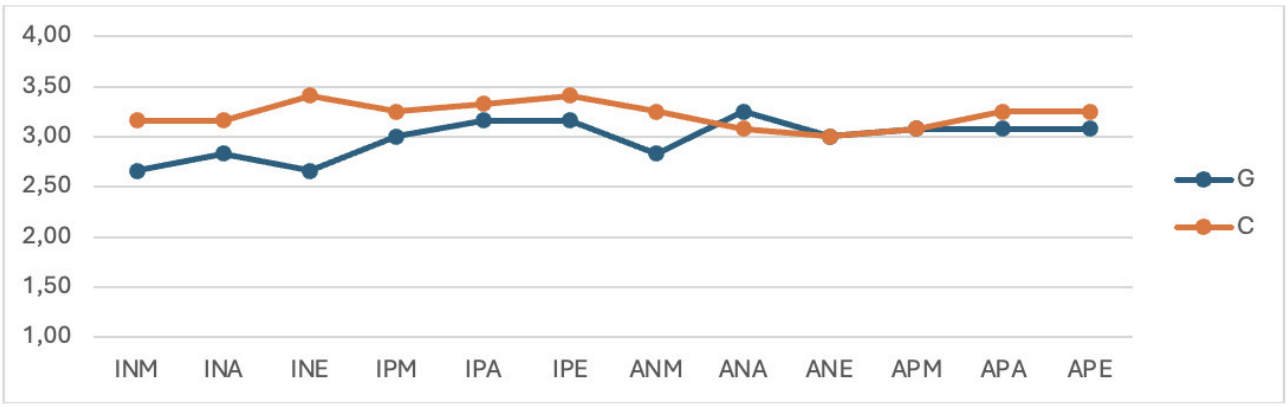


**Figure 2.** Distribution of average responses generated by the implementations

G, Google Gemini Pro Advanced 1.5; C, Chat GPT 4.0; INM, Initial Non-pre-prompt Morning; INA, Initial Non-pre-prompt Afternoon; INE, Initial Non-pre-prompt Evening; IPM, Initial Preprompt Morning; INA, Initial Pre-prompt Afternoon; INE, Initial Pre-prompt Evening; ANM, After One Week Non-pre-prompt Morning; ANA, After One Week Non-pre-prompt Afternoon; ANE, After One Week Non-pre-prompt Evening; APM, After One Week Pre-prompt Morning; ANA, After One Week Pre-prompt Afternoon; ANE, After One Week Pre-prompt Evening

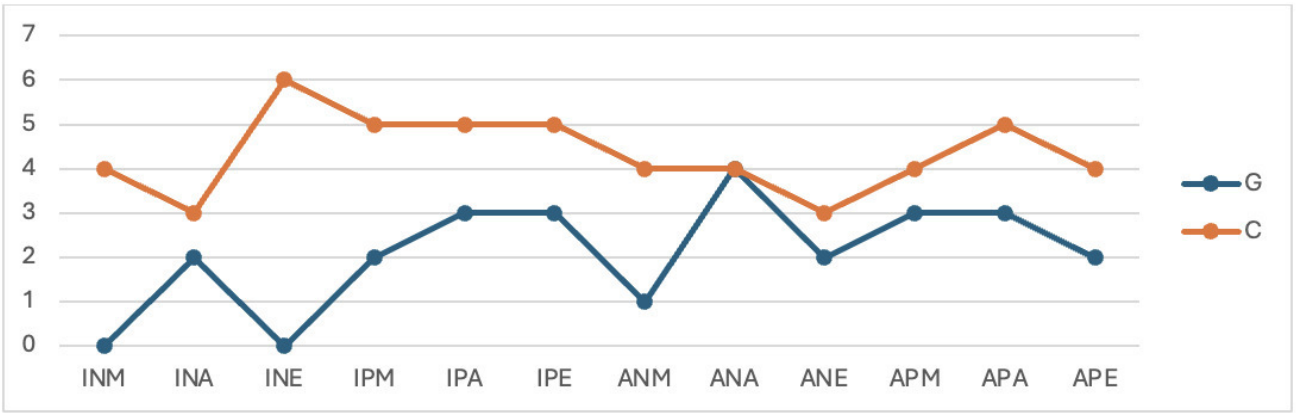


**Figure 3.** Distribution of the number of questions completely correctly answered from the implementations

G, Google Gemini Pro Advanced 1.5; C, Chat GPT 4.0; INM, Initial Non-pre-prompt Morning; INA, Initial Non-pre-prompt Afternoon; INE, Initial Non-pre-prompt Evening; IPM, Initial Preprompt Morning; INA, Initial Pre-prompt Afternoon; INE, Initial Pre-prompt Evening; ANM, After One Week Non-pre-prompt Morning; ANA, After One Week Non-pre-prompt Afternoon; ANE, After One Week Non-pre-prompt Evening; APM, After One Week Pre-prompt Morning; ANA, After One Week Pre-prompt Afternoon; ANE, After One Week Pre-prompt Evening

0.83–0.91), Cohen's kappa ($\kappa$ = 0.85, SE = 0.01, 95% CI: 0.81–0.89), and Fleiss's kappa ($\kappa$ = 0.81, SE = 0.02, 95% CI: 0.76–0.85) coefficients all indicated "almost perfect" agreement based on Gwet's benchmark classification. Krippendorff's alpha yielded a slightly lower coefficient ($\alpha$ = 0.78, SE = 0.06, 95% CI: 0.74–0.82), which still corresponds to a "substantial" level of agreement. These findings confirm that the expert scoring was consistent and repeatable, supporting the internal reliability of the applied holistic rubric.

## Discussion

This study assessed the response performance of chatbots C and G in answering questions related to implant-supported prostheses in Turkish under different conditions. The results showed that the respective accuracies of C and G were 34.7% and 17.4%, while the repeatability of both chatbots ranged from "substantial" to "almost perfect." Although both chatbots were highly reliable and consistent in their responses, there was a significant difference in their accuracy. Therefore, the first null hypothesis was rejected, while the second null hypothesis was accepted.

The Turkish response means C and G changed in different time periods and both pre-prompt and non-pre-prompt situations. For non-pre-prompt situations, C consistently exhibited higher means than G, with a notable performance increase, especially during the afternoon and evening hours. Conversely, G's responses maintained a more stable pattern, demonstrating reliability. For pre-prompt conditions, C's performance showed significant improvement, particularly during the morning and afternoon, although variations were noted depending on the times. However, G maintained consistent performance for pre-prompt conditions. Because pre-prompts present an acceptable alternative, domain-specific training is crucial to improving LLM performance in healthcare, particularly with the development of models including expanded token limitations.[29] In addition, although both C and G are LLMs, their different architectural designs may also cause these differences; thus, even under the same conditions, their output may differ in terms of accuracy. ChatGPT's GPT-4 architecture leverages reinforcement learning from human feedback (RLHF), allowing it to generate more adaptable and nuanced responses. By contrast, Gemini is based on Google's LaMDA architecture, which emphasizes dialogic coherence and contextual understanding. These foundational differences may account for the observed variation in performance across tasks. Whereas ChatGPT uses a method of deep learning that involves fine-tuning specific tasks on large datasets and is trained using the GPT (Generative Pretrained Transformer) architecture,

**Table 5.** Repeatability and coefficients based on expert grading

| Methods | Coefficient | SE | 95%CI (Range) | | Benchmark Scale |
|---|---|---|---|---|---|
| Brennan and Prediger | 0.87 | 0.01 | 0.83 | 0.91 | Almost Perfect |
| Cohen kappa | 0.85 | 0.01 | 0.81 | 0.89 | Almost Perfect |
| Fleiss kappa | 0.81 | 0.02 | 0.76 | 0.85 | Almost Perfect |
| Krippendorff alpha | 0.78 | 0.06 | 0.74 | 0.82 | Substantial |

Benchmark scale: Poor <0.0, Slight 0.0-0.2, Fair 0.2-0.4, Moderate 0.4-0.6, Substantial 0.6-0.8, and Almost Perfect 0.8-1.0. CI, confidence interval; SE, standard error.

Gemini primarily provides a better understanding of context and is built on Google's LaMDA (Language Model for Dialogue Implementation) neural network architecture. Because of the differences in network designs and training data, LLMs may provide completely different results when asked the same questions, illuminating different aspects. Whereas ChatGPT may generate more varied results because its training set is bigger. Gemini uses the most current data for training; however, ChatGPT-3.5 is limited to data up to September 2021, which means that replies to more recent events or developments may not be accurate or appropriate.[20,30] Compared to its previous version, C's 175 billion parameters allow it to provide more accurate and context-sensitive answers to complex medical questions.[15] Sanderson reported that the C does not explain exactly how it works, what data it uses, how the model works, and its enhanced capabilities, although the risk of producing inaccurate information (hallucination) is not completely eliminated.[23] Because hallucinations are still a problem, the need for professional surveillance persists.[8,13] Hallucinations are characterized by confident yet reality-based responses and present significant risks in clinical environments, which may result in misdiagnoses or inappropriate treatment options. Even with advancements in model improvements, the problem of hallucinations remains unsolved, requiring clinician surveillance in the application of LLMs for implant-supported prostheses.

In the present study, C performed better than G in terms of the number of completely correct responses in all time periods and conditions in Turkish responses. In addition, C's performance increased more significantly, especially in the pre-prompt condition. However, G showed lower performance in the non-pre-prompt conditions but a more balanced increase when the pre-prompt was used. In terms of time periods, it was found that both chatbots performed relatively poorly in the morning hours, whereas their performance was higher in the afternoon and evening hours. These findings indicate that pre-prompting improves response quality in both C and G, although this effect is more pronounced in C. In terms of time of day, morning hours have a diminishing effect on performance, whereas chatbots provide higher-quality responses during the afternoon and evening hours. These results provide important clues for understanding how chatbot performance varies with time and conditions.

The accuracy of responses generated by LLMs depends on the quantity, quality, and variety of data utilized during their training process.[14] Using a pre-prompt and prompting with detailed knowledge of the literature enhances this accuracy.[9,29] The results of this study are consistent with the findings of Gheisarifar et al.[8] regarding patients' frequently asked questions in prosthodontics, Chatzopoulos et al.[6] regarding clinically relevant questions in periodontology, Özdemir and Yapici[31] regarding restorative dentistry, and Rokhshad et al.[32] regarding pediatric dentistry.

The AIs may generate different answers to the same questions on different days and at different times of the day. In contrast to the present study, Freire et al.,[12] using open-ended prosthodontic questions, found that C was inadequate in providing accurate (25.6%) and consistent (ranging from substantial to moderate) answers regarding removable prostheses and tooth-supported fixed prostheses.[12] Another study on implant dentistry reported that Gemini showed higher reliability and usefulness grades in comparison to ChatGPT-3.5 and ChatGPT-4.0 in closed questions.[19] On the other hand, ChatGPT-3 achieved 100% accuracy in defining radiographic landmarks.[33] The discrepancies in the findings may result from variations in issues relevant to different dental specialties, the content and type of the questions asked, time-based updates, divergences in the pre-prompts used, and differences in the language used to ask the questions.

Variations were observed even though the responses to the repeated questions in Turkish showed an acceptable level of repeatability. The findings indicate that the "almost perfect" agreement observed in the Brennan and Prediger's, Cohen's kappa, and Fleiss's kappa coefficient results suggests that raters evaluated the questions similarly, demonstrating high repeatability. By contrast, the slightly lower Krippendorff's alpha value may be attributable to this coefficient's sensitivity to different data distributions. However, this does not compromise the overall reliability of the study. In terms of inter-rater reliability, it was found that two graders scored the responses in a previous study[12], and a third specialist grader was consulted when there was a discrepancy between the graders. Therefore, in this study, the responses were scored by two graders, and because no inconsistencies were observed, the opinion of a third expert was not required. This methodological design is consistent with that of other studies.[6,8,12,19]

In terms of repeatability, consistent with the present study, Taymour et al.[18] reported that ChatGPT-3.5, ChatGPT-4.0, and Google Gemini chatbots showed acceptable levels of reliability in dental implant-related questions, and Rokhshad et al.[32] reported that Google Bard, ChatGPT-4.0, Llama, Sage, Claude 2 100k, Claude-instant, Claude-instant-100k, and Google Palm chatbots displayed high levels of consistency in pediatric dentistry questions. This repeatability aligns with findings from other studies, which indicate that AI-generated responses may vary in categories besides maintaining factual accuracy.[8] According to Gheisarifar et al.,[8] consistency evaluation showed that although ChatGPT-3.5 did not meet acceptable consistency levels for patients' commonly asked questions in prosthodontics, the Gemini, ChatGPT-4.0, and Bing chatbots did. Furthermore, ChatGPT-4.0 showed the highest consistency compared to other chatbots. These findings strongly indicate that the holistic rubric used in the present study was well-designed, the rater training was effective, and the grading process was reliable regarding repeatability.

The limitations of this study include the specificity of the subject concerning implant-supported prostheses in Turkish, and the limited number of reviewers. It is not advisable to make assumptions about other topics. Another limitation was that only the most commonly selected chatbots were used, which exhibited differences in architecture, updates, and training data; however, new chatbots have been introduced recently. The limited sample of chatbots, the study's exclusive focus on the Turkish language, and the lack of visual-based question types limit the generalizability of the results. Moreover, even when achieving a high level of agreement among

raters, the potential of human subjectivity may influence the objectivity of the ratings. Future studies should build upon these findings by utilizing larger datasets and incorporating different question types, including images, to develop a more comprehensive understanding of chatbot performance.

## Conclusion

Based on the findings of this study, the following conclusions were drawn:

- 1. The use of pre-prompting improved accuracy and repeatability in both C and G, although this effect was more pronounced in C.
- 2. C gave a more accurate response in the evening, whereas G showed improved accuracy in the afternoon.
- 3. Although the answers generated by G were less accurate than those of C, G has the ability to update faster than C.
- 4. Both C and G have limitations in providing accurate and repeatable responses regarding implant-supported prostheses. Clinicians should be aware of these limitations.

Although LLMs demonstrate promising reliability levels, their limited accuracy in domain-specific tasks such as those relating to implant-supported prostheses underlines the necessity of human verification. Incorporating structured pre-prompting and time-aware usage may enhance their future utility in clinical practice.

## Ethical Approval

This study did not require ethical approval because no human participants were involved, and no personal data were collected.

## Acknowledgements

Not applicable.

## Financial Support

No funding was received for this study.

## Author Contributions

Methodology : All authors
Investigation : All authors
Writing : All authors
Editing : D.Y.

## Conflict of Interest

We have no conflict of interest.

## Authors' ORCID(s)

D.Y. 0000-0003-4570-9067
E.D.C. 0000-0002-5334-2421

## References

1. Eggmann F, Blatz MB. ChatGPT: Chances and Challenges for Dentistry. Compend Contin Educ Dent. 2023;44(4):220–224.
2. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. Ann Intern Med. 2024;177(2):210–220. doi:10.7326/m23-2772.
3. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–180. doi:10.1038/s41586-023-06291-2.
4. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: A systematic review and meta-analysis. J Biomed Inform. 2024;151:104620. doi:10.1016/j.jbi.2024.104620.
5. Khan B, Fatima H, Qureshi A, Kumar S, Hanan A, Hussain J, et al. Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector. Biomed Mater Devices. 2023:1–8. doi:10.1007/s44174-023-00063-2.
6. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: Assessing their performance in clinically relevant questions. J Prosthet Dent. 2024. doi:10.1016/j.prosdent.2024.10.020.
7. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res. 2020;99(7):769–774. doi:10.1177/0022034520915714.
8. Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, Afshari FS, Yuan JC, et al. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. J Prosthet Dent. 2025;134(1):199–206. doi:10.1016/j.prosdent.2025.03.009.
9. Sadowsky SJ. Can ChatGPT be trusted as a resource for a scholarly article on treatment planning implant-supported prostheses? J Prosthet Dent. 2025. doi:10.1016/j.prosdent.2025.03.025.
10. Singi SR, Sathe S, Reche AR, Sibal A, Mantri N. Extended Arm of Precision in Prosthodontics: Artificial Intelligence. Cureus. 2022;14(11):e30962. doi:10.7759/cureus.30962.
11. Revilla-León M, Gómez-Polo M, Vyas S, Barmak AB, Gallucci GO, Att W, et al. Artificial intelligence models for tooth-supported fixed and removable prosthodontics: A systematic review. J Prosthet Dent. 2023;129(2):276–292. doi:10.1016/j.prosdent.2021.06.001.
12. Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. J Prosthet Dent. 2024;131(4):659.e1–659.e6. doi:10.1016/j.prosdent.2024.01.018.
13. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023;15(2):e35179. doi:10.7759/cureus.35179.
14. 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. doi:10.1111/jerd.13046.
15. Stroop A, Stroop T, Zawy Alsofy S, Wegner M, Nakamura M, Stroop R. Assessing GPT-4's accuracy in answering clinical pharmacological questions on pain therapy. Br J Clin Pharmacol. 2025. doi:10.1002/bcp.70036.
16. Hosseini M, Gao CA, Liebovitz DM, Carvalho AM, Ahmad FS, Luo Y, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. PLoS One. 2023;18(10):e0292216. doi:10.1371/journal.pone.0292216.
17. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023;11(6). doi:10.3390/healthcare11060887.
18. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery. 2023;93(5):1090–1098. doi:10.1227/neu.0000000000002551.

19. 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. doi:10.1016/j.prosdent.2024.12.016.
20. Tokgöz Kaplan T, Cankar M. Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini. Dent Traumatol. 2025;41(2):178–186. doi:10.1111/edt.12999.
21. Barrington NM, Gupta N, Musmar B, Doyle D, Panico N, Godbole N, et al. A Bibliometric Analysis of the Rise of ChatGPT in Medical Research. Med Sci (Basel). 2023;11(3). doi:10.3390/medsci11030061.
22. Google. Google Gemini: Next-generation Model [Web Page]; 2024. Available from: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/.
23. Sanderson K. GPT-4 is here: what scientists think. Nature. 2023;615(7954):773. doi:10.1038/d41586-023-00816-5.
24. Dentistry BSfR. Crowns, Fixed Bridges and Dental Implants: Guidelines. United Kingdom: British Society for Restorative Dentistry; 2013.
25. Koçak D. Investigation of Rater Tendencies and Reliability in Different Assessment Methods with Many Facet Rasch Model. lnternational Electronic Journal of Elementary Education. 2020;12:349–358. doi:10.26822/iejee.2020459464.
26. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108–113. doi:10.1111/iej.13985.
27. Suárez A, Jiménez J, Llorente de Pedro M, Andreu-Vázquez C, Díaz-Flores García V, Gómez Sánchez M, et al. Beyond the Scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. Comput Struct Biotechnol J. 2024;24:46–52. doi:10.1016/j.csbj.2023.11.058.
28. Gwet KL. Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters. Advanced Analytics, LLC; 2014.
29. Rewthamrongsris P, Burapacheep J, Trachoo V, Porntaveetus T. Accuracy of Large Language Models for Infective Endocarditis Prophylaxis in Dental Procedures. Int Dent J. 2025;75(1):206–212. doi:10.1016/j.identj.2024.09.033.
30. 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. doi:10.1093/ejo/cjae017.
31. Ozdemir ZM, Yapici E. Evaluating the Accuracy, Reliability, Consistency, and Readability of Different Large Language Models in Restorative Dentistry. J Esthet Restor Dent. 2025;37(7):1740–1752. doi:10.1111/jerd.13447.
32. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. J Dent. 2024;144:104938. doi:10.1016/j.jdent.2024.104938.
33. Mago J, Sharma M. The Potential Usefulness of ChatGPT in Oral and Maxillofacial Radiology. Cureus. 2023;15(7):e42133. doi:10.7759/cureus.42133.