DOI: http://doi.org/10.54617/adoklinikbilimler.1698260

Original Research Article

Benchmarking Different Natural Language Processing Models for Their Responses to Queries on Tooth-supported Fixed Dental Prostheses in Terms of Accuracy and Consistency

Emine Dilara Çolpak [1], **Deniz Yılmaz** [2]

[1]DDS, MSc, Assistant Professor, Department of Prosthodontics, Faculty of Dentistry, Alanya Alaaddin Keykubat University, Antalya, Turkiye. [2] DDS, PhD Assistant Professor, Department of Prosthodontics, Faculty of Dentistry, Alanya Alaaddin Keykubat University, Antalya, Turkiye.

ABSTRACT

Aim: This study aimed to evaluate the accuracy and consistency of responses generated by four different natural language processing (NLP) models to the queries on tooth-supported fixed dental prostheses.

Materials and Method: Twelve open-ended questions in Turkish were created and posed to four different NLPs according to the following models: OpenAI o3 (LRM-O), OpenAI GPT 4.5 (LLM-G), DeepSeek R1 (LRM-R), and DeepSeek V3 (LLM-V) with pre-prompts in the morning, afternoon, and evening. The responses were evaluated with a holistic rubric. For accuracy assessments, the Kruskal–Wallis H test was used. Consistency between the graders' responses was assessed using the Brennan and Prediger coefficient and the Cohen kappa coefficient. Consistency among LLMs was assessed using the Fleiss kappa and Krippendorff alpha coefficients ($p < 0.05$).

Results: There was no statistically significant difference in accuracy between the LRM-O, LLM-G, LRM-R, and LLM-V groups ($p = 0.30$). The respective accuracies of LRM-O, LLM-G, LRM-R, and LLM-V were 77.7%, 50%, 66.6%, and 77.7%. In addition, the consistency among LLMs was found to be almost perfect, whereas that of LRMs was substantial.

Conclusion: Within the limitations of the study, LRMs and LLMs exhibited similar accuracy. However, the consistency among LLMs was higher than that of LRMs.

Received:13.05.2025; Accepted: 02.08.2025
Corresponder author: Dr. Emine Dilara Çolpak
Alanya Alaaddin Keykubat University, Faculty of Dentistry, Department of Prosthodontics, Antalya, Turkiye
E-mail: dilara.colpak@alanya.edu.tr

Keywords: Artificial intelligence; Dental prostheses; Treatment protocols

Citation: Çolpak E. D., Yılmaz D. Benchmarking Different Natural Language Processing Models for Their Responses to Queries on Tooth-supported Fixed Dental Prostheses in Terms of Accuracy and Consistency ADO Kilinik Bilimler Dergisi 2025;14(3):215-223

Editor: Sinem Akgül, Gazi University, Ankara, Türkiye

INTRODUCTION

In recent years, there has been growing interest in artificial intelligence (AI) tools, such as chatbots, within the field of prosthodontics. Particularly in areas with limited specialized training, such as oral health, these technologies may assist dental professionals in diagnosing and developing treatment protocols and predicting treatment outcomes, thereby enhancing workflow efficiency.[1] The launch and development of AI and large language models (LLMs) offer an innovative means to overcome deficiencies in information delivery. LLMs exhibit the ability to generate coherent textual responses to various prompts, are trained on extensive datasets, and utilize natural language processing (NLP) techniques to produce human-like responses.[2] Although the flexibility of LLMs allows for a wide range of applications, such as clinical decision-making and patient education, challenges such as insufficient transparency and neural hallucinations, where seemingly reliable but false information is generated, reveal significant risks to the potential of LLMs.[2] In addition to the scale of the LLM, the performance of LLMs is directly affected by their architectural designs.[2-3] Moreover, the quality, diversity, currency, and frequency of updates to the training data are crucial for accuracy.

ChatGPT (Chat Generative Pre-trained Transformer), developed by OpenAI (OpenAI, San Francisco, California, USA) in 2018 is an LLM capable of generating text by collecting information from various data sources, such as books, journals, articles, or open-access websites; translating between languages; and generating answers to questions based on the information it gathers from these sources.[4] A rival AI model promoting transparency and regenerability, DeepSeek (DeepSeek Artificial Intelligence Co., Beijing, China), was introduced in China in December 2024 and gained worldwide attention due to its low price and high download rates.[5] However, these programs may have limitations, including the potential inaccuracy of information obtained through web scanning and the lack of details.[1,4]

In early 2025, a brand-new type of model evolved: large reasoning models (LRMs). LRMs, which build on ordinary LLM architectures and training methods, have exhibited significant performance in reasoning and planning tasks that were previously outside the capabilities of prior LLMs.[3] LRMs can also be referred to as several “reasoning” LLMs that use “chain of thought” (CoT) reasoning, which outputs an answer using logical steps to mimic the way a human thinks.[6] OpenAI launched o3-mini on January 31, 2025. This appeared only a few days after DeepSeek introduced its own reasoning model, DeepSeek-R1 (DeepSeek Artificial Intelligence Co., Beijing, China), which also specializes in math, coding, and reasoning tasks.[7] On April 16, 2025, OpenAI launched o3 (OpenAI, San Francisco, California, USA), its most recent reasoning model, which proved more accurate than o3-mini.[8] These models have attracted considerable interest for their advanced and sophisticated reasoning capabilities.[9-11] Furthermore, these models are based on prior advancements like ChatGPT 4.5 (OpenAI, San Francisco, California, USA) and DeepSeek V3 (DeepSeek Artificial Intelligence Co., Beijing, China), indicating a new era of competition among LLM developers aiming to enhance performance in scientific domains.[7]

The application of prosthetic treatments mostly depends on in-person communication, clinical and radiographic assessments, and operative treatments. In addition to providing potential clinical support to prosthodontists, chatbots are also used in patient and dental education[12,13]. The accuracy of responses generated by chatbots regarding patient needs is crucial, as incorrect information may result in complications and malpractice during prosthetic treatments[13,14]. Although a number of studies[1,6,13,15-18] have examined LRMs, to the authors’ knowledge, none have compared LLMs to LRMs regarding accuracy and consistency in prosthodontics. Therefore, the aim of this study was to evaluate the accuracy and consistency of responses related to tooth-supported fixed dental prostheses generated by four different chatbots in Turkish. The null hypotheses were that there would be no difference in the (1) accuracy or (2) consistency of responses regarding tooth-supported fixed dental prosthesis information among models.

MATERIAL AND METHODS

Twelve questions were prepared to evaluate the accuracy and consistency of answers generated by four different chatbots (Open AI o3 [LRM-O], DeepSeek R1 [LRM-R], ChatGPT 4.5 [LLM-G], and DeepSeek V3 [LLM-V]), performing with two different model designs (LLMs and LRMs) in tooth-supported fixed dental prostheses. Ethical approval was not sought because no human participants were involved and no personal data was collected.

Two experienced prosthodontists (D.Y. and E.D.Ç.) created 14 open-ended questions based on the Fixed Bridges and Dental Implants Guideline[19] published by the British Society for Restorative Dentistry and subsequently translated them into Turkish. Content validity was confirmed through collaboration with a measurement and evaluation professional (D.K.), who edited the questions. Twelve questions were selected based on expert recommendations, and necessary revisions were made. A small pilot study was conducted to evaluate the clarity of the questions. Ultimately, 12 revised open-ended questions—six general and six specific—were selected for the study (Table 1). The same set of questions was submitted to all chatbots on April 25, 2025, from four different computers connected to the same internet network,

Table 1. Questions used in the study

Question Number	Question Text*
G1	What should be done during the intraoral examination and evaluation of a tooth-supported fixed dental prosthesis?
G2	What should determine the shade of a tooth-supported fixed dental prosthesis?
G3	Once the impression of the tooth-supported fixed dental prosthesis is complete, what should be done with the impression?
G4	What qualities should interocclusal registration materials have in tooth-supported fixed dental prostheses?
G5	What factors must be considered to accurately record the maxillo-mandibular relationship for a tooth-supported fixed dental prosthesis?
G6	What are the general properties that a temporary cement should have when cementing a provisional restoration?
S1	What are the principles of tooth preparations for a tooth-supported fixed dental prosthesis?
S2	If the pulp vitality/integrity of the tooth is likely to be jeopardized by the extent of the preparation required for tooth-supported fixed dental prosthesis, what additional preparation treatments would be indicated?
S3	What is the recommended technique for fabricating a provisional tooth-supported fixed dental prosthesis in situations where the form and function of the tooth must remain unchanged?
S4	Which impression material is generally used for the antagonist arch in a tooth-supported fixed dental prosthesis?
S5	What is the main position of the mandible when a reorganized approach is planned in a tooth-supported fixed dental prosthesis?
S6	What are the success criteria for permanent cementation of tooth-supported fixed dental prostheses?

* The questions presented were translated into Turkish
G: General Question; S: Specific Question.

simultaneously, at three different times during the same day: morning, afternoon, and evening (in the time zone UTC+3). The pre-prompt ("I would like you to answer my questions as a prosthodontist") was used to give the chatbots a perspective before the questions.[20] These questions were posed three times for each model, and to minimize memory retention bias and reset memory, 48 questions were answered using the "new chat" option. Two prosthodontists graded the responses with a holistic rubric (1 point: completely incorrect response; 2 points: correct response containing errors; 3 points: partially correct response; 4 points: completely correct response).[21]

All chatbot performances were evaluated based on their accuracy and consistency. Accuracy was measured as the percentage of completely correct answers compared to the total number of repetitions in the set of questions generated by the LRM-O, LRM-R, LLM-G, and LLM-V models, and 95% confidence intervals were calculated using the Wald binomial technique.[22,23] In the present study, consistency was assessed as the agreement among chatbots in response to questions posed at different times, while inter-rater reliability was assessed as the agreement between different graders' scores for the responses generated by chatbots. The Brennan–Prediger coefficient, Cohen's kappa, Fleiss's kappa, and Krippendorff's alpha were used to examine the agreement among chatbots and the inter-rater reliability for ordinal data, with 95% confidence intervals. The estimated coefficients were classified using Gwet's benchmarking scale[24] :<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.

The data was analyzed using a statistical software program (IBM SPSS Statistics for Windows, v25.0; IBM Corp., Armonk, New York, USA) for accuracy evaluations, while a different statistical software program[25] (R Foundation for Statistical Computing, R Core Team, Vienna, Austria) was used for consistency analysis ($p < 0.05$). The Kolmogorov–Smirnov and Shapiro–Wilk tests were used to analyze the normality of the data distribution. When the analysis revealed a non-normal distribution

Table 2. Statistical assessment of differences among the responses of the four programs to questions and consistency of responses at different times of day

		p			Tests
		All Questions	**General Questions**	**Specific Questions**	
Comparison of inter-chatbot responses (LRM-O, LRM-R, LLM-G, and LLM-V)	Morning	0.61	0.34	0.23	Kruskal–Wallis H Test
	Afternoon	0.31	0.60	0.13	
	Evening	0.55	0.25	0.39	
	Mean	0.30	0.29	0.31	
Comparison of intra-chatbot responses at different times of day	LRM-O	0.11	0.26	0.37	Friedman Test
	LRM-R	0.22	0.37	0.37	
	LLM-G	0.05	1.00	0.05	
	LLM-V	0.26	0.26	1.00	

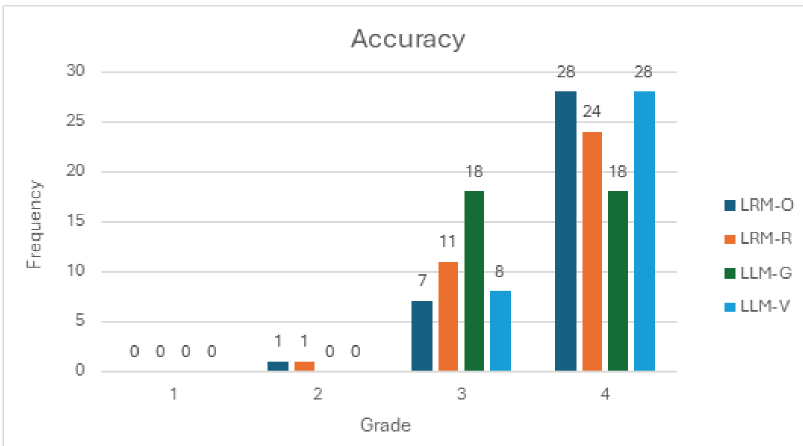

Figure 1. Distribution of grades for responses generated by chatbots

of the data, the Kruskal–Wallis H test was used to evaluate the presence of statistically significant differences among the responses of the four programs to questions as well as to compare the mean grades of the answers generated by each program at different times of day. The Friedman test for repeated measures was employed to determine whether statistically significant differences existed among the scores of responses generated by each program in the morning, afternoon, and evening. Additionally, while the consistency between the graders' responses (inter-rater reliability) was assessed using the Brennan–Prediger coefficient and the Cohen kappa coefficient, the agreement of the responses generated by chatbots was assessed using the Fleiss kappa and Krippendorff alpha coefficients.

RESULTS

Four NLP models (OpenAI o3, DeepSeek R1, ChatGPT 4.5, and DeepSeek V3) responded to 36 questions at three different times of day regarding tooth-supported fixed prostheses, generating a total of 144 responses. Figure 1 shows the distribution of responses generated by chatbots. Figure 2 shows the distribution of the number of questions answered with completely correct responses from the different time periods.

There was no statistically significant difference between morning (p = 0.61), afternoon (p = 0.34), or evening (p = 0.23) hours when questions were posed without classification to the LRM-O, LRM-R, LLM-G, and LLM-V at different times of the day. Furthermore, no statistically significant differences were found in the categorization of questions as general or specific to LRM-O, LRM-R, LLM-G, and LLM-V across different times of day: morning (p = 0.31), afternoon (p = 0.60), or evening (p = 0.13) for general questions, and morning (p = 0.55), afternoon (p = 0.25), or evening (p = 0.39) for specific questions (Table 2).

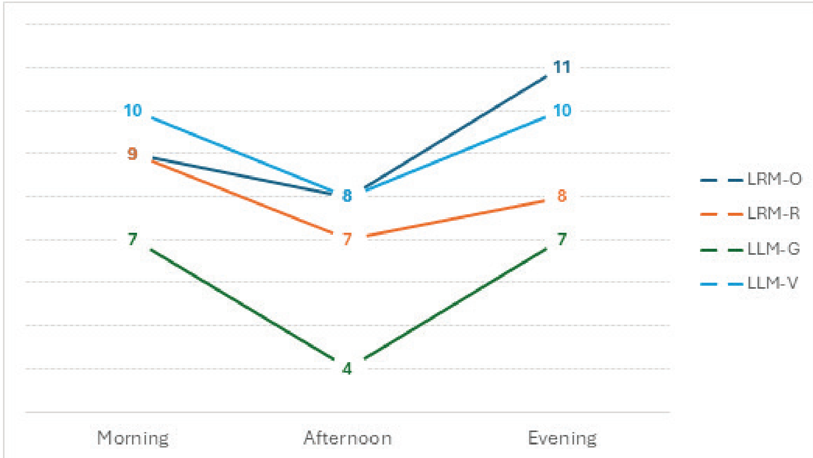

Figure 2. Distribution of the number of questions answered with a completely correct response from different time periods (times are UTC+3)

Table 3. Mean response grades and accuracy of chatbots

	Morning		Afternoon		Evening		Mean		Accuracy
	X ± SE	SD	X ± SE	SD	X ± SE	SD	X ± SE	SD	%
LRM-O	3.66 ± 0.19	0.65	3.66 ± 0.14	0.49	3.91 ± 0.08	0.29	3.75 ± 0.12	0.43	77.7
LRM-R	3.75 ± 0.13	0.45	3.50 ± 0.19	0.67	3.66 ± 0.14	0.49	3.63 ± 0.13	0.46	66.6
LLM-G	3.58 ± 0.15	0.51	3.33 ± 0.14	0.49	3.58 ± 0.14	0.51	3.5 ± 0.13	0.46	50
LLM-V	3.83 ± 0.11	0.39	3.66 ± 0.14	0.49	3.83 ± 0.11	0.39	3.77 ± 0.10	0.36	77.7

SE: Standard Error; SD: Standard Deviation.

When analyzing the responses generated by each program to uncategorized questions throughout the day, it was found that the results were statistically similar in the morning, afternoon, and evening hours for LRM-O ($p = 0.11$), LRM-R ($p = 0.22$), LLM-G ($p = 0.05$), and LLM-V ($p = 0.26$). Additionally, no statistical significance was shown in the responses generated by each model to general questions during the day for LRM-O ($p = 0.26$), LRM-R ($p = 0.37$), LLM-G ($p = 1.00$), and LLM-V ($p = 0.26$), as well as in responses to specific questions for LRM-O ($p = 0.37$), LRM-R ($p = 0.37$), LLM-G ($p = 0.05$), and LLM-V ($p = 1.00$) (Table 2).

When comparing the average responses from each program to all questions without classification ($p = 0.30$), general questions ($p = 0.30$), and specific questions ($p = 0.31$), they were statistically comparable at different times of the day (Table 2). Table 3 shows the means and the related percentages for LRM-O, LRM-R, LLM-G, and LLM-V. The accuracy levels of LRM-O, LRM-R, LLM-G, and LLM-V were calculated as 77.7%, 66.6%, 50%, and 77.7%, respectively (Table 3).

Table 4 shows a significant degree of inter-rater reliability and agreement. The analyses conducted using the Brennan–Prediger and Cohen's kappa coefficients indicated an "excellent" level of agreement (Benchmark Scale: almost perfect) among different graders. Whereas Fleiss's kappa similarly demonstrated "excellent" agreement (Benchmark Scale: almost perfect) for LRM-R, LLM-G, and LLM-V, LRM-O showed slightly lower agreement, classified as "substantial" on the Benchmark Scale. Furthermore, Krippendorff's alpha values revealed an "excellent" level of agreement (Benchmark Scale: almost perfect) for both LLM-G and LLM-V. By contrast, LRM-O and LRM-R showed slightly lower agreement and were classified as "substantial."

Table 4. Inter-rater reliability and agreement of responses by each chatbot

	Methods	Coefficient	SE	95% CI (Range)	Benchmark Scale
Inter-rater Reliability	Brennan and Prediger	0.93	0.02	0.89–0.97	Almost Perfect
	Cohen's kappa	0.95	0.01	0.92–0.98	Almost Perfect
LRM-O	Fleiss kappa	0.80	0.06	0.69–0.91	Substantial
	Krippendorff alpha	0.80	0.06	0.69–0.91	Substantial
LRM-R	Fleiss kappa	0.82	0.05	0.72–0.92	Almost Perfect
	Krippendorff alpha	0.78	0.06	0.65–0.91	Substantial
LLM-G	Fleiss kappa	0.91	0.03	0.86–0.96	Almost Perfect
	Krippendorff alpha	0.89	0.03	0.83–0.95	Almost Perfect
LLM-V	Fleiss kappa	0.95	0.01	0.92–0.98	Almost Perfect
	Krippendorff alpha	0.91	0.03	0.86–0.96	Almost Perfect

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.

DISCUSSION

When four different chatbots (two LRMs and two LLMs) were compared within themselves and with one another regarding tooth-supported fixed prostheses-related queries, no statistically significant difference in terms of accuracy was found among them. Therefore, the first null hypothesis was accepted. When the consistency of the programs was evaluated, the second hypothesis was rejected because the consistency among LRMs was found to be lower than that of LLMs.

The responses generated by each model during the morning, afternoon, and evening hours did not reveal any statistically significant differences. Accuracy was evaluated based on completely correct answers, and although there was no significant difference regarding their overall accuracy, it was found that LRM-O = LLM-V (77.7%) > LRM-R (66.6%) > LLM-G (50%). For responses to tooth-supported fixed prostheses-related queries, LRM-O and LLM-V exhibited similar accuracy, which was higher than those of the other chatbots. This finding should be borne in mind when posing questions about tooth-supported fixed prostheses. Because the chatbots were fed similar information about tooth-supported fixed prostheses, their answers were also alike. The results of this study are inconsistent with previous research using open-ended prosthodontic questions, which showed lower accuracy rates (25.6%) regarding removable prostheses and tooth-supported fixed prostheses.[26] The improved accuracy in our study may be explained by AI companies following scientific research and correcting their deficiencies through subsequent upgrades. Both classes of models are periodically retrained or fine-tuned using datasets and training methods updated by research teams, instead of "automatically" accepting current research articles and performing simultaneous learning. Although both LLMs provided acceptable responses, the DeepSeek-V3 model was shown to generate more accurate results.

The results of the present study were consistent with Özcivelek and Özcan[13]. They reported that ChatGPT-o1 showed superior accuracy compared to DeepSeek-R1 when evaluating responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots in response to patient inquiries about dental and maxillofacial prostheses. However, Shirani [15] reported that when comparing the performance of ChatGPT 4o, DeepSeek R1, and Gemini 2 Pro in multiple-choice and open-ended short-answer fixed prosthodontics questions over time, accuracy regarding open-ended short-answer ranged between 52.8% and 58.3% after four weeks. The results of that study are inconsistent with the present research. The discrepancies between the research findings may be related to the methodological design, which uses different assessment tools such as holistic rubrics or 3-point scales, pre-prompts like "I would like you to answer my questions as a prosthodontist" or "the correct answers are as follows, please learn them," response limitations such as "none" or "the answer should be five words or fewer," as well as the different questions and sources utilized.

On the other hand, Kaygisiz and Teke reported that although both DeepSeek-V3 and ChatGPT-4o provided acceptable responses, the DeepSeek-V3 model was shown to generate more accurate results in diagnosing oral pathologies.[1] Another study on oral lesions reported that DeepSeek-R1 outperformed both ChatGPT-4o (88.9% accuracy) and the journal's readers (47.8% accuracy) with a diagnostic accuracy of 91.6%.[16] Zhou et al.[17] reported that all models obtained a merely "fair" quality grade, indicating the need for changes in citation practices and customization. However, when compared to ChatGPT-o3, both DeepSeek-R1 and ChatGPT-4o generated surgical information that was easier to read and comprehend. The findings may be inconsistent due to variations in topics related to the different medical and dental specialties, the range of questions posed, discrepancies in the pre-prompts utilized, and shifts in the language used when preparing the questions.

Although there was no significant difference between the morning, afternoon, and evening implementations, the chatbots generated fewer completely correct answers in the afternoon than in the morning and evening implementations. The mean accuracy of the answers generated by the LLMs in the afternoon was lower, but they generated more comparable results in the morning and evening hours, while the mean accuracy of the answers generated in the afternoon was lower in the LRMs. LRM-O received higher grades in the evening compared to the morning and afternoon, while LRM-R received higher grades in the evening than in the morning. In assessing LLMs and LRMs regarding consistency, although there is a lack of research on DeepSeek, several studies show that ChatGPT may generate varying responses to the same questions at various intervals.[22,26-27] Although the previous article evaluated the accuracy and consistency of ChatGPT throughout morning, afternoon, and evening hours, there has been a lack of discussion on potential discrepancies between these periods or the fundamental reasons for these variations.[26] The increasing number of Turkish and worldwide users in the UTC (Coordinated Universal Time) +3 time zone connecting to the servers of the two companies—Open AI in the US, which is in the UTC -5 time zone, and DeepSeek in China, which is in the UTC+8 time zone—may be responsible for these discrepancies, perhaps due to infrastructure constraints that cannot match the rising demand from a growing user base. On the other hand, the present study's comparison of NLPs showed that the consistency among LLMs was categorized as "almost perfect," whereas the consistency among LRMs was categorized as "substantial" according to the benchmark scale (LLM > LRM). LRMs divide the problem into multiple logical steps, each step invoking CoT processes and different submodules (such as arithmetic, logic, and code execution) as necessary. They use stochastic methods, including sampling and beam search.[28] These multi-step and random components create minor deviations in each run in LRMs, reducing consistency compared to LLMs.[29]

Accuracy and consistency were crucial parameters in the performance evaluations of chatbots[1,15,26]. Chatbots show promise as educational tools for patients and dental professionals in prosthodontics; nonetheless, their limitations in accuracy and consistency may result in misinformation or inadequate treatment.[30] Using chatbots in prosthodontics requires careful consideration of the potential risks linked to even a small percentage of incorrect responses. Consequently, it is essential for dental professionals to evaluate the responses generated by chatbots regarding tooth-supported fixed prostheses.[1] In addition, the current limitations of these models in terms of their ability to provide clinically trustworthy information are highlighted by the fact that consistent chatbot behavior does not guarantee the accuracy of replies.[15]

Chatbots may generate responses that appear consistent yet lack real meaning. This phenomenon is termed "hallucination."[31] In addition to hallucinations, another limitation of the present study is that chatbots may yield biased results because they respond to individual- and country-specific questions. Because the present study is limited by its focus on the tooth-supported fixed prosthesis specialty in Turkish, as well as the small number of reviewers, its findings should not be generalized to other subjects. To gain more insight into chatbot performance, future studies should use different models, other topics of prosthodontics, larger datasets, and question types that include visuals.

CONCLUSION

When using AI models in dentistry, it is crucial to consider the potential risks that even just a small percentage of inaccurate responses can generate. It is, therefore, essential for clinicians to assess the responses generated by chatbots regarding tooth-supported fixed prostheses. Additionally, in terms of license and access, because the chatbots tested in this study are from companies with DeepSeek open-source licenses as well as OpenAI closed API access (and perform similarly), using free chatbots may be a more user-friendly approach. Based on the findings of this study, the following conclusions were drawn:

1. The accuracy of the LRMs and LLMs models is similar for queries about tooth-supported fixed prostheses.

2. The accuracy of the responses generated by all models at different times of the day is similar.

3. The consistency among LLMs is higher than that of the LRM models for queries on tooth-supported fixed prostheses.

ACKNOWLEDGEMENTS

None.

CONFLICT OF INTEREST STATEMENT

The authors report there are no competing interests to declare.

Farklı Doğal Dil İşleme Modellerinin Doğruluk ve Tutarlılık Açısından Diş Destekli Sabit Protez Sorgulamalarında Karşılaştırılması

ÖZET

Amaç: Bu çalışmanın amacı, dört farklı doğal dil işleme (NLP) modeli tarafından oluşturulan diş destekli sabit diş protezlerine ilişkin yanıtların doğruluğunu ve tutarlılığını değerlendirmektir.

Gereç ve Yöntemler: 12 adet açık uçlu olarak Türkçe dilinde hazırlanan sorular oluşturuldu ve modellere göre 4 farklı NLP'ye yöneltilmiştir: OpenAI o3 (LRM-O), OpenAI GPT 4.5 (LLM-G), DeepSeek R1 (LRM-R) ve DeepSeek V3 (LLM-V). Yanıtlar holistic rubric kullanılarak değerlendirilmiştir. Doğruluk değerlendirmeleri için Kruskal-Wallis H testi kullanılmıştır. Puanlayıcıların yanıtları arasındaki tutarlılık Brennan ve Prediger katsayısı ve Cohen kappa katsayısı kullanılarak değerlendirilmiştir. Tutarlılık ise Fleiss kappa ve Krippendorff alfa katsayıları kullanılarak değerlendirilmiştir ($p < 0.05$).

Bulgular: LRM-O, LLM-G, LRM-R ve LLM-V grupları arasında doğruluk açısından istatistiksel olarak anlamlı bir fark bulunamamıştır ($p = 0.30$). LRM-O, LLM-G, LRM-R ve LLM-V'nin doğruluğu sırasıyla %77.7, %50, %66.6 ve %77.7'dir. Ayrıca, LLM'lerin tutarlılığı neredeyse mükemmel bulunurken, LRM'ler önemli düzeydeydi.

Sonuç: Çalışmanın sınırları dahilinde LRM'ler ve LLM'ler benzer doğruluk sergilemiştir. Ancak, LLM'lerin tutarlılığı LRM'lerden daha yüksek bulunmuştur.

Anahtar Kelimeler: Diş protezi; tedavi protokolleri; yapay zeka

REFERENCES

1. Kaygisiz ÖF, Teke MT. Can DeepSeek and ChatGPT be used in the diagnosis of oral pathologies? BMC Oral Health 2025;25:638.

2. 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;2025:1-10.

3. Kambhampati S, Stechly K, Valmeekam K. (How) Do reasoning models reason? Ann N Y Acad Sci 2025;1547:33–40.

4. Haupt CE, Marks M. AI-generated medical advice: GPT and beyond. JAMA 2023;329:1349–50.

5. Gibney E. China's cheap, open AI model DeepSeek thrills scientists. Nature 2025;638:13–4.

6. Hoyt RE, Knight D, Haider M, Bajwa, M. Evaluating a large reasoning models performance on open-ended medical scenarios. medRxiv 2025;2025:1-15.

7. Jiang Q, Gao Z, Karniadakis GE. DeepSeek vs. ChatGPT vs. Claude: a comparative study for scientific computing and scientific machine learning tasks. Theoretical and Applied Mechanics Letters 2025;15:100583.

8. OpenAI. Introducing OpenAI o3 and o4-mini 2025. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed 13 May 2025.

9. Egger J, De Paiva LF, Luijten G, Krittanawong C, Keyl J, Sallam M, et al. Is DeepSeek-R1 a game changer in healthcare? a seed review. TechRxiv 2025;4:1–21.

10. Sallam M, Kholoud Al-Mahzoum, Sallam M, Mijwil MM. DeepSeek: is it the end of generative AI monopoly or the mark of the impending doomsday? Mesopotamian Journal of Big Data 2025;2025:26–34.

11. Normile D. Chinese firm's large language model makes a splash. Science 2025;387:238.

12. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35:1098-1102.

13. Özcivelek T, Özcan B. Comparative evaluation of responses from DeepSeek-R1, ChatGPT-o1, ChatGPT-4, and dental GPT chatbots to patient inquiries about dental and maxillofacial prostheses. BMC Oral Health. 2025;25:871.

14. Cuevas-Nunez M, Silberberg VIA, Arregui M, Jham B, Baallester-Victoria R, Koptseva I, et al. Diagnostic performance of ChatGPT-4.0 in histopathological description analysis of oral and maxillofacial lesions: a comparative study with pathologists. Oral Surg Oral Med Oral Pathol Oral Radiol. 2025;139:453–61.

15. 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;S0022-3913:00400-7.

16. Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. J Dent Sci Epub 2025.

17. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. Int J Med Inform 2025;198:105871.

18. Hou Y, Patel J, Dai L, Zhang E, Liu Y, Zhan Z, et al. Benchmarking of Large Language Models for the Dental Admission Test. Health Data Sci 2025;5:0250.

19. British Society for Restorative Dentistry. Crowns, fixed bridges and dental implants: guidelines. Woodford Green: British Society for Restorative Dentistry; 2013. p. 8-21.

20. Zaghir J, Naguib M, Bjelogrlic M, Névéol A, Tannier X, Lovis C. Prompt engineering paradigms for medical applications: scoping review. J Med Internet Res 2024;26:e60501.

21. Koçak D. Investigation of rater tendencies and reliability in different assessment methods with the many facet Rasch model. IEJEE. 2020;12:349–58.

22. 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;57:108–13.

23. 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 2023;24:46–52.

24. Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. 4th ed. Gaithersburg, Advanced Analytics LLC; 2014. p. 163-183.

25. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna 2021. https://www.R-project.org. Accessed 13 May 2025

26. 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;131:659.e1–6.

27. Cinar C. Analyzing the performance of ChatGPT about osteoporosis. Cureus 2023;15:e45890.

28. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst 2022;35:24824–37.

29. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, et al. Self-consistency improves chain-of-thought reasoning in language models. ArXiv 2023;1–24.

30. Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, Afshari FS, Yuan JC, Sukotjo C. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. J Prosthet Dent. 2025;134:199-206.

31. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023;15:e35179.