THE JOURNAL OF PROSTHETIC DENTISTRY

RESEARCH AND EDUCATION

Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics

Maryam Gheisarifar, DDS, MS,[a] Marwa Shembesh, BDS, MSc, DScD,[b] Merve Koseoglu, DDS,[c] Qiao Fang, DDS, MSD,[d] Fatemeh Solmaz Afshari, DMD, MS,[e] Judy Chia-Chun Yuan, DDS, MS MAS,[f] and Cortino Sukotjo, DDS, PhD, MMSc, MHPE[g]

ABSTRACT

Statement of problem. Healthcare-related information provided by artificial intelligence (AI) chatbots may pose challenges such as inaccuracies, lack of empathy, biases, over-reliance, limited scope, and ethical concerns.

Purpose. The purpose of this study was to evaluate and compare the validity and consistency of responses to prosthodontics-related frequently asked questions (FAQ) generated by 4 different chatbot systems.

Material and methods. Four prosthodontics domains were evaluated: implant, fixed prosthodontics, complete denture (CD), and removable partial denture (RPD). Within each domain, 10 questions were prepared by full-time prosthodontic faculty members, and 10 questions were generated by GPT-3.5, representing its top frequently asked questions in each domain. The validity and consistency of responses provided by 4 chatbots: GPT-3.5, GPT-4, Gemini, and Bing were evaluated. The chi-squared test with the Yates correction was used to compare the validity of responses between different chatbots (α=.05). The Cronbach alpha was calculated for 3 sets of responses collected in the morning, afternoon, and evening to evaluate the consistency of the responses.

Results. According to the low threshold validity test, the chatbots' answers to ChatGPT's implant-related, ChatGPT's RPD-related, and prosthodontists' CD-related FAQs were statistically different ($P<.001$, $P<.001$, and $P=.004$, respectively), with Bing being the lowest. At the high threshold validity test, the chatbots' answers to ChatGPT's implant-related and RPD-related FAQs and ChatGPT's and prosthodontists' fixed prosthetics-related and CD-related FAQs were statistically different ($P<.001$, $P<.001$, $P=.004$, $P=.002$, and $P=.003$, respectively), with Bing being the lowest. Overall, all 4 chatbots demonstrated lower validity at the high threshold than the low threshold. Bing, Gemini, and ChatGPT-4 chatbots displayed an acceptable level of consistency, while ChatGPT-3.5 did not.

Conclusions. Currently, AI chatbots show limitations in delivering answers to patients' prosthodontic-related FAQs with high validity and consistency. (J Prosthet Dent 2025;134:199-206)

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for this journal and was not involved in the editorial review or the decision to publish this article.

[a]Clinical Assistant Professor, Department of Restorative Dentistry, College of Dentistry, University of Illinois Chicago, Chicago, Ill.
[b]Clinical Assistant Professor, Department of Restorative Dentistry, College of Dentistry, University of Illinois Chicago, Chicago, Ill.
[c]Associate Professor, Department of Prosthodontics, Faculty of Dentistry, University of Sakarya, Sakarya, Turkey; and PhD student, Department of Prosthodontics, Faculty of Dentistry, University of Ataturk, Erzurum, Turkey.
[d]Clinical Assistant Professor, Department of Restorative Dentistry, College of Dentistry, University of Illinois Chicago, Chicago, Ill.
[e]Associate Professor, Department of Restorative Dentistry, College of Dentistry, University of Illinois Chicago, Chicago, Ill.
[f]Professor and Associate Dean for Clinical Affairs, Department of Restorative Dentistry, College of Dentistry, University of Illinois Chicago, Chicago, Ill.
[g]Professor and Chair, Department of Prosthodontics, University of Pittsburgh, School of Dental Medicine, Pittsburgh, PA.

Clinical Implications

AI chatbots demonstrate potential as patient educational tools in prosthodontics, but their limitations in accuracy, consistency, and contextual understanding could lead to misinformation or inequitable care. The findings of the present research underlines the importance of clinician oversight to mitigate the potential risks associated with unverified AI-generated content.

Artificial intelligence (AI) chatbots have transformed digital communication by offering a platform to individuals who seek personalized information and tailored, detailed responses to their specific questions.[1] Large language models (LLMs) are a type of AI that mimics human language comprehension and processing.[2]

Virtual AI methodologies can be categorized into 2 approaches: knowledge-based AI and data-driven AI. Knowledge-based AI seeks to model human expertise, relying on existing concepts and knowledge to develop problem-solving methods.[3,4] Data-driven AI operates by training mathematical models to predict or identify patterns based on data interpretation.[5] Through deep learning algorithms, chatbots can analyze vast amounts of data, enabling them to enhance their responses and improve over time, thus simulating the neural networks of the human brain.[6–8] Various LLMs have been developed to generate text that closely resembles human language.[9,10]

Chat Generative Pretrained Transformer (ChatGPT) is a natural language processing model with 175 billion parameters that uses deep learning to generate human-like responses.[11] OpenAI has not officially revealed the exact number of parameters in GPT-4, but it is widely believed to be considerably larger than GPT-3.5, potentially reaching into the hundreds of billions. This increase in model size has enhanced GPT-4's accuracy, its ability to grasp context more effectively, and its overall reasoning capabilities.[12]

More recently, Google introduced Gemini, its most advanced and powerful AI system to date.[13] Gemini builds on Google's expertise in search and language comprehension. It utilizes Bidirectional Encoder Representations from Transformers (BERT), a transformer-based model specifically designed to understand the context of words within a sentence, to interpret language more precisely.[14,15]

Microsoft launched the new Bing in February 2023, introducing an AI-powered "copilot" with chatbot capabilities. This enhanced version of Bing responds to user queries by conducting web searches and evaluating content based on relevance, quality, credibility, and freshness, among other factors. It then delivers a summarized response grounded in the top search results with links to the original sources for further reference.[16,17]

These chatbots engage interactively with users. Initially, they receive an input prompt from the user, which is processed and converted into a mathematical representation using a deep learning model.[18] The chatbot then predicts the most appropriate response or sequence of responses based on the patterns it has learned from its training data.[18] Tools such as these can assist with patient education by addressing their inquiries, thus improving patient understanding of treatments, potential side effects or complications, prognoses, and outcomes.[19,20] Accurate information empowers patients to provide informed consent, follow recommended treatments, and maintain realistic expectations.[21,22] Conversely, LLMs are not specifically created to provide medical advice and may produce coherent responses that lack factual accuracy, known as "hallucinations."[23] The absence of transparency regarding the source of information and the risk of confidently spreading inaccurate or misleading information are crucial issues that require attention.

In dentistry, the available research on AI chatbot performance is limited. To date, studies have explored its usability in generating scientific content in oral and maxillofacial surgery,[24,25] oral medicine,[26] and endodontics.[22,27] Additionally, the ability of chatbots to respond to questions in board-style dental knowledge has been assessed.[28] Freire et al[29] evaluated the accuracy and repeatability of answers generated by one AI chatbot (GPT-4) to clinicians' procedural questions about removable partial dentures (RPDs) and tooth-supported fixed dental prostheses. Chat GPT-4 showed limitations in delivering accurate and reliable responses to clinicians' questions.[29] In the clinic, patients may have a diverse range of prosthodontic treatment questions. The patient may use chatbots for the answers to their clinical questions. Although AI chatbots are used in various areas of healthcare, the authors are unaware of an evaluation of their performance as public information sources in prosthodontics. This study aimed to assess and compare the validity and consistency of responses provided by 4 chatbot systems (GPT-3.5, GPT-4, Gemini, and Bing) to a set of frequently asked questions (FAQs) from patients on the subject of prosthodontics. The research hypothesis was that the 4 tested artificial intelligence chatbots would be able to provide accurate and reliable answers to patients' FAQs in prosthodontics. The null hypotheses were that no difference would be found in the accuracy and consistency of responses among the 4 chatbots.

MATERIAL AND METHODS

This research did not require ethical approval as human participants were not involved. Patient's FAQs were

Table 1. Patient's dental implant-related FAQ

ChatGPT FAQ	Prosthodontists FAQ
Q1. What are dental implants?	Q1. What are the advantages of dental implants?
Q2. How successful are dental implants?	Q2. What are the disadvantages of dental implants?
Q3. How long do dental implants last?	Q3. Is it painful during dental implant surgery?
Q4. Can dental implants be done in one day?	Q4. After dental implant treatment, how should I care for it?
Q5. Are there any risks or complications with dental implants?	Q5. Do I need to floss around dental implants?
Q6. What are the alternatives to dental implants?	Q6. Can I eat everything with my implant crown?
Q7. Can dental implants fail?	Q7. Is it better to extract my tooth and get implant?
Q8. How long is the recovery time after getting dental implants?	Q8. Can I have infection around my dental implant?
Q9. What materials are used in dental implants?	Q9. Do I need to get bone grafting before implant placement?
Q10. Can dental implants be placed immediately after a tooth extraction?	Q10. How long should I wait till I get my implant crown?

FAQ, frequently asked questions.

Table 2. Patient's fixed prosthodontics-related FAQ

ChatGPT FAQ	Prosthodontists FAQ
Q1. What is a dental crown?	Q1. What is a temporary dental crown?
Q2. What types of dental crowns are available?	Q2. Why do I need a temporary crown? What is the purpose of temporary crown?
Q3. How long does a dental crown last?	Q3. Can I get root canal after the crown is made?
Q4. Are there alternatives to dental crown?	Q4. My tooth has a crack, do I have to crown it?
Q5. How is a dental crown different from a veneer?	Q5. Should I get teeth whitening before or after getting a new crown?
Q6. Can I eat normally with a dental crown?	Q6. Can I get cavity under my crown?
Q7. How do I choose between the different types of dental crown materials?	Q7. Can my tooth get damaged by being crowned?
Q8. Do dental crowns feel like natural teeth?	Q8. Does the crown make my tooth stronger?
Q9. What is the success rate of dental crowns?	Q9. Is the crown a permanent treatment?
Q10. Can a dental crown be done in one visit?	Q10. Will the color of my crown change over time?

FAQ, frequently asked questions.

formulated in 4 prosthodontic domains: implants, fixed prosthodontics, complete dentures (CD), and RPDs. A total of 80 questions were collected through 2 sources. First, 40 patient FAQs, 10 from each domain, were selected from 10 full-time prosthodontists at University of Illinois Chicago, College of Dentistry. Secondly, GPT-3.5 provided a list of the top 30 FAQs for each domain, from which 10 questions per domain were selected. Questions overlapping with those formulated by the expert prosthodontists were excluded. Finally, the top ChatGPT 10 FAQs were chosen after eliminating the overlapping content.

The list of 20 questions for each domain is shown in Tables 1 to 4. These questions encompass areas that included terminology, diagnosis, clinical details of treatments, postoperative care, maintenance, treatment outcomes and prognosis, risks and side effects, prevention, and alternative treatment options. Each of the 80 questions was subsequently presented to the 4 AI chatbots. The main application programming interface (API) of each chatbot was used to simulate real-world interactions during this process. The following approaches were used: for GPT-3.5 and GPT-4, API was accessed using https://chat.openai.com/, for Gemini, API was accessed using https:// gemini.google.com, and for Bing/Copilot, API was accessed using https:// www.bing.com. All the questions were asked on the same day, with a new chat initiated for each question.

Two board-certified prosthodontists (M.G., M.S.) evaluated and scored the answers independently using a 5-point Likert scale. First, the evaluators scored a set of 10 questions and were calibrated; this process was repeated till the evaluators were calibrated to kappa of 0.8. Secondly, the reviewers independently evaluated all responses using a 5-point Likert scale. A modified version of the Global Quality Score (GQS)[27,30] (Table 5) was applied to assign scores based on the "context" and "content" of the responses. Using these criteria, the

Table 3. Patient's complete denture-related FAQ

ChatGPT FAQ	Prosthodontists FAQ
Q1. What are dentures?	Q1. How to clean complete dentures?
Q2. How do I get used to wearing dentures?	Q2. Do I need implants if my denture is not stable?
Q3. Do dentures affect how I speak?	Q3. Can I eat apple with my complete dentures?
Q4. Will dentures change how I look?	Q4. Can I have the denture without covering my palate?
Q5. Can I sleep with my dentures?	Q5. For my lower denture, does it have to go all the way in the back?
Q6. How long do complete dentures last?	Q6. May I use denture adhesive?
Q7. Can I eat normally with dentures?	Q7. How many implants do I need to get my denture stable?
Q8. How do complete dentures stay in place?	Q8. Do I have to come for cleaning after I get my dentures?
Q9. Will dentures affect my taste?	Q9. Can dentures be repaired?
Q10. Do I need to use denture adhesive forever?	Q10. How can I get my dentures more retentive?

FAQ, frequently asked questions.

Table 4. Patient's removable partial denture-related FAQ

ChatGPT FAQ	Prosthodontists FAQ
Q1. What is a partial denture?	Q1. Are partial dentures best way to replace missing teeth?
Q2. How do partial dentures stay in place?	Q2. How to remove partial dentures with clasp?
Q3. Can I eat normally with partial denture?	Q3. How can I clean partial dentures?
Q4. How long do partial dentures last?	Q4. Will partial denture affect my other teeth?
Q5. Do I need to remove my partial denture at night?	Q5. Does partial dentures have to cover my palate?
Q6. Are there different types of partial dentures?	Q6. Does partial dentures have to be metal?
Q7. Is it difficult to speak with partial dentures in?	Q7. Why do you have to grind my teeth away when making partial denture?
Q8. Do I need to use denture adhesive with partial denture?	Q8. Can I have a small partial denture just to cover my missing teeth, instead of a big metal plate?
Q9. Can I sleep with my partial dentures in?	Q9. Can you make partial denture with pink color and flexible?
Q10. How does the cost of partial denture compare to other tooth replacement option?	Q10. Is partial denture a better option than a bridge or an implant?

FAQ, frequently asked questions.

Table 5. Five-point Likert scale - modified version of global quality score (GQS)

Score	Rating	Description
5	Strongly Agree	Answer correct, and content comprehensive.
4	Agree	Answer correct and most of content correct, but lacks information or contains minor errors.
3	Neutral	Answer somewhat correct, but details primarily incorrect, missing, or irrelevant.
2	Disagree	Answer incorrect, but content includes some correct elements.
1	Strongly Disagree	Answer and entire content incorrect or irrelevant.

chatbot responses were evaluated in a detailed manner, considering both the accuracy of the answers (context) and the thoroughness and precision of the information provided (content). This approach allowed for an assessment of the responses' validity. Additionally, each question was asked 3 times in a day (morning, afternoon, and evening) to analyze the consistency of the chatbots' answers.

After independently completing the scoring, the 2 reviewers (a total of 960 answers each) compared their assessments and discussed any discrepancies. Differences were resolved through evidence-based discussions focused on the context and content of the responses. Ultimately, a single consolidated scoresheet was created for all 960 answers to be used in the statistical analyses. The obtained final scores were classified into 2 categories: "valid" and "invalid."[27] To determine the validity of each response (that is, the extent to which the responses from each chatbot accurately addressed the intended query), 2 different validity assessments were used: a low-threshold validity test and a high-threshold validity test. The threshold score for the low-threshold test was 4. A chatbot's answer was considered valid if all 3 responses to a question scored 4 or higher. Any response with a score below 4 was deemed invalid for that chatbot's answer. For the high-threshold test, the score threshold was set at 5. A chatbot's response was determined to be valid only if all 3 replies received a score of 5. If any response received a score below 5, the chatbot's answer was regarded as invalid.

Descriptive statistics determined the mean of obtained scores for the chatbot responses to FAQs within each domain. The chi-squared test with the Yates correction was used to compare the validity of responses between different chatbots (α=.05). Consistency refers to the extent to which the chatbot generates consistent responses when used frequently under consistent circumstances. The Cronbach alpha was calculated for 3 sets of responses collected in the morning, afternoon, and evening to evaluate the consistency and thus consistency of the responses. The Cronbach alpha measures the level of consistency on a standardized scale from 0 to 1, where 1 indicates complete consistency in responses. Consistency levels are classified as excellent (≥0.9), good (≥0.8 to <0.9), acceptable (≥0.7 to <0.8), or questionable (≥0.6 to <0.7).[31] All statistical analyses were performed using a software program (IBM SPSS Statistics, v27; IBM Corp).

RESULTS

Table 6 displays the mean score for the responses to each question regarding dental implants, fixed prosthodontics, CDs, and RPDs. Scores ranged from 3 to 5 on dental implant-related, 3.7 to 5 on fixed prosthodontics-related, 2 to 5 on CD-related, and 3 to 5 on RPD-related questions.

According to the low threshold validity test (Table 7), the chatbots' answers to ChatGPT's implant-related FAQs were statistically different (χ^2=46.91, P<.001). All of ChatGPT-3.5's (100%) answers were valid, while 33.3% of Bing's answers were considered valid. Chat GPT-3.5's and ChatGPT-4's valid answer rates were higher and statistically different from those of Bing and Gemini.

Low threshold test results showed that all chatbots' responses to implant-related prosthodontists' questions

Table 6. Mean and standard deviation (SD) ratings for chatbots' answers to 10 frequently asked questions

Topic	Chatbot	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total
ChatGPT's Implant-Related FAQ	Bing	3 ±0.1	3 ±0.1	3 ±0.2	3 ±0.1	3.3 ±0.2	4 ±0.2	3 ±0.1	4.3 ±0.3	4.3 ±0.2	3 ±0.2	3.4 ±0.2
	Gemini	4 ±0.3	3.3 ±0.2	3.3 ±0.2	3 ±0.2	4 ±0.2	4 ±0.3	3 ±0.1	3.7 ±0.2	4.7 ±0.3	3 ±0.1	3.6 ±0.2
	ChatGPT 3.5	4.7 ±0.2	5 ±0.0	4.7 ±0.2	4 ±0.3	4.3 ±0.3	4.7 ±0.2	4.7 ±0.2	4 ±0.3	4.7 ±0.2	4.3 ±0.3	4.5 ±0.2
	ChatGPT 4	5 ±0.0	4.7 ±0.3	4.3 ±0.2	4 ±0.2	4.3 ±0.3	4 ±0.1	5 ±0.0	3.7 ±0.2	4.7 ±0.2	4.3 ±0.3	4.4 ±0.2
Prosthodontists' Implant-Related FAQ	Bing	4 ±0.2	4.3 ±0.3	5 ±0.0	3.7 ±0.3	5 ±0.0	3 ±0.2	5 ±0.0	5 ±0.0	4 ±0.1	4 ±0.1	4.3 ±0.1
	Gemini	5 ±0.0	4.7 ±0.2	5 ±0.0	4 ±0.1	5 ±0.0	3.7 ±0.2	4.3 ±0.3	5 ±0.0	3.3 ±0.3	4 ±0.2	4.4 ±0.1
	ChatGPT 3.5	4 ±0.2	4.7 ±0.3	4.7 ±0.3	5 ±0.0	5 ±0.0	4 ±0.2	4.7 ±0.2	3 ±0.3	4.3 ±0.2	4 ±0.3	4.3 ±0.2
	ChatGPT 4	5 ±0.0	5 ±0.0	5 ±0.0	5 ±0.0	5 ±0.0	4 ±0.3	4.7 ±0.3	3 ±0.2	4 ±0.3	4.3 ±0.2	4.5 ±0.1
ChatGPT's Fixed Prosthesis-Related FAQ	Bing	5 ±0.0	4 ±0.3	4.7 ±0.2	3.7 ±0.3	4.3 ±0.3	5 ±0.0	4 ±0.3	4.7 ±0.1	4.3 ±0.2	4.3 ±0.2	4.4 ±0.2
	Gemini	4 ±0.2	4 ±0.2	5 ±0.0	5 ±0.0	4.3 ±0.3	5 ±0.0	4 ±0.3	4 ±0.3	5 ±0.0	4.3 ±0.1	4.5 ±0.1
	ChatGPT 3.5	4 ±0.2	5 ±0.0	4.3 ±0.3	5 ±0.0	5 ±0.0	5 ±0.0	4.3 ±0.3	4.3 ±0.3	4.7 ±0.1	5 ±0.0	4.7 ±0.1
	ChatGPT 4	4.7 ±0.2	5 ±0.0	4.7 ±0.2	5 ±0.0	5 ±0.0	5 ±0.0	4.7 ±0.2	4.7 ±0.2	5 ±0.0	5 ±0.0	4.9 ±0.1
Prosthodontists' Fixed Prosthesis-Related FAQ	Bing	4.3 ±0.3	4.3 ±0.2	4 ±0.3	4.3 ±0.2	5 ±0.0	4 ±0.3	4 ±0.3	4.3 ±0.2	3.7 ±0.3	4.7 ±0.2	4.3 ±0.2
	Gemini	4.3 ±0.2	4.3 ±0.2	5 ±0.0	5 ±0.0	5 ±0.0	3.7 ±0.3	4 ±0.3	5 ±0.0	4 ±0.3	4.3 ±0.3	4.5 ±0.2
	ChatGPT 3.5	5 ±0.0	4.3 ±0.2	5 ±0.0	4.3 ±0.2	5 ±0.0	4.7 ±0.2	4.7 ±0.2	4.7 ±0.2	5 ±0.0	4.3 ±0.3	4.7 ±0.1
	ChatGPT 4	5 ±0.0	5 ±0.0	4.3 ±0.2	5 ±0.0	5 ±0.0	4 ±0.3	4.7 ±0.2	5 ±0.0	4.3 ±0.3	5 ±0.0	4.7 ±0.1
ChatGPT's CD-Related FAQ	Bing	4 ±0.2	5 ±0.0	3.3 ±0.3	4.7 ±0.2	4.7 ±0.2	3.7 ±0.3	4.3 ±0.2	4 ±0.2	3.3 ±0.3	3.3 ±0.3	4 ±0.2
	Gemini	4 ±0.3	5 ±0.0	4 ±0.3	5 ±0.0	4 ±0.3	5 ±0.0	4.3 ±0.2	4 ±0.3	4 ±0.3	4.3 ±0.2	4.4 ±0.2
	ChatGPT 3.5	4.3 ±0.2	5 ±0.0	4.3 ±0.2	5 ±0.0	4.7 ±0.2	4 ±0.3	4.7 ±0.2	4.7 ±0.2	4.7 ±0.2	4.3 ±0.2	4.6 ±0.2
	ChatGPT 4	3.3 ±0.3	5 ±0.0	5 ±0.0	5 ±0.0	4.7 ±0.2	5 ±0.0	4.3 ±0.2	4.7 ±0.2	4 ±0.3	4.3 ±0.2	4.5 ±0.1
Prosthodontists' CD-Related FAQ	Bing	4.7 ±0.2	4.3 ±0.3	3.7 ±0.3	2.7 ±0.3	3 ±0.3	4 ±0.3	3 ±0.3	2.3 ±0.3	3.7 ±0.3	3.3 ±0.3	3.5 ±0.3
	Gemini	4.7 ±0.2	3.7 ±0.3	4 ±0.3	3.3 ±0.3	3 ±0.3	4.7 ±0.2	4 ±0.3	4 ±0.3	4.7 ±0.2	5 ±0.0	4.1 ±0.2
	ChatGPT 3.5	4.7 ±0.2	4.7 ±0.2	5 ±0.0	3.3 ±0.3	4.3 ±0.3	3.7 ±0.3	3.7 ±0.3	5 ±0.0	5 ±0.0	5 ±0.0	4.4 ±0.2
	ChatGPT 4	4.3 ±0.2	2 ±0.0	5 ±0.0	4.3 ±0.3	4 ±0.3	4.7 ±0.2	4 ±0.3	5 ±0.0	4.3 ±0.3	5 ±0.0	4.3 ±0.2
ChatGPT's RPD-Related FAQ	Bing	4 ±0.3	5 ±0.0	3.3 ±0.3	4.7 ±0.3	4.7 ±0.3	3.7 ±0.3	4.3 ±0.3	4 ±0.3	3.3 ±0.3	3.3 ±0.3	4 ±0.3
	Gemini	4 ±0.3	5 ±0.0	4 ±0.3	5 ±0.0	4 ±0.3	5 ±0.0	4.3 ±0.2	4 ±0.3	4 ±0.3	4.3 ±0.2	4.3 ±0.2
	ChatGPT 3.5	4.3 ±0.3	5 ±0.0	4.3 ±0.3	5 ±0.0	4.7 ±0.2	4 ±0.3	4.7 ±0.2	4.7 ±0.2	4.7 ±0.2	4.3 ±0.3	4.6 ±0.2
	ChatGPT 4	3.3 ±0.3	5 ±0.0	5 ±0.0	5 ±0.0	4.7 ±0.2	5 ±0.0	4.3 ±0.3	4.7 ±0.2	4 ±0.3	4.3 ±0.3	4.5 ±0.2
Prosthodontists' RPD-Related FAQ	Bing	4.7 ±0.2	4.3 ±0.3	3.7 ±0.3	2.7 ±0.2	3 ±0.3	4 ±0.3	3 ±0.3	2.3 ±0.3	3.7 ±0.3	3.3 ±0.3	3.5 ±0.3
	Gemini	4.7 ±0.2	3.7 ±0.3	4 ±0.3	3.3 ±0.3	3 ±0.3	4.7 ±0.2	4 ±0.3	4 ±0.3	4.7 ±0.2	5 ±0.0	4.1 ±0.2
	ChatGPT 3.5	4.7 ±0.2	4.7 ±0.2	5 ±0.0	3.3 ±0.3	4.3 ±0.3	3.7 ±0.3	3.7 ±0.3	5 ±0.0	5 ±0.0	5 ±0.0	4.4 ±0.2
	ChatGPT 4	4.3 ±0.3	2 ±0.0	5 ±0.0	4.3 ±0.3	4 ±0.3	4.7 ±0.2	4 ±0.3	5 ±0.0	4.3 ±0.3	5 ±0.0	4.3 ±0.2

CD, complete denture; FAQ, frequently asked questions; RPD, removable partial denture.

reached a relatively high level of validity and that no significant difference (χ2=.84, *P*=.840) was found. All chatbots' responses to ChatGPT's and prosthodontists' fixed prosthodontics-related questions reached a relatively high level of validity, and no significant differences (χ^2=2.03, *P*=.565 and χ^2=2.05, *P*=.568, respectively) were found.

Based on the low threshold validity assessment, all chatbots' answers to ChatGPT's CD-related FAQ reached a high level of validity, and were statistically similar (χ^2=2.41, *P*=.098). The chatbots' answers to prosthodontists' CD-related questions were significantly different (χ^2=13.33, *P*=.004). A total of 83% of answers from Gemini, ChatGPT-3.5, and ChatGPT-4 were valid, whereas half of Bing's answers were considered valid. Bing's valid answer rates were lower and statistically different from those of ChatGPT-3.5, ChatGPT-4, and Gemini (*P*<.05).

In accordance with the low threshold validity, the chatbots' answers to ChatGPT's RPD-related FAQs were statistically different (χ^2=55.44, *P*<.001). ChatGPT-3.5's and ChatGPT-4's answers were valid, whereas 36.7% of Bing's answers were considered to be invalid. Bing's valid answer rates were lower and statistically different from those of ChatGPT-3.5, ChatGPT-4, and Gemini (*P*<.05). All chatbots' responses to prosthodontists' RPD-related questions displayed a relatively high level of validity, but there was no significant difference (χ^2=2.12, *P*=.123).

According to the high threshold validity test results (Table 7), the chatbots' answers to ChatGPT's implant-related FAQ were statistically different (χ^2=24.89, *P*<.001). Half of ChatGPT-3.5's answers were valid, while 6.7% of Bing's and Gemini's answers were considered valid. Chat GPT-3.5's and ChatGPT-4's valid answer rates were higher and statistically different from those of Bing and Gemini (*P*<.05). The high threshold validity of chatbots' responses to implant-related prosthodontists' questions ranged from 43.3% to 60%, but there was no significant difference (χ^2=1.97, *P*=.579).

High threshold test results showed that the chatbots' answers to ChatGPT's fixed prosthetics-related FAQs were statistically different (χ^2=12.33, *P*=.004). Chat GPT-3.5's and ChatGPT-4's valid answer rates were higher and statistically different from those of Bing and Gemini. The chatbots' answers to prosthodontists' fixed prosthetics-related FAQ were statistically different (χ^2=14.71, *P*=.002). A total of 73.3% of ChatGPT-4's answers were valid, and 35% of Bing's answers were considered valid. Chat GPT-3.5's and ChatGPT-4's valid answer rates were higher and statistically different from those of Bing and Gemini (*P*<.05).

Based on the high threshold validity assessment, the chatbots' answers to ChatGPT's CD-related FAQ were statistically different (χ^2=8.69, *P*=.034). A total of 60% of ChatGPT-4's answers were valid, while 26.7% of Bing's answers were considered valid. Chat GPT-3.5's and ChatGPT-4's valid answer rates were higher and statistically different from those of Bing and Gemini (*P*<.05). The chatbots' answers to prosthodontists' CD-related FAQs were statistically different (χ^2=15.67, *P*=.003). A total of 60% of ChatGPT-3.5's answers were valid, while 13.3% of Bing's answers were considered valid. Chat GPT-3.5's and ChatGPT-4's valid answer rates were higher and statistically different from those of Bing and Gemini (*P*<.05).

In accordance with the high threshold validity, the chatbots' answers to ChatGPT's RPD-related FAQs were statistically different (χ^2=22.39, *P*<.001). A total of 66.7% of ChatGPT-3.5's answers were valid, while 10% of Bing's answers were considered valid. Bing's valid answer rates were lower and statistically different from those of ChatGPT-3.5, ChatGPT-4, and Gemini (*P*<.05). The high threshold validity of chatbots' responses to

Table 7. Different chatbots' valid response rate percentages (%). Low threshold (LT), High threshold (HT). Similar superscript letters indicate no statistically significant difference between columns (*P*>.05)

Topic	Validity Test	Chatbots' Valid Response Rates (%)			
		Bing	Gemini	GPT-3.5	GPT-4
ChatGPT's Implant-Related FAQ	LT	33.3[a]	53.3[a]	100[b]	96.7[b]
	HT	6.7[a]	6.7[a]	50[b]	43[b]
Prosthodontists' Implant-Related FAQ	LT	83.3[a]	86.7[a]	90[a]	90[a]
	HT	46.7[a]	53.3[a]	43.3[a]	60[a]
ChatGPT's Fixed Prosthesis-Related FAQ	LT	96.7[a]	100[a]	96.7[a]	100[a]
	HT	43.3[a]	46.7[a]	70.0[b]	86.7[b]
Prosthodontists'Fixed Prosthesis-Related FAQ	LT	96.7[a]	96.7[a]	100[a]	100[a]
	HT	35[a]	50[a]	70[b]	73.3[b]
ChatGPT's CD-Related FAQ	LT	83.3[a]	86.7[a]	100[a]	93.3[a]
	HT	26.7[a]	40[a]	56.7[b]	60[b]
Prosthodontists' CD-Related FAQ	LT	50[a]	83.3[b]	83.3[b]	83.3[b]
	HT	13.3[a]	33.3[a]	60[b]	53.3[b]
ChatGPT's RPD-Related FAQ	LT	36.7[a]	90[b]	100[b]	100[b]
	HT	10[a]	50[b]	66.7[b]	56.7[b]
Prosthodontists' RPD-Related FAQ	LT	86.7[a]	96.7[a]	100[a]	100[a]
	HT	43.7[a]	56.7[a]	60[a]	60[a]

CD, complete denture; FAQ, frequently asked questions; RPD, removable partial denture.

prosthodontists' RPD-related questions ranged from 43.7% to 60%, and there was no statistically significant difference (χ^2=4.55, P=.208).

According to the results of the consistency assessment, Gemini, ChatGPT-4, and Bing showed acceptable consistency (Cronbach alpha of.798,.718, and.707 respectively). However, ChatGPT-3.5 displayed questionable consistency (Cronbach alpha of.646).

DISCUSSION

The purpose of this study was to evaluate the validity and consistency of AI chatbots in responding to patients' FAQs in prosthodontics; the null hypotheses that no difference would be found in the accuracy and consistency of responses among the 4 chatbots were rejected for both validity and consistency. This study's findings revealed both strengths and limitations across these models, particularly regarding accuracy, consistency, and patient education.

AI chatbots, particularly large language models, are reshaping digital communication and patient interaction by delivering responses tailored using advanced data-driven methodologies.[6,7] The data used to train these AI models varies, which can account for the differences in their responses. These variations can result in differing narratives when the chatbots generate human-like responses.[32,33] Notably, Bing was unique in providing reference links, thus allowing users to verify the information source.[16,17] This feature aligns with the importance of transparency in healthcare communication, as verifiable sources enhance credibility and support informed patient decision-making.[1,21] However, despite its transparency, Bing showed some limitations in delivering clinically accurate information when compared with ChatGPT-4, which performed better in high-validity response assessments.

Gemini's use of images presented a novel approach to assisting patient comprehension by providing visual explanations for complex prosthodontic terms,[13] aligning with studies that emphasize the value of visual aids in healthcare education.[5] However, Gemini's response accuracy varied, especially in clinically nuanced questions, indicating the need for more contextually relevant responses to complement visual aids.[22] Features like Bing's reference links[16,17] and Gemini's visual aids[13] may enhance patient education, yet inconsistencies in clinical accuracy underscore the necessity for professional oversight.[27,33] ChatGPT-3.5 provided accurate but often overly verbose responses, which may overwhelm patients. AI "hallucinations" remain a concern, emphasizing the need for clinician oversight.[20,28]

Over time the chatbot's accuracy may improve, remain stable, or detoriate.[11] In the present study, the consistency assessment indicated that the Gemini, ChatGPT-4, and Bing chatbots achieved acceptable consistency levels, whereas ChatGPT-3.5 did not. Despite acceptable consistency in the answers across repeated questions, variations were observed. However, the core information provided remained largely consistent. This consistency aligned with findings from other studies reporting that AI-generated responses can differ in expression while maintaining factual accuracy.[2] Given this, minor variability in chatbot outputs may not necessarily impact clinical applicability, provided the essential content remains intact and accurate.[27]

Overall, ChatGPT-4 demonstrated both the highest accuracy and acceptable consistency. However, response validity ranged from 43% to 86.7% under high-threshold and 83.3% to100% under low-threshold validity. Bing's responses were generally lower in accuracy and ChatGPT-3.5 and Gemini performed moderately well but often lacked the depth observed in ChatGPT-4. While AI chatbots, particularly GPT-4, demonstrate potential as patient educational tools in prosthodontics, the effective deployment of chatbots in patient education and clinical applications requires both accuracy for evidence-based information and consistency to maintain reliability.[27,29]

Limitations of the current study included lack of generalizability across a diverse patient population, biases in training data, and subjective evaluation methods with a limited number of reviewers. As prosthodontics knowledge evolves, chatbot responses may not consistently align with current clinical guidelines. Future research should incorporate diverse expert panels, real-world testing, and error analysis to enhance chatbot efficacy. Integrating AI tools into clinician feedback mechanisms could further improve their utility as supplementary resources in patient care.[11,29]

CONCLUSIONS

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

1. At the low threshold, Bing showed the lowest validity in ChatGPT's implant-related, RPD-related, and prosthodontists' CD-related questions. All other chatbots' responses to all questions demonstrated a high level of validity.
2. At the high threshold, all 4 chatbots exhibited a lower overall validity compared with the low threshold. ChatGPT-3.5 and 4 showed a higher level of validity compared with Bing and Gemini, with Gemini performing better than Bing.
3. Gemini, ChatGPT-4, and Bing chatbots displayed an acceptable level of reliability, whereas ChatGPT-3.5

did not. Among the 4 chatbots, Gemini had the highest overall consistency.
4. Currently, chatbots have limitations in providing accurate and consistent responses regarding prosthodontics-related FAQs.

REFERENCES

1. Deiana G, Dettori M, Arghittu A, et al. Artificial intelligence and public health: Evaluating ChatGPT responses to vaccination myths and misconceptions. *Vaccines*. 2023;11:1217.
2. Iannantuono GM, Bracken-Clarke D, Floudas CS, et al. Applications of large language models in cancer care: Current evidence and future perspectives. *Front Oncol*. 2023;13:1–6.
3. Janiesch C, Zschech P, Heinrich K. Machine learning and deep learning. *Electron Market*. 2021;31:685–695.
4. Steels L, Lopez de Mantaras R. The Barcelona declaration for the proper development and usage of artificial intelligence in Europe. *AI Commun*. 2018;31:485–494.
5. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: Chances and challenges. *J Dent Res*. 2020;99:769–774.
6. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521: 436–444.
7. Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks*. 2015;61:85–117.
8. Sarker IH. Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. *SN Comput Sci*. 2021;2:154.
9. Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmol Sci*. 2023;3:100324.
10. Sallam M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11:887.
11. Giansanti D. Artificial intelligence in public health: Current trends and future possibilities. *Int J Environ Res Public Health*. 2022;19:11907.
12. OpenAI. GPT-4 Technical Report.https://openai.com/research/gpt-4. Accessed January 15, 2024.
13. Anon. Introducing Gemini: Google's most capable AI model yet. https:// blog.google/technology/ai/google-gemini-ai/#sundar-note. Accessed January 15, 2024.
14. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv* 2018. 1810.04805.
15. Pichai, S. An important next step on our AI journey. https:// blog.google/intl/en-africa/products/explore-get-answers/an- important-next-step-on-our-ai-journey/.Accessed January 15, 2024.
16. Birkun AA, Gautam A. Large language model-based chatbot as a source of advice on first aid in heart attack. *CurrProblCardiol*. 2024;49:102048.
17. Birkun AA, Gautam A. Large language model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehosp Disaster Med*. 2023;38:757–763.
18. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409. 3215* 2014.
19. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183:589–596.
20. Safi Z, Abd-Alrazaq A, Khalifa M, Househ M. Technical aspects of developing chatbots for medical applications: Scoping review. *J Med Internet Res*. 2020;22:19127.
21. Austvoll-Dahlgren A, Helseth S. What informs parents' decision-making about childhood vaccinations? *JAdv Nurs*. 2010;66:2421–2430.
22. Suárez A, Díaz-Flores García V, Algar J, et al. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *Int Endod J*. 2024;57:108–113.
23. Danesh A, Pazouki H, Danesh K, et al. The performance of artificial intelligence language models in board-style dental knowledge assessment: A preliminary study on ChatGPT. *J Am Dent Assoc*. 2023;154:970–974.
24. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. *Cureus*. 2023;15:42133.
25. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? *J Stomatol Oral Maxillofac Surg*. 2023;124:8–11.
26. Diniz-Freitas M, Rivas-Mundiña B, García-Iglesias JR, et al. How ChatGPT performs in oral medicine: The case of oral potentially malignant disorders. *Oral Dis*. 2024;30:1912–1918.
27. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, et al. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int Endod J*. 2024;57:305–314.
28. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15:2–5.
29. Freire Y, Santamaría Laorden A, Orejas Pérez J, et al. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *J Prosthet Dent*. 2024;131:659.e1–659.e6.
30. Bernard A, Langille M, Hughes S, et al. A systematic review of patient inflammatory bowel disease information resources on the world wide web. *Am J Gastroenterol*. 2007;102:2070–2077.
31. Rokhshad R, Zhang P, Mohammad-Rahimi H, et al. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. *JDent*. 2024;144:104938.
32. Bhardwaz S, Kumar J. An extensive comparative analysis of chatbot technologies - ChatGPT, Google BARD and Microsoft Bing. 2nd international conference on applied artificial intelligence and computing ((ICAAIC)). Salem, India: IEEE,; 2023:673–679.
33. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. *JMed Syst*. 2023;47:33.

Corresponding author:
Dr Cortino Sukotjo
Department of Prosthodontics
University of Pittsburgh
School of Dental Medicine
240 South 40th Street
Pittsburgh, PA 19104
Email: Cos162@pitt.edu

https://doi.org/10.1016/j.prosdent.2025.03.009