This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

Evaluating the Evidence Base of Large Language Models in Answering Clinical Questions Related to Denture Care and Maintenance

Savvas N. Kamalakidis, DDS, CAGS, MPH, PhD[1,2] */ Konstantinos Michalakis, DDS, CAGS, MSc, MSc, PhD*[3,4] */Kostis Giannakopoulos, DDS, PhD5 /Eleftherios G. Kaklamanos, DDS, MSc, MA, PhD*[5–7]

[1]*Department of Prosthodontics, School of Dentistry, Faculty of Health Sciences, Aristotle University of Thessaloniki, Greece*

[2]*Department of Prosthodontics, Tufts University School of Dental Medicine, Boston, Massachusetts, USA*

[3]*Department of Restorative Sciences and Biomaterials, Boston University Henry M. Goldman School of Dental Medicine, Boston, Massachusetts, USA*

[4]*Boston University Center for Multiscale and Translational Mechanobiology, Boston, Massachusetts, USA*

[5]*School of Dentistry, European University Cyprus, Nicosia, Cyprus.*

[6]*Department of Preventive Dentistry, Periodontology and Implant Biology, School of Dentistry, Faculty of Health Sciences, Aristotle University of Thessaloniki, Greece*

[7]*Hamdan bin Mohammed College of Dental Medicine, Mohammed bin Rashid University of Medicine and Health Sciences (MBRU), Dubai, United Arab Emirates*

Correspondence to: Dr Savvas N. Kamalakidis, drkamalakidis@gmail.com

Submitted September 25, 2025; accepted December 31, 2025.

Abstract

Purpose: Large language models (LLMs) have gained significant attention and are increasingly considered as decision-support tools in healthcare. Nevertheless, their accuracy in relation to established prosthodontic guidelines remains underexplored. The purpose of this study was to evaluate and compare the evidence-based potential of answers provided by 4 LLMs to common clinical questions regarding denture care and maintenance. Material and Methods: A total of 10

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

open-ended questions pertinent to denture care and maintenance were posed to 4 distinct LLMs, namely ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3. The answers were evaluated independently by 2 prosthodontists against established guidelines for comprehensiveness, scientific accuracy, clarity, and relevance. Differences were analyzed using Friedman and Wilcoxon signed-rank tests. To assess intra-evaluator reliability, a reevaluation of the LLM responses was performed after 4 weeks, and Cronbach's α and interclass correlation coefficient (ICC) were used (α=.05). Results: ChatGPT 4o and Google Gemini Advanced recorded the highest mean scores (8.5 out of 10), followed by DeepSeek V3 (8.4 out of 10) and Microsoft Copilot (8.0 out of 10). No statistically significant differences were observed among the models. Conclusion: In this limited set of denture-care questions, LLMs often provided high-quality responses that aligned with ACP denture care guidelines, although occasional inaccuracies were observed. Their use shows potential as additional decision-support tools, but insights are limited to routine denture hygiene and maintenance questions. Caution and expert supervision are still crucial, as LLMs can't replace dental professionals in prosthodontic treatment or patient care. *Int J Prosthodont 2026. doi: 10.11607/ijp.9646*

Introduction

Digital transformation in healthcare has rapidly progressed over the past decade, with artificial intelligence (AI) and large language models (LLMs) becoming increasingly popular in both medical and dental fields.[1] According to the glossary of digital dental terms,[2] AI is defined as the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. LLMs are advanced types of AI-based applications that are designed to

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

process vast amounts of data and focus on understanding, predicting, and generating human-like text.[3] They benefit from integrating natural language processing (NLP) algorithms, which combine both natural language understanding (NLU) and generation (NLG), and apply them in chatbots to interact with end users.[4]

Historically, in healthcare, AI has been integrated into clinical decision support (CDS) systems that provide clinicians with real-time advice, patient-specific information, and evidence-based knowledge to assist in treatment decisions.[5] Deep learning (DL) techniques, particularly convolutional neural networks (CNNs), have been highly effective in analyzing images for diagnosis across various medical fields.[6] These machine learning (ML) applications require minimal human supervision to process input data into their continuously improving and calibrated statistical models, delivering validated clinical conclusions.[7] In dentistry, specific-purpose AI applications have been used for implant type recognition, dental caries diagnosis, identification of root fractures and periapical lesions, as well as for automated design of dental restorations and shade matching.[8] Additionally, advancements in AI technology have improved orthodontic treatment planning and analysis, robotic-assisted surgery in oral and maxillofacial surgery, victim identification in forensic odontology, radiographic image recognition, and electronic patient record management.[9]

Currently, conversational LLMs are used in medical and dental fields, including clinical, educational, and research applications. The rapid development of these chatbots shows how a general-purpose technology can quickly become part of specific healthcare-related areas.[10] Although they are not designed for dentistry-related content, clinicians, dental students, and the general public regularly use them, raising several ethical concerns related to “algorithm bias” and “hallucinations”.[11] The primary limitations of LLMs in dental education include their inability to

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

specify information sources and their tendency to generate fabricated citations, which poses a challenge for less experienced individuals.[12] The accuracy of chatbots in answering clinically relevant inquiries has been evaluated in various dental disciplines.[13–28] The mode of evaluation also varied, from open-ended question prompts to closed-ended ones.

The published literature on prosthodontics shows that the lowest accuracy rates in responses, regardless of the chatbot type, were related to removable dental prostheses (RDPs).[19,25] Daily care and maintenance of RDPs are critically important for the oral and systemic health of the wearer, with evidence-based guidelines provided by the American Dental Association (ADA) and the American College of Prosthodontists (ACP).[29] Since LLMs can serve as free sources of information for denture care guidelines used by clinicians and denture wearers alike, it is crucial to assess their performance in this area.

The purpose of the present study was to evaluate and compare the evidence-based potential of answers provided by 4 LLMs to common clinical questions regarding denture care and maintenance. The null hypothesis was that the generated answers would not differ in comprehensiveness, scientific accuracy, clarity, and relevance among various LLMs, compared to the scientific evidence and guidelines used as the standard.

Material and Methods

The current study did not involve human participants, so no ethical approval was obtained. Four different LLMs, ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3 were asked 10 questions related to denture care and maintenance (Table 1). The questions, created and approved by 2 board-certified prosthodontists with over 20 years of experience (S.K., K.M.), were supported by evidence from guidelines issued by the American College of

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

Prosthodontists,[29] which served as the 'reference standard' for comparing the LLMs' responses. Questions and prompts were formulated with appropriate terminology and were open-ended, requiring text-based responses. Each question was asked once to each LLM by one author (E.G.K.), with no follow-ups, rewording, or additional explanations if the LLM could not answer. The responses obtained were stored in a spreadsheet. All models were accessed through their official web interfaces in April 2025 and were queried with default settings. Parameters like temperature and randomness, which affect output variability, were not adjustable by users in the standard web versions of these systems. Therefore, all responses reflect outputs under default conditions.

The 2 evaluators independently assessed each answer generated from the 4 LLMs for every question. They graded each response on a scale from 0 (minimum) to 10 (maximum) based on a predefined rubric for comprehensiveness, scientific accuracy, clarity, and relevance.14-17 To ensure blinding, answers were anonymized by assigning a letter to each LLM, so evaluators were unaware of which LLM they were grading. The correct answer or 'reference standard,' which served as the basis for evaluation, was awarded the full score of 10 out of 10. Evaluations were scheduled at consistent times of day to minimize fatigue-related variability. Only one evaluation session was performed per weekday, and evaluators did not have access to their evaluations after completion. Four weeks later, the LLMs' answers were reevaluated and graded again, following the same protocol, to assess intra-evaluator reliability.

Computing measures of central tendency and variability summarized the data. To assess reliability, Cronbach's α and the intraclass correlation coefficient (ICC) were calculated. Additionally, to examine differences between grades, the Friedman, Kruskal-Wallis, and Wilcoxon signed-rank tests were used. All statistical analyses were performed using a statistical

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

software program (IBM SPSS Statistics, v.29.0), enhanced with the Exact Tests module for Monte Carlo simulations. The significance level for all hypothesis tests was set at α=.05.

Results

The generated responses of the 4 LLMs to the 10 clinical questions are shown in Supplemental Table 1 (available online). The descriptive statistics for the scores assigned by the 2 evaluators to the responses provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3 on 2 separate occasions, one month apart, are depicted in Table 2. The scores for ChatGPT 4o, Google Gemini Advanced, and DeepSeek V3 averaged around 8.5/10, while Microsoft Copilot's average was 8.05/10.

Table 1 Open-ended questions answered using ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3.

Question Number	Question Description
1	Assume that you are an expert prosthodontist. Please answer the following question. How should patients who wear dentures clean them?
2	Assume that you are an expert prosthodontist. Please answer the following question. Should dentures be cleaned by dentists?
3	Assume that you are an expert prosthodontist. Please answer the following question. Are there any risks associated with the use of denture cleansers?
4	Assume that you are an expert prosthodontist. Please answer the following question. Should dentures be brushed with denture creams?
5	Assume that you are an expert prosthodontist. Please answer the following question. Are there any methods to clean dentures alternative to brushing?

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

6	Assume that you are an expert prosthodontist. Please answer the following question. Should patients use denture adhesives?
7	Assume that you are an expert prosthodontist. Please answer the following question. Are there any risks associated with the use of denture adhesives?
8	Assume that you are an expert prosthodontist. Please answer the following question. How should denture adhesives be used correctly?
9	Assume that you are an expert prosthodontist. Please answer the following question. How often should dentures be relined or rebased?
10	Assume that you are an expert prosthodontist. Please answer the following question. How often should the denture wearers be checked by the dentist?

Table 2 Descriptive statistics for scores assigned by 2 evaluators to answers provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3 on 2 different times.

	Score 1								**Score 2**							
	ChatGPT 4o		**Google Gemini Advanced**		**Microsoft Copilot**		**DeepSeek V3**		**ChatGPT 4o**		**Google Gemini Advanced**		**Microsoft Copilot**		**DeepSeek V3**	
Evaluator	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Min	5	5	5.5	5	7	5.75	4.5	5	5	5.25	7	5.25	6.8	5.5	4	5.25
Median	9.4	8.9	9.3	8.6	9.0	7.8	8.7	8.5	9.0	8.9	8.80	8.8	8.0	7.8	8.4	8.3
Max	9.5	9.5	9.8	9.8	10	9	9.5	10	9.3	9.5	9.8	9.8	9.8	9	9.8	10
Mean	8.5	8.5	8.7	8.4	8.6	7.7	8.4	8.5	8.4	8.5	8.6	8.5	8.2	7.7	8.3	8.4
SEM	0.48	0.42	0.43	0.43	0.29	0.30	0.46	0.48	0.41	0.40	0.31	0.43	0.25	0.28	0.51	0.44
SD	1.51	1.31	1.34	1.36	0.93	0.96	1.46	1.50	1.28	1.26	0.99	1.37	0.78	0.90	1.62	1.40

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

CV	17.8%	15.4 %	15.4 %	16.2 %	10.8%	12.5 %	17.4 %	17.7 %	15.3 %	14.8 %	11.5 %	16.1 %	9.5 %	11.7 %	19.5 %	16.7 %

CV: Coefficient of Variation; Max: maximum; Min: minimum; SD: Standard Deviation; SEM: Standard Error of Mean

Overall, Cronbach's α and the ICC indicated high reliability for the assessments provided by the evaluators to the LLMs' answers. Cronbach's α values were above 0.8, and all ICCs were statistically significant for all LLMs (Table 3). Supporting evidence was provided by Friedman and Wilcoxon signed-rank tests, which did not indicate any statistically significant differences between the scores assigned by the 2 evaluators on both dates to the responses given by the 4 LLMs (Table 4).

Table 3 Cronbach α and Intraclass Correlation Coefficient (ICC) for scores assigned by 2 evaluators to answers provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3 on 2 different occasions, one month apart.

Large Language Models (LLMs)	**Pooled Scores 1 and 2**		
	Cronbach α	**Interclass Correlation Coefficient**	
		Single	**Average**
ChatGPT 4o	0.965	0.87 (*P*<.001)	0.97 (*P* <.001)
Google Gemini Advanced	0.813	0.52 (*P* <.001)	0.81 (*P* <.001)
Microsoft Copilot	0.804	0.51 (*P* <.001)	0.80 (*P* <.001)
DeepSeek V3	0.979	0.92 (*P* <.001)	0.98 (*P* <.001)

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

Table 4 Wilcoxon signed-rank test for differences in scores given by 2 evaluators to answers provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3, on 2 separate scoring sessions, one month apart. Friedman test for pooled scores given by 2 evaluators.

Large Language Models (LLMs) [Evaluator 1-2]	Wilcoxon Signed-rank Test		Friedman Test Pooled Scores 1 and 2
	Score1	Score 2	
ChatGPT 4o	.937	.848	.444
Google Gemini Advanced	.476	.649	.732
Microsoft Copilot	.060	.062	.060
DeepSeek V3	.986	.895	.946

As a result, an average score was calculated for each LLM based on the scores given by both evaluators across both dates, to be used in subsequent Kruskal-Wallis and Wilcoxon signed-rank tests. Figure 1 depicts the average scores of the answers to each question provided by the 4 LLMs. Table 5 presents the descriptive statistics for the average scores of the answers from the 4 LLMs. ChatGPT 4o and Google Gemini Advanced answers were rated as the best, followed by DeepSeek V3 and Microsoft Copilot. Kruskal-Wallis (P=.238) and Wilcoxon signed-rank tests showed no statistically significant differences between the average scores of the 4 LLMs (Table 6).

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

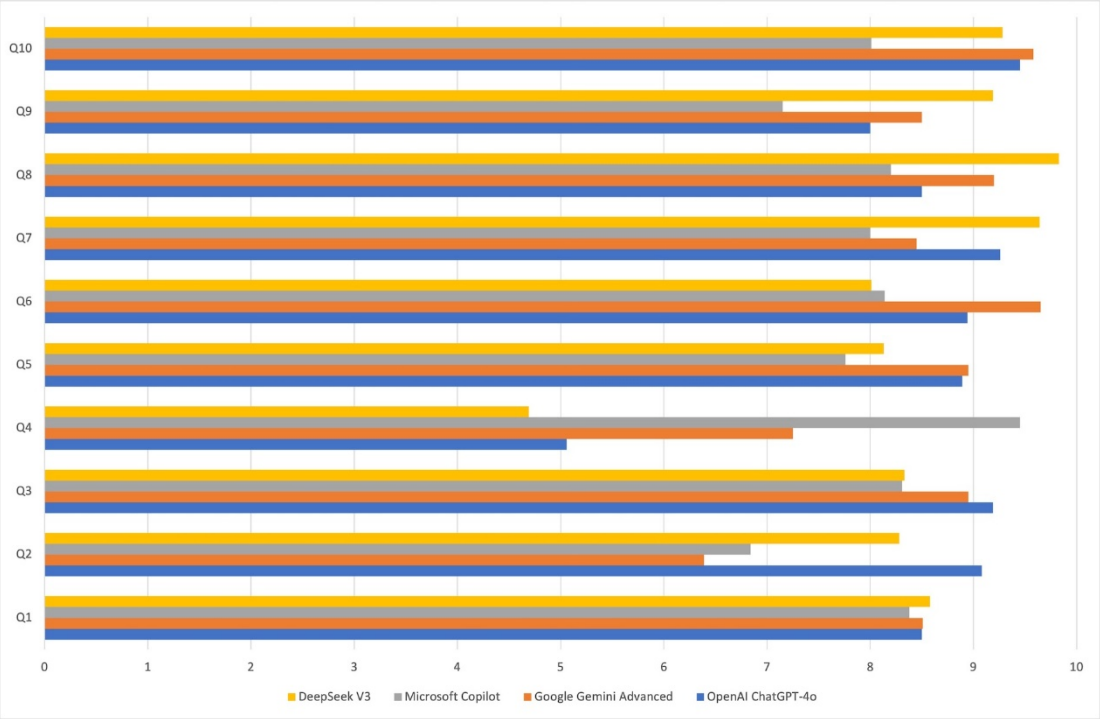

Fig 1 Average scores for answers to each question provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3.

Table 5 Descriptive statistics for average scores of answers provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3.

Average score	ChatGPT 4o	Google Gemini Advanced	Microsoft Copilot	DeepSeek V3
Min	5.1	6.4	6.8	4.7
Median	8.9	8.7	8.1	8.5
Max	9.5	9.7	9.5	9.8
Mean	8.5	8.5	8.0	8.4
SEM	0.40	0.32	0.22	0.46
SD	1.28	1.02	0.80	1.45
CV	15.1%	12.0%	10.0%	17.3%

CV: Coefficient of Variation; Max: maximum; Min: minimum; SD: Standard Deviation; SEM: Standard Error of Mean

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

Table 6 Wilcoxon signed-rank test for average scores of answers provided by ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3.

Large Language Models (LLMs) [Average Scores]	**Wilcoxon Signed-rank Test (*P*-value)**
ChatGPT 4o vs Google Gemini Advanced	.627
ChatGPT 4o vs Microsoft Copilot	.085
ChatGPT 4o vs DeepSeek V3	.713
Google Gemini Advanced vs Microsoft Copilot	.141
Google Gemini Advanced vs DeepSeek V3	.864
Microsoft Copilot vs DeepSeek V3	.131

Discussion

The present investigation evaluated the ability of 4 LLMs to answer clinically relevant questions regarding denture care and maintenance. Although small numerical differences were found between the examined LLMs, the null hypothesis could not be rejected because the statistical tests indicated no significant differences among the models.

Specifically, this study assessed the performance of 4 modern large language models (LLMs)—ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3,—on 10 open-ended clinical questions related to denture care and maintenance. Using a predefined, multidimensional 10-point rubric (covering comprehensiveness, factual accuracy, clarity, and relevance), the models consistently achieved high average scores with no statistically significant differences among them, indicating practical equivalence in single-turn prompts for the tested questions. Reliability measures were strong across both rating sessions: internal consistency and

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

inter-rater agreement (Cronbach's α and ICC) confirmed the stability of the scoring process and the reliability of the conclusions.

The overall average score, ranging from 8.0 to 8.5 out of 10, indicates that all tested LLMs were able to produce reasonably accurate, comprehensive, and clinically relevant responses. The slightly higher performance of ChatGPT 4o and Google Gemini Advanced may be due to their broader multimodal pretraining and larger training datasets. In contrast, Microsoft Copilot's marginally lower score might relate to its optimization for productivity tasks rather than clinical reasoning.[16] DeepSeek V3, despite being an open-source model, showed performance comparable to the proprietary models, demonstrating the rapid progress of open-source AI in healthcare applications. The results are consistent with previous investigations that used the same methodology in the field of periodontology.[15–17]

The LLMs' responses were scored by 2 experienced prosthodontists, on 2 different occasions, one month apart. The high values for Cronbach's α and ICCs indicated a firm consistency and reproducibility of the evaluator's ratings. Repeated scoring after a four-week wash-out period confirmed intra-evaluator reliability, demonstrating that the evaluation rubric was robust and applicable to diverse responses. In a recent investigation into answering multiple-choice and short-answer questions in fixed prosthodontics over a 4-week period, the author concluded that the model's performance improved when provided with the ground truth responses.[20] In the present investigation, the answers were benchmarked against the ADA/ACP guidelines in effect from 2011 and updated by ADA in 2024, practically eliminating the possible confounding factor of knowledge cut-off dates employed by all tested LLMs.

Our results, which include high-quality, mostly guideline-aligned answers and no significant differences among contemporary LLMs under single-turn prompting, align with recent

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

prosthodontic evaluations that show strong—but not perfect—performance on maintenance and exam-style questions. A recent study investigating ChatGPT's performance in the fields of removable and fixed prosthodontics,[19] found acceptable accuracy with moderate to substantial repeatability, while also highlighting key gaps, especially as case complexity increased. This reflects our finding that performance is strongest on routine, well-codified tasks. Similarly, research on board-style assessments has shown that newer models, along with contextual prompting or fine-tuning, can lead to improved results. This correlates with our findings on modern systems, which display tight score dispersion.[26] Beyond exams, creating patient-facing materials demonstrates that LLMs can generate clear and understandable content relevant to the field of dentistry. However, this content still requires expert review, supporting our recommendation that clinicians should supervise outputs before being used with patients.[27] Recent reviews suggest that dentistry should adopt standardized rubrics and transparent reporting to benchmark multiple LLMs—a method our design employs—while remaining cautious of hallucinations and domain-specific blind spots.[1] Lastly, research on prosthesis frequently asked questions mirrors our finding that guidance is generally helpful, but occasionally inaccurate, highlighting the importance of verifying information against trusted sources.[28]

The authors are unaware of any previous study that has systematically evaluated and compared the responses of 4 widely used LLMs in the field of removable prosthodontics, with specific emphasis on denture care and maintenance. This focus is novel, as most prior investigations have assessed a single model, concentrated on fixed prosthodontics or examination-style questions, or limited their evaluation to accuracy alone. In contrast, the present study employed a carefully designed, multifaceted rubric that assessed not only correctness but also clarity, relevance, and completeness, which are qualities that closely align with the needs of

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

clinical practice. Two experienced prosthodontists independently evaluated the responses in blinded, repeated sessions, yielding high reliability and supporting the robustness of the outcome measures. The single-turn prompting strategy mirrors how clinicians, students, or patients are most likely to use such systems in practice, avoiding artificial improvements associated with prompt engineering. Benchmarking against established ACP denture-care recommendations ensured that responses were judged against authoritative, evidence-based standards. Together, these design elements provide a novel and methodologically rigorous contribution to the growing literature on AI in prosthodontics, highlighting both the potential and the risks of LLMs as decision-support tools in removable prosthodontics.

Since only 10 questions were tested, this limits the extent to which the results can be applied to the full spectrum of prosthodontics. Future studies should expand the question set across prosthodontics and restorative dentistry, evaluate the impact of conversational depth and iterative prompting, and explore the integration of domain-specific context or fine-tuned dental LLMs. Another limitation of this study concerns the stochastic nature of LLM outputs. Each question was asked only once per model using default settings, without repeated runs. Since these systems rely on probabilistic sampling methods, which include temperature and randomness, different outputs may appear in separate sessions. While documenting default access conditions enhances transparency, it does not eliminate variability. Future studies should ask each model multiple times per question, calculate the average and standard deviation of evaluator scores, and examine how much outputs vary. Additionally, some proprietary systems (Microsoft Copilot and Google Gemini Advanced) are continuously updated without public versioning, which further affects reproducibility over time. Equally important is assessing patient-facing usability, including readability, accessibility, and the potential for misunderstandings. As shown in Supplemental

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

Table 1, Question 4, Microsoft Copilot incorrectly identified denture adhesives as denture cleansers, which could mislead patients into using incorrect cleaning methods. Similarly, a response in Question 5 recommended vinegar soaks without warning about possible damage to metal-based dentures. These are potentially harmful inaccuracies, unlike harmless omissions such as not advising to clean dentures over a towel or in a water-filled sink (ChatGPT-4o, Question 1). Ethical concerns such as algorithmic bias, lack of transparent sourcing, and the risk of overreliance on AI-generated responses also require further investigation.

Finally, although repeated questioning would give a more reliable estimate of stochastic variability, this method is not currently practical for proprietary LLMs like Copilot or Gemini. These systems are updated constantly without notice, so repeated queries over time would reflect model updates rather than true random variation. As a result, the current study provides only a single snapshot of performance at a single time point. This limitation highlights the difficulty of reproducibility when assessing commercial, cloud-based models, compared to open-source systems, which can be version-controlled for research. When used correctly, LLMs can act as helpful tools for clinicians. They can help with immediate recall of standard protocols, ensure consistent phrasing of patient instructions, and quickly summarize standard recommendations for common maintenance issues. Additionally, they can translate technical content into easy-to-understand language, making patient education materials clearer and more consistent. However, their role should be strictly supportive. As a result, model outputs intended for patient use should be checked against expert guidance and reviewed by a clinician, with a focus on individualized risk factors.

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

Conclusions

Based on the findings of the present investigation, it could be concluded that all 4 tested LLMs: ChatGPT 4o, Google Gemini Advanced, Microsoft Copilot, and DeepSeek V3, generally provided responses on denture care and maintenance that were rated highly in terms of comprehensiveness, accuracy, clarity, and relevance, though occasional inaccuracies of clinical relevance were observed. Although performance was promising, the results are limited to a small set of routine hygiene and maintenance questions. The study design did not account for stochastic variability, and safety analysis was not comprehensive; therefore, the outputs cannot be assumed to apply broadly across all areas of prosthodontics. These models should not replace dental professionals, as their improper use could negatively impact patient care, leading to potentially life-threatening conditions. Finally, the results of the study should not be generalized for every aspect of the prosthodontic field.

Supplemental Materials

Supplemental materials will be available in the final version of this article.

Acknowledgments

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Grammarly in order to improve the grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

None

Declaration of Interest:

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

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.

References

1. Umer F, Batool I, Naved N. Innovation and application of large language models (LLMs) in dentistry - a scoping review. BDJ Open 2024;10:90.

2. Glossary of Digital Dental Terms, 2nd Edition: American College of Prosthodontists and ACP Education Foundation. J Prosthodont 2021;30:172-181.

3. 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.

4. Büttner M, Leser U, Schneider L, Schwendicke F. Natural language processing: Chances and challenges in dentistry. J Dent 2024;141:104796.

5. Ayorinde A, Mensah DO, Walsh J, Ghosh I, Ibrahim SA, Hogg J, et al. Health care professionals' experience of using AI: Systematic review with narrative synthesis. J Med Internet Res 2024;26:e55766.

6. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Clinical applications of artificial intelligence in periodontology: A scoping review. Medicina (Kaunas) 2025;61:1066.

7. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury 2023;54:S69-S73.

8. 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;129:276-292.

9. Samaranayake L, Tuygunov N, Schwendicke F, Osathanon T, Khurshid Z, Boymuradov SA, et al. The transformative role of artificial intelligence in dentistry: A comprehensive overview. Part 1: Fundamentals of AI, and its contemporary applications in dentistry. Int Dent J 2025;75:383-396.

10. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). NPJ Digit Med 2024;7:183.

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

11. Tuygunov N, Samaranayake L, Khurshid Z, Rewthamrongsris P, Schwendicke F, Osathanon T, et al. The transformative role of artificial intelligence in dentistry: A comprehensive overview Part 2: The promise and perils, and the International Dental Federation Communique. Int Dent J 2025;75:397-404.

12. Alhazmi N, Alshehri A, BaHammam F, Philip M, Nadeem M, Khanagar S. Can large language models serve as reliable tools for information in dentistry? A systematic review. Int Dent J 2025;75:100835.

13. Giannakopoulos K, Kavadella A, Salim AA, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: Comparative mixed methods study. J Med Internet Res 2023;25:e51580.

14. Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evidence-based potential of generative artificial intelligence large language models in orthodontics: A comparative study of ChatGPT, Google Bard, and Microsoft Bing. Eur J Orthod 2024:cjae017.

15. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: Assessing their performance in clinically relevant questions. J Prosthet Dent 2024:S0022-3913(24)00714-00715.

16. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Evaluation of large language model performance in answering clinical questions on periodontal furcation defect management. Dent J (Basel) 2025;13:271.

17. Koidou VP, Chatzopoulos GS, Tsalikis L, Kaklamanos EG. Large language models in peri-implant disease: How well do they perform? J Prosthet Dent 2025:S0022-3913(25)00102-00107.

18. Dermata A, Arhakis A, Makrygiannakis MA, Giannakopoulos K, Kaklamanos EG. Evaluating the evidence-based potential of six large language models in paediatric dentistry: A comparative study on generative artificial intelligence. Eur Arch Paediatr Dent 2025;26:527-535.

19. 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.e1-659.e6.

20. 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(25)00400-00407.

21. Camargo ES, Quadras ICC, Garanhani RR, de Araujo CM, Stuginski-Barbosa J. A comparative analysis of three large language models on bruxism knowledge. J Oral Rehabil 2025;52:896-903.

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.

This peer-reviewed, accepted manuscript will undergo final editing and production prior to print publication.
Any blinded information will be available then.

22. Salem M, Karasan D, Revilla-León M, Barmak AB, Sailer I (2025). Performance of artificial intelligence-based chatbots (ChatGPT-3.5 and ChatGPT-4.0) answering the International Team of Implantology exam questions. J Esthet Restor Dent, https://doi.org/10.1111/jerd.13496.

23. Tussie C, Starosta A. Comparing the dental knowledge of large language models (2024). Br Dent J, https://doi.org/10.1038/s41415-024-8015-2.

24. Almalki A, Althubaitiy RO, Alkhtani F, Anadioti E, Abozaed W. Assessment of ChatGPT's performance on the ACP 2024 National Prosthodontics Resident Exam (NPRE) (2025). Eur J Dent Educ, https://doi.org/10.1111/eje.70045.

25. Eraslan R, Ayata M, Yagci F, Albayrak H. Exploring the potential of artificial intelligence chatbots in prosthodontics education. BMC Med Educ 2025;25:321.

26. Dashti M, Khosraviani F, Azimi T, Hefsi D, Ghasemi S, Fahimipour A, et al. Assessing ChatGPT-4's performance on the US prosthodontic exam: Impact of fine-tuning and contextual prompting vs. base knowledge, a cross-sectional study. BMC Med Educ 2025;25:761.

27. Sivaramakrishnan G, Almuqahwi M, Ansari S, Lubbad M, Alagamawy E, Sridharan K. Assessing the power of AI: A comparative evaluation of large language models in generating patient education materials in dentistry. BDJ Open 2025;11:59.

28. Esmailpour H, Rasaie V, Babaee Hemmati Y, Falahchai M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. BMC Oral Health 2025;25:574.

29. Felton D, Cooper L, Duqum I, Minsley G, Guckes A, Haug S, et al. Evidence-based guidelines for the care and maintenance of complete dentures: a publication of the American College of Prosthodontists. J Prosthodont 2011;Suppl 1:S1-S12.

© 2026 BY QUINTESSENCE PUBLISHING CO, INC. PRINTING OF THIS DOCUMENT IS RESTRICTED TO PERSONAL USE ONLY.
NO PART MAY BE REPRODUCED OR TRANSMITTED IN ANY FORM WITHOUT WRITTEN PERMISSION FROM THE PUBLISHER.