THE JOURNAL OF PROSTHETIC DENTISTRY

RESEARCH AND EDUCATION

Evidence-based evaluation of large language models in advanced clinical decision-making for removable prosthodontics

Savvas N. Kamalakidis, DDS, CAGS, MPH, PhD,[a] Damian J. Lee, DDS, MS,[b] Konstantinos Vazouras, DDS, DMD, MDS,[c] Kostis Giannakopoulos, DDS, PhD,[d] and Eleftherios G. Kaklamanos, DDS, MSc, MA, PhD[e]

ABSTRACT

Statement of problem. Artificial intelligence is widely used to answer questions about prosthodontic treatments, but responses regarding removable prosthodontics remain limited with gaps.

Purpose. Large language models (LLMs) have been increasingly used by clinicians and trainees to access clinical information. However, their ability to make accurate, evidence-based decisions in removable prosthodontics remains unclear. This study aimed to evaluate and compare the performance of 5 LLMs in responding to advanced clinical questions related to removable complete and partial dentures.

Material and methods. A curated set of 10 advanced, evidence-based clinical questions covering key domains in removable prosthodontics was developed based on current consensus statements, systematic reviews, and professional guidelines. Each question was posed independently to 5 LLMs using single-turn prompts and default settings. Two board-certified prosthodontists independently evaluated each response using a predefined 10-point rubric assessing scientific accuracy, comprehensiveness, clarity, and clinical relevance, with guideline-based reference answers serving as the standard. Intra-rater reliability was evaluated by repeat scoring after a 4-week interval. Statistical analyses included reliability testing, a linear mixed-effects model, and nonparametric comparisons among models (α=.05).

Results. DeepSeek V3 recorded the highest mean scores, averaged across both evaluators and both scoring sessions (8.1/10), followed by ChatGPT-4o (7.2/10), Google Gemini Advanced (6.7/10), Microsoft Copilot (6.3/10), and ChatGPT-4.0 (6.0/10). Statistically significant differences were observed among the models ($P<.001$). The Cronbach α and intraclass correlation coefficients (ICC) indicated high internal consistency and reliability.

Conclusions. Contemporary LLMs can provide clinically relevant responses to advanced questions in removable prosthodontics, extending into higher-level diagnostic and treatment-planning domains, with DeepSeek V3 demonstrating the highest performance. However, their output should be interpreted with caution and used only as adjunctive decision-support tools under expert clinician oversight. (J Prosthet Dent xxxx;xxx:xxx-xxx)

The digitalization of healthcare has advanced considerably over the past decade, with artificial intelligence (AI) and large language models (LLMs) attaining widespread recognition within diverse scientific domains.[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 AI systems designed and trained to process large volumes of data from diverse sources, with the goal of understanding, predicting, and generating responses that resemble human-like communication.[3] They can provide

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.
[a]Assistant Professor, Department of Prosthodontics, School of Dentistry, Aristotle University of Thessaloniki, Greece; and Adjunct Assistant Professor, Department of Prosthodontics, Tufts University School of Dental Medicine, Boston, MA.
[b]Associate Professor and Chair, Department of Prosthodontics, Tufts University School of Dental Medicine, Boston, MA.
[c]Associate Professor, Department of Prosthodontics, Tufts University School of Dental Medicine, Boston, MA.
[d]Associate Professor, School of Dentistry, European University Cyprus, Nicosia, Cyprus.
[e]Associate Professor, Department of Preventive Dentistry, Periodontology and Implant Biology, School of Dentistry, Aristotle University of Thessaloniki, Greece; Associate Professor, School of Dentistry, European University Cyprus, Nicosia, Cyprus; and Adjunct Associate Professor,Hamdan bin Mohammed College of Dental Medicine, Mohammed bin Rashid University of Medicine and Health Sciences (MBRU), Nicosia, Cyprus.

Clinical Implications

While contemporary LLMs can provide information to patients regarding removable prosthodontics, clinicians should be aware that answers provided by AI may vary significantly.

specific information, respond to questions, and participate in discussions on various topics.[4] They gain advantages by integrating natural language processing (NLP) algorithms that combine natural language understanding (NLU) and natural language generation (NLG), especially when used in chatbots for user interaction.[5]

In healthcare, AI has been integrated into clinical decision support (CDS) systems that provide clinicians with real-time guidance, offering personalized responses and feedback that are crucial for individual learning and information acquisition to support treatment decisions.[6] Deep learning (DL) methods, particularly convolutional neural networks (CNNs), have demonstrated substantial efficacy in image analysis for diagnostic purposes across various medical fields.[7] These machine learning (ML) applications require minimal human oversight to convert input data into their continuously refined and calibrated statistical models, providing validated clinical conclusions.[8] In dentistry, AI applications with specific purposes have been utilized for recognizing implant types, diagnosing dental caries, identifying root fractures and periapical lesions, and automating the design of dental restorations and shade matching.[9] Furthermore, developments in AI have enhanced orthodontic treatment planning and analysis, enabled robotic-assisted procedures in oral and maxillofacial surgery, improved victim identification in forensic odontology, and optimized electronic patient records management.[10]

At present, conversational LLMs have been employed within the medical and dental fields, serving clinical, educational, and research functions. The broad accessibility of AI-powered chatbots has revolutionized how patients access and use health-related information. More people are turning to conversational AI tools for guidance on symptoms, treatment choices, and postoperative care, frequently doing so before or instead of consulting a healthcare professional.[11] Although these models are not specifically trained on dentistry-related content, they are often used by clinicians, dental students, and the general public. This common usage has prompted ethical concerns, especially about algorithmic bias and the production of incorrect or fabricated information ("hallucinations").[12] In dental education, a primary limitation of LLMs has been their inability to consistently cite sources and their tendency to generate fictitious references, which raises significant concerns.[13] The performance of chatbot-generated responses to clinically relevant questions has been assessed across multiple dental disciplines with a wide range of evaluation frameworks and using open-ended to closed-ended prompts.[14–30]

The published literature on prosthodontics indicates that the lowest response accuracy, across various chatbot types, was associated with removable dental prostheses (RDPs).[20,26] Additionally, when evaluating chatbots' validity in responses to frequently asked prosthodontic questions, those related to removable complete and partial dentures received lower scores than those for dental implants and fixed prosthodontics.[26] While these tools offer convenience and accessibility, their use raises concerns regarding misinformation, oversimplification of complex clinical concepts, and potential deviation from established professional guidelines. Clinicians increasingly depend on clinical guidelines, systematic reviews, and professional recommendations for diagnosis, treatment planning, and long-term care. As patients access more health information online, gaps between evidence-based advice and publicly available content can shape patient expectations and compliance. It is crucial to ensure the accuracy, reliability, and transparency of information sources, particularly given emerging technologies such as AI-driven tools that support clinical decision-making and patient education.

This study aimed to evaluate and compare the evidence-based performance of 5 LLMs in responding to advanced clinical questions related to removable complete and partial dentures. The null hypotheses were that no significant differences would be observed among the tested LLMs in terms of comprehensiveness, scientific accuracy, clarity, and clinical relevance when compared with the scientific evidence and professional guidelines used as the reference standard.

MATERIAL AND METHODS

The present investigation did not involve human participants, and, therefore, no ethical approval was required. Ten advanced clinical questions on removable complete and partial dentures were submitted to 5 LLMs: Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0 (Table 1). These questions were developed by 3 board-certified prosthodontists (S.K., D.L., and K.V.). The questions were designed to assess evidence-based clinical reasoning and decision-making rather than factual recall, requiring integration of current guidelines, systematic evidence, and clinical judgment. The LLM responses were compared with evidence-based guidelines published by the American College of Prosthodontists,[31,32] systematic

reviews of selected scientific literature,[33–35] best-evidence consensus statements,[36,37] and published papers in prosthodontic journals,[38–41] which served as the reference standard. All 3 prosthodontists agreed, following discussion, on the source and validity of the reference standard. The questions were open-ended, employed appropriate professional terminology, and required narrative text responses. Each question was posed once to each LLM by a single investigator (E.G.K.), without follow-up queries, rephrasing, or additional clarification when an answer could not be generated. All responses were recorded in a spreadsheet. All LLMs were accessed in April 2025 via their official web-based platforms using default system settings. User-controlled parameters, such as temperature or randomness, that may have influenced response variability were not adjustable in the standard web interfaces. Consequently, all outputs reflected responses generated under default conditions.

Two evaluators (S.K., K.V.) assessed each response produced by the LLMs for all questions. Responses were scored on a 0 to 10 scale using a predefined rubric evaluating comprehensiveness, scientific accuracy, clarity, and relevance. This rubric had been used in peer-reviewed studies examining LLM performance in orthodontics, periodontology, peri-implant disease, and dental education.[14–19] Responses perfectly aligned with the reference standard received a score of 10, while a score of 0 indicated complete deviation from the reference standard. Final scores for each response were computed as the mean of the 4 equally weighted evaluation domains. To maintain blinding, each LLM was assigned a letter identifier so that evaluators were unaware of the source model. Evaluations were conducted at standardized times to reduce fatigue-related bias, with only 1 evaluation session scheduled per weekday. Evaluators did not have access to their previous assessments after completion. Four weeks later, all responses were re-evaluated using the same protocol to assess intra-evaluator reliability.

Measures of central tendency and variability were used to summarize the data. To evaluate reliability, the Cronbach α, Pearson r, Spearman rho, and the intraclass correlation coefficient (ICC) were calculated. Furthermore, a linear mixed-effects model was used to analyze differences between grades, accounting for the hierarchical data structure. The Friedman test and the Wilcoxon signed-rank test were also employed. All statistical analyses were conducted using a statistical software program (IBM SPSS Statistics, v.29.0; IBM Corp), enhanced with the Exact Tests module for Monte Carlo simulations (α=.05 for all hypothesis tests).

RESULTS

The responses generated by the 5 LLMs to the 10 clinical questions and the rubric are presented in Supplemental Tables 1 and 2 (available online). Table 2 presents descriptive statistics for the scores assigned by the 2 evaluators to the responses of the 5 LLMs. DeepSeek V3 scored higher on both evaluation sessions, followed by ChatGPT-4o, Google Gemini Advanced, Microsoft Copilot, and then ChatGPT-4.0.

Table 3 presents the correlation between the scores given by the 2 evaluators. Both the Pearson *r* and Spearman rho revealed statistically significant, strong, and very strong correlations among evaluators' scores across all tested LLMs in both scoring sessions, indicating that the 2 evaluators rated the LLMs' answers similarly. Additionally, the Cronbach α values exceeded 0.8, and all ICCs were statistically significant, indicating high internal consistency and reliability (Table 4). The Wilcoxon signed rank tests and Friedman tests were performed on scores from both evaluators across both scoring sessions for each LLM. These tests revealed no overall statistically significant differences in evaluators' scores across any of the LLMs, supporting high inter-evaluator reliability (Table 5).

Table 1. Open-ended questions answered using Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0

Question Number	Question Description
1	Assume that you are an expert prosthodontist. Please answer the following question: Is bilateral balanced occlusal scheme better than lingualized for Complete Dentures?
2	Assume that you are an expert prosthodontist. Please answer the following question: Is there a difference regarding the accuracy of fit between conventional and digitally fabricated Removable Partial Dentures?
3	Assume that you are an expert prosthodontist. Please answer the following question: Is there a difference in bond strength regarding denture liners between conventional and CAD/CAM dentures?
4	Assume that you are an expert prosthodontist. Please answer the following question: How often and under what conditions should removable complete dentures and removable partial dentures be replaced?
5	Assume that you are an expert prosthodontist. Please answer the following question: Is altered cast technique beneficial for Kennedy class I RPD?
6	Assume that you are an expert prosthodontist. Please answer the following question: What is the most efficient oral care regimen for removable prostheses hygiene?
7	Assume that you are an expert prosthodontist. Please answer the following question: What is the most effective treatment for Denture Stomatitis?
8	Assume that you are an expert prosthodontist. Please answer the following question: Are digital impression procedures for Complete Dentures more accurate than conventional ones?
9	Assume that you are an expert prosthodontist. Please answer the following question: Do removable prostheses increase the risk for aspiration pneumonia?
10	Assume that you are an expert prosthodontist. Please answer the following question: How does the clinician establish the correct Vertical Dimension of Occlusion for an edentulous patient?

Table 2. Descriptive statistics for scores assigned by 2 evaluators to answers provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0 on 2 different sessions

	Microsoft Copilot		Google Gemini Advanced		DeepSeek v3		ChatGPT-4o		ChatGPT-4.0	
	Score 1									
Evaluator	1	2	1	2	1	2	1	2	1	2
Min	3	3	3	2	6	7	5	5	4	3
Median	7.0	7.0	7.0	6.5	8.0	8.5	7.5	7.0	6.0	6.0
Max	8	8	9	10	9	9	8	8	7	8
Mean	6.2	6.1	6.8	6.6	8.1	8.2	7.2	7.1	5.8	6.0
SEM	0.57	0.51	0.61	0.75	0.31	0	0.33	0.28	0.42	0.52
SD	1.81	1.62	1.92	2.37	0.99	0.92	1.03	0.88	1.32	1.63
CV	29.3%	26.1%	28.3%	35.8%	12.3%	11.2%	14.3%	12.3%	22.7%	27.2%
	Score 2									
Evaluator	1	2	1	2	1	2	1	2	1	2
Min	3	3	3	3	6	7	5	6	3	4
Median	6.5	6.5	7.5	6.5	8.0	8.5	7.5	7.0	6.0	6.0
Max	8	8	9	10	10	9	8	9	8	9
Mean	6.3	6.2	6.8	6.7	8.0	8.1	7.2	7.3	5.9	6.2
SEM	0.58	0.47	0.66	0.73	0.39	0.31	0.33	0.30	0.48	0.51
SD	1.83	1.49	2.10	2.31	1.25	0.99	1.03	0.95	1.52	1.62
CV	29.0%	23.7%	30.9%	34.5%	15.6%	12.3%	14.3%	13.0%	25.8%	26.1%

CV, coefficient of variation; Max, maximum; Min, minimum; SD, standard deviation; SEM, standard error of mean.

Table 3. Correlation between the scores given by 2 evaluators to answers provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0 on 2 different sessions

Large Language Models (LLMs) [Evaluator 1-2]	Score 1		Score 2	
	Pearson Correlation (*P* value)	Spearman Rho (*P* value)	Pearson Correlation (*P* value)	Spearman Rho (*P* value)
Microsoft Copilot	0.931 (<.001)	0.824 (.003)	0.939 (<.001)	0.942 (<.001)
Google Gemini Advanced	0.953 (<.001)	0.962 (<.001)	0.949 (<.001)	0.972 (<.001)
DeepSeek V3	0.705 (.023)	0.660 (.038)	0.896 (<.001)	0.910 (<.001)
ChatGPT-4o	0.836 (.003)	0.743 (.014)	0.726 (.017)	0.688 (.028)
ChatGPT-4.0	0.874 (.001)	0.847 (.002)	0.910 (<.001)	0.892 (.001)

Statistically significant values shown in bold ($P<.05$).

To account for the hierarchical structure of the data, a linear mixed-effects model was applied with LLM as a fixed effect and evaluator, question, and session included as random effects. The analysis revealed a statistically significant effect of LLM on the assigned scores ($P<.001$), indicating differences in performance among the evaluated models. Pairwise comparisons with Bonferroni adjustment demonstrated that DeepSeek V3 achieved significantly higher scores compared to all other LLMs. ChatGPT-4o also showed significantly higher scores compared to Microsoft Copilot ($P=.012$) and ChatGPT-4.0 ($P=.011$). Google Gemini Advanced performed significantly better than ChatGPT-4.0 ($P=.039$), while no statistically significant differences were observed between Google Gemini Advanced and ChatGPT-4o ($P=.680$) or between Microsoft Copilot and Google Gemini Advanced ($P=.180$). Additionally, no significant difference was found between Microsoft Copilot and ChatGPT-4.0 ($P=.401$) (Table 6).

As a result, an average score (Fig. 1) was calculated for each LLM based on the scores given by both evaluators across both scoring sessions to be used in the nonparametric Friedman and Wilcoxon signed rank tests for comparison purposes. Table 7 presents descriptive statistics for the average scores of the answers provided by the 5 LLMs. Notably, DeepSeek V3 scored higher (8.1/10), followed by ChatGPT-4o (7.2/10), Google Gemini Advanced (6.7/10), Microsoft Copilot (6.3/10), and then ChatGPT-4.0 (6.0/10). Based on Friedman's test, statistically significant differences in the average scores of the 5 LLMs were observed ($P=.001$). DeepSeek V3 outperformed all other tested LLMs, while ChatGPT-4o was distinct from Microsoft Copilot and ChatGPT-4.0. Google Gemini Advanced also differed from ChatGPT-4.0 (Table 8). These findings were consistent with the linear mixed-effects model, supporting their robustness.

DISCUSSION

The present investigation evaluated the ability of 5 LLMs to respond to advanced clinical questions related to removable complete and partial dentures. The null hypotheses that no significant differences would be observed among the tested LLMs in terms of comprehensiveness, scientific accuracy, clarity, and clinical relevance when compared with the scientific evidence and professional guidelines used as the reference standard were rejected because the statistical tests indicated significant differences among the models.

Specifically, this study assessed the performance of 5 contemporary LLMs, namely Microsoft Copilot, Google

Table 4. Cronbach α and Intraclass Correlation Coefficient (ICC) for scores assigned by 2 evaluators to answers provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0 on 2 different sessions, 1 month apart

Large Language Models (LLMs)	Score 1			Score 2			Pooled Scores 1 & 2		
	Cronbach a	ICC (*P*-value)		Cronbach a	ICC (*P*-value)		Cronbach a	ICC (*P*-value)	
		Single	Average		Single	Average		Single	Average
Microsoft Copilot	0.961	0.93 (<.001)	0.96 (<.001)	0.959	0.92 (<.001)	0.96 (<.001)	0.978	0.92 (<.001)	0.98 (<.001)
Google Gemini Advanced	0.966	0.93 (<.001)	0.97 (<.001)	0.971	0.94 (<.001)	0.97 (<.001)	0.985	0.94 (<.001)	0.99 (<.001)
DeepSeek V3	0.826	0.70 (.008)	0.83 (.008)	0.932	0.87 (<.001)	0.93 (<.001)	0.937	0.79 (<.001)	0.94 (<.001)
ChatGPT-4o	0.904	0.82 (.001)	0.90 (.001)	0.839	0.72 (.006)	0.84 (.006)	0.925	0.75 (<.001)	0.93 (<.001)
ChatGPT-4.0	0.924	0.86 (<.001)	0.92 (<.001)	0.952	0.91 (<.001)	0.95 (<.001)	0.975	0.91 (<.001)	0.98 (<.001)

Statistically significant values shown in bold (*P*<.05).

Table 5. Wilcoxon signed-rank test for differences in scores given by 2 evaluators to answers provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0, on 2 separate sessions, 1 month apart

Large Language Models (LLMs) [Evaluator 1–2]	Wilcoxon Signed-Rank Test		Friedman Test Pooled Scores 1 and 2
	Score1	Score 2	
Microsoft Copilot	0.898	0.952	0.875
Google Gemini Advanced	0.683	0.975	0.908
DeepSeek V3	0.974	0.967	0.702
ChatGPT-4o	1.000	0.923	0.906
ChatGPT-4.0	0.671	0.360	0.378

Friedman test for pooled scores given by 2 evaluators.

Table 6. Pairwise comparisons between Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0

Large Language Models (LLMs) [Average Scores]	Mixed-Effects Model, Bonferroni Adjusted (*P*-value)
Microsoft Copilot vs Google Gemini Advanced	.180
Microsoft Copilot vs DeepSeek V3	**<.001**
Microsoft Copilot vs ChatGPT-4o	**.012**
Microsoft Copilot vs ChatGPT-4.0	.410
Google Gemini Advanced vs DeepSeek V3	**.021**
Google Gemini Advanced vs ChatGPT-4o	.680
Google Gemini Advanced vs ChatGPT-4.0	**.039**
DeepSeek V3 vs ChatGPT-4o	**.017**
DeepSeek V3 vs ChatGPT-4.0	**<.001**
ChatGPT-4o vs ChatGPT-4.0	**.011**

Statistically significant values shown in bold (*P*<.05).

Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0, on 10 open-ended clinical questions related to removable prosthodontics. Using a predefined, multidimensional 10-point rubric covering comprehensiveness, scientific accuracy, clarity, and relevance, the models consistently achieved high average scores with statistically significant differences among them, indicating non-equivalence in single-turn prompts for the tested questions. Reliability measures demonstrated robustness across both rating sessions, with internal consistency and inter-rater agreement (Cronbach α and ICC) affirming the stability of the scoring process and the reliability of the conclusions.

The overall average score, ranging from 6.0 to 8.1 out of 10, indicated that all tested LLMs produced reasonably accurate, comprehensive, and clinically relevant responses. Although DeepSeek V3 is an open-source model, it achieved the highest performance compared with proprietary models, demonstrating the rapid advances of open-source AI in healthcare. The superior performance of ChatGPT-4o relative to Microsoft Copilot may be attributed to its broader multimodal pretraining and larger datasets training.[4] In contrast, Microsoft Copilot's lower score might be related to its optimization for productivity tasks.[17] Additionally, the better performance of Google Gemini Advanced compared to ChatGPT-4.0 could be attributed to its continuous learning capabilities, which allow it to

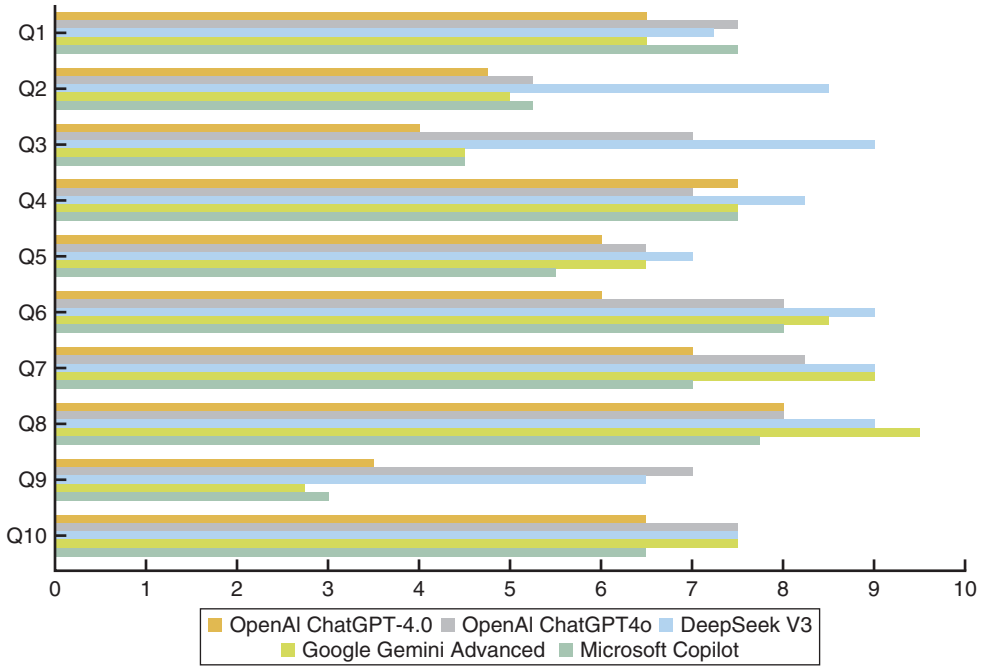

Figure 1. Average scores for answers to each question provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0.

Table 7. Descriptive statistics for average scores of answers provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0

Average Score	Microsoft Copilot	Google Gemini Advanced	DeepSeek V3	ChatGPT-4o	ChatGPT-4.0
Min	3.0	2.8	6.5	5.3	3.5
Median	6.8	7.0	8.4	7.3	6.3
Max	8.0	9.5	9.0	8.3	8.0
Mean	6.3	6.7	8.1	7.2	6.0
SEM	0.52	0.68	0.30	0.28	0.47
SD	1.64	2.14	0.96	0.88	1.47
CV	26.0%	31.9%	11.9%	12.2%	24.6%

CV, coefficient of variation; Max, maximum; Min, minimum; SD, standard deviation; SEM, standard error of mean.

Table 8. Wilcoxon signed-rank test for average scores of answers provided by Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0

Large Language Models (LLMs) [Average Scores]	Wilcoxon Signed-Rank Test (*P*-value)
DeepSeek V3 versus ChatGPT-4o	**.026**
DeepSeek V3 versus Google Gemini Advanced	**.030**
DeepSeek V3 versus Microsoft Copilot	**.004**
DeepSeek V3 versus ChatGPT-4.0	**.001**
ChatGPT-4o versus Google Gemini Advanced	.716
ChatGPT-4o versus Microsoft Copilot	**.047**
ChatGPT-4o versus ChatGPT-4.0	**.016**
Google Gemini Advanced versus Microsoft Copilot	.194
Google Gemini Advanced versus ChatGPT-4.0	**.046**
Microsoft Copilot versus ChatGPT-4.0	.401

Statistically significant values shown in bold (*P*<.05).

incorporate real-time knowledge updates and produce more current responses, as shown in a recently published study related to implant dentistry.[30] Finally, ChatGPT-4.0 showed the lowest mean score among the tested models in the present study, which may be attributed to ongoing advancement in AI model capabilities. The AI models released in 2024 or 2025 benefited from much larger training datasets and enhanced architectures, reflecting a general scaling trend in which bigger models trained on more data tend to yield better performance.[26]

The current results, which included high-quality, mostly guideline-aligned answers under single-turn prompting, were consistent with recent prosthodontic evaluations that reported satisfactory pooled accuracy in answering patients' questions.[4] Furthermore, the inclusion of DeepSeek V3, an open-sourced model, facilitated discussion not only of performance but also of reproducibility, transparency, and future academic validation, which are significant concerns when evaluating proprietary systems. A recent study[20] investigating ChatGPT's performance in the fields of removable and fixed prosthodontics reported acceptable accuracy with moderate to substantial repeatability while also highlighting key gaps, particularly as the complexity increased. This was reflected in the findings of the present investigation. However, the content still require expert review, supporting the recommendation that clinicians

should supervise outputs before being used with patients.[28] Recent reviews suggested that dentistry should adopt standardized rubrics and transparent reporting to benchmark multiple LLMs, a method the current design employs, while remaining cautious of hallucinations and domain-specific blind spots.[1] Additionally, 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 clinical practice. The single-turn prompting strategy mirrored how clinicians, students, or patients were most likely to use such systems in practice, avoiding artificial improvements associated with prompt engineering. Benchmarking against established ACP recommendations and consensus reports, rather than a single guideline or textbook, ensured that responses were judged against authoritative, evidence-based standards. Together, these design elements provided 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. Lastly, as concluded in a recent study, the use of open-ended questions may have contributed positively to the overall performance of the tested AI chatbots, as responses generated by open-ended questions, compared to closed-ended ones, exhibit higher reliability.[30]

The present study focused on advanced clinical questions in removable prosthodontics, with only 10 questions administered. This limitation constrained the extent to which the results can be generalized to broader topics within prosthodontics. Although the questions were expert-curated and based on high-level evidence sources, they were not designed as a psychometric questionnaire or scale and did not constitute a standardized or officially validated instrument. Future research should consider expanding the question set to include fixed prosthodontics and restorative dentistry, assessing the influence of conversational depth and iterative prompting, and exploring the integration of domain-specific context or finely tuned dental LLMs. Additionally, the present study aimed to evaluate LLM performance under realistic, user-driven conditions rather than using optimized or academically refined prompts. For this reason, overly structured or leading phrasing was intentionally avoided, as it could have artificially influenced the model toward a predetermined response and overestimated its performance. Instead, the questions were designed to mimic natural clinical inquiry, where ambiguity, terminology overlap, and simplified language are common. Another notable limitation concerned the stochasticity of LLM outputs. Each question was posed only once per model, using default settings, without performing repeated runs. Since these systems rely on probabilistic sampling methods that incorporate parameters such as temperature and randomness, different outputs may emerge across sessions. While documenting default access conditions promoted transparency, it did not eliminate variability. Future investigations should involve querying each model multiple times per question, calculating the mean and standard deviation of evaluator scores, and analyzing the extent of output variability. Additionally, some proprietary systems, such as Microsoft Copilot and Google Gemini Advanced, are continually updated without public version control, which further complicates reproducibility over time. The assessment of patient-facing usability, including aspects such as readability, accessibility, and the potential for misunderstandings, has been equally vital. Moreover, ethical considerations, such as algorithmic bias, lack of transparent sourcing, and the risks associated with overreliance on AI-generated responses, merit further examination. Another limitation was the rapid development of LLMs. Data collection was conducted in April 2025 using models available through standard user interfaces. As newer models continue to emerge, performance may improve over time. Still, testing widely accessible tools in real-world conditions offers clinically relevant insights for current users.

While repeated questioning could produce a more accurate estimate of stochastic variability, this approach is currently impractical for proprietary LLMs like Microsoft Copilot or Google Gemini Advanced. These models are updated continuously without prior notice, so repeated queries over time would capture updates rather than actual random variation. Consequently, the present study provided only a single performance snapshot at a single point in time. This limitation highlighted the challenge of reproducibility with commercial, cloud-based models, compared to open-source systems that can be version-controlled for research. Future studies should incorporate multi-run designs with variance analysis. When used appropriately, LLMs serve as valuable tools for clinicians, aiding instant recall of standard protocols, maintaining consistent phrasing in patient instructions, and summarizing common dental maintenance advice. They can also translate technical language into simpler terms, improving patient education materials. Nonetheless, their outputs should be purely supportive and should not be used for diagnostic or treatment-planning decisions, particularly in a discipline as biomechanically and biologically complex as removable prosthodontics.

CONCLUSIONS

Based on the findings of the present investigation, the following conclusions were drawn:

1. All 5 tested LLMs: Microsoft Copilot, Google Gemini Advanced, DeepSeek V3, ChatGPT-4o, and ChatGPT-4.0 generally provided responses on removable complete and partial dentures that were rated highly in terms of comprehensiveness, accuracy, clarity, and relevance. However, occasional inaccuracies of clinical relevance were observed.
2. Although performance was promising, particularly for DeepSeek V3, which scored the highest, the results were limited to a small set of removable prosthodontics questions.
3. These models should not replace dental professionals, as misuse could harm patient care and cause adverse effects.
4. The study's results should not be generalized across the entire prosthodontic field or extrapolated to diagnostic decision-making or treatment planning.

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.

APPENDIX A. SUPPORTING INFORMATION

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.prosdent.2026.06.018.

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. Zhang R, Pan YU, Liu Y, Deng Y, Pow EHN. Leveraging large language models for patient instructions in dentistry-A systematic review and meta-analysis. *J Prosthodont* 2025:1–10.
5. Büttner M, Leser U, Schneider L, Schwendicke F. Natural language processing: Chances and challenges in dentistry. *J Dent*. 2024;141:104796.
6. Ayorinde A, Mensah DO, Walsh J, et al. Health care professionals' experience of using AI: Systematic review with narrative synthesis. *J Med Internet Res*. 2024;26:e55766.
7. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Clinical applications of artificial intelligence in periodontology: A scoping review. *Medicina (Kaunas)*. 2025;61:1066.
8. Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. *Injury*. 2023;54:S69–S73.
9. Revilla-León M, Gómez-Polo M, Vyas S, et al. Artificial intelligence models for tooth-supported fixed and removable prosthodontics: A systematic review. *J Prosthet Dent*. 2023;129:276–292.
10. Samaranayake L, Tuygunov N, Schwendicke F, 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.
11. 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.
12. Tuygunov N, Samaranayake L, Khurshid Z, 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.
13. 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.
14. 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.
15. 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.
16. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Large language models in periodontology: Assessing their performance in clinically relevant questions. *J Prosthet Dent*. 2025;134:2328–2336.
17. 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.
18. Koidou VP, Chatzopoulos GS, Tsalikis L, Kaklamanos EG. Large language models in peri-implant disease: How well do they perform? *J Prosthet Dent*. 2025;134:2435–2442.
19. 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.
20. 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.
21. 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;135:571–576.
22. 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.
23. Salem M, Karasan D, Revilla-León M, Barmak AB, Sailer I. 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*. 2025;37:2412–2416.
24. Tussie C, Starosta A. Comparing the dental knowledge of large language models. *Br Dent J* 2024:1–3.
25. Almalki A, Althubaitiy RO, Alkhtani F, Anadioti E, Abozaed W. Assessment of ChatGPT's performance on the ACP 2024 National Prosthodontics Resident Exam (NPRE). *Eur J Dent Educ 2025* 2025:1–5.
26. Gheisarifar M, Shembesh M, Koseoglu M, et al. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. *J Prosthet Dent*. 2025;134:199–206.
27. Dashti M, Khosraviani F, Azimi T, 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.
28. 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.
29. 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.
30. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. *J Prosthet Dent*. 2025;134:2427–2434.
31. Felton D, Cooper L, Duqum I, et al. Evidence-based guidelines for the care and maintenance of complete dentures: A publication of the American College of Prosthodontists. *J Prosthodont*. 2011;20:S1–S12.
32. The frequency of denture Replacement; American College of Prosthodontists Position Statement. 2015. Accessed 30 April 2025. https://www.gotoapro.org.
33. Mohamedin SM, Komiha AM, Aboelroos AE, Mosaad MM, Swelem AA. Can digital scans replace conventional impressions for complete denture fabrication? A scoping review. *J Prosthet Dent*. 2026;135:319–330.

34. Lim TW, Li KY, Burrow MF, McGrath C. Association between removable prosthesis-wearing and pneumonia: A systematic review and meta-analysis. *BMC Oral Health*. 2024;24:1061.
35. Cagna DR, Donovan TE, McKee JR, et al. Annual review of selected scientific literature: A report of the Committee on Scientific Investigation of the American Academy of Restorative Dentistry. *J Prosthet Dent*. 2024;132:1133–1214.
36. Goldstein G, Kapadia Y, Campbell S. Complete denture occlusion: Best evidence consensus statement. *J Prosthodont*. 2021;30:72–77.
37. Goldstein G, Goodacre C, MacGregor K. Occlusal vertical dimension: Best evidence consensus statement. *J Prosthodont*. 2021;30:12–19.
38. Frank R, Brudvik J, Noonan C. Clinical outcome of the altered cast impression procedure compared with use of one-piece cast. *J Prosthet Dent*. 2004;91:468–476.
39. Turrell AJW. Clinical assessment of vertical dimension. *J Prosthet Dent*. 2006;96:79–83.
40. Abduo J. Occlusal schemes for complete dentures: A systematic review. *Int J Prosthodont*. 2013;26:26–33.
41. Felton DA. Edentulism and comorbid factors. *J Prosthodont*. 2009;18:88–96.

Corresponding author:
Dr Damian J. Lee
1 Kneeland Street, DHS 220
Boston, MA 02111
Email: Damian.Lee@tufts.edu

CRediT authorship contribution statement

Savvas N. Kamalakidis: Conceptualization, Formal analysis, Investigation, Data curation, Methodology, Visualization, Writing-original draft. **Damian J. Lee**: Methodology, Writing – review and editing. **Konstantinos Vazouras**: Investigation, Methodology, Writing – review and editing. **Kostis Giannakopoulos**: Visualization, Writing – review and editing. **Eleftherios G. Kaklamanos**: Formal analysis, Data curation, Writing – review and editing, Project administration, Methodology, Supervision.

https://doi.org/10.1016/j.prosdent.2026.06.018