DOI: https://doi.org/10.54393/pbmj.v9i8.1434

PAKISTAN BIOMEDICAL JOURNAL
https://www.pakistanbmj.com/journal/index.php/pbmj/index
ISSN (E): 2709-2798, (P): 2709-278X
Volume 9, Issue 08 (August 2026)



OPEN ACCESS
Check for updates

Original Article

Accuracy and Consistency of Three Large Language Models on Fixed Prosthodontics Questions

Anam Fayyaz Bashir[1*], Moeen Ud Din Ahmad[1], Ussamah Waheed Jatala[2], Aisha Arshad Butt[1], Saira Waheed[1] and Saima Razzaq Khan[1]

[1]Department of Operative Dentistry, Lahore Medical and Dental College, Lahore, Pakistan
[2]Department of Prosthodontics, Lahore Medical and Dental College, Lahore, Pakistan

ARTICLE INFO

Keywords:
Artificial Intelligence, Large Language Models, Prosthodontics, Fixed Prosthodontics

How to Cite:
Bashir, A. F., Ahmad, M. U. D., Jatala, U. W., Butt, A. A., Waheed, S., & Khan, S. R. (2026). Accuracy and Consistency of Three Large Language Models on Fixed Prosthodontics Questions: Consistency of Large Language Models on Fixed Prosthodontics Questions. Pakistan BioMedical Journal, 9(8), 35-40. https://doi.org/10.54393/pbmj.v9i8.1434

***Corresponding Author:**
Anam Fayyaz Bashir
Department of Operative Dentistry, Lahore Medical and Dental College, Lahore, Pakistan
dranam.fayyaz@gmail.com

Received Date: 24th June, 2026
1st Revision Received: 5th August, 2026
Acceptance Date: 27th August, 2026
Published Date: 31st August, 2026

ABSTRACT

Large language models (LLMs) are increasingly used in dental education and clinical settings, but their accuracy and consistency in fixed prosthodontics remain uncertain. **Objectives:** To compare the accuracy (using a strict three-attempt criterion) and repeated response consistency of ChatGPT-5.4, Gemini 3.1 Pro, and Claude Opus 4.8 in answering binary fixed prosthodontics questions across six clinical domains. **Methods:** Forty binary questions covering six core fixed prosthodontic domains were developed by the principal investigator. Two experienced faculty members independently answered the questions, and disagreements were resolved through discussion to establish the gold standard. Each question was submitted thrice to each LLM in a new chat session on the same day. Accuracy was the percentage of correct responses compared with the gold standard across all three attempts, while consistency was the percentage of identical responses across three attempts, regardless of correctness. Accuracy and consistency were expressed as percentages with 95% confidence intervals. Cochran's Q test, Cohen's kappa, and Fleiss' kappa were used for statistical analysis. **Results:** Pre-consensus examiner agreement was moderate, 77.5% (Cohen's κ=0.55). Claude Opus 4.8 achieved the highest accuracy (95% and consistency (97.5%, followed by ChatGPT-5.4 87.5% accuracy and 92.5% consistency and Gemini 3.1 Pro (85% accuracy and 87.5% consistency. No significant difference was observed in accuracy among models (Q=5.20, p=0.074). Intra-model consistency for three models was almost perfect (κ=0.83–0.97). **Conclusions:** The evaluated LLMs demonstrated high accuracy and consistency for structured fixed prosthodontics questions, with Claude Opus 4.8 highest, though not significant.

INTRODUCTION

Artificial intelligence (AI) has entered health care rapidly. In prosthodontics, convolutional neural networks (CNNs) now assist with diagnostic support, treatment planning, and fabrication workflow, tasks which were previously entirely dependent on the operator's eye and hand [1]. These technologies have yielded a measurable increase in accuracy and chairside efficiency [2]. With these design-oriented advances, large language models (LLMs), which are text-producing AI systems that are trained on large data sets and capable of producing human-like responses, have gained extensive attention for their roles in clinical decision support, patient education, and dental training.[3] While AI is a broad field in computer science, LLMs are a subgroup of AI designed for natural language processing. AI chatbots, such as ChatGPT and Gemini, are the user applications powered by these LLMs. Beyond conversational chatbot applications, AI-based tools have also been applied to structured clinical tasks within prosthodontic workflows. An umbrella review of 11 systematic reviews reported substantial AI performance for caries and fracture detection (accuracy approximately 82%–89%) and for identifying implant types on radiographs (approximately 95.6% pooled accuracy), though the majority of included reviews were rated low or critically low methodological quality [2]. Similarly, a scoping review of AI applications in dental crown prosthesis reported use in

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i8.1434

finish-line detection, shade matching, preparation assessment, and crown design, but noted that studies evaluating AI-designed crowns typically involved small patient samples and did not always disclose their underlying algorithms [4]. LLMs have been assessed across multiple dental domains with varied results. In endodontics, ChatGPT models 3.5 and 4 consistently outperformed Gemini variants 1.5 Pro and 1.5 Flash on binary endodontic questions, yet agreement with expert consensus was weak-to-moderate, highlighting limits of current models in some domains [5]. Studies assessing LLM performance in restorative dentistry and endodontics multiple-choice questions showed that models such as ChatGPT-4 and Claude 3 reached only moderate accuracy, with inaccuracies when questions were rephrased [6]. These findings question the reliability of LLMs when applied to clinical dental scenarios. In prosthodontics, research has given various performances across models and subtopics. Kuşçu and Görüş reported a range of performance for chatbots when 167 Turkish Dentistry Specialization Examination (DUS) questions were put in AI, from 70% accuracy for DeepSeek-V2 to the lowest at 32% for Claude 4, stressing the mixed responses among models [7]. Sozen Yanik *et al.* stated that evaluation of LLMs in maxillofacial prosthetics questions found no statistically significant difference among ChatGPT-4o, Gemini 2.5 Flash, Claude Sonnet 4, and DeepSeek V3, with accuracies ranging from approximately 68.9% to 81.1%, whether Turkish or English was used [8]. LLMs are known to produce hallucinations, defined as false or fabricated information presented in a plausible form, which arise from limitations in training data quality and insufficient domain-specific fine-tuning [3, 9]. These risks are particularly concerning in a clinical field such as fixed prosthodontics, where inaccurate responses could reinforce erroneous clinical reasoning. Furthermore, response consistency, the ability of a model to provide consistent answers across repeated queries, represents an equally critical dimension of reliability that has received limited attention [10]. The variability and sometimes inadequate accuracy of current LLMs in handling specialized prosthodontic queries create uncertainty regarding their safe integration into clinical or educational settings.

In spite of growing evidence on LLM performance in dentistry and maxillofacial prosthodontics, the subdomain of fixed prosthodontics remains limited. No previous study has directly compared accuracy and response consistency of the latest versions of ChatGPT, Gemini and Claude on binary fixed prosthodontics questions using a repeated-measures design. This gap leaves clinicians and educators without clear evidence on the reliability of these tools for knowledge-based queries in crown and bridge dentistry, raising concerns about their safe integration into clinical decision support and dental education. This study aimed to compare the accuracy (using a strict three-attempt criterion) and repeated response consistency of ChatGPT-5.4, Gemini 3.1 Pro, and Claude Opus 4.8 in answering binary fixed prosthodontics questions across six clinical domains.

METHODS

This repeated-measures study was conducted at the Department of Operative Dentistry, Lahore Medical and Dental College, following IRB approval (LMDC/FD/1096/26 dated 25th March 2026). The study duration was from March 2026 to April 2026. Forty binary, Yes/No, questions were developed by the principal author, adapted from the methodology of Çakar et al. and grounded in Contemporary Fixed Prosthodontics by Rosenstiel [5, 11]. All questions were followed by: Answer only Yes or No. Clinical image and scenario questions were excluded for objective scoring. The questions covered six fixed prosthodontic domains: tooth preparation principles, impression techniques and provisionalization, pontic and bridge design, material selection and cementation, management of endodontically treated teeth, and clinical outcomes and decision-making. Domain sample sizes ranged from 3 questions (management of endodontically treated teeth) to 10 questions (pontic and bridge design); the exact number of questions per domain is shown (Table 1).

Table 1: Distribution of the 40 Binary Fixed Prosthodontics Questions Across Six Domains, with Number of Gold-Standard Yes/No Responses

Domains	Number of Questions (n)		Brief Description of Questions
Tooth Preparation	7		Finish line designs (chamfer, shoulder, feather-edge), occlusal and axial reduction, and total occlusal convergence for all-ceramic and metal-ceramic crowns
	Yes	No	
	5	2	
Impression Techniques and Provisionalization	6		Accuracy of impression materials, use of custom trays, provisional materials and occlusal adjustment of provisional restorations
	Yes	No	
	2	4	
Pontic and Bridge Design	10		Hygienic and esthetic pontic designs, occlusal schemes, connector size, pier abutments and cantilever fixed partial dentures
	Yes	No	
	4	6	

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i8.1434

Material Selection and Cementation	6		Indications for lithium disilicate and zirconia, choice of luting cements and light-cure resin cement limitations
	Yes	No	
	1	5	
Management of Endodontically Treated Teeth	3		Need for post-and-core, fibre versus cast posts and importance of ferrule
	Yes	No	
	1	2	
Clinical Outcomes and Decision-Making	8		Shade selection, biological width violation, secondary caries, porcelain fracture, ridge preservation and long-term recall
	Yes	No	
	8	0	

As no previous study had compared these three LLMs in repeated inquiry fixed prosthodontics questions, the forty-question set method was adopted from Çakar *et al.* [5]. Two fixed prosthodontic faculty members with more than 15 years of clinical and academic expertise filled out the questionnaire. Their responses were quantified using percentage agreement and Cohen's kappa. Then both were asked to reach a consensus after discussion to make a unanimous gold standard answer. This gold standard was used for scoring chatbot responses as correct or incorrect. Three LLMs were evaluated: ChatGPT-5.4 (OpenAI Inc., San Francisco, CA, USA), Gemini 3.1 Pro (Google LLC, Mountain View, CA, USA), and Claude Opus 4.8 (Anthropic, San Francisco, CA, USA). All models were questioned on their publicly available websites on the same day so as to avoid LLM update/system modifications. Every setting was left at default on the website. A designated person typed each question to keep entries uniform. A new conversation was opened for every question, and all 40 questions were run three separate times per model on that same day. A question was marked as correctly answered when all three responses matched the gold standard answer for all models. Response accuracy was judged strictly: the LLM answer was marked correct only when the response matched the gold standard on all three times. If even one response was incorrect in any attempt, the answer was marked incorrect. This minimized the possibility of correct responses from chance rather than reliable knowledge of the LLM. For Cohen's kappa calculation (agreement between each model and the gold standard), any item for which a model's three repeated responses were not unanimous was coded as a disagreement with the gold standard, irrespective of the majority response. This strict criterion was chosen because unreliable, non-reproducible responses are of limited practical value in clinical or educational use, regardless of majority correctness. This minimized the possibility of correct responses from chance rather than reliable knowledge of the LLM. Consistency was scored separately from accuracy, with three matching responses called consistent, regardless of correct/incorrect answer. Any mix of yes or no was called an inconsistent response

Data were analyzed in SPSS version 25.0. Accuracy and consistency were reported as proportions with 95% confidence intervals, calculated using the Clopper-Pearson exact method. Cochran's Q tested whether all three models differed in accuracy. If a significant difference was found, paired models would be compared with McNemar's test with the Bonferroni method. Fleiss' kappa was used to show consistency of responses by each model, whereas Cohen's kappa measured agreement with the gold standard, which were according to Landis and Koch (≤0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect). 95% confidence intervals for Cohen's kappa and Fleiss' kappa were calculated using the large-sample normal approximation method. Significance was set at 0.050. Domain accuracy was defined as the percentage within that domain answered correctly three times, and domain consistency was the percentage of all three answers matching one another irrespective of correctness. The resulting patterns were presented as a colored heatmap.

RESULTS

Pre-consensus agreement between examiners was 77.5%, 31 of 40 items, with a Cohen's kappa of 0.55 (95% CI 0.28–0.81), indicating moderate agreement according to Landis and Koch criteria. The nine items on which evaluators differed were resolved through discussion until full consensus was reached to make the final gold standard. Claude Opus 4.8 achieved the highest accuracy at 95%, followed by ChatGPT-5.4 at 87.5% and Gemini 3.1 Pro at 85% (Table 2). Cochran's Q test revealed no statistically significant difference in accuracy among the three models ($Q = 5.20$, $df = 2$, $p=0.074$). As the overall comparison was non-significant, pre-specified post-hoc pairwise McNemar comparisons were not performed. As a sensitivity analysis, accuracy was also calculated using a majority-vote criterion (≥2 of 3 attempts matching the gold standard). Results were nearly identical to the strict criterion: ChatGPT-5.4, 35/40 (87.5%); Gemini 3.1 Pro, 35/40 (87.5%); Claude Opus 4.8, 38/40 (95.0%); only one additional item was reclassified as correct for Gemini 3.1 Pro. This indicates that model accuracy was not substantially inflated by chance agreement in a minority of attempts. Response consistency was high for all three models (Table 2). Claude Opus 4.8 was the most consistent, producing identical responses on 97.5% of items, followed by ChatGPT-5.4 at 92.5% and Gemini 3.1 Pro at 87.5%.

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i8.1434

Cochran's Q test revealed no statistically significant difference in consistency among the three models (Q = 3.43, df = 2, p=0.180)(Table 2).

Table 2: Response Accuracy and Response Consistency of the Three Large Language Models Across 40 Fixed Prosthodontics Binary Questions

LLM	Accuracy, n (%)	95% CI	Consistency, n (%)	95% CI
ChatGPT-5.4	35 (87.5%)	73.2-95.8	37 (92.5%)	79.6-98.4
Gemini 3.1 Pro	34 (85%)	70.2-94.3	35 (87.5%)	73.2-95.8
Claude Opus 4.8	38 (95%)	83.1-99.4	39 (97.5%)	86.8-99.9

Intra-model consistency and agreement with the expert standard

Intra-model consistency across the three repeated administrations, quantified using Fleiss' kappa, was almost perfect for all three models: Claude Opus 4.8 (κ = 0.97), ChatGPT-5.4 (κ = 0.90), and Gemini 3.1 Pro (κ = 0.83). Agreement between each model's responses and the expert gold standard, assessed using Cohen's kappa, was almost perfect for Claude Opus 4.8 (κ = 0.90) and substantial for ChatGPT-5.4 (κ = 0.75) and Gemini 3.1 Pro (κ = 0.70), interpreted according to the Landis and Koch criteria (Table 3).

Table 3: Intra-model Response Consistency (Fleiss' κ) and Agreement with the Expert Gold Standard (Cohen's κ)

LLM	Fleiss' κ	Fleiss' κ 95% CI	Cohen's κ	Cohen's κ 95% CI	Interpretation (Landis & Koch)
ChatGPT-5.4	0.90	0.72–1.00	0.75	0.54–0.95	Substantial
Gemini 3.1 Pro	0.83	0.65–1.00	0.70	0.48–0.92	Substantial
Claude Opus 4.8	0.97	0.79–1.00	0.90	0.77–1.00	Almost perfect

κ interpreted per Landis and Koch: 0.61–0.80, substantial; 0.81–1.00, almost perfect

Domain accuracy and consistency are presented. Claude Opus 4.8 answered every question correctly on all three attempts in four of the six domains; however, the smaller domains should be interpreted with caution given the limited number of items. The lowest accuracy was recorded by ChatGPT-5.4 and Gemini 3.1 Pro in the management of endodontically treated teeth domain (66.7%); as this domain contained only three questions, this estimate should be interpreted with caution. A similar pattern was observed for consistency, with Claude Opus 4.8 producing the most identical responses across all attempts in five of the six domains (Figure 1).

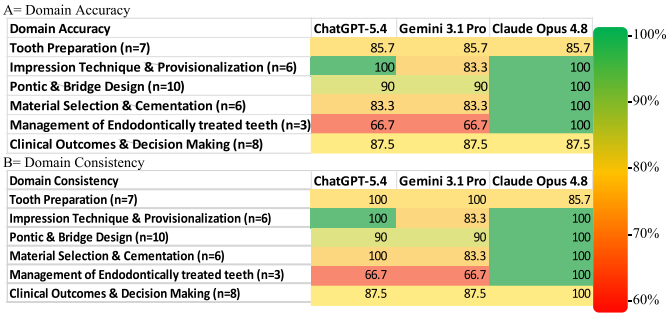


Figure 1: Domain Accuracy (A) and Consistency (B) of ChatGPT-5.4, Gemini 3.1 Pro and Claude Opus 4.8 Across the Six Fixed Prosthodontic Domains

Values represent the percentage of questions per domain answered correctly on all three attempts (accuracy) or answered identically across all three attempts (consistency). Cell colour indicates performance, with green representing higher and red representing lower values.

DISCUSSION

The present study evaluated the accuracy and response consistency of three LLMs, Claude Opus 4.8, ChatGPT-5.4, and Gemini 3.1 Pro, in answering 40 binary fixed prosthodontics questions using a repeated-measures design with three-attempt scoring criteria. Claude Opus 4.8 achieved the highest accuracy 95% and consistency 97.5%, followed by ChatGPT-5.4 at 87.5% accuracy and 92.5% consistency, and Gemini 3.1 Pro had 85% accuracy and 87.5% consistency. Although Claude numerically achieved the highest accuracy and consistency of the other two models, the overall difference in accuracy was not statistically significant (Cochran's Q test, p=0.074). No comparable studies from Pakistan were identified for direct local comparison. The accuracy observed in this study was higher than what has been reported for LLMs in prosthodontics. Kuşçu and Görüş tested five models on 167 prosthodontics items from the DUS, reporting that DeepSeek-V2 achieved the highest 70.06% accuracy while Claude 4 had the lowest at 32.34% [7]. In contrast, Geduk *et al.* used DUS in Turkey, finding that ChatGPT-4.0 reached 91.3% accuracy, but overall performance was varied across models [12]. Erdal *et al.* evaluated 128 DUS prosthodontics questions, 2012–2021, and reported high performance for selected models under independent querying, with GPT-5, GPT-4o, and Gemini 2.5 Pro achieving 91%, 86%, and 88%, respectively [13]. Their findings are consistent with the current study, suggesting that LLMs may perform well with structured and objectively scored questions. The current study's higher response accuracy may be attributable to the use of binary questions, a narrow and well-defined fixed prosthodontics domain, and the stringent repeated-query methodology that reduced the impact of random responses. In contrast, studies involving more open-ended or complex clinical tasks have reported lower and more variable performance. Sears *et al.* comparing four models

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i8.1434

on the generation of maxillofacial prosthetic treatment options, found differences between models, with Copilot being weakest and Gemini and DeepSeek scoring highest in accuracy [14]. Gheisarifar *et al.* reported reduced validity when chatbots were faced with patient-generated frequently asked questions, especially under stricter validity thresholds, with Bing performing the worst [15]. Koyuncuoglu *et al.* evaluated four chatbots on periodontology questions at two time points separated by one month, further emphasizing the value of our repeated-measures scoring approach [16]. Domain analysis revealed comparatively uniform performance across most of the six fixed prosthodontic domains. The lowest accuracy was recorded in the management of endodontically treated teeth, where both ChatGPT-5.4 and Gemini 3.1 Pro achieved 66.7%. However, this domain contained only three questions, and findings should be interpreted cautiously. More broadly, domain sizes ranged from only 3 to 10 questions across the six domains; domain-level accuracy and consistency estimates should therefore be considered exploratory rather than definitive and would benefit from confirmation in a larger question bank. The lower performance is clinically plausible because questions on endodontically treated teeth require integration of multiple factors including remaining tooth structure, ferrule design, post indication, fracture risk and long-term prognosis, rather than recall of a single fact. Kong and Kim reported use of AI applications in crown prosthetics, with finish-line detection, shade matching, preparation assessment, and crown design, but highlighted that current evidence remained inadequate due to small studies, and few did any algorithm reporting [4]. Faculty and student perspectives align with the current results. Chutia *et al.* found that faculty favored AI use in interdisciplinary dental education, while also finding training gaps, weak infrastructure, and ethical concerns [17]. Similarly, Joseph *et al.* conducted a cross-sectional survey on 700 dental students, reporting strong support for adopting LLMs, but highlighted the importance of governance and training [18]. Umer *et al.* in a review, stated that LLMs were exploding in dentistry, but that standards were needed to evaluate them [19]. Najeeb and Islam reinforced this in their review that, despite high diagnostic accuracy across AI in restorative dentistry, challenges included data biases, ethical concerns, and the absence of standardized protocols [20].

The limitations of this study were that the question bank was modest and the binary format, although ideal for objective scoring, cannot probe the depth of clinical reasoning. In addition, a single-day evaluation window captures a snapshot rather than the longer-term behavior of models that are updated frequently. Some domains, e.g. management of endodontically treated teeth, contained few questions, so domain-specific findings should be interpreted cautiously. The gold standard established by a larger expert panel would have provided a more robust reference standard and reduced the influence of individual examiner bias on the gold-standard answers. All models were queried using default settings on their public web interfaces; parameters such as temperature, which can influence response variability, could not be adjusted or standardized across platforms. Future studies should expand the question bank to include open-ended and case-based scenarios and evaluate longitudinal performance as models are updated. Multimodal assessments that incorporate radiographs and clinical photographs would further strengthen the evidence base for the safe educational and clinical use of LLMs in fixed prosthodontics.

CONCLUSIONS

The evaluated LLMs demonstrated high accuracy and consistency for structured fixed prosthodontics questions, with Claude Opus 4.8 achieving the numerically highest values, though differences among models were not statistically significant. LLMs can be used as supplementary aids in dental education and early clinical decision support, but they are not ready to be used unsupervised. This study adds to growing evidence that the need for continued benchmarking and establishment of universal standards regarding LLMs is a must.

Authors' Contribution

Conceptualization: AFB
Methodology: AFB, AAB
Formal analysis: MUDA, UWJ, UWJ
Writing and Drafting: AFB, MUDA, SW, SRK
Review and Editing: AFB, MUDA, UWJ, AAB, SW, SRK

All authors approved the final manuscript and take responsibility for the integrity of the work

Conflicts of Interest

All the authors declare no conflict of interest.

Source of Funding

The authors received no financial support for the research, authorship and/or publication of this article.

REFERENCES

[1] Al Hendi KD, Alyami MH, Alkahtany M, Dwivedi A, Alsaqour HG. Artificial Intelligence in Prosthodontics. Bioinformation. 2024 Mar; 20(3): 238. doi: 10.6026/973206300200238.

[2] Alfaraj A, Limones Á, Ahmad S, Aljubairah F, Albalaw S, Albesher M *et al.* Harnessing AI in Prosthodontics and Implant Dentistry: An Umbrella Review of Systematic Evidence. Journal of Prosthodontics. 2026 Feb; 35(2): 127-42. doi: 10.1111/jopr.70091.

[3] Roustan D and Bastardot F. The Clinicians' Guide to Large Language Models: A General Perspective with

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)

DOI: https://doi.org/10.54393/pbmj.v9i8.1434

a Focus on Hallucinations. Interactive Journal of Medical Research. 2025 Jan 28; 14(1): e59823. doi: 10.2196/59823.

[4] Kong HJ and Kim YL. Application of Artificial Intelligence in Dental Crown Prosthesis: A Scoping Review. BioMed Central Oral Health. 2024 Aug; 24(1): 937. doi: 10.1186/s12903-024-04657-0.

[5] Çakar M, Avcı AT, Düzgün S, Aslan T, Hekimoğlu KN. Assessment of the Accuracy of Modern Artificial Intelligence Chatbots in Responding to Endodontic Queries. Australian Endodontic Journal. 2025 Dec; 51(3): 732-9. doi: 10.1111/aej.70012.

[6] Lafourcade C, Kerouredan O, Ballester B, Richert R. Accuracy, Consistency, and Contextual Understanding of Large Language Models in Restorative Dentistry and Endodontics. Journal of Dentistry. 2025 Jun; 157: 105764. doi: 10.1016/j.jdent.2025.105764.

[7] Kuşçu HY and Görüş Z. Performance of Large Language Models on Prosthodontics Questions of the Dentistry Specialization Examination: A Comparative Analysis (2014–2024). Anatolian Current Medical Journal; 7(6):893-9. doi: 10.38053/acmj.1789931.

[8] Sozen Yanik I, Sahin Hazir D, Bilgin Avsar D. Cross-Lingual Performance of Large Language Models in Maxillofacial Prosthodontics: A Comparative Evaluation. BioMed Central Oral Health. 2025 Oct; 25(1): 1630. doi: 10.1186/s12903-025-07035-6.

[9] Pulkundwar P, Dhanawade V, Yadav R, Sonkar M, Asurlekar M, Rathod S. A Concise Review of Hallucinations in LLMs and their Mitigation. 2025 Dec.

[10] Madfa AA, Alshammari AF, Anazi BA, Alenezi YE, Alkurdi KA. Accuracy and Reliability of Manus, ChatGPT, and Claude in Case-Based Dental Diagnosis. Frontiers in Oral Health. 2025; 6: 1686090. doi: 10.3389/froh.2025.1686090.

[11] Rosenstiel SF. Contemporary Fixed Prosthodontics 6E, South Asia Edition-E-Book: Contemporary Fixed Prosthodontics 6E, South Asia Edition-E-Book. Elsevier Health Sciences. 2022 Nov.

[12] Geduk G, Hasırcı UC, Kusay DD, Aras RÇ, Çapar İ, Altın E *et al.* A Comparative Analysis of the Performance of Large Language Models in the Dentistry Specialty Examination. Scientific Reports. 2026 Jan; 16(1): 6739. doi: 10.1038/s41598-026-37800-8.

[13] Erdal SG, Güdül AB, Köroğlu A. The Effectiveness of Large Language Models in Dental Specialty Questions: A Comparative Study in the Field of Prosthodontics. BioMed Central Medical Education. 2026 Feb; 26(1): 455. doi: 10.1186/s12909-026-08808-5.

[14] Sears LM, Koseoglu M, Antonopoulou S, Lee SK, Kase M, Colebeck A *et al.* Assessing the Ability of Large Language Models to Summarize and Generate Maxillofacial Prosthetic Treatment Options. Journal of Prosthodontics. 2026 Apr. doi: 10.1111/jopr.70136.

[15] Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, Afshari FS, Yuan JC *et al.* Evaluating the Validity and Consistency of Artificial Intelligence Chatbots in Responding to Patients' Frequently Asked Questions in Prosthodontics. The Journal of Prosthetic Dentistry. 2025 Jul; 134(1): 199-206. doi: 10.1016/j.prosdent.2025.03.009.

[16] Koyuncuoglu CZ, Selcuker AH, Ozyilmaz E. The Reliability of Answers from Four Different AI Chatbots on Periodontology Theoretical Exam Questions: An Evaluation in Dental Education. BioMed Central Oral Health. 2025 Dec; 26(1): 114. doi: 10.1186/s12903-025-07387-z.

[17] Chutia J, Rupali R, Jain M, Angel A, Agarwal P, Sawhney C. Faculty Perspectives on Integrating Artificial Intelligence into Orthodontic and Interdisciplinary Dental Education: Opportunities, Challenges, And Strategies. Cureus. 2025 Sep; 17(9). doi: 10.7759/cureus.92877.

[18] Joseph AM, Almutairi RE, Alrashidi WA, Almutairi RM, Theeban Y, Aldhuwayhi S *et al.* Factors Influencing Large Language Model Adoption among Dental Students: A Cross-Sectional Study. Scientific Reports. 2026 Apr. doi: 10.1038/s41598-026-47512-8.

[19] Umer F, Batool I, Naved N. Innovation and Application of Large Language Models (LLMs) in Dentistry – A Scoping Review. British Dental Journal Open. 2024; 10(1): 90. doi: 10.1038/s41405-024-00277-6.

[20] Najeeb M and Islam S. Artificial Intelligence (AI) in Restorative Dentistry: Current Trends and Future Prospects. BioMed Central Oral Health. 2025 Apr; 25(1): 592. doi: 10.1186/s12903-025-05989-1.

Copyright© 2026 The Authors. Published by CrossLinks International Publishers LLC, USA
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0)