Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Clinical Reliability and Consistency of Large Language Models in Prosthodontics: An Expert Evaluation of ChatGPT and Gemini

Wijdane EL HAWARI

wijdane.elhawari@usmba.ac.ma

Sidi Mohamed Ben Abdellah University

Noura qarmiche

Sidi Mohamed Ben Abdellah University

samira el fakir

Sidi Mohamed Ben Abdellah University

Nada Othmani

Sidi Mohamed Ben Abdellah University

ossama bentahar

Sidi Mohamed Ben Abdellah University

Research Article

Keywords: Artificial Intelligence, Large Language Models, ChatGPT, Gemini, Prosthodontics, Dental Education, Generative AI, Accuracy Assessment

Posted Date: June 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9992064/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Oral Health on August 25th, 2026. See the published version at https://doi.org/10.1186/s12903-026-09666-9.

Abstract

Background:
The integration of artificial intelligence in medicine, both for education and clinical practice, has become widespread. In dentistry, these tools may be beneficial in terms of learning, clinical practice, and access to information. Nevertheless, their accuracy in specialized fields of dental medicine such as prosthodontics remains insufficiently explored.

This study aims to evaluate the accuracy of the responses generated by ChatGPT and Gemini in prosthodontics, compare their performances in removable partial prosthodontics, complete removable prosthodontics, fixed prosthodontics, occlusodontics, and maxillofacial prosthodontics. Moreover, it aims at analyzing the precision, consistency and reproducibility of the generated responses while identifying the main limitations observed during the use of these models.

Methods:
Seventy questions covering several fields of prosthodontics were developed based on recognized scientific recommendations and references. The questions were asked in French to ChatGPT (GPT-4.0) and Gemini Pro 1.5. Each question generated 30 responses per model. A total of 4,200 responses were analyzed. The evaluation was performed by three examiners using a five-point Likert scale assessing the factual accuracy of the responses.

Results:
Gemini demonstrated overall superior performance compared with ChatGPT in the majority of the studied subfields. In maxillofacial prosthodontics, 50% of the questions achieved more than 90% correct responses with Gemini compared to 21.48% with ChatGPT. Similar findings were observed in complete removable prosthodontics, removable partial prosthodontics, and fixed prosthodontics. In occlusodontics, the performances of both models were more comparable. Gemini showed better response stability with a concordance coefficient of $\kappa = 0.62$ compared with $\kappa = 0.46$ for ChatGPT. Although Gemini achieved a slightly higher mean accuracy, the difference was not statistically significant ($p = 0.109$). Both models showed difficulties with specialized terminology and complex concepts, with the occasional presence of hallucinations and erroneous or incomplete responses.

Conclusion:
Conversational artificial intelligence models show promising potential as complementary tools for learning and access to information in prosthodontics. However, their level of accuracy remains insufficient for autonomous use in such a specialized field. Gemini appears to provide more stable and consistent responses than ChatGPT, without however guaranteeing constant scientific accuracy. The development of models specifically trained in dentistry, as well as regular evaluations of their performance, appears necessary to improve their reliability and integration into education and clinical practice.

Introduction

Artificial intelligence involves creating machines by combining linguistics and computer science. These machines would be capable of performing tasks that normally require human intelligence (1). Artificial intelligence in the healthcare sector is being used more and more, especially with the emergence of large-scale linguistic models such as chatbots (2).The fine and targeted tuning of these models allows the generation of a number of generative artificial intelligence chatbots including ChatGPT, Gemini and others (3).The application of these principles in the health sector aims to transform patient care, treatment plans, and medical training (4).

The use of AI tools in dentistry, particularly orthodontics, periodontology, endodontics, and prosthodontics, is becoming increasingly widespread. These tools can now analyze images, interpret radiographs, synthesize data, and provide information on clinical materials and techniques (5).

The accuracy of this generated information remains to be assessed, as there is a risk that it may be erroneous, raising ethical concerns (6). This could potentially compromise clinical confidence as well as academic reliability (7).

In this perspective and in the field of Prosthodontics, the use of these LLMs by healthcare professionals, dental students and patients has become commonplace; the reliability of the information obtained should be put under the microscope, especially since the results of the queries can be exploited and form the basis of several attitudes for both healthcare professionals and patients.

The objective of the present study is to evaluate the accuracy and relevance of the responses obtained by ChatGpt and Gemini in the field of fixed prosthetics, resin partial dentures, metal partial dentures and maxillofacial prosthetics.

Materials and methods

The main aim was to interrogate two AI language models (ChatGPT and Gimini) on the subject of prosthodontics, and to compare them in order to assess the accuracy and reliability of data obtained from these two chatbots. Additionally, the study aimed to provide recommendations based on the findings. Ethical approval was not required, as this study did not involve any direct intervention or interaction with human participants. (Fig. 1).

Questions were primarily formulated based on established guidelines such as: "Principles, Concepts, and Practices in Prosthodontics" by the Academy of Prosthodontics (1994), "Guides to Standards in Prosthetic Dentistry - Complete and Partial Dentures" published by the British Society for the Study of Prosthetic Dentistry (1995, 2003), "Guidelines for Crowns, Fixed Bridges, and Implants" by the British Society for Restorative Dentistry (2017), and "Guidelines for the Management of Patients with Orofacial Pain and Temporomandibular Disorders" from the Journal of Oral and Maxillofacial Surgery (2024). The study encompassed various aspects of prosthodontics including: Complete Dentures, Partial Dentures,

Maxillofacial Prosthetics, Fixed Prosthetics, and Occlusodontics. 70 questions were selected for the study and are provided in the Appendix.

Three evaluators (W.E, N.Q, and N.O) posed questions to the two AI models, ChatGPT and Gimini, across three daily sessions (Morning, Afternoon, and Evening). Each question elicited 10 responses in the morning, 10 in the afternoon, and 10 in the evening. The prompts used were standardized among evaluators and focused on obtaining concise responses for each question.

The version of GPT utilized was "GPT Open AI 4.0", while Gimini operated on "Gimini Pro 1.5". The questions and responses were collected in French over a single day, while the analysis of the results was conducted starting from July 2025. Responses were compiled into a Microsoft Excel file. In total, the dataset comprised 4,200 responses, evenly split between CHATGPT and Gimini (2,100 each).

The responses generated by ChatGPT and Gemini were independently evaluated by two prosthodontics experts (W.E, O.B), who independently scored each response using a 5-point likert scale (1: Very inaccurate, 2: Inaccurate, 3: Somewhat accurate, 4: Accurate, 5: Very accurate). Responses were categorized accordingly: Scale 1 for incorrect or significantly off-topic responses, Scale 2 for minor errors or ambiguities, Scale 3 for partially correct responses with omissions or lack of clarity, Scale 4 for entirely correct responses, and Scale 5 for responses with precise references or additional relevant information.

Accuracy was defined as the proportion of correct responses obtained among all questions generated by ChatGPT and Gemini. Overall accuracy and its 95% confidence interval were estimated using the Wald binomial method. The 5-point Likert scale responses were transformed into a binary variable, where scores of 1 and 2 were classified as incorrect, while scores of 3, 4, and 5 were classified as correct. The comparison between the two artificial intelligence systems focused on their accuracy, expressed as the percentage of correct responses; Fleiss' kappa coefficient was used to assess the temporal reproducibility of AI-generated responses (16) by measuring the agreement between responses generated at different times of the day. Fleiss' kappa values between 0.41 and 0.60 were considered indicative of moderate agreement, values between 0.61 and 0.80 reflected substantial agreement, and values between 0.81 and 1.00 indicated almost perfect agreement.For each questioning iteration, the percentage of correct responses obtained for the 70 questions was calculated separately for each AI system, and the mean percentages were then compared according to the type of artificial intelligence used. The comparison of the two means was performed using Student's t-test

Results

Both AI systems achieved relatively low accuracy rates, with approximately one-fifth of the responses being correct. Gemini achieved a slightly higher mean accuracy (M = 21.43%, SD = 3.29%) compared to ChatGPT (M = 19.76%, SD = 4.53%). However, the difference was not statistically significant (p = 0.109), suggesting that despite observable trends in favor of Gemini, the overall accuracy remains low for both

models (approximately 1 in 5 responses correct) and cannot be confidently distinguished on a purely statistical basis.

In the domain of Maxillofacial Prosthodontics (Q1–Q14), Gemini demonstrated a substantial advantage, with 50% of the questions showing over 90% correct responses, compared to only 21.48% for ChatGPT. Both models presented a similar rate of highly incorrect answers (over 90% incorrect), with one question affected in each case (Q1, 7.14%). In Complete Denture Prosthodontics (Q15–Q28), Gemini achieved high precision in 4 out of 14 questions (28.57%), while ChatGPT did not exceed 90% correct responses for any question. Both models exhibited one question each with more than 90% incorrect responses. In Partial Denture Prosthodontics (Q29–Q42), Gemini again outperformed ChatGPT, with 3 questions exceeding 90% accuracy (21.42%) compared to only 2 for ChatGPT (14.28%). However, Gemini also showed a higher number of systematically incorrect responses (3 questions with > 90% incorrect), indicating instability in specific areas. In Fixed Prosthodontics (Q43–Q56), Gemini maintained its lead (4 correct questions vs. 2 for ChatGPT), although both models showed an equal number of systematically incorrect answers (2 each). In Occlusodontics (Q57–Q70), performance between models was more balanced: Gemini achieved high precision in 4 questions (28.57%) and ChatGPT in 3 (21.42%), with both models failing similarly in 3 questions each (21.42%) (Fig. 2).

Gemini obtained a Kappa coefficient of 0.62 [0.61–0.63], representing good agreement, while ChatGPT scored 0.46 [0.45–0.47], indicating only moderate agreement. These results suggest that Gemini is more temporally reliable, generating more consistent outputs regardless of the time of day.

Visual analysis of response accuracy across 30 selected questions was presented in a line graph (Fig. 3). The Gemini curve (blue) showed higher stability and remained consistently close to 60%, whereas the ChatGPT curve (orange) displayed marked fluctuations with significant drops in performance for several questions (e.g., Q8 and Q30), indicating lower reliability. While ChatGPT occasionally outperformed Gemini (e.g., Q11 and Q24), these occurrences were isolated and did not alter the overall trend favoring Gemini.

Discussion

The objective of this study was to compare the accuracy and reproducibility of responses from two Chatbots: Gemini and ChatGPT, the questions proposed relate to prosthodontics.

In this study, the language of interaction with both chatbots was French, a choice motivated by the fact that it is the second language of our country. Studies conducted in this area, such as those by Schwendicke et al. (2023) (9), have shown that these models generally perform better in English. However, Moroccan learners are trained in faculties of dentistry in French, which suggests that this would be the language used. The free and publicly accessible versions were used in the present study, allowing for a more realistic analysis given their accessibility.

The sensitivity of the models to slight syntactic or semantic variations in question wording may account for differences in their responses. This aspect was not addressed in the present study. In addition, the nature of the question format may also influence the performance of these models.

The specific terminology of the different subsections of prosthodontics sometimes posed difficulties for assimilation by the chatbots. This can be explained by the fact that generalized models lack precision and sometimes even understanding when dealing with specialized biomedical language, hence the idea of developing specialized dental models. This difficulty has already been reported by Kung et al. (2023) (8).

When comparing Gemini with ChatGPT, the latter produced less higher response coherence (Fleiss' κ = 0.62) than Gemini (κ = 0.46). The overall level of accuracy was moderate; the mean proportion of correct answers was 20%, a value that was not statistically significant ($p = 0.109$). The proportion of incorrect responses was high, which may be explained by the limited way in which the models interpret the provided prompts. Moreover, the repetition of such errors may contribute to the consolidation of erroneous concepts among users.

Despite the fact that the questions provided to both chatbots were clearly defined, occasional hallucinations were detected, a limitation that has been widely documented in the literature (7).

We noted a difficulty in analyzing certain terms specific to prosthodontics, including removable prostheses, fixed prostheses, maxillofacial prostheses, and occlusodontics. This indicates the limited capabilities of ChatGPT and Gemini when dealing with this highly specialized area of dentistry. This conclusion is consistent with the findings reported by Sismanoglu S et al (2025) (12). Consequently, their use by dental students should be controlled and supervised.

Our results generally show that Gemini is superior, tending to provide more stable responses than ChatGPT. This should not be interpreted as a guarantee of accuracy; indeed, a stable but incorrect response can be more misleading than a variable one.

Certain limitations must be acknowledged for this study; namely, the successive updates to these chatbots due to their evolving nature. Thus, the results and conclusions drawn at a given time may change with the evolution of the language model. Therefore, their periodic re-evaluation is necessary (10,13).

Another limitation lies in the use of free versions for querying these models, which does not take into account the capabilities of paid or specialized systems. Furthermore, the evaluation was conducted in a single linguistic context, which is not representative of other languages.

Future research could use identical questions posed in multiple languages to assess the influence of linguistic context on the responses obtained. It would also be relevant to examine whether paid versions or domain-specific models provide more appropriate answers. Finally, implementing teacher-developed tutoring strategies could help mitigate some of the limitations observed in this study.

Regarding the refinement of the suggested prompts aimed at improving answer accuracy (14), the questions were designed using reliable sources and reviewed by several authors. LLMs should be viewed as complementary tools that support structured, methodical learning (11).

Conclusion

The results of this study show that ChatGPT and Gemini possess interesting potential as complementary tools for learning and rapid access to information in prosthodontics. However, their level of accuracy remains variable depending on the prosthodontic disciplines evaluated more than this, it remains insufficient to guarantee autonomous and reliable use in such a specialized field of dental medicine.

Gemini generally demonstrated better performance in terms of stability, consistency, and reproducibility of responses compared with ChatGPT. Nevertheless, despite this relative superiority, both models showed significant limitations regarding specific terminology and specialized clinical reasoning, with the occasional presence of incorrect or incomplete responses.

These findings highlight the necessity of academic validation and human supervision before any use of the responses generated by these tools.

Thus, artificial intelligence models should currently only be considered complementary tools and cannot replace the clinician's expertise or specialized university education. The future development of dentistry-specific models, combined with regular evaluations and updates of their performance, ultimately appears to be an essential step.

Declarations

Author Contributions:

- Author 1: Conceptualization, study design, data collection, investigation, manuscript drafting, and manuscript revision.
- Author 2: Statistical analysis, participation in the study, and manuscript review.
- Author 3: Statistical analysis and participation in the study.
- Author 4: Participation in the study and manuscript revision.
- All authors have read and agreed to the published version of the manuscript.

Funding:

- This research received no external funding.

Institutional Review Board Statement:

- Not applicable.

Informed Consent Statement:

- Not applicable.

Data Availability Statement:

- The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments:

- The authors would like to thank all individuals who contributed to this study.

Conflicts of Interest:

- The authors declare no conflict of interest.

References

1. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023 Mar 19;11(6):887. doi:10.3390/healthcare11060887. PMID: 36981544; PMCID: PMC10048148.
2. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, Fries JA, Wornow M, Swaminathan A, Lehmann LS, Hong HJ, Kashyap M, Chaurasia AR, Shah NR, Singh K, Tazbaz T, Milstein A, Pfeffer MA, Shah NH. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. JAMA. 2025 Jan 28;333(4):319–328. doi:10.1001/jama.2024.21700. PMID: 39405325; PMCID: PMC11480901).
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023 Aug;29(8):1930–1940. doi:10.1038/s41591-023-02448-8. Epub 2023 Jul 17. PMID: 37460753.
4. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. J Biomed Sci. 2025 May 7;32(1):45. doi:10.1186/s12929-025-01131-z. PMID: 40335969; PMCID: PMC12057020.
5. Giannakopoulos K, Kavadella A, Aaqel Salim A, 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 Dec 28; 25:e51580. doi:10.2196/51580. PMID: 38009003; PMCID: PMC10784979.
6. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, Raynier JL, Clowez G, Boileau P, Ruetsch-Chelli C. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. J Med Internet Res. 2024 May 22;26:e53164. doi:10.2196/53164. PMID: 38776130; PMCID: PMC11153973.

7. Jain A, Nimonkar P, Jadhav P. Citation integrity in the age of AI: evaluating the risks of reference hallucination in maxillofacial literature. J Craniomaxillofac Surg. 2025 Oct;53(10):1871–1872. doi: 10.1016/j.jcms.2025.08.004. Epub 2025 Aug 19. PMID: 40830000.
8. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. *Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models*. PLoS Digit Health. 2023; 2(2): e0000198. doi: 10.1371/journal.pdig.0000198
9. Falk Schwendicke, Akhilanand Chaurasia, Thomas Wiegand, Sergio E. Uribe, Margherita Fontana, Ilze Akota, Olga Tryfonos, Joachim Krois; IADR e-oral health network and the ITU/WHO focus group AI for health. *Artificial intelligence for oral and dental healthcare: Core education curriculum.* Journal of Dentistry. 2023;128:104363. doi:10.1016/j.jdent.2022.104363
10. Herrera-Poyatos D, Peláez-González C, Zuheros C, Herrera-Poyatos A, Tejedor V, Herrera F, Montes R. An overview of model uncertainty and variability in LLM-based sentiment analysis: challenges, mitigation strategies, and the role of explainability. Front Artif Intell. 2025 Aug 11;8:1609097. doi: 10.3389/frai.2025.1609097. PMID: 40860717; PMCID: PMC12375657.
11. Ng FYC, Thirunavukarasu AJ, Cheng H, Tan TF, Gutierrez L, Lan Y, Ong JCL, Chong YS, Ngiam KY, Ho D, Wong TY, Kwek K, Doshi-Velez F, Lucey C, Coffman T, Ting DSW. Artificial intelligence education: An evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med. 2023 Oct 17;4(10):101230. doi: 10.1016/j.xcrm.2023.101230. PMID: 37852174; PMCID: PMC10591047.
12. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and gemini advanced achieve comparable results to humans? BMC Med Educ. 2025 Feb 10;25(1):214. doi: 10.1186/s12909-024-06389-9. PMID: 39930399; PMCID: PMC11809121.
13. Chen, L., Zaharia, M., & Zou, J. (2023). *How Is ChatGPT's Behavior Changing Over Time?* arXiv preprint arXiv:2307.09009.
14. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, Kiuchi T. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024 Jul 25;26:e60807. doi: 10.2196/60807. PMID: 39052324; PMCID: PMC11310649.
15. Aljamaan F, Temsah MH, Altamimi I, Al-Eyadhy A, Jamal A, Alhasan K, Mesallam TA, Farahat M, Malki KH. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. JMIR Med Inform. 2024 Jul 31;12:e54345. doi: 10.2196/54345. PMID: 39083799; PMCID: PMC11325115.
16. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276 – 82. PMID: 23092060; PMCID: PMC3900052.

Figures

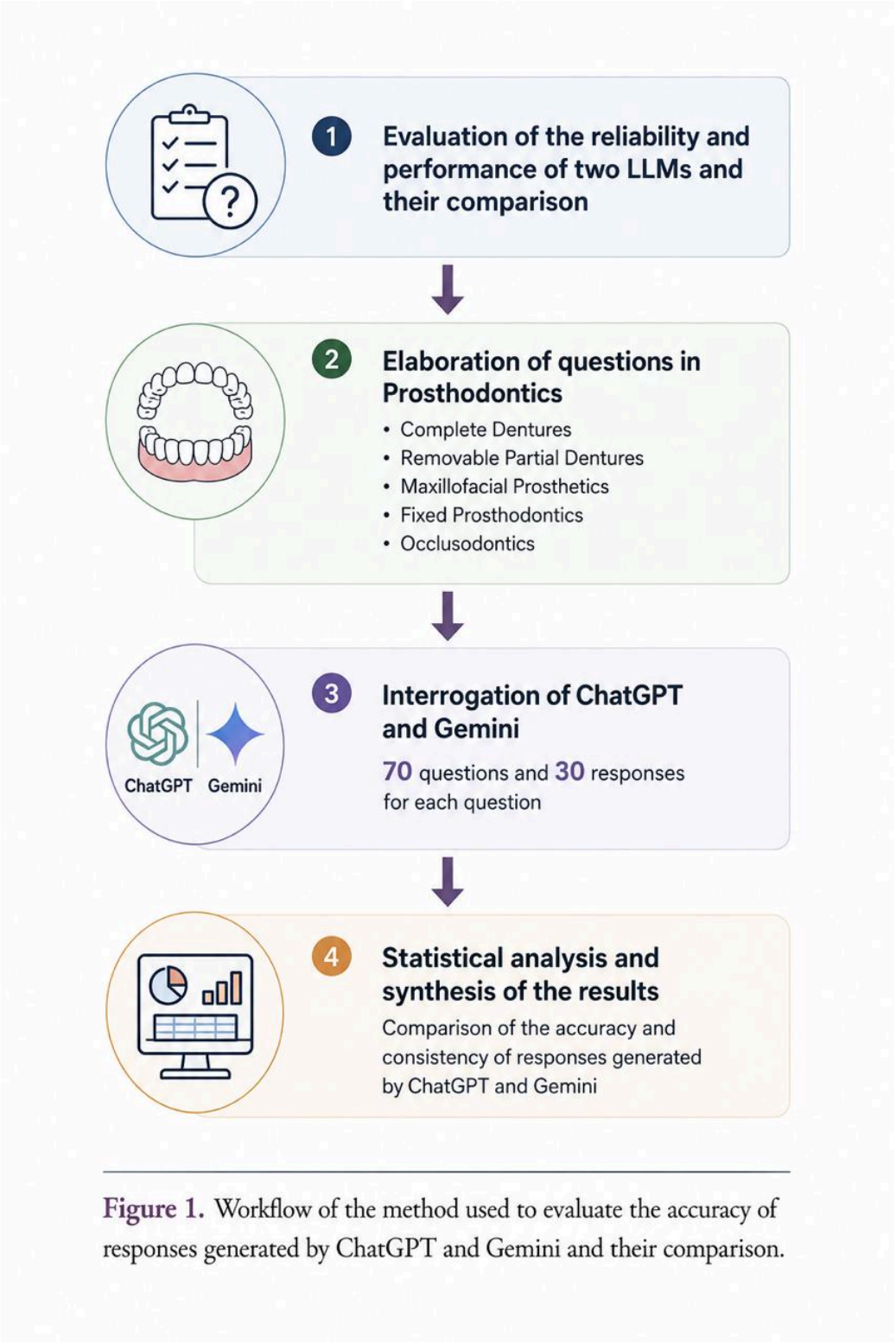
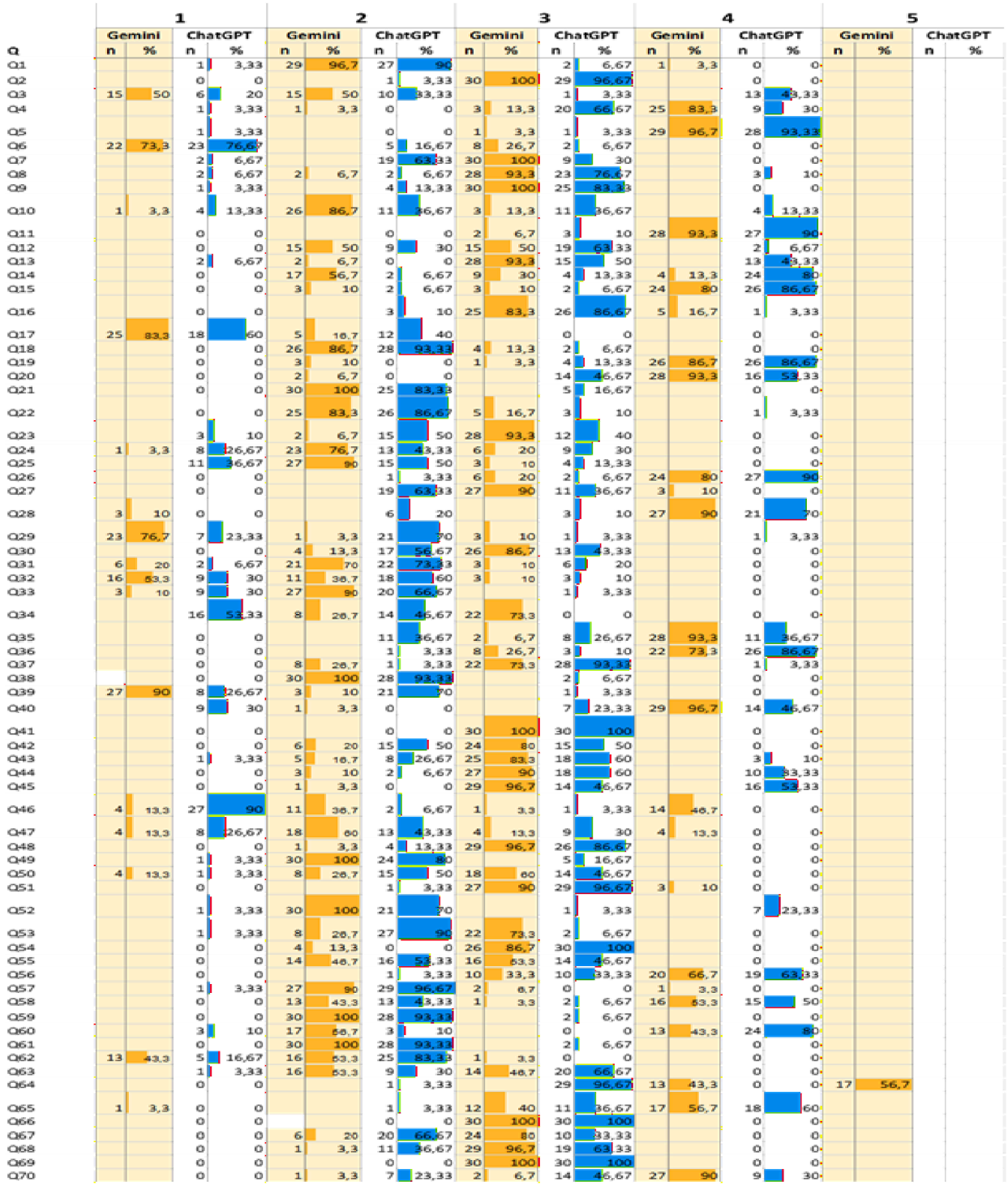


Figure 1. Workflow of the method used to evaluate the accuracy of responses generated by ChatGPT and Gemini and their comparison.

Figure 1

Workflow of the method used to evaluate the accuracy of responses generated by ChatGPT and Gemini and their comparison



Q	1				2				3				4				5			
	Gemini		ChatGPT		Gemini		ChatGPT		Gemini		ChatGPT		Gemini		ChatGPT		Gemini		ChatGPT	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Q1			1	3,33	29	96,7	27	90			2	6,67	1	3,3	0	0				
Q2			0	0			1	3,33	30	100	29	96,67			0	0				
Q3	15	50	6	20	15	50	10	33,33			1	3,33			13	43,33				
Q4			1	3,33	1	3,3	0	0	3	13,3	20	66,67	25	83,3	9	30				
Q5			1	3,33			0	0	1	3,3	1	3,33	29	96,7	28	93,33				
Q6	22	73,3	23	76,67			5	16,67	8	26,7	2	6,67			0	0				
Q7			2	6,67			19	63,33	30	100	9	30			0	0				
Q8			2	6,67	2	6,7	2	6,67	28	93,3	23	76,67			3	10				
Q9			1	3,33			4	13,33	30	100	25	83,33			0	0				
Q10	1	3,3	4	13,33	26	86,7	11	36,67	3	13,3	11	36,67			4	13,33				
Q11			0	0			0	0	2	6,7	3	10	28	93,3	27	90				
Q12			0	0	15	50	9	30	15	50	19	63,33			2	6,67				
Q13			2	6,67	2	6,7	0	0	28	93,3	15	50			13	43,33				
Q14			0	0	17	56,7	2	6,67	9	30	4	13,33	4	13,3	24	80				
Q15			0	0	3	10	2	6,67	3	10	2	6,67	24	80	26	86,67				
Q16			0	0			3	10	25	83,3	26	86,67	5	16,7	1	3,33				
Q17	25	83,3	18	60	5	16,7	12	40			0	0			0	0				
Q18			0	0	26	86,7	28	93,33	4	13,3	2	6,67			0	0				
Q19			0	0	3	10	0	0	1	3,3	4	13,33	26	86,7	26	86,67				
Q20			0	0	2	6,7	0	0			14	46,67	28	93,3	16	53,33				
Q21			0	0	30	100	25	83,33			5	16,67			0	0				
Q22			0	0	25	83,3	26	86,67	5	16,7	3	10			1	3,33				
Q23			3	10	2	6,7	15	50	28	93,3	12	40			0	0				
Q24	1	3,3	8	26,67	23	76,7	13	43,33	6	20	9	30			0	0				
Q25			11	36,67	27	90	15	50	3	10	4	13,33			0	0				
Q26			0	0			1	3,33	6	20	2	6,67	24	80	27	90				
Q27			0	0			19	63,33	27	90	11	36,67	3	10	0	0				
Q28	3	10	0	0			6	20			3	10	27	90	21	70				
Q29	23	76,7	7	23,33	1	3,3	21	70	3	10	1	3,33			1	3,33				
Q30			0	0	4	13,3	17	56,67	26	86,7	13	43,33			0	0				
Q31	6	20	2	6,67	21	70	22	73,33	3	10	6	20			0	0				
Q32	16	53,3	9	30	11	36,7	18	60	3	10	3	10			0	0				
Q33	3	10	9	30	27	90	20	66,67			1	3,33			0	0				
Q34			16	53,33	8	26,7	14	46,67	22	73,3	0	0			0	0				
Q35			0	0			11	36,67	2	6,7	8	26,67	28	93,3	11	36,67				
Q36			0	0			1	3,33	8	26,7	3	10	22	73,3	26	86,67				
Q37			0	0	8	26,7	1	3,33	22	73,3	28	93,33			1	3,33				
Q38			0	0	30	100	28	93,33			2	6,67			0	0				
Q39	27	90	8	26,67	3	10	21	70			1	3,33			0	0				
Q40			9	30	1	3,3	0	0			7	23,33	29	96,7	14	46,67				
Q41			0	0			0	0	30	100	30	100			0	0				
Q42			0	0	6	20	15	50	24	80	15	50			0	0				
Q43			1	3,33	5	16,7	8	26,67	25	83,3	18	60			3	10				
Q44			0	0	3	10	2	6,67	27	90	18	60			10	33,33				
Q45			0	0	1	3,3	0	0	29	96,7	14	46,67			16	53,33				
Q46	4	13,3	27	90	11	36,7	2	6,67	1	3,3	1	3,33	14	46,7	0	0				
Q47	4	13,3	8	26,67	18	60	13	43,33	4	13,3	9	30	4	13,3	0	0				
Q48			0	0	1	3,3	4	13,33	29	96,7	26	86,67			0	0				
Q49			1	3,33	30	100	24	80			5	16,67			0	0				
Q50	4	13,3	1	3,33	8	26,7	15	50	18	60	14	46,67			0	0				
Q51			0	0			1	3,33	27	90	29	96,67	3	10	0	0				
Q52			1	3,33	30	100	21	70			1	3,33			7	23,33				
Q53			1	3,33	8	26,7	27	90	22	73,3	2	6,67			0	0				
Q54			0	0	4	13,3	0	0	26	86,7	30	100			0	0				
Q55			0	0	14	46,7	16	53,33	16	53,3	14	46,67			0	0				
Q56			0	0			1	3,33	10	33,3	10	33,33	20	66,7	19	63,33				
Q57			1	3,33	27	90	29	96,67	2	6,7	0	0	1	3,3	0	0				
Q58			0	0	13	43,3	13	43,33	1	3,3	2	6,67	16	53,3	15	50				
Q59			0	0	30	100	28	93,33			2	6,67			0	0				
Q60			3	10	17	56,7	3	10			0	0	13	43,3	24	80				
Q61			0	0	30	100	28	93,33			2	6,67			0	0				
Q62	13	43,3	5	16,67	16	53,3	25	83,33	1	3,3	0	0			0	0				
Q63			1	3,33	16	53,3	9	30	14	46,7	20	66,67			0	0				
Q64			0	0			1	3,33			29	96,67	13	43,3	0	0	17	56,7		
Q65	1	3,3	0	0			1	3,33	12	40	11	36,67	17	56,7	18	60				
Q66			0	0			0	0	30	100	30	100			0	0				
Q67			0	0	6	20	20	66,67	24	80	10	33,33			0	0				
Q68			0	0	1	3,3	11	36,67	29	96,7	19	63,33			0	0				
Q69			0	0			0	0	30	100	30	100			0	0				
Q70			0	0	1	3,3	7	23,33	2	6,7	14	46,67	27	90	9	30				

Figure 2

Presentation of responses according to Likert scale.

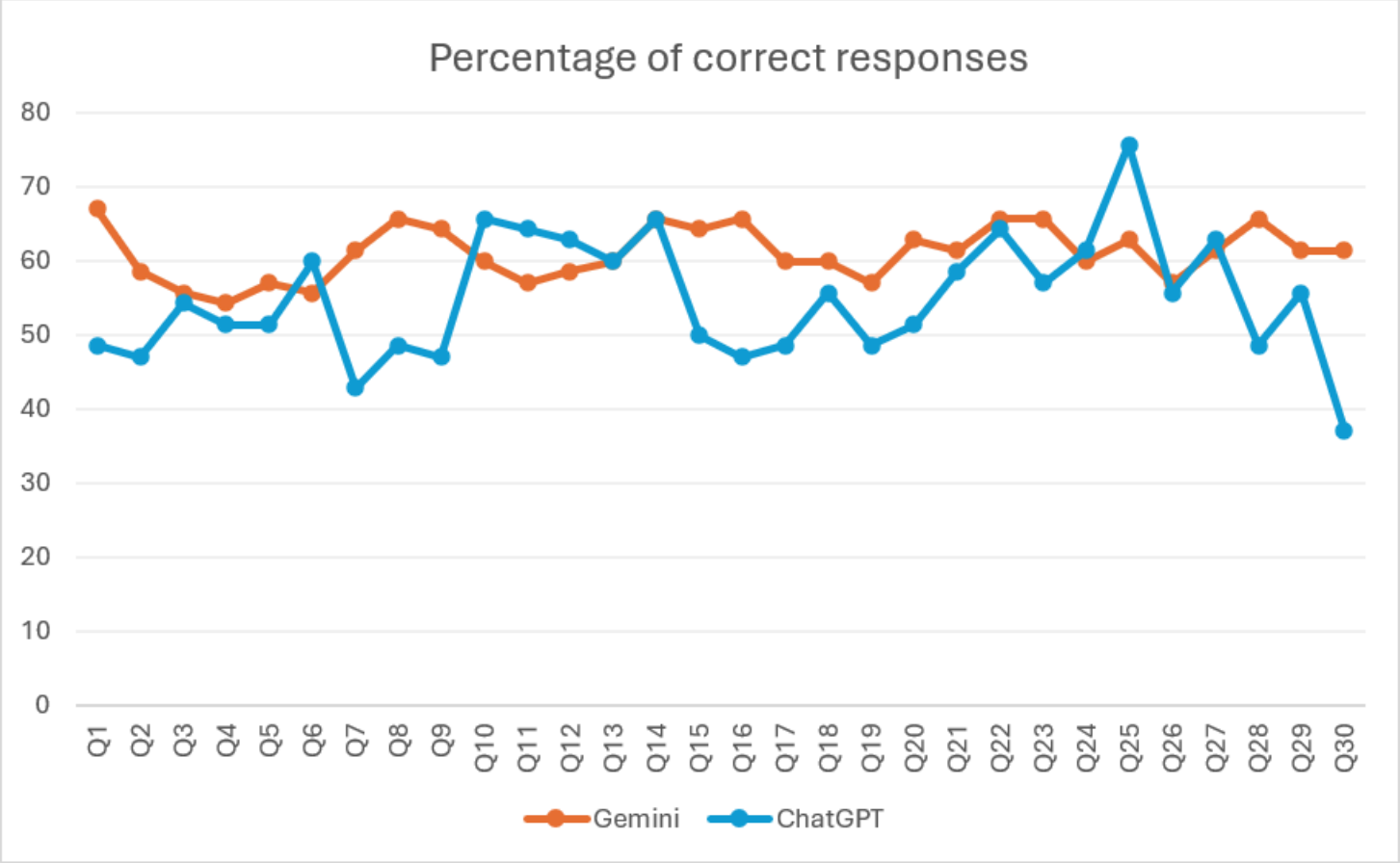


Figure 3

Precision of ChatGPT and Gemini

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- APPENDIX.docx