Received: 26 March 2024 | Revised: 24 June 2024 | Accepted: 25 June 2024

DOI: 10.1111/jre.13323

CLINICAL RESEARCH

Journal of PERIODONTAL RESEARCH WILEY

Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education

Hamoun Sabri[1,2] | Muhammad H. A. Saleh[1] | Parham Hazrati[1] | Keith Merchant[3] | Jonathan Misch[1,4] | Purnima S. Kumar[1] | Hom-Lay Wang[1] | Shayan Barootchi[1,2,5]

[1]Department of Periodontics and Oral Medicine, School of Dentistry, University of Michigan, Ann Arbor, Michigan, USA

[2]Center for Clinical Research and Evidence Synthesis in Oral Tissue Regeneration (CRITERION), Ann Arbor, Michigan, USA

[3]Naval Post-Graduate Dental School, Bethesda, Maryland, USA

[4]Private Practice, Ann Arbor, Michigan, USA

[5]Division of Periodontology, Department of Oral Medicine, Infection, and Immunity, Harvard School of Dental Medicine, Boston, Massachusetts, USA

Correspondence

Hamoun Sabri, Department of Periodontics and Oral Medicine, University of Michigan School of Dentistry, 1011 N University Ave, Ann Arbor, MI 48109, USA.
Email: hsabri@umich.edu

Shayan Barootchi, Division of Periodontology, Department of Oral Medicine, Infection, & Immunity, Harvard School of Dental Medicine, Boston, MA, USA; Department of Periodontics and Oral Medicine, School of Dentistry, University of Michigan, Ann Arbor, MI, USA.
Email: shbaroot@umich.edu

Abstract

Introduction: The emerging rise in novel computer technologies and automated data analytics has the potential to change the course of dental education. In line with our long-term goal of harnessing the power of AI to augment didactic teaching, the objective of this study was to quantify and compare the accuracy of responses provided by ChatGPT (GPT-4 and GPT-3.5) and Google Gemini, the three primary large language models (LLMs), to human graduate students (control group) to the annual in-service examination questions posed by the American Academy of Periodontology (AAP).

Methods: Under a comparative cross-sectional study design, a corpus of 1312 questions from the annual in-service examination of AAP administered between 2020 and 2023 were presented to the LLMs. Their responses were analyzed using chi-square tests, and the performance was juxtaposed to the scores of periodontal residents from corresponding years, as the human control group. Additionally, two sub-analyses were performed: one on the performance of the LLMs on each section of the exam; and in answering the most difficult questions.

Results: ChatGPT-4 (total average: 79.57%) outperformed all human control groups as well as GPT-3.5 and Google Gemini in all exam years ($p < .001$). This chatbot showed an accuracy range between 78.80% and 80.98% across the various exam years. Gemini consistently recorded superior performance with scores of 70.65% ($p = .01$), 73.29% ($p = .02$), 75.73% ($p < .01$), and 72.18% ($p = .0008$) for the exams from 2020 to 2023 compared to ChatGPT-3.5, which achieved 62.5%, 68.24%, 69.83%, and 59.27% respectively. Google Gemini (72.86%) surpassed the average scores achieved by first- (63.48% ± 31.67) and second-year residents (66.25% ± 31.61) when all exam years combined. However, it could not surpass that of third-year residents (69.06% ± 30.45).

Conclusions: Within the confines of this analysis, ChatGPT-4 exhibited a robust capability in answering AAP in-service exam questions in terms of accuracy and reliability while Gemini and ChatGPT-3.5 showed a weaker performance. These findings underscore the potential of deploying LLMs as an educational tool in periodontics and oral implantology domains. However, the current limitations of these models such as inability to effectively process image-based inquiries, the propensity for generating inconsistent responses to the same prompts, and achieving high (80% by GPT-4) but not absolute accuracy rates should be considered. An objective comparison of their capability versus their capacity is required to further develop this field of study.

KEYWORDS
American Academy of Periodontology, artificial intelligence, ChatGPT, ChatGPT-3.5, ChatGPT-4, dental education, Gemini, Google Bard, Google Gemini, periodontal education

1 | INTRODUCTION

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that is designed to enable computers to understand text and spoken words in the same sense that humans can. In contrast to traditional computing, which learns from data, neural networks in AI are based on pattern recognition, thereby providing more accurate decision-making. NLP algorithms are steadily entering the mainstream in daily life; the most straightforward and publicly accessible forms being smart home devices, virtual assistance tools, and chatbots.[1,2] The broad applicability of these platforms is also rapidly being harnessed in the medical and dental research arenas, as well as in healthcare settings.[3,4] One of the most fascinating forms of NLPs is the large language model (LLM).[5] LLMs are fundamentally deep machine-learning models trained on a wide range of data in an unsupervised fashion[5,6] to predict relationships between the elements that create a language and generate a meaningful response.

In the medical field, LLMs have two significant applications. Firstly, patients use search engines to seek information regarding their health issues.[7,8] Secondly, medical professionals and researchers have been exploring the capability of these models in answering research questions and clinical decision making.[9,10] However, these AI models may lead to possible harm stemming from inaccurate medical recommendations given to patients or false data provided to professionals.[11–13] Concerns like this are especially relevant to healthcare academia, where AI-based training and testing curricula are being developed.

One method of measuring the ability of these models to provide accurate information is by estimating their performance on a standardized test. This approach has been used in several medical specialties[14–16] since 2021. The consensus is that the performance is based on (a) the amount of information available for the model to access and train on, (b) the types of questions (single vs. multiple choice), and (c) the verbosity of the question or the need to analysis graphics or images to return an answer.[17,18]

The most commonly used LLMs are ChatGPT (GPT-4 and GPT-3.5, OpenAI, San Francisco, CA, USA) and Google Gemini (previously named as Google Bard; Alphabet Inc., Mountain View, CA, USA).[19–21] ChatGPT reached 100 million users within 3 months of its introduction in November 2022, with roughly 1.6 billion visits per month[21,22] reflecting the extensive worldwide interest in LLMs. ChatGPT operates based on the Generative Pretrained Transformer-3.5 (GPT-3.5) and GPT-4 models, which implement transformer architecture and reinforcement learning from human feedback. On March 14, 2023, OpenAI introduced GPT-4, which is reported to be more reliable, creative, and able to handle many more nuanced instructions.[23,24] OpenAI has announced that GPT-4 could perform well in academic and specialized fields.[24] ChatGPT operates based on the Generative Pretrained Transformer-3.5 (GPT-3.5) and GPT-4 models, which implement transformer architecture and reinforcement learning from human feedback. On the other hand, developed by Google and released in March of 2023, Gemini performs based on the language model for dialogue applications. This method allows Google Gemini to generate comprehensive and informative responses to questions, whereas ChatGPT provides creative and coherent text. Both ChatGPT and Gemini benefit from using real-time human feedback to improve the accuracy and creativity of their responses.[21,22] This indicates interactive learning for these models, which refines and adjusts their responses, thus leading to enhanced performance over time. ChatGPT has the advantage of training based on much larger data frames, leading to more diverse knowledge and a wider range of abilities, while Google Gemini benefits from accessing the most recent and updated data. Consequently, in theory, Gemini has the potential to provide users with more up-to-date and contextually relevant information.[19–21]

Since periodontics is an evidence-based specialty in which knowledge of biology, anatomy, and pathology is as essential as surgical knowledge, we sought to investigate the ability of two LLMs to accurately respond to standardized multiple-choice questions from the annual in-service examination organized by the American Academy of Periodontology (AAP). This exam comprises 10 sections spanning clinical and foundational sciences, diagnosis, and treatment planning, with 95% text-based questions and a few graphical questions. The responses are assessed in a multiple-choice (MCQ) format. As this exam covers a wide range of oral and craniofacial domains, ranging from biochemistry to dento-maxillofacial anatomy to critical thinking as applied to clinical diagnosis and treatment planning, it is an ideal means to assess the performance of LLMs in the field of dentistry. Based on this rationale, the current study aimed to explore the performance of Google Gemini and ChatGPT (via GPT-4 and GPT-3.5) LLMs in answering the AAP in-service examination compared to human exam participants, that is, periodontal residents.

2 | MATERIALS AND METHODS

2.1 | Study design/Ethical considerations

This study was conceptualized within a comparative cross-sectional design. The most recent guidelines of The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement via the EQUATOR network were implemented (Appendix 1). The study was concluded in May 2024 via ChatGPT (GPT3.5: last accessed on July 15, 2023, GPT-4: last accessed on April 20, 2024) and Google Gemini (last accessed on July 20, 2023). Since the study data did not involve human or animal research subjects, the study was exempt from ethical approval from the institutional review board. In-service exam data and statistics for the years 2020–2023 were accessed through a formal request to the AAP via the AAP In-Service Examination Committee.

2.2 | Data source, inclusion, and exclusion criteria

AAP in-service exam questions from 2020 to 2023 (four exam sets) were obtained through password-protected member access (via www.perio.org). This exam generally consists of 300–400 MCQs. All questions and corresponding answer keys were considered and reviewed for eligibility for the study. Upon screening the questions by one examiner (H.S.), the following exclusion criteria were employed:

- Illustration-based questions: Questions containing a graphical component (clinical images, medical photography, and graphs) either in the question stem or among answer choices were excluded due to the inability of AI language models to analyze graphical content.
- Within the 2022 and 2023 exams, questions with any reference to an article published after October 1, 2021, were excluded due to the inability of GPT-3.5 to access data published after September 2021. Special care was given to ensure the exclusion of questions with content indexed on Google before October 2021.
- Questions that were dropped from the final scoring report by the AAP were excluded.

All other remaining questions were included in the study.

The AAP in-service exam board classifies each year's exam questions into 10 sections: (1) Embryology and Anatomy, (2) Biochemistry-Physiology, (3) Microbiology and Immunology, (4) Periodontal Etiology and Pathogenesis, (5) Pharmacology and Therapeutics, (6) Biostatistics, Experimental Design, and Data Analysis, (7) Diagnosis, (8) Treatment Planning and Prognosis, (9) Therapy, and (10) Oral Pathology/Oral Medicine. Based on this, each question's category was also noted and imported into the dataset. Figure 1 presents bar graphs visualizing the proportion of every exam section within each year.

2.3 | Periodontal in-service exam data sets

The AAP prepares an annual report that summarizes the exam results. The report includes the average scores of all residents, categorized by the year of residency (first, second, and third). Moreover, the report provides statistics for each of the 10 sections of the exam, along with mean percentage scoring for each question. All of these variables are accompanied by their corresponding standard deviation (SD) and range.

2.4 | Primary outcome

The primary outcome of this study was the performance of ChatGPT (GPT-4 and 3.5) and Gemini reported as the correct answer percentage (%) of each exam year (2020–2023) calculated as:

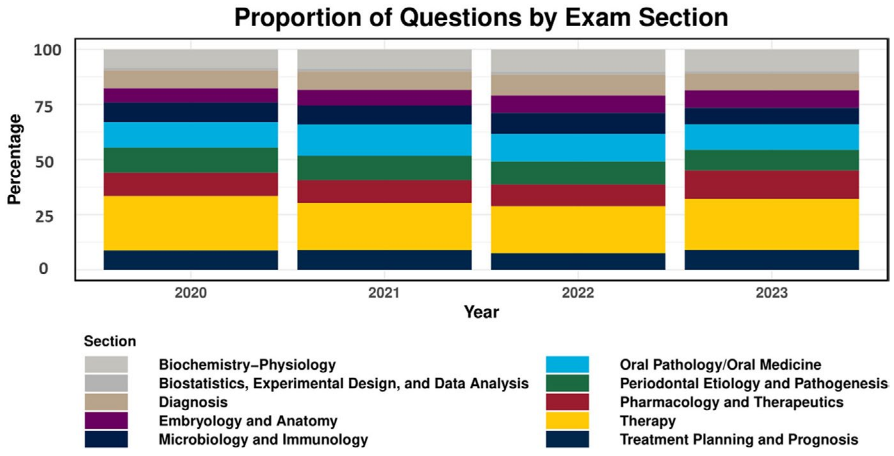


FIGURE 1 The proportion of each section in American Academy of Periodontology's in-service exam, divided by each exam year (2020–2023).

$$\text{Performance} = \frac{\text{Total number of correct answers}}{\text{Total number of questions}} \times 100$$

Further comparison to the main test-takers of this exam, periodontics residents from accredited residency programs (postgraduate year 1, 2, and 3 [PGY-1, -2, and -3] residents).

2.5 | Questions' difficulty level definition and evaluation

Based on the correct answer percentage of each question, the questions with a percentage of less than 50% correctness among all residents were considered as "difficult" questions. The rationale was to detect the most challenging questions within each exam and sub-analyze the performance of the AI language models in answering those as the study's secondary outcome. It should be noted that special care was given to exclude revoked questions or those that were removed from scoring for any reason. Only the questions included in each exam year's final question bank were analyzed based on the previously mentioned exclusion criteria. Figure A2 depicts two examples of difficult questions included in the analysis.

2.6 | Artificial intelligence model testing: Prompt engineering and encoding

The freely available public version of ChatGPT (GPT-3.5), its premium version (GPT-4), and Google Gemini, were employed. One investigator (P.H.) who was blinded to the year of the exam questions and answers, encoded, and input all MCQs with respective answer choices into the AI models using a copy-paste function. Subsequently, the responses were collected by the same blinded examiner and input into a pre-organized electronic spreadsheet. To reduce bias, the models were prompted three times for each question, and the answer choice with at least 2-time (2/3) repetition was selected as the final answer. Multiple other "regeneration" attempts were performed in the rare case of indecisiveness by the models after three attempts, until it reached 75% of consistency in generating the same output. In addition, the consistency of the AI models in generating the same output was tested and at least 75% of consistency was achieved (Google Gemini: 86.67% ± 6.49, ChatGPT-4: 94.17% ±6.67 and ChatGPT-3.5: 76.33% ±11.74).

We followed a multiple-choice single-answer without forced justification (MC-NJ) prompting approach, reproducing the original in-service exam question verbatim and asking for the best answer choice without further justification and reasoning.[25] A new ChatGPT or Gemini session was started for each entry in order to avoid crossover learning and memory retention leading to potential bias. Figure A1 illustrates the portal environment of these bots as well as a sample prompt along with the response generated by the models.

2.7 | Statistical analysis

See Appendix 2.

3 | RESULTS

3.1 | Included questions and descriptive characteristics

Initially, 1365 questions were imported and screened, of which 1312 were included in this study. The reasons for the exclusion of 53 questions were: illustrated content (n = 25), content pertinent to materials published after September 2021 (n = 14), and items that were dropped from scoring by the AAP (n = 14). The distribution of the questions among the exam years (2020–2023) was 368, 337, 305, and 302, respectively.

Table 1 provides a detailed description of the questions included by year and exam category. The "Therapy" section had the highest number of questions (n = 298), while "Biostatistics, experimental design and data analysis" presented with the lowest number of total questions (n = 15).

A total of 2375 periodontal residents took the exam between 2020 and 2023. With the exception of 2021 (n = 479), the number of participants was between 500 and 600. In all exam years, the second-year residents outnumbered the third- and first-year residents. The average raw score on all four exams combined was 63.48% ± 31.67 (PGY-1 residents), 66.25% ± 31.61 (PGY-2 residents), and 69.06% ± 30.45 (PGY-3 residents), while the pooled average raw score was 66.39% ± 32.77. The complete descriptive statistics of

TABLE 1 Descriptive statistics of included questions.

Included questions to the analysis					
Exam Section	**2020**	**2021**	**2022**	**2023**	**Total**
Biochemistry-Physiology	31	30	31	30	122
Biostatistics, Experimental Design, and Data Analysis	4	4	4	3	15
Diagnosis	30	28	29	23	110
Embryology and Anatomy	24	24	24	24	96
Microbiology and Immunology	33	29	29	23	114
Oral Pathology/Oral Medicine	42	48	38	35	163
Periodontal Etiology and Pathogenesis	42	37	32	28	139
Pharmacology and Therapeutics	39	35	30	39	143
Therapy	91	72	65	70	298
Treatment Planning and Prognosis	32	30	23	27	112
Total	368	337	305	302	1312

scores achieved by residents divided by each exam year and experience level is presented in Table 2 (top).

3.2 | 2020–2023 Exam performance analysis: Human versus AI

When comparing the performance of AI models to periodontal residents (human control groups), in the exam year of 2020, all three bots, Gemini (70.65%) and GPT-3.5 (62.5%), and GPT-4 (78.80%) scored higher than PGY-1 (58.39% ± 29.94), PGY-2 (60.53% ± 30.17), PGY-3 (62.22% ± 28.75), and all residents combined (60.35% ± 30.14). However, in the 2021 exam, although GPT-4 (78.93%) and Gemini (73.29%) outperformed all resident classes and their combined average, GPT-3.5 (68.24%) did not surpass the performance of PGY-3 (68.99% ± 35.31) residents. Similarly, regarding 2022 exam questions, while Gemini (75.73%) and GPT-4 (80.98%) performed better than all resident classes, GPT-3.5 (69.83%) only surpassed the PGY-1 residents' average score (66.57% ± 28.76). Furthermore, in the 2023 exam, although the PGY-3 resident performance (74.18% ± 28.95) was the highest amongst the other resident groups, Gemini and GPT-3.5 models, GPT-4 scored the highest among all included groups (79.80%). Moreover, GPT-3.5 (59.27%) failed to outperform any of the other human groups, Gemini (72.18%) or GPT-4. Whereas Gemini achieved a higher score compared to PGY-1 residents (63.92% ± 34.56), PGY-2 residents (68.45% ± 33.27), and the average of all residents (68.35% ± 34.76). Table 2 and Figure 2 provide a comprehensive comparison and description of scores by each resident year, AI model, and year of the exam. Finally, when all exam questions were combined (from 2020 to 2023, $N = 1312$), GPT-4 (79.57%) ranked first followed by Gemini (72.86%), PGY-3 residents ($n = 640$, 69.06% ± 30.45), the combination of all residents ($n = 2375$, 66.39% ± 32.77), PGY-2 residents ($n = 753$, 66.25% ± 31.61), GPT-3.5 (64.93%), and PGY-1 residents ($n = 706$, 63.48% ± 31.67).

3.3 | 2020–2023 Exam performance analysis: Comparison of AI models

When comparing the performance of the three LLMs, with regard to the achieved score each year and the total of all questions combined, GPT-4 outperformed the other models consistently while similarly, Gemini performed superior to GPT-3.5 in all years. Performance of GPT-4 versus the other models reached a statistical significance in all 4 years and all-years combined ($p < .001$). Moreover, the difference between scores from Gemini versus GPT-3.5 reached a statistically significant level in 2020 (70.65% to 62.5%; $p = .01$), 2023 (72.18% to 59.27%; $p = .0008$) and all years combined (72.86% to 64.93%; $p < .001$). Figure 4a depicts the results of this comparison.

3.4 | Section analysis

The sub-analysis on the performance of the three AI language models in each exam section revealed a significantly ($p < .001$) higher score by GPT-4 versus GPT-3.5 in all exam sections except "Biostatistics, experimental design and data analysis" (93.33% vs. 86.7%; $p = .26$). When compared to Gemini, similar results were observed where GPT-4 scored significantly higher in all sections except "Biostatistics, experimental design and data analysis" (93.33% vs. 73.3%; $p = .58$) and "Diagnosis" (70% vs. 60%; $p = .32$). Furthermore, significantly greater scores were noted by Gemini in the "Treatment planning and prognosis" and "Therapy" sections of the exam compared to GPT-3.5 (61.6% vs. 37.5% and 61.7% vs. 47%, respectively; $p < .001$). Also, Gemini outperformed GPT-3.5 in all other sections except in "Biostatistics, experimental design and data analysis" "biochemistry-physiology," and "Microbiology and immunology," nevertheless, none of these differences were statistically significant. Table 2 (middle) and Figure 3 summarize the exam section sub-analysis.

3.5 | Difficult questions

After the application of the aforementioned inclusion criteria, 127 questions were designated as "difficult" questions that presented the lowest correct response rate, averaging a 40.52% accuracy among residents. Regarding the AI models' performance on the same subset of difficult questions, GPT-4 (62.99%) surpassed the other two models (with statistically significant difference when compared to GPT-3.5; $p = .02$). Moreover, Gemini (57.48%) outperformed GPT-3.5 (53.33%) nonetheless, without a statistically significant difference ($p = .09$), and all three models surpassed the pooled performance of the residents. Table 2 (bottom) and Figure 4b outline the details of these sub-analyses.

4 | DISCUSSION

This comparative cross-sectional study primarily aimed to evaluate the performance of three of the most commonly used LLM chatbots on the American Academy of Periodontology in-service examination. To serve as a human control group in comparison to AI model performance, postgraduate resident scores across more than 50 North American training programs were analyzed. First and foremost, this study was the first of its kind in dental research to incorporate human control groups in the testing of AI models. While similar studies have been conducted in other research disciplines, none have directly compared the three models. Najafali et al.[16] compared Google Bard (now known as Gemini) performance on the American Society of Plastic Surgeon's (ASPS) in-service examination to the performance among resident year groups. The results indicated that Gemini performed better than

TABLE 2 (Top) Descriptive statistics of periodontal residents' performance divided by year of training. (Middle) Results of sub-analyses. Section analysis. Performance of each large language model divided by exam sections and respective *p* value comparing the difference in performances. (Bottom) Results of performance analysis of artificial intelligence models on the most difficult periodontal in-service exam questions.

	Exam Year				
	2020 (368)	**2021 (337)**	**2022 (305)**	**2023 (302)**	**2020–2023**
PGY-1 Residents					
N	174 (33.14%)	158 (33.05%)	182 (32.21%)	192 (36.16%)	706
Avg Score	214.86 ± 29.94 (58.39% ± 29.94)	216.04 ± 33.43 (64.06% ± 33.43)	203.13 ± 28.76 (66.57% ± 28.76)	193.18 ± 34.56 (63.92% ± 34.56)	206.43 ± 32.84 (63.48% ± 31.67)
PGY-2 Residents					
N	182 (34.67%)	170 (35.56%)	203 (35.92%)	198 (37.29%)	753
Avg Score	222.58 ± 30.17 (60.53% ± 30.17)	229.51 ± 33.04 (68.02% ± 33.04)	216.57 ± 28.97 (71.00% ± 28.97)	206.82 ± 33.27 (68.45% ± 33.27)	218.86 ± 31.11 (66.25% ± 31.61)
PGY-3 Residents					
N	169 (32.19%)	150 (31.38%)	180 (31.85%)	141 (26.55%)	640
Avg Score	229.07 ± 28.75 (62.22% ± 28.75)	232.61 ± 35.31 (68.99% ± 35.31)	219.29 ± 28.81 (71.84% ± 28.81)	223.91 ± 28.95 (74.18% ± 28.95)	224.32 ± 30.32 (69.06% ± 30.45)
All Residents					
N	525	479	565	531	2375
Avg Score	222.11 ± 30.14 (60.35% ± 30.14)	226.03 ± 34.57 (67.04% ± 34.57)	213.11 ± 29.63 (69.81% ± 29.63)	206.43 ± 34.76 (68.35% ± 34.76)	214.62 ± 32.71 (66.39% ± 32.77)
Google Gemini	260 (70.65)	247 (73.29)	231 (75.73)	218 (72.18)	956 (72.86%)
GPT-3.5	230 (62.5)	230 (68.24)	213 (69.83)	179 (59.27)	852 (64.93%)
GPT-4	(290) 78.80%	(266) 78.93%	(247) 80.98%	(241)79.80	1044 (79.57%)
GPT-3.5 vs. Bard	**<0.01**[b]	0.15	0.1	**<0.001**[b]	**<0.001**[b]
GPT4 vs. Bard	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
GPT4 vs. GPT-3.5	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

				***p* Value**		
Section	**GPT-4 score**	**Gemini score**	**GPT-3.5 score**	**GPT4 vs. Gemini**	**GPT-4 vs. GPT-3.5**	**GPT-3.5 vs. Gemini**
Embryology and Anatomy	80 (83.33%)	77 (80.2%)	67 (69.8%)	**.02**	**<.01**	.09
Biostatistics, Experimental Design, and data analysis	14 (93.33%)	11 (73.3%)	13 (86.7%)	.58	.26	.37
Biochemistry-Physiology	114 (93.44%)	103 (84.4%)	104 (85.2%)	**<.001**	**<.001**	.85
Microbiology and Immunology	101 (88.59%)	92 (80.7%)	94 (82.5%)	**<.001**	**<.001**	.73
Periodontal Etiology and Pathology	109 (78.41%)	97 (69.8%)	83 (59.7%)	**<.001**	**<.001**	.07
Pharmacology and Therapeutics	131 (91.60%)	123 (86.0%)	118 (82.5%)	**<.001**	**<.001**	.41
Diagnosis	77 (70%)[b]	66 (60.0%)	61 (55.5%)	.32	**.01**	.49
Treatment Planning and Prognosis	79 (70.53%)	69 (61.6%)	42 (37.5%)	**<.001**	**.03**	**<.001**
Therapy	206 (69.12%)	184 (61.7%)	140 (47.0%)	**<.001**	**<.001**	**<.001**
Oral Pathology/Oral Medicine	148 (90.79%)	134 (82.2%)	130 (79.8%)	**<.001**	**<.001**	.57

Difficult Questions				
Model	**Correct**	**Total**	**Percentage**	***p*-Value**
GPT-4	80	127	62.99%	**.02**[a], .09[b]
GPT-3.5	69	127	54.33%	**.02**[a], .70[c]
Gemini	73	127	57.48%	.09[b], .70[c]
Residents	52	127	40.52%	–

Abbreviations: Avg, Average; *N*, Number of residents who participated in the exam. Bold values indicate a statistically significant *p* value.

[a] *p* Value for chi-square test, comparing GPT-4 versus GPT-3.5.

[b] *p* Value for chi-square test, comparing GPT-4 versus Bard.

[c] *p* Value for chi-square test, comparing GPT-3.5 versus Bard.

FIGURE 2 Comparison of the exam performance of test groups (ChatGPT-4, ChatGPT-3.5, and Google Gemini) and human control groups (The United States periodontics residents).

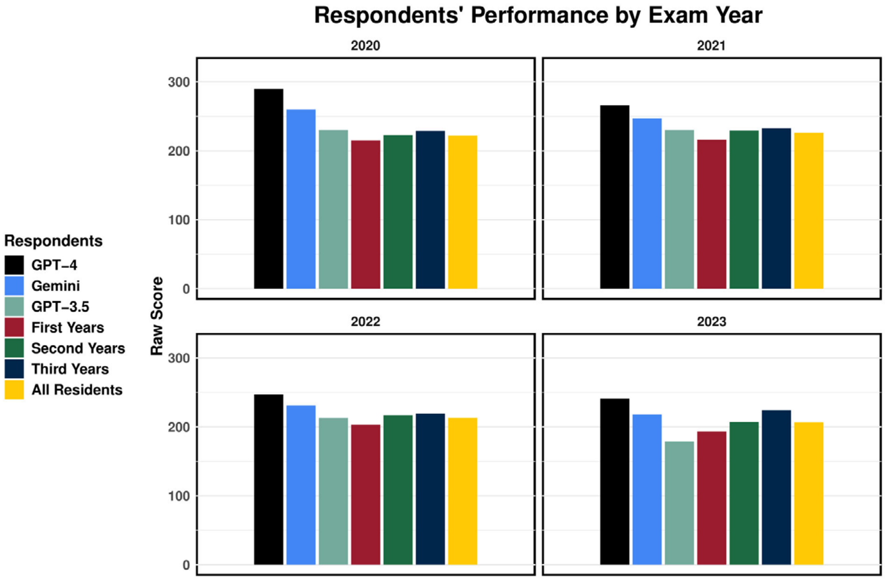
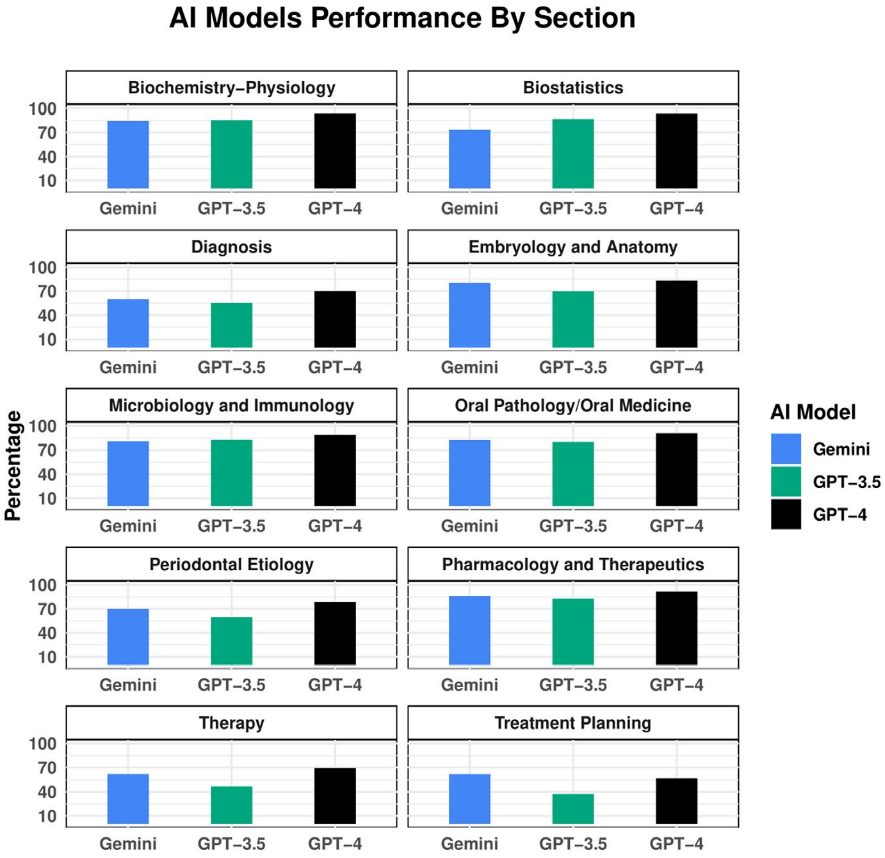


FIGURE 3 Comparison of the performance of ChatGPT-4, ChatGPT-3.5, and Google Gemini in each category of the exam.



56% of PGY-1 residents, but only surpassed 23% and 16% of PGY-2 and PGY-3 residents, respectively.[16] A similar study analyzing ChatGPT's (GPT-3.5) performance on the same ASPS in-service examination questions showed that ChatGPT outperformed 49% of the PGY-1 and 13% of PGY-2 residents.[15] A more recent experiment by Massey et al.[26] tested ChatGPT-3.5 and ChatGPT-4 on the American Board of Orthopedic Surgery written examination with the residents' scores serving as the control group. ChatGPT-3.5 performed poorly compared to GPT-4 similar to our study. ChatGPT-3.5, GPT-4, and orthopedic surgery residents scored 29.4%, 47.2%, and 74.2%, respectively. These outcomes may be attributed to differences in exam structure and content between the ASPS and AAP in-service examinations. It should be noted that the latter comprises more basic science and literature knowledge and less advanced clinical scenarios compared to the former, thus explaining the disparity in AI model performance.

The summary of our results showed that all three AI models achieved an acceptable score in answering the AAP in-service exam

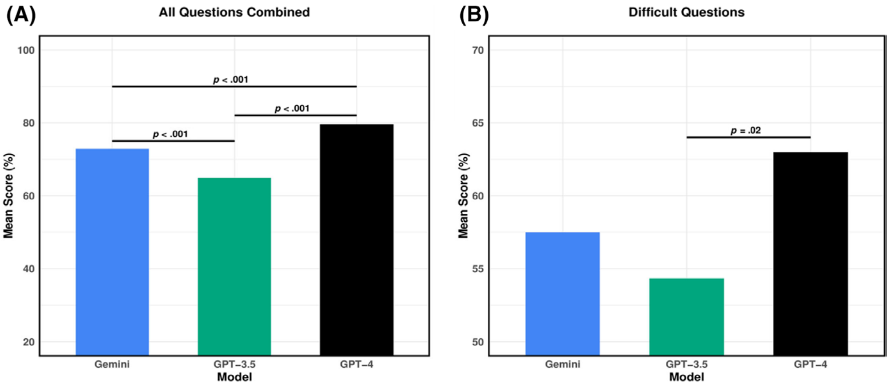


FIGURE 4 Inter-AI models' comparison. (A) All questions (2020–2023) combined. (B) The most difficult questions.

questions, while GPT-4 with scoring approximately 80% correct responses showed promising results compared to the other two. Moreover, GPT-4 and Gemini outperformed all PGY resident groups in all exam years with the exception of PGY-3 resident scoring higher than Gemini in 2023 exam. ChatGPT-3.5, on the other hand, displayed a weaker performance, especially on 2023 exam questions, with a correct response rate of 59.27%, scoring below the PGY-1 resident average (63.92% ± 34.56). When all exam years were combined, Google Gemini (72.86%), ChatGPT-3.5 (64.93%) and GPT-4 (79.57%) all performed better than PGY-1 residents (63.48% ± 31.67), and ChatGPT-3.5 failed to score higher than other resident cohorts (PGY-2: 66.25% ± 31.61, PGY-3: 69.06% ± 30.45, and all PGY groups combined: 66.39% ± 32.77).

Based on these findings, ChatGPT (via GPT-4) would be suitable for educational applications as a result of the models' high accuracy in challenging the AAP in-service examination (80% overall and 63% in most challenging questions). One application could be as a tool for residents to prepare for the exam. Given the proven reliability of this AI resource and the potential to study and learn exquisitely specific details beyond the scope of a single exam question, the benefits of integrating AI models in the postgraduate educational environment are worthy of exploration. The results of this study indicate that GPT-4 would serve as a superior choice compared to Google Gemini and GPT-3.5 with Gemini also outperforming GPT-3.5. GPT-4's superior performance can be attributed to its advanced architecture, larger parameter count, and extensive fine-tuning on domain-specific texts, which enhance its ability to understand and generate accurate responses to complex questions. Additionally, GPT-4 benefits from improved contextual processing and error reduction mechanisms. Moreover, Gemini's advantage over GPT-3.5 can be linked to its access to real-time online information, which provides it with updated and relevant knowledge. This continuous learning from real-time data allows Gemini to offer more current and precise answers compared to GPT-3.5, which relies on a static dataset. Therefore, in examinations where very recent research papers would be used as resources, and when accessibility of real-time knowledge is of great importance one might prefer using Gemini. Conversely, pertinent to more general scientific prompts without referring to a specific research paper (such as basic science and biostatistics), the results of this study suggest that GPT-4 would be the model of choice. Furthermore, GPT-4 and Gemini models proved their reliability in answering specific questions about periodontal disease and implant dentistry, highlighting the AI resources as a potential tool for both resident and patient education. However, future studies are needed to confirm the potential benefit of AI models in the context of clinical care, especially from the perspective of the dental patient.

Two sub-analyses were performed in this comparative study. Firstly, each exam section's sub-analysis was performed, indicating a 70%–93% average score for GPT-4, 60%–86% for Gemini and 37.5%–86.7% for GPT-3.5. GPT-3.5's lowest scores were recorded in the "Treatment Planning and Prognosis" and "Therapy" sections. All three models showed more than 50% accuracy in other exam sections. When only GPT-4 and Gemini are considered, both models showed more than 60% performance in all sections and more than 70% in most of the sections. Fundamentally, these results also indicate a high reliability of these two models in the assessed topics. Considering that approximately half of the AAP in-service examination consists of basic sciences topics (anatomy, physiology, biochemistry, and biostatistics), the possible benefit and utility of these models may extend into other dental specialties and perhaps, beyond the boundaries of dentistry and medicine. In this regard, conflicting reports such as Farajollahi and Modaberi state that only 40% of correct answers were noted when GPT-3.5 was used for Endodontic board exam-style questions (with 70% being a pass rate).[27] Conversely, Danesh et al.[28] reported between 61.3% and 76.9% proportion of correct answers in 3 American dental board exam-style questions when GPT-3.5 and GPT-4 were used, respectively. Overall, it can be concluded that these LLMs can serve as a "go-to" tool for dental students, specialty residents, and academicians, as well as providing a very user-friendly and easily accessible source of data for studying and/or educational pursuits. Nevertheless, caution is required in implementing these tools, since the highest accuracy achieved was around 93% (GPT-4 in the "Biochemistry-Physiology" section). Likewise, there is still debate on whether these AI models can aid researchers in scientific projects and within different related tasks such as writing, data analysis, and referencing.

A second sub-analysis was conducted to pursue a more nuanced understanding of the LLMs' performance, specifically focusing on their abilities to tackle the most challenging questions within the in-service exam. To perform so, we included the questions with a correct response rate below 50%. While the control group of residents displayed an average score of 40.52%, GPT-4 (62.99%) outperformed Gemini and GPT-3.5 outperformed the other groups by 57.48% ($p<.001$) correct responses, while ChatGPT-3.5 scored 53.33%. What truly stood out was the remarkable performance of Gemini in this challenging scenario.

When it comes to comparing the performance of the three LLMs implemented in this study, the findings revealed a significantly higher performance by GPT-4 versus the other two bots in all exam years as well as all questions combined ($p<.001$). Moreover, Gemini performed superior to GPT-3.5 in 2020 ($p<.01$), 2023 ($p<.001$), and all years combined ($p<.001$). A recent study by Danesh et al.,[29] comparing GPT-4 to GPT-3.5 in answering AAP's 2023 in-service exam questions, indicated similar results (57.9% vs. 73.6%) to that of our study (59.27% vs. 79.80%). The slight difference in results would likely to have occurred due to Danesh's study including more questions in their dataset (311 vs. 302 in our study). Nevertheless, similar to our study, they indicated a significantly higher performance by GPT-4 (difference 15.8%; 95% CI, 8.3%–23.0%; $\chi^2=17.12$; df = 1; $p<.001$). However, it should be noted that no human control group was included in that study and only 2023 exam questions were assessed in contrast to our study. Another recent study, in Prosthodontics, assessed ChatGPT-4's accuracy in answering short questions with regards to removable prostheses. The responses showed a reliability of 25.6%, with a confidence range between 22.9% and 28.6%.[30] Furthermore, several studies have been published in medical research comparing LLMs used in our study. Ali et al. reported 62.4% accuracy for ChatGPT (−3.5) and 44.2% for Gemini when prompted to answer oral board questions sourced from a neurosurgery test bank.[20] Similarly, testing the LLMs in solving physiology case vignettes, Dhanvijay et al. reported the superiority of ChatGPT (GPT-3.5) compared to Gemini (3.19 ± 0.3/4 vs. 2.91 ± 0.5/4). On the other hand, the findings of Rahsepar et al. were consistent with our results, revealing 80.8% accuracy for Gemini in answering common lung cancer questions as compared to 70.8% for GPT-3.5. It should be noted that Rahsepar et al. used essay-type questions in contrast to the multiple-choice questions analyzed by our study. The cumulative findings of these studies along with our report suggest variability in the performance of the LLMs among different topics and disciplines. More specifically, the variability is likely due to the differences in question types (e.g., multiple-choice vs. essay), question content (e.g., clinical vignette vs. basic sciences), and the recency of knowledge (e.g., current literature vs. seminal literature). For these reasons, the overall superiority of Gemini or ChatGPT cannot be established. Interpretation of results from these studies emphasizes on the fact that despite promising results achieved in our study on periodontics in-service exam questions, the performance of these LLMs still needs improvement, especially when it comes to short answer or essay questions versus MCQs. Additionally, major limitations of these chatbots should be acknowledged. These include but are not limited to, their inability to interpret image-based prompts and the potential for generating different responses in multiple attempts. Currently, the creators of these models are working to standardize responses across multiple attempts and expand the types of supported prompts, such as images, datasets, and tables.[31]

The integration of AI models with dental education has the potential to transform the learning landscape. For teachers who construct exams, AI platforms can support the creation of customized content, including new questions with answers and reasonable distractors. Teachers can also supplement their approach to learners' achievement of objectives by utilizing the detailed justification provided by AI in certain sections, this can provide insights into effective teaching methods for those topics. For examiners, examining the AI's performance on difficult questions can help determine if these questions are accurately assessing desired skills and knowledge, potentially offering new ways to prepare students. For students, interactive AI-driven sessions, like mock exams or Q&A activities, can be incorporated into their studying routines for a more engaging learning experience. Moreover, AI can offer personalized tutoring, focusing on areas where individual students struggle, and enhancing their learning journey. However, the fact that the maximum achieved accuracy, by GPT-4, in our study was 80%, and thereby, caution in the use of chatbots (especially GPT-3.5) is required. Finally, to improve the exam quality, several strategies can be adopted. AI performance can serve as a benchmark for the exam's difficulty level, indicating whether some sections are too easy or too hard. Second, AI's performance can be used to assess the validity and reliability of the exam, ensuring it effectively measures the intended skills.

4.1 | Limitations

Despite providing valuable insight regarding educational and academic applications of AI in Periodontics, Oral Implantology, and Dentistry, this investigation was subject to certain constraints. Readers should bear in mind that with the daily advances in the world of AI, newer and likely enhanced versions of LLMs are becoming available, which would indeed result in their increased performance, compared to the models implemented in this study. Nonetheless, it should be acknowledged that our research did not solely aim to test the accuracy of these models, but rather provide a broader insight into applications of AI and LLMs in dental and periodontal education and explore how these models can improve educational quality. This might in fact limit the direct clinical translation of our research findings to routine provider clinical. Nonetheless. As these emerging technologies have indeed profound impact in our field, their exploration is crucial. As an example of many clinical implications that AI can have, patients are soon likely to cross-examine dental treatment plans with these LLMs after a consultation visit with a practitioner, or they can refer to such chatbots for asking questions regarding post-operative care, etc. In addition, from the standpoint of translation into direct clinical/surgical care by providers, these tools can act as yet another means of dental education platform, which clinicians or aspiring providers may refer to for dental education. These are only some of the many examples of how AI can and will certainly impact our field. Lastly, it should be emphasized that the performance of these models should also be tested in other question formats such as open-ended or essay questions, as herein we queried this only on MCQ items. Additionally, the models' inability to analyze figures or tables limits their potential use, and the study's sample size of 1312 questions may not be comprehensive enough to draw definitive conclusions.

5 | CONCLUSION

Within the limitations of this comparative study, the reliability of three AI-based LLMs was demonstrated through accuracy in selecting correct answers to prompts sourced from the AAP in-service examination (2020–2023). ChatGPT-4's performed superior to all other human and AI groups with achieving approximately 80% of correct response rate. While also Google Gemini's scores outperformed ChatGPT-3.5 in most exam years and content categories while surpassing several periodontal resident cohorts. Within a broader scope, the findings of this study suggest a future application for utilizing large language models in periodontics and oral implantology as a study reference or easy-access tool for academic activities, particularly as they are being enhanced, and the models are actively improved. However, the need for further improvement in these models should be recognized to meet the minimum requirements for applicability in clinical practice, patient management, and decision-making.

AUTHOR CONTRIBUTIONS

H.S.: conceptualization, statistical analysis, manuscript writing, final approval. **M.H.A.S.:** Conceptualization, data interpretation, critical reading of final version. **P.H.:** data acquisition, data management. **P.S.K.:** Conceptualization, critical reading of final version, final approval. **J.M:** Conceptualization, manuscript writing. **K.M.:** Conceptualization, critical reading of final version, final approval. **H-.L.W.:** Conceptualization, critical reading of final version, and final approval. **S.B.:** Conceptualization, critical reading of final version, and final approval.

ACKNOWLEDGMENTS

The authors express their sincere gratitude to the American Academy of Periodontology (AAP) for graciously granting permission to undertake this study utilizing the in-service examination materials.

CONFLICT OF INTEREST STATEMENT

Dr. Saleh is an appointed committee member of the AAP In-Service Examination Committee. Dr. Merchant is the appointed Co-Chairman of the AAP In-Service Examination Committee.

DATA AVAILABILITY STATEMENT

The data used in this study would be provided upon reasonable request from the first author or corresponding author.

ORCID

Hamoun Sabri https://orcid.org/0000-0001-6581-2104
Parham Hazrati https://orcid.org/0000-0002-8362-3208
Purnima S. Kumar https://orcid.org/0000-0001-5844-1341
Hom-Lay Wang https://orcid.org/0000-0003-4238-1799
Shayan Barootchi https://orcid.org/0000-0002-5347-6577

REFERENCES

1. Kopytko V, Shevchuk L, Yankovska L, Semchuk Z, Strilchuk R. Smart home and artificial intelligence as environment for the implementation of new technologies. *Traekt Nauki=Path Sci*. 2018;4:2007-2012.
2. Xu L, Sanders L, Li K, Chow JC. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021;7:e27850.
3. Teng M, Ren Y. *Application and development prospect of artificial intelligence in the daily life of the elderly and disabled*. In: 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE; 2021:767-770.
4. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol*. 2023;13:1219326.
5. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29:1930-1940.
6. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. *Nat Rev Phys*. 2023;5:277-280.
7. Cocci A, Pezzoli M, Lo Re M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis*. 2023;27(1):103-108. doi:10.1038/s41391-023-00705-y.
8. Jin JQ, Dobry AS. ChatGPT for healthcare providers and patients: practical implications within dermatology. *J Am Acad Dermatol*. 2023;89:870-871.
9. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330:866-869.
10. Harris E. Large language models answer medical questions accurately, but can't match clinicians' knowledge. *JAMA*. 2023;330:792.
11. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern promethean dilemma. *Croat Med J*. 2023;64:1-3.
12. Wach K, Duong CD, Ejdys J, et al. The dark side of generative artificial intelligence: a critical analysis of controversies and risks of ChatGPT. *Entrep Bus Econ Rev*. 2023;11:11-30.
13. Lee H. The rise of ChatGPT: exploring its potential in medical education. *Anat Sci Educ*. 2023;17:926-931.
14. Huang RS, Lu KJQ, Meaney C, Kemppainen J, Punnett A, Leung FH. Assessment of resident and AI Chatbot performance on the University of Toronto Family Medicine Residency Progress Test: comparative study. *JMIR Med Educ*. 2023;9:e50514.
15. Humar P, Asaad M, Bengur FB, Nguyen V. ChatGPT is equivalent to first year plastic surgery residents: evaluation of ChatGPT on the plastic surgery in-service exam. *Aesthet Surg J*. 2023;43:NP1085-NP1089.
16. Najafali D, Reiche E, Araya S, et al. Bard versus the 2022 American Society of Plastic Surgeons in-service examination: performance on the examination in its intern year. *Aesthetic Surg J Open Forum*. 2023;6:ojad066.
17. Vaishya R, Iyengar KP, Patralekh MK, et al. Effectiveness of AI-powered Chatbots in responding to orthopaedic postgraduate exam questions—an observational study. *Int Orthop*. 2024. Epub ahead of print. doi:10.1007/s00264-024-06182-9
18. Dwivedi YK, Kshetri N, Hughes L, et al. Opinion Paper: "so what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int J Inf Manage*. 2023;71:102642.
19. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. *Ann Biomed Eng*. 2023;51:2647-2651.
20. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question Bank. *Neurosurgery*. 2023;93(5):1090-1098.
21. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. *Radiology*. 2023;307:e230922.

22. Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions on lung cancer and lung cancer screening. *AJR Am J Roentgenol*. 2023;221:701-704.
23. Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. *arXiv preprint arXiv:230308774* 2023.
24. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ*. 2023;9:e48002.
25. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2:e0000198.
26. Massey PA, Montgomery C, Zhang AS. Comparison of ChatGPT-3.5, ChatGPT-4, and Orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg*. 2023;31:1173-1179.
27. Farajollahi M, Modaberi A. Can ChatGPT pass the "Iranian Endodontics Specialist Board" exam? *Iran Endod J*. 2023;18:192.
28. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: a preliminary study on ChatGPT. *J Am Dent Assoc*. 2023;154:970-974.
29. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. *J Periodontol*. 2024. Epub ahead of print. doi:10.1002/JPER.23-0514.
30. 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.
31. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901.

How to cite this article: Sabri H, Saleh MHA, Hazrati P, et al. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. *J Periodont Res*. 2025;60:121-133. doi:10.1111/jre.13323

APPENDIX 1

STROBE STATEMENT—CHECKLIST OF ITEMS THAT SHOULD BE INCLUDED IN REPORTS OF CROSS-SECTIONAL STUDIES

	Item No	Recommendation	Page Number
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract	1,2
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/Rationale	2	Explain the scientific background and rationale for the investigation being reported	3
Objectives	3	State-specific objectives, including any prespecified hypotheses	3,4
Methods			
Study design	4	Present key elements of study design early in the paper	5
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	5
Participants	6	(a) Give the eligibility criteria and the sources and methods of selection of participants	5,6
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	5,6
Data sources/ Measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	5,6
Bias	9	Describe any efforts to address potential sources of bias	6
Study size	10	Explain how the study size was arrived at	5
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	5,6
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding	6
		(b) Describe any methods used to examine subgroups and interactions	6
		(c) Explain how missing data were addressed	5,6
		(d) If applicable, describe analytical methods taking account of sampling strategy	NA
		(e) Describe any sensitivity analyses	NA
Results			
Participants	13	(a) Report numbers of individuals at each stage of study—for example, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed	6,7
		(b) Give reasons for non-participation at each stage	7
		(c) Consider use of a flow diagram	NA

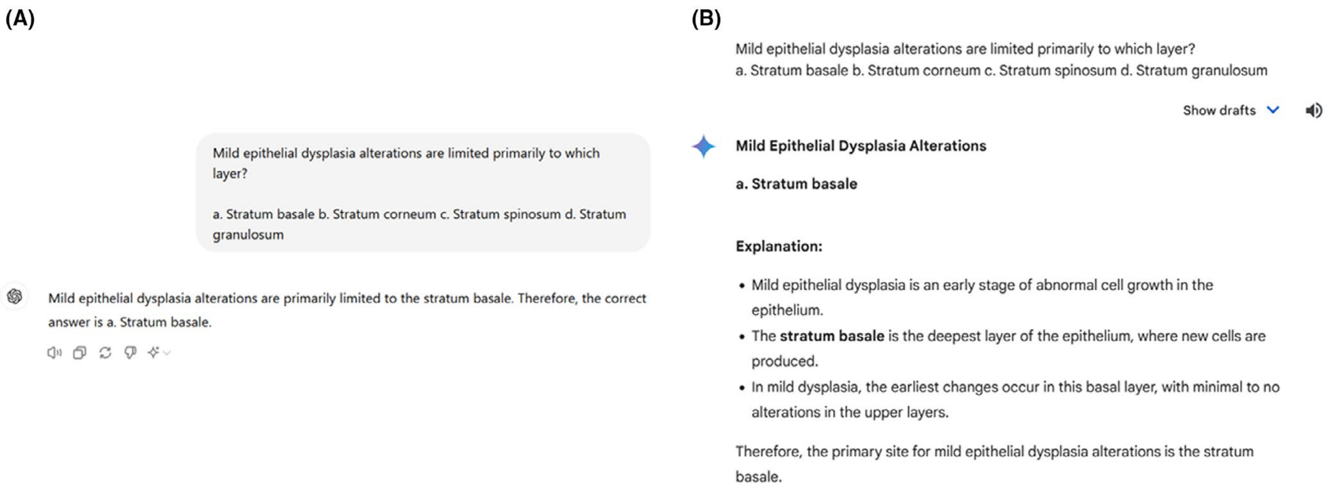
	Item No	Recommendation	Page Number
Descriptive data	14	(a) Give characteristics of study participants (e.g., demographic, clinical, and social) and information on exposures and potential confounders	6,7
		(b) Indicate number of participants with missing data for each variable of interest	6,7
Outcome data	15	Report numbers of outcome events or summary measures	7
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	NR
		(b) Report category boundaries when continuous variables were categorized	7,8
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	NR
Other analyses	17	Report other analyses done—for example, analyses of subgroups and interactions, and sensitivity analyses	8
Discussion			
Key results	18	Summarize key results with reference to study objectives	8
Limitations	19	Discuss the limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	10
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	8,9
Generalizability	21	Discuss the generalizability (external validity) of the study results	10
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	NR

APPENDIX 2

MATERIALS AND METHODS

Statistical analysis

Firstly, descriptive statistics were reported as means and standard deviation (SD) for continuous variables, whereas count and percentage were utilized for categorical variables. A Chi-square test was employed to compare the performance of Gemini and ChatGPT bots. To minimize bias, the statistician (H.S) who performed the data analysis was blinded to the allocated group of LLMs (i.e., Gemini and ChatGPT). All statistical analyses were conducted using a software package (RStudio software, 2023.06.2 Build 561, R Studio, PBC, Boston, MA, USA). A *p*-value of .05 was set for statistical significance level.



(A)

Mild epithelial dysplasia alterations are limited primarily to which layer?

a. Stratum basale b. Stratum corneum c. Stratum spinosum d. Stratum granulosum

Mild epithelial dysplasia alterations are primarily limited to the stratum basale. Therefore, the correct answer is a. Stratum basale.

(B)

Mild epithelial dysplasia alterations are limited primarily to which layer?
a. Stratum basale b. Stratum corneum c. Stratum spinosum d. Stratum granulosum

Mild Epithelial Dysplasia Alterations

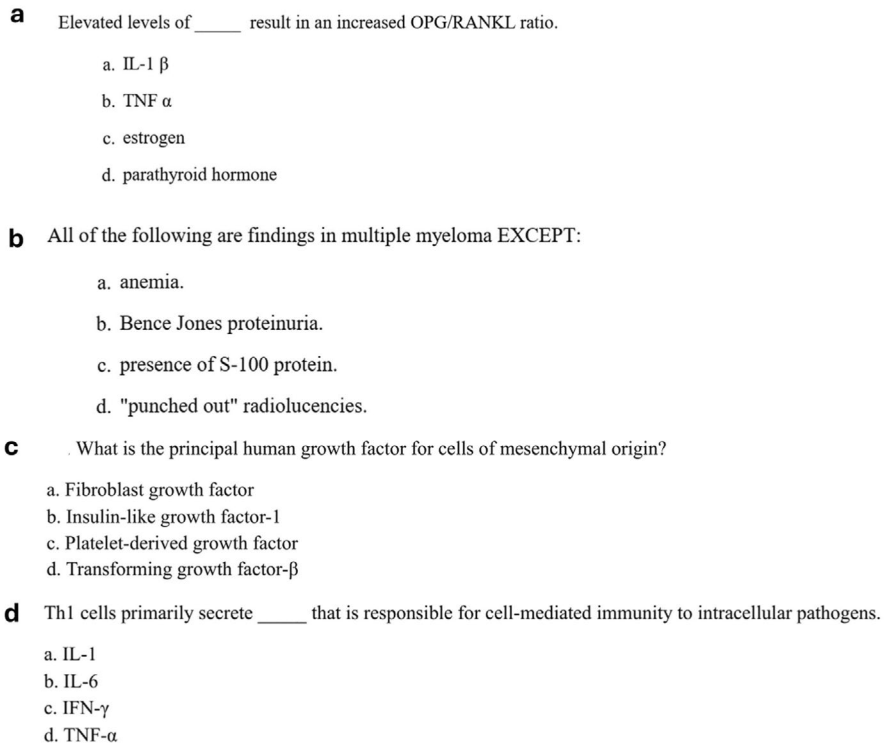
a. Stratum basale

Explanation:

- Mild epithelial dysplasia is an early stage of abnormal cell growth in the epithelium.
- The **stratum basale** is the deepest layer of the epithelium, where new cells are produced.
- In mild dysplasia, the earliest changes occur in this basal layer, with minimal to no alterations in the upper layers.

Therefore, the primary site for mild epithelial dysplasia alterations is the stratum basale.

FIGURE A1 Illustration of ChatGPT (A) and Google Gemini (B) portal environments, and an example of prompts presented to this models with the corresponding answer.



a Elevated levels of _____ result in an increased OPG/RANKL ratio.

a. IL-1 β
b. TNF α
c. estrogen
d. parathyroid hormone

b All of the following are findings in multiple myeloma EXCEPT:

a. anemia.
b. Bence Jones proteinuria.
c. presence of S-100 protein.
d. "punched out" radiolucencies.

c What is the principal human growth factor for cells of mesenchymal origin?

a. Fibroblast growth factor
b. Insulin-like growth factor-1
c. Platelet-derived growth factor
d. Transforming growth factor-β

d Th1 cells primarily secrete _____ that is responsible for cell-mediated immunity to intracellular pathogens.

a. IL-1
b. IL-6
c. IFN-γ
d. TNF-α

FIGURE A2 Two examples (A–D) of difficult questions included in the study sub-analysis.