Revista
ESTOMATOLÓGICA
Herediana
E-ISSN 2225-7616

ORIGINAL ARTICLE
ARTÍCULO ORIGINAL / ARTIGO ORIGINAL

DOI: https://doi.org/10.20453/reh.v35i4.6253

Can artificial intelligence-based large language models pass the National Dentistry Examination in Peru?

¿Pueden los grandes modelos de lenguaje de inteligencia artificial aprobar el Examen Nacional de Odontología en el Perú?

Os grandes modelos de linguagem de inteligência artificial conseguem ser aprovados no Exame Nacional de Odontologia no Peru?

Miguel Á. Saravia-Rojas[1],
Carlos Mendiola-Aquino[1],
Francisco Orejuela-Ramirez[1],
Wanderley Tunquipa-Chacón[1],
Rocio Geng-Vivanco[1]

[1] Universidad Peruana Cayetano Heredia, School of Stomatology. Lima, Peru.

ABSTRACT

Objective: To determine which artificial intelligence (AI) large language model demonstrates the highest accuracy in answering the 2023 National Dentistry Examination (ENAO, by its acronym in Spanish) in Peru, compared with the official answer key. **Material and methods:** The 100 multiple-choice questions from the 2023 ENAO were tested using ChatGPT-3.5, ChatGPT-4, Gemini, and Copilot. Responses were categorized by subject area and scored as correct or incorrect. Data were analyzed using the chi-square test ($\alpha = 0.05$). **Results:** ChatGPT-4 achieved the highest overall accuracy (90.00%), followed by Gemini (82.00%), Copilot (79.00%), and ChatGPT-3.5 (76.00%). Across most models, the highest accuracy was observed in Public Health, Research, Health Services Management, and Ethics, whereas lower performance was observed in Anatomy and in Oral Medicine and Pathology. Pairwise comparisons revealed that ChatGPT-4 performed significantly better than ChatGPT-3.5 (difference: 14%; $p = 0.0084$) and Copilot (difference: 11%; $p = 0.0316$); no significant differences were found among the remaining model comparisons ($p > 0.05$). **Conclusion:** All AI language models demonstrated effectiveness in answering the 2023 ENAO questions, with ChatGPT-4 achieving the highest accuracy.

Keywords: artificial intelligence; dental education; educational assessment; large language models.

Received: 2025-01-17
Accepted: 2025-11-13
Online: 2025-12-30

Cite as:
Saravia-Rojas MÁ, Mendiola-Aquino C, Orejuela-Ramirez F, Tunquipa-Chacón W, Geng-Vivanco R. Can artificial intelligence-based large language models pass the National Dentistry Examination in Peru? Rev Estomatol Herediana. 2025; 35(4): 305-311. DOI: 10.20453/reh.v35i4.6253

RESUMEN

Objetivo: Determinar qué modelo de lenguaje de gran tamaño basado en inteligencia artificial (IA) presenta mayor precisión al responder el Examen Nacional de Odontología (ENAO) de 2023 en Perú, en comparación con el banco de respuestas oficiales. **Materiales y métodos:** Las 100 preguntas de opción múltiple del examen se probaron en ChatGPT-3.5, ChatGPT-4, Gemini y Copilot, y las respuestas se clasificaron por materias. Cada respuesta se marcó como correcta o incorrecta, y los datos se analizaron mediante la prueba de chi-cuadrado ($\alpha = 0,05$). **Resultados:** ChatGPT-4 alcanzó la mayor precisión global (90,00 %), seguido de Gemini (82,00 %), Copilot (79,00 %) y ChatGPT-3.5 (76,00 %). Por área temática, Salud Pública, Investigación, Gestión de Servicios de Salud y Ética mostraron las mayores tasas de acierto en la mayoría de los modelos, mientras que Anatomía y Medicina Oral y Patología mostraron un desempeño inferior. Las comparaciones pareadas revelaron que ChatGPT-4 tuvo un rendimiento significativamente superior al de ChatGPT-3.5 (diferencia: 14 %; $p = 0,0084$) y al de Copilot (diferencia: 11 %; $p = 0,0316$), mientras que no se encontraron diferencias estadísticamente significativas entre los demás modelos ($p > 0,05$). **Conclusión:** Todos los modelos de lenguaje de gran tamaño basados en IA demostraron su eficacia al responder las preguntas del ENAO de 2023, siendo que ChatGPT-4 mostró la mayor precisión.

Palabras clave: inteligencia artificial; educación odontológica; evaluación educativa; modelos de lenguaje de gran tamaño.

RESUMO

Objetivo: Determinar qual modelo de linguagem de inteligência artificial (IA) apresenta maior precisão ao responder ao Exame Nacional de Odontologia (ENAO) de 2023 no Peru, em comparação com o banco de respostas oficial. **Materiais e métodos:** As 100 perguntas de múltipla escolha do exame foram testadas no ChatGPT-3.5, ChatGPT-4, Gemini e Copilot, e as respostas foram classificadas por matéria. Cada resposta foi marcada como correta ou incorreta, e os dados foram analisados através do teste qui-quadrado ($\alpha = 0,05$). **Resultados:** O ChatGPT-4 alcançou a maior precisão geral (90,00%), seguido pelo Gemini (82,00%), Copilot (79,00%) e ChatGPT-3.5 (76,00%). Por assunto, Saúde Pública, Pesquisa, Gestão de Serviços de Saúde e Ética apresentaram a maior precisão na maioria dos modelos, enquanto um desempenho inferior foi observado em Anatomia e Medicina Oral e Patologia. Comparações pareadas revelaram que o ChatGPT-4 teve um desempenho significativamente melhor do que o ChatGPT-3.5 (diferença: 14%; $p = 0,0084$) e o Copilot (diferença: 11%; $p = 0,0316$), enquanto não foram encontradas diferenças significativas entre os demais modelos ($p > 0,05$). **Conclusão:** Todos os modelos de linguagem da IA demonstraram a sua eficácia ao responder às perguntas da ENAO de 2023, sendo que o ChatGPT-4 apresentou a maior precisão.

Palavras-chave: inteligência artificial; educação odontológica; avaliação educacional; modelos de linguagem de grande porte.

INTRODUCTION

Artificial intelligence (AI) has revolutionized how we interact with the world. At the forefront of this transformation are language models, a subcategory category of AI designed to understand and respond to human language in a natural and contextually relevant manner (1). These models use deep learning algorithms to analyze large datasets, enabling them to generate coherent text based on user inputs (2). Language models have become key tools across various fields. In education, they assist with tutoring, generate practice questions, and provide feedback and explanations for complex concepts (3). Additionally, they can simulate real-life scenarios, serving as valuable supplementary resources for students, teachers and professionals (4).

Chatbots, powered by these advanced language models, are increasingly used to enhance user experience across various platforms. In education, chatbots can answer exam questions, provide study assistance, and offer insights on various topics, thus enhancing the learning process and supporting academic success (3, 4). The accuracy, precision, and reliability of chatbots in solving

exams in various medical and dental sciences have been studied (5-11). Chatbots demonstrate a satisfactory level of accuracy, positioning them as potential interactive tools in education to support learning (5, 7-9, 11). However, studies reveal varying accuracy among chatbots, often due to outdated data and difficulties in addressing complex situational queries or questions with multiple correct answers (7, 8). In decision-making scenarios, chatbots may provide incorrect or incomplete information. Therefore, caution is advised when considering them as reliable support (9, 10).

Initially launched as free versions, chatbots have recently introduced paid versions that offer enhanced inference capabilities and improved accuracy (6, 9). As chatbots continue to evolve, ongoing attention and evaluation are essential. Studies also emphasize the need to reassess their performance and efficacy across different languages (10, 12).

The National Dentistry Examination (ENAO, by its acronym in Spanish) in Peru, administered by the Peruvian Association of Dental Schools (ASPEFO, by its acronym in Spanish), is a standardized test aimed at evaluating the knowledge and competencies of dentistry graduates and is a requirement for practicing dentistry in Peru. The exam is conducted annually and covers a wide range of topics pertinent to dental practice. Its purpose is to ensure that practitioners meet the national standards for professional competence.

This study aimed to determine which AI language model exhibits the highest accuracy in answering the 2023 ENAO, compared to the official answer bank provided by ASPEFO. By evaluating the performance of different AI models, research seeks to assess their efficacy as educational tools. The null hypothesis tested was that there would be no difference among the language models in their performance on the ENAO exam.

METHODS

This was a comparative, cross-sectional, observational study evaluating the accuracy of AI language models in answering the 2023 ENAO in Peru. The 2023 ENAO consisted of two tests (Part 1 and Part 2), each containing 100 multiple-choice questions. For practical reasons, only Part 1 was included in this study, as it encompasses core dental topics and offers a representative sample for assessing the accuracy of AI language models. Ethical approval was not required, as the analysis exclusively involved publicly available data from an official examination, without involving human participants, personal data, or clinical interventions. The complete set of questions and official answers for Part 1 is accessible through the ASPEFO at: https://aspefo.org/wp-content/uploads/2023/03/Prueba-1-Solucionario.pdf

The 100 questions from Part 1 were administered to four AI language models: ChatGPT-3.5 (developed by OpenAI), ChatGPT-4 (by OpenAI), Gemini (by Google), and Copilot (by Microsoft). These platforms were selected based on their widespread use and accessibility during the study period. Prior to data collection, a training phase was conducted to standardize the process of generating prompts for the AI models. The prompt used was: "What is the correct answer to the following 2023 ENAO question?" followed by the question and its multiple-choice options. Each model's answers were captured via screenshots and transcribed into an Excel database.

The responses were then categorized by subject area: Embryology and Histology, Anatomy, Oral Medicine and Pathology, Oral and Maxillofacial Surgery, Public Health, Research, Health Services Management, and Ethics. Each response was marked as correct or incorrect by comparison with the official answer key provided by ASPEFO.

The correct response rate was calculated as:

$$\text{Correct response rate (\%)} = \frac{\text{Number of correct answers}}{\text{Total number of questions}} \times 100$$

Descriptive statistics were used to summarize the number and percentage of correct answers per subject and language model. Incorrect answers from Gemini, ChatGPT-3.5, and Copilot were compared against correct responses from ChatGPT-4 by subject to quantify differences in performance. Percentages of incorrect answers were calculated relative to the number of correct answers of ChatGPT-4. Pairwise differences in the percentage of correct answers between models were calculated, and their statistical significance was assessed using Z tests for proportions ($p<0.05$). All analyses were performed using SPSS Statistics v. 29.

RESULTS

Table 1 presents the number and percentage of correct answers provided by each AI model across different subjects in response to the ENAO exam. ChatGPT-4 achieved the highest overall accuracy, correctly answering 90 out of 100 questions (90.00%), followed by Gemini (82.00%) and Copilot (79.00%), while ChatGPT-3.5 showed the lowest performance (76.00%). By subject, Public Health, Research, Health Services Management, and Ethics was the area with the highest accuracy for ChatGPT-3.5 (n = 35/40; 87.50%), Copilot (n = 35/40; 87.50%), and ChatGPT-4 (n = 37/40; 92.50%). In contrast, Embryology and Histology was the area with the highest accuracy for Gemini (n = 28/30; 93.33%). The lowest scores were observed in Oral Medicine and

Pathology for Gemini (n = 6/10; 60.00%); in Anatomy (n = 6/10; 60.00%) and Oral Medicine and Pathology (n = 6/10; 60.00%) for ChatGPT-3.5; in Embryology and Histology (n = 21/30; 70.00%) and Anatomy (n = 7/10; 70.00%) for Copilot; and in Embryology and Histology for ChatGPT-4 (n = 26/30; 86.67%).

Table 1. Number and percentage of correct answers by subject from the language models in response to the ENAO Exam.

Subject	Total No. of questions	Gemini		ChatGPT-3.5		Copilot		ChatGPT-4	
		No. of correct answers	Correct response rate (%)	No. of correct answers	Correct response rate (%)	No. of correct answers	Correct response rate (%)	No. of correct answers	Correct response rate (%)
Embryology and Histology	30	28	93.33	22	73.33	21	70.00	26	86.67
Anatomy	10	7	70.00	6	60.00	7	70.00	9	90.00
Oral Medicine and Pathology	10	6	60.00	6	60.00	8	80.00	9	90.00
Oral and Maxillofacial Surgery	10	9	90.00	7	70.00	8	80.00	9	90.00
Public Health, Research, Health Services Management, and Ethics	40	32	80.00	35	87.50	35	87.50	37	92.50
Total	100	82	82.00	76	76.00	79	79.00	90	90.00

Table 2 shows the number and percentage of incorrect answers from Gemini, ChatGPT-3.5, and Copilot relative to the correct answers provided by ChatGPT-4, by subject. Overall, Gemini had the fewest incorrect answers (8.90%), followed by Copilot (12.20%) and ChatGPT-3.5 (15.60%). By subject, Gemini made no errors in Oral and Maxillofacial Surgery (0.00%), two errors in Anatomy (22.20%), three errors in Oral Medicine and Pathology (33.30%), and five errors in Public Health, Research, Health Services Management, and Ethics (13.50%). Notably, in Embryology and Histology, Gemini outperformed ChatGPT-4, providing two additional correct answers (-7.70%). ChatGPT-3.5 had four incorrect answers in Embryology and Histology (15.40%), three errors each in Anatomy (33.30%) and Oral Medicine and Pathology (33.30%), and two errors in both Oral and Maxillofacial Surgery (22.20%) and Public Health, Research, Health Services Management, and Ethics (5.40%). Copilot made five errors in Embryology and Histology (19.20%), two errors each in Anatomy (22.20%) and Public Health, Research, Health Services Management, and Ethics (5.40%), and one error each in Oral Medicine and Pathology (11.10%) and Oral and Maxillofacial Surgery (11.10%).

Table 2. Number and percentage of incorrect answers from Gemini, ChatGPT-3.5, and Copilot relative to correct answers by ChatGPT-4, by subject.

Subject	No. of incorrect answers by Gemini (%)	No. of incorrect answers by ChatGPT-3.5 (%)	No. of incorrect answers by Copilot (%)
Embryology and Histology	-2* (-7.70%)	4 (15.40%)	5 (19.20%)
Anatomy	2 (22.20%)	3 (33.30%)	2 (22.20%)
Oral Medicine and Pathology	3 (33.30%)	3 (33.30%)	1 (11.10%)
Oral and Maxillofacial Surgery	0 (0.00%)	2 (22.20%)	1 (11.10%)
Public Health, Research, Health Services Management, and Ethics	5 (13.50%)	2 (5.40%)	2 (5.40%)
Total	8 (8.90%)	14 (15.60%)	11 (12.20%)

Percentages represent the proportion of incorrect answers from each model relative to the number of correct answers generated by ChatGPT-4 for the same subject. *Negative value indicates that Gemini provided more correct answers than ChatGPT-4 in this subject.

Table 3 presents the pairwise differences in the percentage of correct answers across the language models. ChatGPT-4 achieved significantly higher accuracy compared with ChatGPT-3.5 (difference: 14%; $p = 0.0084$) and Copilot (difference: 11%; $p = 0.0316$). The remaining comparisons did not show statistically significant differences ($p > 0.05$).

Table 3. Pairwise differences in percentage of correct answers between language models.

		Difference (% of correct answers)	p
ChatGPT-4	Gemini	8.0	0.103
	ChatGPT-3.5	14.0	0.0084
	Copilot	11.0	0.0316
Gemini	ChatGPT-3.5	6.0	0.2976
	Copilot	3.0	0.5924
ChatGPT-3.5	Copilot	-3.0	0.6115

Differences calculated as (% correct of first model – % correct of second model). Statistical significance assessed with Z test of proportions ($p < 0.05$).

DISCUSSION

Language models have emerged as tools capable of understanding and responding to queries in a human-like manner. These models have recently shown promise as useful aids in preparing for licensing exams in healthcare (11, 13). However, their performance varies depending on factors such as the quality of their training data, the design of their algorithms, and their validation against real-world clinical scenarios (6, 7, 10, 11, 13). This study aimed to evaluate the accuracy of responses generated by four AI language models on the 2023 ENAO exam, revealing significant differences among the models and leading to the rejection of the null hypothesis.

One of the most well-known language models is ChatGPT, developed by OpenAI. Trained on an extensive dataset, it can generate responses across various languages and subjects. Initially available as a free version, ChatGPT-4, a paid version, was released in February 2023. This version is claimed to offer improved accuracy (5, 14), as evidenced by this study and a previous research by Gilson et al. (5). ChatGPT-4 demonstrated a significant improvement in responding to medical questions, achieving 60% higher accuracy compared to the free version, ChatGPT-3.5 (5). In our study, ChatGPT-4 provided the highest number of correct responses on the ENAO exam, attaining a correct response rate of 90% or more in all areas except for Embryology and Histology.

In another study, Takagi et al. (6) compared the performance of both language models on the Japanese Medical Licensing Examination, finding that the paid version achieved greater accuracy, especially on challenging and specific questions. Language models have shown consistent improvements in areas where they previously faced challenges. Notably, ChatGPT-4 demonstrates significant advancement in tasks where ChatGPT-3.5 showed weaknesses (5).

In this study, ChatGPT-3.5 exhibited the lowest performance among the language models. Previous research has highlighted that ChatGPT-3.5 struggles with certain exams, likely due to outdated data and insufficient advancements (7, 8, 15). It often fails to fully grasp the context of questions, make logical deductions, and provide accurate answers (15). Additionally, ChatGPT-3.5 faces challenges with multiple-choice questions, which could explain our results (6, 7). Compared to the correct responses of ChatGPT-4, ChatGPT-3.5 had the most errors, especially in areas that were more difficult for the language models. Users should not rely solely on its responses for specialized consultations (15).

Few studies have assessed the performance of Gemini or its predecessor, Bard, developed by Google, on exams in the health field. Patil et al. (11) compared ChatGPT-4 and Bard on radiology questions, finding that ChatGPT-4 demonstrated greater accuracy than Bard, outperforming it in some areas while showing no significant difference in others. Our study supports this finding, as ChatGPT-4 outperformed Gemini in most areas, although their performances were comparable in Oral and Maxillofacial Surgery. However, Gemini excelled in Embryology and Histology, showing two more correct answers than ChatGPT-4, likely due to recent updates. Gemini has shown significant advancements in processing large volumes of information and undergoes extensive ethical and safety testing to identify biases, ensuring its safe use (16). Our study found that Gemini's overall performance on the ENAO exam was comparable to ChatGPT-4, with both models exhibiting limitations and occasionally providing incorrect responses. This highlights the need for cautious use of these tools, as they can potentially mislead users without prior knowledge.

Another study that compared ChatGPT-3.5, ChatGPT-4, and Bard found that ChatGPT-4 consistently outperformed the others, followed by Bard and then ChatGPT-3.5. Nevertheless, only Bard answered the most specific questions correctly (17), suggesting it may have particular strengths. This was also evident in our study, where Gemini demonstrated superior performance in Embryology and Histology.

On the other hand, Copilot, developed by Microsoft, integrates with Microsoft 365 applications to provide real-time assistance and enhance efficiency in daily tasks. Designed to boost productivity and creativity, Copilot can assist with various tasks, including document writing, data analysis, trend identification, converting Word documents into presentations, and drafting emails (18). In this study, Copilot demonstrated intermediate performance, showing more mistakes compared to ChatGPT-4 in all subjects. However, its overall performance was comparable to Gemini and ChatGPT-3.5.

There are currently no studies evaluating Copilot's performance on health-related exams. Nonetheless, Kaftan et al. (13) compared the accuracy of ChatGPT-3.5, Copilot, and Gemini in interpreting biochemical data. Copilot demonstrated the highest accuracy, followed by Gemini, while ChatGPT-3.5 performed the least effectively. The authors noted that different language models use varying algorithms, with Copilot leveraging data from a wide range of sources, including reputable medical websites, research articles, and clinical guidelines.

Language models exhibit variable performance across different subjects, each demonstrating particular strengths in specific domains (19). Our results demonstrated that subjects such as Oral Medicine and Pathology, Anatomy, and Embryology and Histology had greater challenges, resulting in more mistakes across all language models, with the exception of Gemini. Notably, Gemini outperformed the other models in Embryology and Histology, highlighting its strength in handling specific questions in these areas. Conversely, topics such as Public Health, Research, Health Services Management, and Ethics were less challenging, leading to higher accuracy. This is likely due to the more straightforward nature of these questions, which require less complex reasoning. However, Gemini struggled with some of these questions.

As demonstrated in this study, advanced language models have significant potential for supporting learning across various languages and subjects (5, 6). However, caution is necessary, as these models may frequently provide incorrect answers, potentially leading users to acquire inaccurate information if not verified. Additionally, there is a risk that students might become overly reliant on these language models, which could impact their development of critical thinking and problem-solving skills (20).

Integrating these tools into the educational process is inevitable; however, ongoing research into their applications and updates, as well as comparisons with other AI technologies in education, remains essential (20). Future studies could explore metrics such as response speed and character count, and evaluate the performance of these technologies across different exams and languages. As AI-assisted chatbots continue to evolve, more robust models are expected to emerge, improving accuracy and reliability for educational purposes.

This study has several limitations. First, the accuracy of the language models depends on their training data, which may not be fully updated or specifically tailored to the ENAO exam. Additionally, while the models can identify correct answers, they lack the ability to provide clinical reasoning or contextual understanding. The interpretation of questions may also vary due to ambiguity or differences in language processing. Another limitation is that only a single standardized prompt was used, without exploring whether variations in phrasing could influence responses. Lastly, the study was limited to 100 multiple-choice questions from a single exam, and the findings may not be generalizable to other standardized tests or languages.

CONCLUSION

All language models demonstrated efficacy in answering the 2023 ENAO questions, each with distinct expertise. ChatGPT-4, the paid version, showed the highest accuracy compared to the others.

Conflict of interest:
The authors declare no conflict of interest.

Funding:
Self-funded.

Authorship contribution:
MASR: conceptualization, investigation, project administration, supervision, writing – original draft.
CMA: conceptualization, investigation, supervision, validation.
FOR: data curation, formal analysis, investigation, validation.
WTC: data curation, methodology.
RGV: formal analysis, validation, writing – review & editing.

Corresponding author:
Miguel Á. Saravia-Rojas
✉ miguel.saravia@upch.pe

REFERENCES

1. 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 [Internet]. 2021; 7(4): e27850. Available from: https://doi.org/10.2196/27850
2. Chung J, Kim D, Choi J, Yune S, Song KD, Kim S, et al. Prediction of oxygen requirement in patients with COVID-19 using a pre-trained chest radiograph xAI model: efficient development of auditable risk prediction models via a fine-tuning approach. Sci Rep [Internet]. 2022; 12: 21164. Available from: https://doi.org/10.1038/s41598-022-24721-5
3. Kim D, Chung J, Choi J, Succi MD, Conklin J, Longo MG, et al. Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. Nat Commun [Internet]. 2022; 13: 1867. Available from: https://doi.org/10.1038/s41467-022-29437-8
4. O'Shea A, Li MD, Mercaldo ND, Balthazar P, Som A, Yeung T, et al. Intubation and mortality prediction in hospitalized COVID-19 patients using a combination of convolutional neural network-based scoring of chest radiographs and clinical data. BJR Open [Internet]. 2022; 4(1): 20210062. Available from: https://doi.org/10.1259/bjro.20210062
5. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ [Internet]. 2023; 9: e45312. Available from: https://doi.org/10.2196/45312
6. 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 [Internet]. 2023; 9: e48002. Available from: https://doi.org/10.2196/48002
7. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation study. JMIR Nurs [Internet]. 2023; 6: e47305. Available from: https://doi.org/10.2196/47305
8. Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc [Internet]. 2023; 86(7): 653-658. Available from: https://doi.org/10.1097/JCMA.0000000000000942
9. Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Accuracy of ChatGPT-generated information on head and neck and oromaxillofacial surgery: a multicenter collaborative analysis. Otolaryngol Head Neck Surg [Internet]. 2024; 170(6): 1492-1503. Available from: https://doi.org/10.1002/ohn.489
10. Suárez A, Díaz-Flores V, Algar J, Gómez M, Llorente M, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. Int Endod J [Internet]. 2024; 57(1): 108-113. Available from: https://doi.org/10.1111/iej.13985
11. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative performance of ChatGPT and Bard in a text-based radiology knowledge assessment. Can Assoc Radiol J [Internet]. 2024; 75(2): 344-350. Available from: https://doi.org/10.1177/08465371231193716
12. Morishita M, Fukuda H, Muraoka K, Nakamura T, Hayashi M, Yoshioka I, et al. Evaluating GPT-4V's performance in the Japanese national dental examination: a challenge explored. J Dent Sci [Internet]. 2024; 19(3): 1595-1600. Available from: https://doi.org/10.1016/j.jds.2023.12.007
13. Kaftan AN, Hussain MK, Naser FH. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. Sci Rep [Internet]. 2024; 14: 8233. Available from: https://doi.org/10.1038/s41598-024-58964-1
14. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: differences in the amount of information per medical field. Int J Med Inform [Internet]. 2023; 180: 105283. Available from: https://doi.org/10.1016/j.ijmedinf.2023.105283
15. Farajollahi M, Modaberi A. Can Chatgpt pass the "Iranian endodontics specialist board" exam? Iran Endod J [Internet]. 2023; 18(3): 192. Available from: https://doi.org/10.22037/iej.v18i3.42154
16. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye [Internet]. 2024; 38(13): 2530-2535. Available from: https://doi.org/10.1038/s41433-024-03067-4
17. Ohta K, Ohta S. The performance of GPT-3.5, GPT-4, and Bard on the Japanese national dentist examination: a comparison study. Cureus [Internet]. 2023; 15(12): e50369. Available from: https://doi.org/10.7759/cureus.50369
18. Spataro J. Introducing Microsoft 365 Copilot - Your copilot for work [Internet]. Official Microsoft Blog; 2023, March 16. Available from: https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/
19. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding [preprint en Internet]. arXiv; 2021. Available from: https://doi.org/10.48550/arXiv.2009.03300
20. Memarian B, Doleck T. ChatGPT in education: methods, potentials, and limitations. Comput Hum Behav Artif Humans [Internet]. 2023; 1(2): 100022. Available from: https://doi.org/10.1016/j.chbah.2023.100022