CeuEM Revista Española de Educación Médica

UNIVERSIDAD DE MURCIA

Is GPT-4 capable of passing MIR 2023? Comparison between GPT-4 and ChatGPT-3 in the MIR 2022 and 2023 exams

¿Es capaz GPT-4 de aprobar el MIR 2023? Comparativa entre GPT-4 y ChatGPT-3 en los exámenes MIR 2022 y 2023

Álvaro Cerame* [1], Juan Juaneda* [2], Pablo Estrella-Porter [3], Lucía de la Puente [4], Joaquín Navarro [5], Eva García [6], Domingo A. Sánchez [7,8], Juan Pablo Carrasco [9]**

1 Comprehensive Care Plan for the Sick Health Professional, Madrid Health Service, Madrid, Spain. ORCID: 0000-0001-9137-7775
2 Preventive Medicine and Public Health Service, Hospital Universitari i Politècnic La Fe, Valencia, Spain. ORCID: 0000-0002-6048-2457.
3 Preventive Medicine Service, Valencia University Clinical Hospital, Valencia, Spain. ORCID: 0000-0003-4137-7691
4 Department of Primary Care, Hospital Universitari i Politècnic La Fe, Valencia, Spain. ORCID: 0009-0007-3263-5691
5 Intensive Care Service, North Huelva Health Management Area, Huelva, Spain. ORCID: 0000-0002-7983-7289.
6 Cardiology Service, Toledo University Hospital Complex, Toledo, Spain. ORCID: 0000-0001-8962-6023
7 General Council of Official Colleges of Physicians of Spain, Madrid, Spain.
8 Medical Oncology Service Morales Meseguer University Hospital, Clinical and Translational Oncology Group IMIB-Arrixaca, Murcia, Region of Murcia, Spain. dsanchez@cgcom.es . ORCID: 0000-0003-2073-0679.
9 Psychiatry Service, Castellón Provincial Hospital, Castellón, Spain. ORCID: 0000-0001-9137-7775
*Co-main authors of the study.
**Corresponding author.

Received: 8/2/24; Accepted: 2/20/24; Posted: 2/26/24

Summary:

Introduction: Artificial intelligence (AI) is generating new controversies, opportunities and risks in medical education. This study evaluates the capacity of the artificial intelligence (AI) versions ChatGPT-3 and GPT-4 to answer the questions of the entrance exam to specialized medical training MIR in Spain, comparing performance between the 2022 and 2023 calls.
Methodology: A cross-sectional descriptive study was carried out, using GPT-4 to answer the 210 questions of the MIR 2023 exam, comparing the results with those of ChatGPT-3 in the MIR 2022 exam. Statistical analysis was used to determine the percentage of correctness in depending on the specialty, type of question and its content.
Results: GPT-4 achieved 173 correct answers out of a total of 210 questions, a higher performance than ChatGPT-3, which obtained 108 correct answers in the previous exam session. Notable improvement was seen in specialties such as Rheumatology, Pediatrics, Geriatrics and Oncology, although some fields such as Pulmonology and Ophthalmology showed less progress or even inferior results.
Conclusion: GPT-4 demonstrated better performance compared to ChatGPT-3, indicating advances in AI's processing and analysis of data, as well as its contextual understanding and application of medical knowledge. However, the importance of recognizing the limitations of AI and the need for a critical approach in its use in medical education is emphasized.

Keywords: Artificial Intelligence, ChatGPT-3, GPT4, medical education, MIR, resident doctor

Resumen:

Introducción: La inteligencia artificial (IA) está generando nuevas controversias, oportunidades y riesgos en la educación médica. Este estudio evalúa la capacidad de las versiones de inteligencia artificial (IA) ChatGPT-3 y GPT-4 para responder a las preguntas del examen de acceso a la formación médica especializada MIR en España, comparando el rendimiento entre las convocatorias de 2022 y 2023.

Metodología: Se realizó un estudio descriptivo transversal, utilizando GPT-4 para responder a las 210 preguntas del examen MIR 2023, comparando los resultados con los de ChatGPT-3 en el examen MIR 2022. Se utilizó análisis estadístico para determinar el porcentaje de acierto en función de la especialidad, tipo de pregunta y contenido de la misma.

Resultados: GPT-4 consiguió 173 aciertos de un total de 210 preguntas, rendimiento superior al de ChatGPT-3, que obtuvo 108 aciertos en el examen de la convocatoria anterior. Se observó una mejora notable en especialidades como Reumatología, Pediatría, Geriatría y Oncología, aunque algunos campos como Neumología y Oftalmología mostraron menos progreso o incluso resultados inferiores.

Conclusión: GPT-4 demostró un mejor rendimiento en comparación con ChatGPT-3, indicando avances en el procesamiento y análisis de datos por parte de la IA, así como en su comprensión contextual y aplicación de conocimientos médicos. Sin embargo, se enfatiza la importancia de reconocer las limitaciones de la IA y la necesidad de un enfoque crítico en su uso en educación médica.

Palabras clave: Inteligencia Artificial, ChatGPT-3, GPT4, educación médica, MIR, médico residente

1. Introduction

The rapid evolution of artificial intelligence (AI) in the 21st century has led to notable innovation in numerous fields of knowledge and professional practice, including medical education (1). Large multimodal language models (LLMs), characterized by their ability to learn, adapt and perform complex tasks, are transforming the landscape of teaching and learning in medicine. In this context, natural language processing tools such as ChatGPT (2) are at the center of debate and innovation in the field of learning and teaching in health sciences. This AI model has the ability to interact in human language, provide detailed explanations, and potentially solve questions that require the integration of several levels of analysis, especially in the field of evaluations and examinations of medical knowledge (3).

Thus, these tools have begun to be used to generate educational content of great interest. Despite their brief history, there is already literature on their use as complements and assistants for teaching, personalized learning, quick access to various sources of information, the generation of clinical cases and exam questions, while they can perform translations. almost immediate to different languages (4).

At the same time, multiple risks associated with the use of AI in medical education have been described. On the one hand, they offer us information whose veracity, ethics and professionalism is not verified or reviewed by a professional, which could pose a risk to students and patients (5). In this sense, it is striking that various examples of factual errors, invented content (known as hallucination in the field of AI), as well as gender, racial and political biases have been observed (6-7). On the other hand, some authors warn that it could foster dynamics in which students have fewer incentives to develop and integrate their own reflective processes; causing a dependency on the use of these tools and consequently reducing their learning abilities (8). However, a considerable number of publications show an optimistic attitude regarding the use of this technology as a tool to be implemented in learning processes (9).

One of the applications of tools like ChatGPT that has sparked interest in the scientific and educational community is its application when responding to evaluation tests and exams for medical professionals. One of these examples would be the entrance exam to

Specialized Health Training in Spain, known as the MIR exam. In one of the first studies published in this area (10), it was determined that ChatGPT-3 was capable of passing the MIR exam (11), with approximately 51% correct answers. Subsequently, with GPT-4, higher percentages of correct answers have been observed in exams carried out in Spain (12-13), reaching between 80-90% correct answers.

Despite this, the number of studies and information published is limited, and no study has been published so far with the results of the MIR 2023 exam, held in January 2024.

International literature published in other countries such as the United States (14), Japan (15), China (16), Germany (17) or Italy (18) also suggests that the GPT-4 version offers higher success rates, with the majority of studies being between 70 and 90% correct with the GPT-4 version and between 50 and 70% with the ChatGPT-3 version. In addition to this, studies have been carried out that try to understand the competence of this tool in specific areas or in question formats. Studies have been carried out in specialties such as traumatology (19), radiology (20) or anatomy (21), obtaining very high percentages of success. However, as in the situation described in the Spanish bibliography, studies have only begun to appear 1 year ago and from the analysis of the existing literature we conclude the need to increase our knowledge in this area.

For all of the above, the main objective of this study is to analyze the capacity of GPT-4 to correctly answer the questions of the call for the MIR 2023 exam, held in January 2024. As secondary objectives, it is intended first Firstly, carry out a specific analysis of the response capacity of the tool based on the specialty, content and typology of the different questions and, secondly, carry out a comparison of the ability to answer correctly with the ChatGPT-3 version in the MIR exam. 2022, carried out in January 2023.

2. Methods

A cross-sectional descriptive study was carried out that evaluated the capacity of the tool based on artificial intelligence, GPT-4, to answer the questions of the MIR 2023 exam, comparing its performance with the results of the previous study (MIR 2022 exam) where the ChatGPT model (also known as GPT-3) (2). To do this, the 210 questions were introduced in a standardized way into GPT-4 in blocks of 50 questions.

A separate analysis of each of the image questions was subsequently carried out in order to assign the task of answering the question together with the corresponding visual content. Two databases were created, one with the GPT-4 answers and the other with the official answers published by the Ministry of Health, classifying as correct when a match occurred between the two. Each question on the MIR 2023 exam was classified with the same variable used in the previous study, using specialty type, question type and question content. To do this, the methodology described in previous articles (11) was used. The percentage of correctness of each variable was calculated through a comparison between the performances of the two versions of ChatGPT of each MIR exam. The results of each exam and each tool were compared using radar charts. The statistical analysis was performed with R version 4.3.0 and specialized libraries.

3. Results

The GPT-4 tool was able to correctly answer 173 questions out of a total of 210 questions on the MIR 2023 exam, which is 65 more correct answers than those obtained in the MIR 2022 exam with ChatGPT-3 (108/210) (table 1). In the comparison by specialties (figure 1), GPT-4 showed greater accuracy in most specialties, with a special difference in Rheumatology, Dermatology, Pediatrics and Neurology. However, in some specialties such as

Pulmonology, Maxillofacial and Otorhinolaryngology (ENT), the increase in correct answers was less pronounced. The same success rates were observed in Nephrology, Legal and Ethical Medicine and Intensive Care Unit (ICU) and a worse performance was only observed in Ophthalmology.

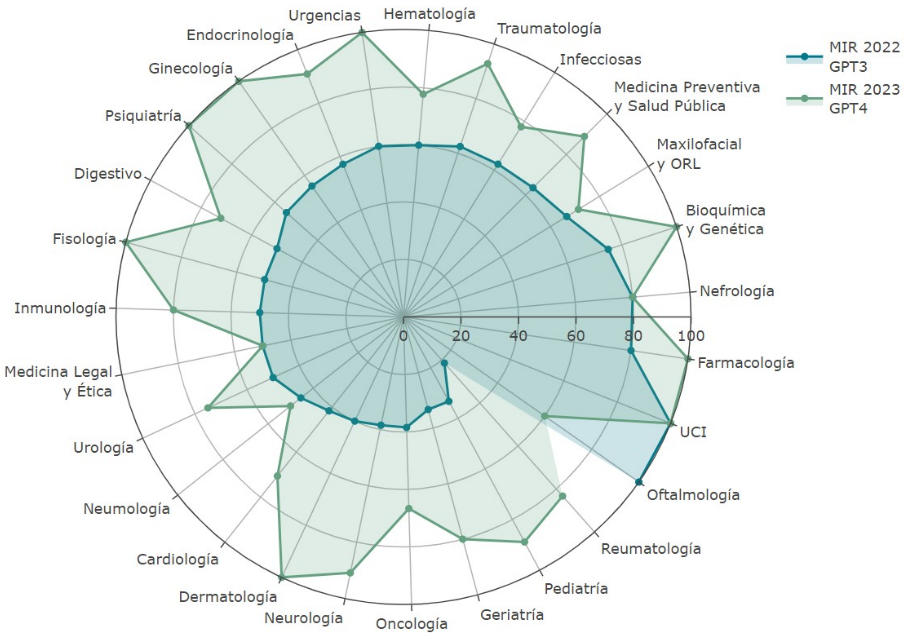

Figure 1. Percentage of correct answers by specialty according to MIR exam and ChatGPT version.

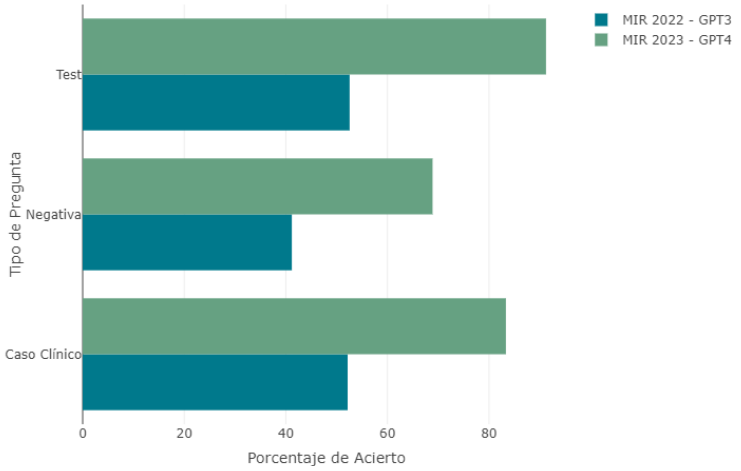

Figure 2. Percentage of correct answers by question type according to MIR exam and ChatGPT version.

Performance by question type (figure 2) showed a substantial improvement with GPT-4 in all question categories, being higher in clinical case type questions and lower in those formulated in a negative format. When comparing the results between the tools by the type of content of the questions (figure 3), GPT-4 outperformed ChatGPT-3 especially in areas such as Pathophysiology, Treatment, Statistics and Diagnosis.

When transposing the number of correct questions to the net result of answers subtracting the proportional percentage of failed questions, GPT-4 obtained 153.3 net correct answers. By including the information in the different MIR position calculators, GPT-4 obtained a position compared to the opponents of the MIR 2023 exam of between 1,100 and 1,300, which would mean a percentile between 90 and 92.

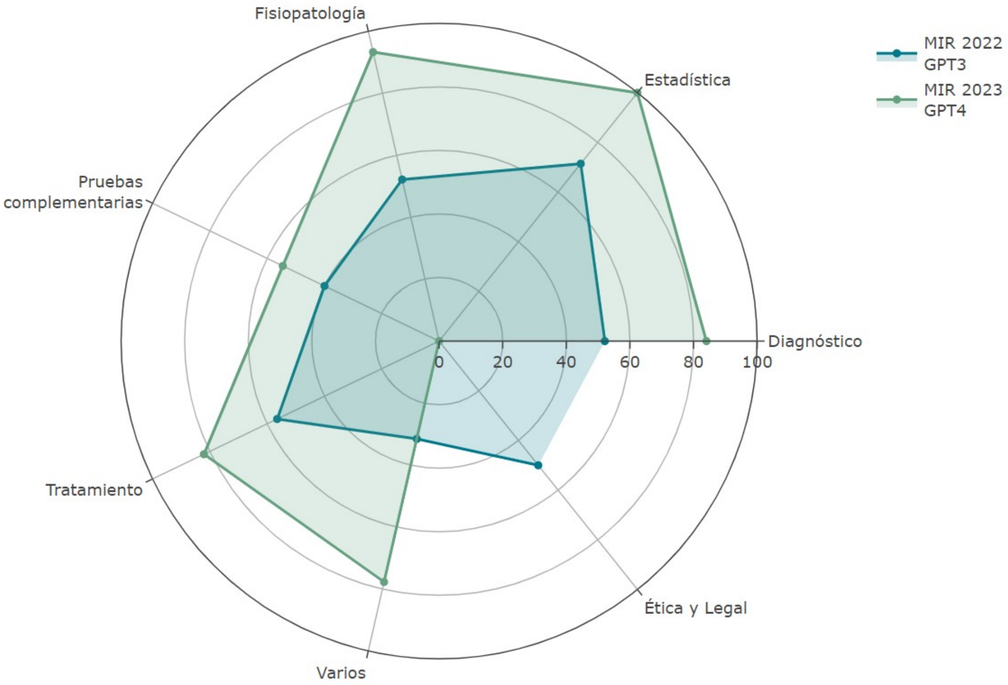

Figure 3. Comparison of percentages of correct answers in each MIR exam by ChatGPT by type of question content.

Table 1 . Total and correct answers by specialty, question type and content according to the MIR exam and ChatGPT version.

	MIR 2022 GPT3		**MIR 2023 GPT4**	
	Total N = 210 [1]	**Correct** N = 108 [1]	**Total** N = 210 [1]	**Correct** N = 173 [1]
Specialty				
Biochemistry and Genetics	4 (1.9%)	3 (2.8%)	3 (1.4%)	3 (1.7%)
Cardiology	12 (5.7%)	5 (4.6%)	17 (8.1%)	12 (6.9%)
Dermatology	5 (2.4%)	2 (1.9%)	8 (3.8%)	8 (4.6%)
Digestive	14 (6.7%)	7 (6.5%)	18 (8.6%)	13 (7.5%)
Endocrinology	14 (6.7%)	8 (7.4%)	11 (5.2%)	10 (5.8%)
Pharmacology	5 (2.4%)	4 (3.7%)	6 (2.9%)	6 (3.5%)
Physiology	4 (1.9%)	2 (1.9%)	5 (2.4%)	5 (2.9%)
Geriatrics	6 (2.9%)	2 (1.9%)	5 (2.4%)	4 (2.3%)
Gynecology	9 (4.3%)	5 (4.6%)	10 (4.8%)	10 (5.8%)
Hematology	5 (2.4%)	3 (2.8%)	9 (4.3%)	7 (4.0%)
Infectious	8 (3.8%)	5 (4.6%)	9 (4.3%)	7 (4.0%)
Immunology	4 (1.9%)	2 (1.9%)	5 (2.4%)	4 (2.3%)
Maxillofacial and ENT	3 (1.4%)	2 (1.9%)	7 (3.3%)	5 (2.9%)
Legal and Ethical Medicine	6 (2.9%)	3 (2.8%)	2 (1.0%)	1 (0.6%)

Preventive Medicine and Public Health	11 (5.2%)	7 (6.5%)	9 (4.3%)	8 (4.6%)
Nephrology	5 (2.4%)	4 (3.7%)	5 (2.4%)	4 (2.3%)
Pneumology	11 (5.2%)	5 (4.6%)	10 (4.8%)	5 (2.9%)
Neurology	13 (6.2%)	5 (4.6%)	11 (5.2%)	10 (5.8%)
Ophthalmology	4 (1.9%)	4 (3.7%)	5 (2.4%)	3 (1.7%)
Oncology	13 (6.2%)	5 (4.6%)	9 (4.3%)	6 (3.5%)
Pediatrics	9 (4.3%)	3 (2.8%)	9 (4.3%)	8 (4.6%)
Psychiatry	11 (5.2%)	6 (5.6%)	7 (3.3%)	7 (4.0%)
Rheumatology	14 (6.7%)	3 (2.8%)	6 (2.9%)	5 (2.9%)
Traumatology	8 (3.8%)	5 (4.6%)	14 (6.7%)	13 (7.5%)
ICU	3 (1.4%)	3 (2.8%)	2 (1.0%)	2 (1.2%)
Emergencies	5 (2.4%)	3 (2.8%)	4 (1.9%)	4 (2.3%)
Urology	4 (1.9%)	2 (1.9%)	4 (1.9%)	3 (1.7%)
Question type				
Clinical case	115 (55%)	60 (56%)	108 (51%)	90 (52%)
Negative	17 (8.1%)	7 (6.5%)	45 (21%)	31 (18%)
Test	78 (37%)	41 (38%)	57 (27%)	52 (30%)
Content				
Diagnosis	71 (34%)	37 (34%)	69 (33%)	58 (34%)
Statistics	7 (3.3%)	5 (4.6%)	5 (2.4%)	5 (2.9%)
Ethics and Legal	6 (2.9%)	3 (2.8%)	1 (0.5%)	0 (0%)
Pathophysiology	23 (11%)	12 (11%)	30 (14%)	28 (16%)
Supplementary tests	15 (7.1%)	6 (5.6%)	11 (5.2%)	6 (3.5%)
Treatment	69 (33%)	39 (36%)	67 (32%)	55 (32%)
Several	19 (9.0%)	6 (5.6%)	27 (13%)	21 (12%)

[1] n (%) (Percentages calculated by columns)

4. Discussion

The results obtained in the MIR 2023 exam by GPT-4 mark a milestone in the evolution of artificial intelligence in the field of medical education. This advancement represents a notable increase in performance, with 65 more hits compared to the previous version, ChatGPT-3, in the MIR 2022 exam (11). This progress is indicative not only of improvements in AI's data processing and analysis capabilities, but also of a refinement in contextual processing and the application of databases and medical knowledge. The variability in performance by medical specialties, with considerable improvements in areas such as Rheumatology, Pediatrics, Geriatrics and Oncology, possibly reflects a greater ability of GPT-4 to integrate and apply complex knowledge in these disciplines. These improvements can be attributed to larger and more diversified data sets, the inclusion of question banks and information sources related to assessments such as the MIR exam, as well as optimized algorithms that allow for deeper analysis of questions and more precise selection. of the responses (9).

The percentage of correct answers was similar to that of articles published in other countries with the GPT-4 version (14-16). This confirms the trend that new versions of AI have a better capacity to correctly answer medical questions and challenges, exceeding a success rate of more than 75% in the vast majority of cases. A significant finding of the present study lies in illustrating what type of questions present the greatest difficulty for GPT-4, which has not been described in previous studies. In this sense, negative questions stand out (formulated from a denial), on complementary tests and those that integrate content from different categories, which are, for the second consecutive year, the questions with the greatest difficulty in being answered correctly by an AI tool (11).

The implications of these advances for medical education are significant. GPT-4 could serve as a complementary tool for teaching and learning, offering medical students an interactive and adaptive way to reinforce their clinical knowledge and skills (3). It also emerges as a tool for trainers, with which to generate questions, clinical cases, tools and learning exercises with which to enrich the training of their students (2).

However, the fascination generated by an AI that passes a MIR exam deserves critical reflection. It is essential to recognize that, although this achievement highlights the evolution and potential of LMMs, it should not be interpreted as a direct comparison to human clinical competence. An exam, by its nature, evaluates knowledge under specific conditions and formats, which differs substantially from the complexity and dynamism and unpredictability of real medical practice. An AI's ability to navigate structured questions and provide data-driven answers contrasts with the clinical judgment, empathy, and decision-making in uncertain contexts that define clinical medicine. This distinction underlines the importance of not transferring AI's ability to process and recognize text patterns to clinical contexts (8).

An essential part of learning in medical education lies in the integration of complex knowledge, pattern recognition, establishing a strong doctor-patient relationship, and ethical and contextual analysis of clinical situations. The use of technological tools, such as AI, that simplify exam solving can create an illusion of competence, potentially disincentivizing students from developing these critical skills. This approach could divert medical training from its fundamental objective: to prepare professionals with a deep and multidimensional understanding of medicine, capable of practicing with empathy, critical judgment and adaptability in the complex clinical environment (4).

The integration of emerging technologies, especially in the field of information and communication, has not yet permeated the training curricula in undergraduate teaching and Specialized Health Training. It is crucial that health professionals, in general and doctors in particular, acquire basic knowledge and skills about how these technologies operate, their benefits, failures and risks in order to develop minimum competence to perform a risk-benefit analysis, understand the responsibility of the professional and the potential harm to third parties when implementing them. This becomes more necessary given the current debate on their use as support tools in clinical decision making, a reality that is already present in some specialties and that is providing enormous help, but that also suggests a risk that must be evaluated. continuously.

Limitations

Given the tremendous speed with which these types of technologies evolve, a limitation of the present study is the short period of time in which the results of GPT-4 are likely to be surpassed by those of other AI tools or by new versions of the same. In this sense, the results of this study need to be contextualized in the time in which they were published. On the other hand, given that the ChatGPT tool has been developed mainly in English, it is possible that the results of our study, having been presented to the tool in the language of the MIR exam (Spanish), underestimate the ability of this tool to correctly medical content questions. However, contrary to this possibility, the results generally coincide with those that have been published in English-speaking countries. Finally, the comparison made between

ChatGPT-3 and GPT-4 corresponds to two different exams respectively, the MIR 2022-2023 exam and the 2023-2024 exam. However, given the representativeness that these examinations show of general medical knowledge and the marked coincidence with the international bibliography in the percentage of correct answers of both tools, we consider that the comparison continues to be valid and interesting for the scientific debate in this area.

5. Conclusions

- The present study reveals that GPT-4 not only outperforms its predecessor ChatGPT-3 in the MIR exam, but also sets a new standard in the ability of AIs to process and analyze highly complex medical information.
- These results underscore the continued evolution and improvement of AI tools in the field of medical education. However, it is crucial to maintain a critical approach and be aware of the inherent limitations of AI, especially in comparison to human clinical competence and decision-making in real medical situations.
- This advance in AI technology opens a promising path towards more effective integration in medical education, enhancing the learning and preparation of future doctors, although always complementing and not replacing human judgment and experience in training processes.

Financing: There has been no financing.

Declaration of conflict of interest: The authors declare that they have no conflict of interest.

Author contributions : ACC wrote the final version of the article and coordinated the project; JJ coordinated and carried out the data analysis and the writing of the final version of the article; JPC has written the first draft and participated in data analysis; DAS has participated in the data analysis and has supervised the final writing of the article; PE has participated in writing the final version and in data analysis; LDLP, EG and JCP have carried out the bibliographic search, participated in the data analysis and in writing the first draft.

References

1. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? Med Educ Online. 2023 Dec; 28(1): 2181052. https://doi.org/10.1080/10872981.2023.2181052
2. OpenAI. Introducing ChatGPT. Available at: https://openai.com/blog/chatgpt/
3. Temsah O, Khan SA, Chaiah Y, Senjab A, Alhasan K, Jamal A, Aljamaan F, Malki KH, Halwani R, Al-Tawfiq JA, Temsah MH, Al-Eyadhy A. Overview of Early ChatGPT's Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts. Cureus. 2023 Apr 8; 15(4): e37281. https://doi.org/10.7759/cureus.37281
4. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, Househ M. The Pros and Cons of Using ChatGPT in Medical Education: A Scoping Review. Stud Health Technol Inform. 2023 Jun 29; 305: 644-647. https://doi.org/10.3233/shti230580
5. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biol Sport. 2023 Apr; 40(2): 615-622. https://doi.org/10.5114/biolsport.2023.125623
6. Davidson T, Bhattacharya D, Weber I. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In Proceedings of the Third Workshop on Abusive Language Online. 2019, pages 25–35, Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/W19-3504/
7. Hadas K, Rikker D, David D. Gender bias and stereotypes in Large Language Models. In Proceedings of The ACM Collective Intelligence Conference (CI '23). Association for Computing Machinery, New York, NY, USA. 2023. 12–24. https://doi.org/10.1145/3582269.3615599
8. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol. 2023 Sep 20; 13(4): 170-178. https://doi.org/10.5662/wjm.v13.i4.170

9. Sahu PK, Benjamin LA, Singh Aswal G, Williams-Persad A. ChatGPT in research and health professions education: challenges, opportunities, and future directions. Postgrad Med J. 2023 Dec 21; 100(1179): 50-55. https://doi.org/10.1093/postmj/qgad090
10. Scaioli G, Lo Moro G, Conrado F, Rosset L, Bert F, Siliquini R. Exploring the potential of ChatGPT for clinical reasoning and decision-making: a cross-sectional study on the Italian Medical Residency Exam. Ann Ist Super Sanita. 2023 Oct-Dec;59(4):267-270. https://doi.org/10.4415/ann_23_04_05
11. Carrasco JP, García E, Sánchez DA, Estrella-Porter P, Puente LDL, Navarro J, et al. Is "ChatGPT" capable of passing the 2022 MIR exam? Implications of artificial intelligence in medical education in Spain. Rev Esp Educ Médica 2023 Feb 16; 4(1). https://revistas.um.es/edumed/article/view/556511/337361
12. Madrid-García, A., Rosales-Rosado, Z., Freites-Nuñez, D. et al. Harnessing ChatGPT and GPT-4 for evaluating the rheumatology questions of the Spanish access exam to specialized medical training. Sci Rep 13, 22129 (2023). https://doi.org/10.1038/s41598-023-49483-6
13. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, Alas-Brun R, Onambele L, Ortega W, Montejo R, Aguinaga-Ontoso E, Barach P, Aguinaga-Ontoso I. Evaluating the Efficacy of ChatGPT in Navigating the Spanish Medical Residency Entrance Examination (MIR): Promising Horizons for AI in Clinical Medicine. Clinics and Practice. 2023; 13(6):1460-1487. https://doi.org/10.3390/clinpract13060130
14. Mihalache A, Huang RS, Popovic MM, Muni RH. GPT-4: An assessment of an upgraded artificial intelligence chatbot in the United States Medical Licensing Examination. MedTeach. 2023 Oct 15:1-7. https://doi.org/10.1080/0142159x.2023.2249588
15. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of ChatGPT.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ 2023;9:e48002. https://doi.org/10.2196/4800
16. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: Pave the way for medical AI. Int J Med Inform. 2023 Sep; 177: 105173. https://doi.org/10.1016/j.ijmedinf.2023.105173
17. Jung LB, Gudera JA, Wiegand TLT, Allmendinger S, Dimitriadis K, Koerte IK. ChatGPT Passes German State Examination in Medicine With Picture Questions Omitted. Dtsch Arztebl Int. 2023 May 30;120(21):373-374. https://doi.org/10.3238/arztebl.m2023.0113
18. Alessandri Bonetti M, Giorgino R, Gallo Afflitto G, De Lorenzi F, Egro FM. How Does ChatGPT Perform on the Italian Residency Admission National Exam Compared to 15,869 Medical Graduates? Ann Biomed Eng. 2023 Jul 25. https://doi.org/10.1007/s10439-023-03318-7
19. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT Performance on the Orthopedic In-Training Examination. JBJS Open Access 8(3): e23.00056, July-September 2023. https://doi.org/10.2106/jbjs.oa.23.00056
20. Kufel J, Paszkiewicz I, Bielówka M, Bartnikowska W, Janik M, Stencel M, Czogalik Ł, Gruszczyńska K, Mielcarska S. Will ChatGPT pass the Polish specialty examination in radiology and diagnostic imaging? Insights into strengths and limitations. Pol J Radiol. 2023 Sep 18;88:e430-e434. https://doi.org/10.5114%2Fpjr.2023.131215
21. Mehrabanian M, Zariat Y. ChatGPT passes anatomy exam. Br Dent J. 2023 Sep; 235(5): 295. https://doi.org/10.1038/s41415-023-6286-7