



ORIGINAL BRIEF

Performance of the ChatGPT tool in solving residency exams

Desempeño de la herramienta ChatGPT en la resolución de exámenes de residencia

Javier Gonzalez-Argote¹  , William Castillo-González^{1,2}  

¹Fundación Salud, Ciencia y Tecnología. Ciudad Autónoma de Buenos Aires, Argentina.

²Universidad de Ciencias Empresariales y Sociales. Ciudad Autónoma de Buenos Aires, Argentina.

Cite as: Gonzalez-Argote J, Castillo-González W. Performance of ChatGPT tool in the resolution of residency exams in Argentina. *Seminars in Medical Writing and Education*. 2024; 3:56. <https://doi.org/10.56294/mw202456>

Submitted: 08-09-2023

Revised: 15-11-2023

Accepted: 20-01-2024

Published: 22-01-2024

Editor: Dr. José Alejandro Rodríguez-Pérez 

*Translated by: Cristhian Alejandro Pérez Pacheco 

ABSTRACT

Introduction: artificial intelligence is classified as a tool of interest at the present time. Through its application, organizational processes and decision-making are transformed; while promoting innovative development.

Objective: to describe the performance of the ChatGPT tool in solving residency exams.

Method: an observational, descriptive, retrospective study was carried out. As the universe of the study, all test-type questions (single selection) were analyzed. The variables analyzed were each of the questions belonging to each exam (correct or incorrect answers). Descriptive statistics were applied.

Results: Syllabus B stood out within group 1 with the highest number of correct answers (209 for 69,66 %). For its part, within group 2, syllabus D was predominant with 141 correct answers (70,5 %). The exams related to nursing stood out.

Conclusions: the use of artificial intelligence tools such as ChatGPT is variable in the field of medical sciences. Their performance in solving scientific questions is heterogeneous. It may vary with respect to the format of the question and the topic addressed.

Keywords: Science, Technology and Society; Medical Education; Professional education; Artificial intelligence; Technology.

RESUMEN

Introducción: la inteligencia artificial se cataloga como una herramienta de interés en los momentos actuales. Mediante su aplicación se transforman los procesos organizativos y toma de decisiones; a la par que fomenta el desarrollo innovador.

Objetivo: describir el desempeño de la herramienta ChatGPT en la resolución de exámenes de residencia.

Método: se realizó un estudio observacional, descriptivo, retrospectivo. Como universo del estudio se analizaron todas las preguntas de tipo test (selección única). Las variables analizadas fueron cada una de las preguntas pertenecientes a cada examen (respuestas correctas o incorrectas). Se aplicó la estadística descriptiva.

Resultados: Sobresalió el temario B dentro del grupo 1 con el mayor número de acierto (209 para un 69,66 %). Por su parte, dentro del grupo 2 resulto predominante el temario D con 141 aciertos (70,5 %). Destacaron los exámenes referentes a enfermería.

Conclusiones: la utilización de las herramientas de inteligencia artificial como ChatGPT es variable en el ámbito de las ciencias médicas. Su desempeño en la resolución de interrogantes científicas es heterogéneo. Puede variar con respecto al formato de la interrogante y la temática abordada.

Palabras clave: Ciencia, Tecnología y Sociedad; Educación Médica; Educación Profesional; Inteligencia Artificial; Tecnología.

INTRODUCTION

The scientific and technological advancement of humanity has found ways and means for its standardization and widespread application across every sector of society. From automated processes in large industries, or the necessary (and indisputable) use of information and communication technologies, to the most sophisticated diagnostic and therapeutic procedures and means (such as biotechnology therapies, nanotechnologies, among others) are present in the daily work of humankind.

Artificial Intelligence (AI) is classified as a noteworthy tool at the present time. Its application leads to the transformation of organizational processes and decision-making; while fostering innovative development. Moreover, it is a service within everyone's grasp.^(1,2) In this context, the ChatGPT (Generative Pre-training Transformer) tool emerges as a generative artificial intelligence system, constructed upon a foundation of over 175 million parameters, and mastering information from more than 8 million documents and sources, which enables its capability to generate coherent responses.^(3,4)

The application of these tools in the university context is immeasurable. They have the potential to enhance the learning environment for students, particularly in their interaction with virtual spaces. Additionally, they expedite the acquisition of information.^(5,6) However, it is imperative to continuously assess the quality of processes to prevent academic errors or other issues that could impact the effectiveness of the educational teaching process.⁽⁷⁾

In the field of health sciences, the performance of AI in this sector has been documented. One approach has concentrated on evaluating its functionality in theoretical exams in medical sciences, yielding varied results in each study, although the average accuracy of responses is found between 50 % and 70 % correct answers.⁽⁷⁾ Moreover, assessments have been made on aspects regarding its application in the enhancement of diagnostic methods, especially in imaging.⁽⁸⁾ Furthermore, in the realm of scientific research within this sector, its application is advocated for the optimization of time in bibliographic searches and data analysis processing; simultaneously, it is imperative to monitor potential biases in the analysis and protection of information which could result in moral implications.⁽⁹⁾

Undoubtedly, technological advancement is evident in every sector of society, particularly in health sciences.⁽¹⁰⁾ The diversity of applications of AI underscore the need for systematic controls to consistently analyze its benefits and prevent potential complications. Hence, the objective of this study is to describe the performance of the ChatGPT tool in solving residency exams.

METHODS

An observational, descriptive, retrospective study was conducted to assess the performance of ChatGPT in solving residency exams from the year 2022. As the universe of the study, all test-type questions (single selection) were analyzed; no sampling techniques were employed, so the analysis encompassed the entirety of the universe. The variables analyzed were each of the questions belonging to each exam (correct or incorrect answers).

To collect information, the ChatGPT tool was employed using the following question: "Can you answer the following multiple-choice questions about medicine with solely the correct items?" Subsequently, exams pertaining to medicine, nursing, biochemistry, and mathematics (divided into four syllabuses) were administered. The analysis focused solely on the correct answers in relation to the total. Descriptive statistics were applied.

Ethical standards for research development in health sciences and the II Declaration of Helsinki were adhered to throughout the study.

RESULTS

Syllabus B stood out within group 1 with the highest number of correct answers (209 for 69,66 %). Similarly, nursing exams were distinctive within the same group, registering 147 correct answers (24,5 %), compared to the rest. For its part, within group 2, syllabus D was predominant with 141 correct answers (70,5 %); with a higher representation from nursing (142 correct answers; 35,5 %).

DISCUSSION

The implementation of new technologies across various branches of science enhances productive processes.⁽¹¹⁾ Similarly, it creates room for an expanding debate regarding their potential limitations or implications in their utilization.

Table 1. Distribution of responses according to the syllabuses and specialties

Specialty	Grupe 1				Grupe 2				T
	Syllabus A		Syllabus B		Syllabus C		Syllabus D		
	Correct answers	%	Correct answers	%	Correct answers	%	Correct answers	%	
Biochemistry	67/100	67	69/100	69	-	-	-	-	136
Nursing	72/100	72	75/100	75	68/100	68	74/100	74	289
Medicine	68/100	68	65/100	65	68/100	68	67/100	67	268
Total	207/300	69	209/300	69,66	136/200	68	141/200	70,5	693

Concerning health sciences and their diverse facets (care, education, and scientific-research), the performance of artificial intelligence, particularly in the specific case of the ChatGPT tool, is both variable and wide. Authors such as Castillo-González⁽¹²⁾ acknowledge its utility across different stages of the editorial process, especially through the correction of writing styles of articles to enhance scientific coherence. However, they underscore the vital role of human creativity as a supervisor in the editorial and investigative processes within medical sciences. This perspective finds support from Vega-Jiménez et al.⁽¹³⁾, who underscore the significance of authors, signatories of investigations, declaring the use of any AI tool in research preparation to prevent potential future conflicts related to authorship and content of the works.

Concerning undergraduate and postgraduate medical education, digital content-generating tools like ChatGPT contribute to the didactics of learning through the optimization of time and quick and effective access to necessary information.⁽¹⁴⁾ Regarding the results presented, Carrasco et al.⁽¹⁵⁾ demonstrate congruent findings regarding the percentages of correct answers per exam. In turn, these authors highlight in their study that questions analyzing multiple elements for their response tend to accumulate a higher percentage of errors. Additionally, Alfertshofer et al.⁽¹⁶⁾ concur with the results of the current study by analyzing the tool's performance in similar exams across different countries, with an average of correct answers ranging from 22 % for exams administered in France to 73 % for those in Italy.

These results can serve as a foundation for the implementation of protocols in different academic institutions, aiming to promote the use of artificial intelligence tools solely for the improvement of educational processes, with the objective of preserving the integrity of the educational teaching process. These criteria align with those presented by Vega-Jiménez et al.⁽¹⁷⁾

Furthermore, it is worth highlighting that the medical care process involves various facets to reach a medical diagnosis. This process integrates theoretical knowledge (demonstrated through several theoretical-practical exams throughout years of training) and practical skills such as listening, feeling, and the ability to effectively communicate with the patient. These skills are key elements of the appropriate questioning and physical examination that medical personnel must conduct; supplemented by the experiential insights of healthcare professionals. Likewise, the execution of specific diagnostic procedures (based on the suspected condition) is imperative to attain an accurate clinical diagnosis.

The saying “there are no diseases, only sick people” supports these criteria and emphasizes the analysis of the patient as a biopsychosocial being on whom multiple processes interact, not only the acute condition but also aggravating factors, triggers, among others; the concept of health is thus supported as a state of physical, mental, and social well-being, not merely the absence of disease.⁽¹⁸⁾ These aspects must be taken into consideration for the diagnosis of different pathologies and their subsequent therapeutic behavior. Therefore, the utilization of these tools should be viewed as auxiliary means for diagnosis rather than the principal element. Similar criteria are presented by Gutiérrez-Cirlos et al.⁽¹⁹⁾

CONCLUSIONS

The utilization of artificial intelligence tools such as ChatGPT is variable within the field of medical sciences. Its performance in solving scientific questions is heterogeneous. It may vary with respect to the format of the question and the topic addressed.

REFERENCES

1. Castillo-González W. The importance of human supervision in the use of ChatGPT as a support tool in scientific writing. *Metaverse Basic and Applied Research* 2023;2:29-29. <https://doi.org/10.56294/mr202329>.
2. Castillo-Gonzalez W. ChatGPT and the future of scientific communication. *Metaverse Basic and Applied Research* 2022;1:8-8. <https://doi.org/10.56294/mr20228>.
3. González LN. EL IMPACTO DE LA INTELIGENCIA ARTIFICIAL EN LOS NEGOCIOS. *Difusiones* 2023;25:153-61.

4. López KMG. Inteligencia artificial generativa: Irrupción y desafíos. *Enfoques* 2023;4:57-82.
5. Espinosa RDC, Caicedo-Erazo JC, Londoño MA, Pitre IJ. Inclusive Innovation through Arduino Embedded Systems and ChatGPT. *Metaverse Basic and Applied Research* 2023;2:52-52. <https://doi.org/10.56294/mr202352>.
6. Ferrer-Benítez M. Online dispute resolution: can we leave the initial decision to Large Language Models (LLM)? *Metaverse Basic and Applied Research* 2022;1:23-23. <https://doi.org/10.56294/mr202223>.
7. Toro-Espinoza MF, Montalván-Espinoza JA, Masabanda-Vaca MA. Aplicación de la inteligencia artificial en el aprendizaje universitario. *Reicomunicar* 2023;6:153-72. <https://doi.org/10.46296/rc.v6i12edespoct.0168>.
8. Ruibal-Tavares E, Calleja-López JR, Rivera-Rosas CN, Aguilera-Duarte LJ. Inteligencia artificial en medicina: panorama actual. *REMUS* 2023. <https://doi.org/10.59420/remus.10.2023.178>.
9. T S, Arumugam T, Pandurangan H, Panjaiyan K. Adopción de la Inteligencia Artificial en la Atención Sanitaria: Una perspectiva enfermera. *Salud, Ciencia y Tecnología* 2023;3:510. <https://doi.org/10.56294/saludcyt2023510>.
10. Cano CAG, Castillo VS, Gallego TAC. Unveiling the Thematic Landscape of Generative Pre-trained Transformer (GPT) Through Bibliometric Analysis. *Metaverse Basic and Applied Research* 2023;2:33-33. <https://doi.org/10.56294/mr202333>.
11. Luna GJJ. Study on the impact of artificial intelligence tools in the development of university classes at the school of communication of the Universidad Nacional José Faustino Sánchez Carrión. *Metaverse Basic and Applied Research* 2023;2:51-51. <https://doi.org/10.56294/mr202351>.
12. Castillo-González W, Lepez CO, Bonardi MC. Chat GPT: a promising tool for academic editing. *Data and Metadata* 2022;1:23. <https://doi.org/10.56294/dm202223>.
13. Jiménez JV, Leyva LLL, Leon AM. ChatGPT e inteligencia artificial, señal de alerta para el proceso editorial de revistas médicas. *Revista Cubana de Información en Ciencias de la Salud* 2023;34.
14. Ledo MJV, Olite FMD, Vera IA, Suárez I del RM, Domínguez AMA, Pedro JYP. Chat en la educación médica. *Educación Médica Superior* 2023;37.
15. Carrasco JP, García E, Sánchez DA, Porter E, De La Puente L, Navarro J, et al. ¿Es capaz “ChatGPT” de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. *Rev Esp Edu Med* 2023;4. <https://doi.org/10.6018/edumed.556511>.
16. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the Seven Seas: A Multinational Comparison of ChatGPT’s Performance on Medical Licensing Examinations. *Ann Biomed Eng* 2023. <https://doi.org/10.1007/s10439-023-03338-3>.
17. Jiménez JV, Gomez EEB, Álvarez PJR. ChatGPT e inteligencia artificial: ¿obstáculo o ventaja para la educación médica superior? *Educación Médica Superior* 2023;37.
18. Colectivo de autores. *Medicina general integral*. Tomo I. Salud y medicina. vol. Vol 1. 4ta ed. La Habana: Editorial Ciencias Medicas; 2022.
19. Gutiérrez-Cirlos C, Carrillo-Pérez DL, Bermúdez-González JL, Hidrogo-Montemayor I, Carrillo-Esper R, Sánchez-Mendiola M. ChatGPT: oportunidades y riesgos en la asistencia, docencia e investigación médica. *GMM* 2023;159:11757. <https://doi.org/10.24875/GMM.230001671>.

FUNDINGS

None.

CONFLICT OF INTEREST

None.

AUTORSHIP CONTRIBUTION

Conceptualization: Javier González Argote.

Research: Javier González Argote, William Castillo-González.

Methodology: Javier González Argote.

Original drafting: Javier González Argote, William Castillo-González.

Proofreading and editing: Javier González Argote, William Castillo-González.