Ciencia y Salud, Vol. 8, No. 2, abril-junio, 2024 • ISSN (impreso): 0378-7680 • ISSN (en línea): 2613-8751
DOI: https://doi.org/10.22206/cysa.2024.v8i2.2828

EL DESEMPEÑO DEL ChatGPT EN LA RESOLUCIÓN DE UN EXAMEN DE RESIDENCIA MÉDICA: ¿UN INDICADOR DE LA EVOLUCIÓN DE INTELIGENCIA ARTIFICIAL EN EDUCACIÓN MÉDICA?

The performance of ChatGPT in solving a medical residency exam: an indicator of the evolution of artificial intelligence in medical education?

Alexander Valdez Disla[1], Vahid Nouri Kandany[2], Pascual Valdez[3]

Recibido: 18 de mayo, 2023 ● Aprobado: 28 de febrero, 2024

Cómo citar: Valdez Disla A., Nouri Kandany V., & Valdez, P. (2024). El desempeño de ChatGPT en la resolución de un examen de residencia médica: ¿un indicador de la evolución de inteligencia artificial en educación médica? *Ciencia y Salud*, *8*(2), 47-55. https://doi.org/10.22206/cysa.2024.v8i2.2828

Resumen

Introducción: ChatGPT (Generative Pre-trained Transformeres) una herramienta de procesamiento de lenguaje natural desarrollada por OpenAI que utiliza el modelo de lenguaje GPT para generar respuestas similares al lenguaje humano natural. Esta tecnología ha demostrado su capacidad para completar tareas complejas y ha atraído la atención en el ámbito educativo, especialmente en la medicina. El objetivo de este estudio es evaluar el desempeño de ChatGPT en la resolución de preguntas del examen de residencia médica para optar por una especialidad (ENURM) en la República Dominicana en 2023.
Métodos: Se ingresaron las 100 preguntas del examen ENURM de 2023 en formato de preguntas de selección múltiple en ChatGPT 3.5, con la instrucción de "seleccionar la respuesta correcta a la siguiente pregunta del examen ENURM 2023". Se realizó un estudio descrip-

Abstract

Introduction: ChatGPT (Generative Pre-trained Transformers) is a natural language processing tool developed by OpenAI that utilizes the GPT language model to generate human-like natural language responses. This technology has proven its capability in completing complex tasks and has garnered attention in the educational field, especially in medicine. The aim of this study is to evaluate the performance of ChatGPT in solving questions from the medical residency exam to opt for a specialty (ENURM) in the Dominican Republic in 2023.
Methods: The 100 questions from the 2023 ENURM exam in multiple-choice question format were entered into ChatGPT 3.5, with the instruction to "select the correct answer to the following ENURM 2023 exam

[1] Universidad Autónoma de Santo Domingo, Instituto de Investigación en Salud. Medico Nutriologo. ORCID: https://orcid.org/0000-0002-4075-155X, email: avaldez68@uasd.edu.do
[2] Universidad Autónoma de Santo Domingo, Centro de investigación biomédica y clínica (CINBIOCLI). Médico Internista. ORCID: https://orcid.org/0000-0001-6361-5529, email: vnouri66@uasd.edu.do
[3] Hospital Vélez Sarsfield, Buenos Aires, Argentina, Médico Internista. ORCID: https://orcid.org/0000-0002-4309-5420, email: pvaldez@fmed.uba.ar

Esta obra está bajo una licencia internacional Creative Commons Atribución-NoComercial-CompartirIgual 4.0. CC BY-NC-SA

tivo transversal para evaluar el desempeño de la herramienta.
Resultados: ChatGPT logró una precisión del 77% en las respuestas proporcionadas, mientras que el 23% de las preguntas no fueron respondidas correctamente. Al desglosar el rendimiento por tipo de pregunta, ChatGPT mostró una eficacia del 74.6% en preguntas directas y del 88.2% en casos clínicos. las especialidades en las cuales se identificaron respuestas incorrectas incluyen hematología, gastroenterología, cardiología, anatomía, genética, cirugía, pediatría, ginecología e infectología. A pesar de estas limitaciones, es relevante destacar que el desempeño de ChatGPT superó el promedio general de los aspirantes a residencias médicas en términos de precisión de respuestas.
Conclusiones: ChatGPT demostró un buen desempeño en la respuesta a preguntas de examen ENURM. Esta herramienta puede ser útil para el procesamiento del lenguaje natural en la educación médica aún con sus limitaciones y no puede reemplazar la enseñanza tradicional y la experiencia clínica.

Palabras clave: ChatGPT, inteligencia artificial, educación médica, lenguaje natural, examen de residencia médica.

question." A cross-sectional descriptive study was conducted to assess the tool's performance.
Results: ChatGPT achieved a 77% accuracy in the responses provided, while 23% of the questions were not answered correctly. When breaking down performance by question type, ChatGPT showed an effectiveness of 74.6% in direct questions and 88.2% in clinical cases. The specialties in which incorrect answers were identified include hematology, gastroenterology, cardiology, anatomy, genetics, surgery, pediatrics, gynecology, and infectious diseases. Despite these limitations, it is relevant to highlight that ChatGPT's performance exceeded the overall average of medical residency applicants in terms of response accuracy.
Conclusions: ChatGPT demonstrated good performance in answering ENURM exam questions. This tool can be useful for natural language processing in medical education despite its limitations and cannot replace traditional teaching and clinical experience.

Keywords: ChatGPT, artificial intelligence, medical education, natural language, medical residency exam.

Introducción

La educación médica es el pilar fundamental en la formación de profesionales de la salud, abarcando desde conocimientos teóricos hasta habilidades prácticas. Integrando un enfoque holístico y tecnologías avanzadas, incluida la Inteligencia Artificial (IA) generativa, busca formar profesionales capaces de ofrecer una atención centrada en el paciente y preparados para los desafíos futuros[1].

La IA mejora el aprendizaje mediante simulaciones clínicas realistas y personalización del contenido educativo, promoviendo una formación médica más efectiva y eficiente[2].

El ChatGPT (Chat Generative Pre-trained Transformer) es una herramienta pública desarrollada por OpenAI basada en la tecnología del modelo de lenguaje GPT[3], OpenAI es un laboratorio de investigación que se ha destacado por su rápido progreso en el desarrollo de tecnologías de IA[4].

ChatGPT emplea su gran capacidad de almacenamiento de datos y su diseño eficiente para entender y dar sentido a las solicitudes de los usuarios, generando respuestas adecuadas que se asemejan al lenguaje humano natural. Su habilidad para generar lenguaje similar al humano y completar tareas complejas, lo convierte en una innovación destacada en el campo de la IA y el procesamiento de lenguaje natura[5–7].

Desde su introducción, ChatGPT ha tenido una gran aceptación en generar respuestas automáticas y tareas complejas, incluyendo resúmenes, poesía, programación, problemas matemáticos, sugerencias metodologías y estadísticas. Además, el

procesamiento del lenguaje natural, como el utilizado en ChatGPT, está atrayendo la atención en ámbito de la educación en especial la médica[8].

La Asociación Médica Mundial y el Comité Permanente de Médicos Europeos recomiendan revisar los planes de estudio y oportunidades educativas para incluir la comprensión de diferentes aspectos de IA en la educación médica, así como en la atención sanitaria[9].

Diferentes estudios confirman la capacidad de ChatGPT en responder preguntas de exámenes en el área de medicina sin ser una herramienta especializada para tales fines.

Por ejemplo, en el caso de exámenes de licencia USMLE Step 1 y Step 2, más del 60% de las preguntas fueron respondidas por chatGPT[10].

Estos resultados por parte de ChatGPT subraya su potencial disruptivo en la educación médica, evidenciando la capacidad de la inteligencia artificial para comprender y aplicar conocimientos médicos complejos.

En otros estudios, en preguntas más específicas sobre parasitología y oftalmología, el porcentaje de respuestas acertadas alcanzaron a un 50% y 60% respectivamente[11].

El Examen Nacional Único de Residencias Médicas (ENURM) se realiza anualmente en el mes de marzo en la República Dominicana. La presente investigación tiene como objetivo evaluar el desempeño de ChatGPT para generar respuestas precisas y coherentes a todas las preguntas de (ENURM) en la República Dominicana en el año 2023. Este estudio comparará el rendimiento de ChatGPT con la respuesta ofrecida por el comité organizador de la prueba, la cual se publicó inmediatamente después del examen[12].

Material y método

Se ha hecho un estudio de corte transversal y descriptivo para evaluar el desempeño de ChatGPT en la resolución de preguntas del examen ENURM de 2023 en la República Dominicana. La redacción de este artículo se basó en los criterios de la guía STROBE para mejorar la calidad y transparencia en la presentación de estudios observacionales en epidemiología.

Para el análisis estadístico de los datos, se aplicaron métodos descriptivos. Las tasas de acierto de ChatGPT se calcularon como el porcentaje de respuestas correctas sobre el total de preguntas introducidas y se compararon con las respuestas oficiales del ENURM. Se empleó el software estadístico SPSS versión 25.0 para realizar análisis de frecuencias y medir la media de las puntuaciones obtenidas por ChatGPT, diferenciadas por categorías y grupos de especialidades.

Con la indicación: "¿seleccionar la respuesta correcta a la siguiente pregunta del examen ENRUM 2023?". Las preguntas del examen fueron ingresadas secuencialmente en ChatGPT, manteniendo el orden original. Las respuestas obtenidas se documentaron primero en una hoja en Word, asignando la letra correspondiente a la respuesta correcta (A, B, C, D). Luego, tanto las respuestas generadas por ChatGPT como las oficiales se catalogaron en hojas distintas de un fichero de Excel para facilitar la comparación. Los datos en hoja de Excel fueron utilizados para realizar procedimiento estadísticos.

Cabe destacar que no había figuras ni gráficas en el examen. Las preguntas se introdujeron en el ChatGPT 3.5 en la fecha 08/03/2023 a las 5:00 pm. Se estableció como criterio comparativo las respuestas oficiales proporcionadas por el comité organizador del ENURM para evaluar el desempeño de ChatGPT. Se optó por este enfoque para asegurar una

evaluación objetiva y precisa del rendimiento de ChatGPT, comparando sus respuestas con el estándar oficial y reconocido del examen ENURM de 2023[12].

Para mejor análisis de los resultados las preguntas fueron clasificadas de las siguientes maneras:

1. En 21 (categorías Cirugía, Ginecobstetricia, Pediatría, Infectología, Hematología, Cardiología, Anatomía, Genética, Gastroenterología, Nefrología, Neurología, Farmacología, Traumatología, Embriología, Salud Pública, Endocrinología, Dermatología, Fisiología, Neumología, Anestesiología y Oncología (La especialidad de oncología se incluyó al grupo quirúrgico, y la anestesiología y epidemiología al grupo clínico- comunitario).

2. En cinco grupos: Quirúrgica, Clínico-Comunitaria, Ginecológica, Pediátrica y Ciencias Básicas. La especialidad de Oncología fue incluida en el grupo Quirúrgico, mientras que Anestesiología y Epidemiología fueron incluidas en el grupo Clínico-Comunitario.

3. La inclusión de oncología en el grupo quirúrgico se justifica por su enfoque en intervenciones quirúrgicas para el tratamiento del cáncer, mientras que anestesiología y epidemiología se agruparon en el sector clínico-comunitario debido a su impacto en la salud poblacional y el manejo clínico.

Las preguntas de casos clínicos en el examen de ENRUM se centran en la aplicación de conocimientos y habilidades de razonamiento clínico en escenarios que simulan situaciones reales de atención médica, exigiendo diagnóstico y manejo de pacientes. En contraste, las preguntas directas apuntan a evaluar el conocimiento teórico del estudiante sobre conceptos médicos sin necesidad de análisis clínico, a través de formatos como selección múltiple.

Se establecieron protocolos específicos para abordar situaciones en las que ChatGPT proporcionara múltiples respuestas o dejara preguntas sin responder en el contexto del ENURM. La emisión de varias opciones por parte de ChatGPT se clasificó sistemáticamente como incorrecta, reflejando el requisito de una única respuesta correcta por pregunta. Asimismo, la omisión de una selección se interpretó como una falta, evidenciando la limitación de ChatGPT para identificar una respuesta precisa dentro del ámbito del examen.

En este estudio no hubo interacciones previas entre los autores y ChatGPT antes del estudio, aunque la precisión de las respuestas de ChatGPT puede estar influenciada por su entrenamiento previo.

Resultados

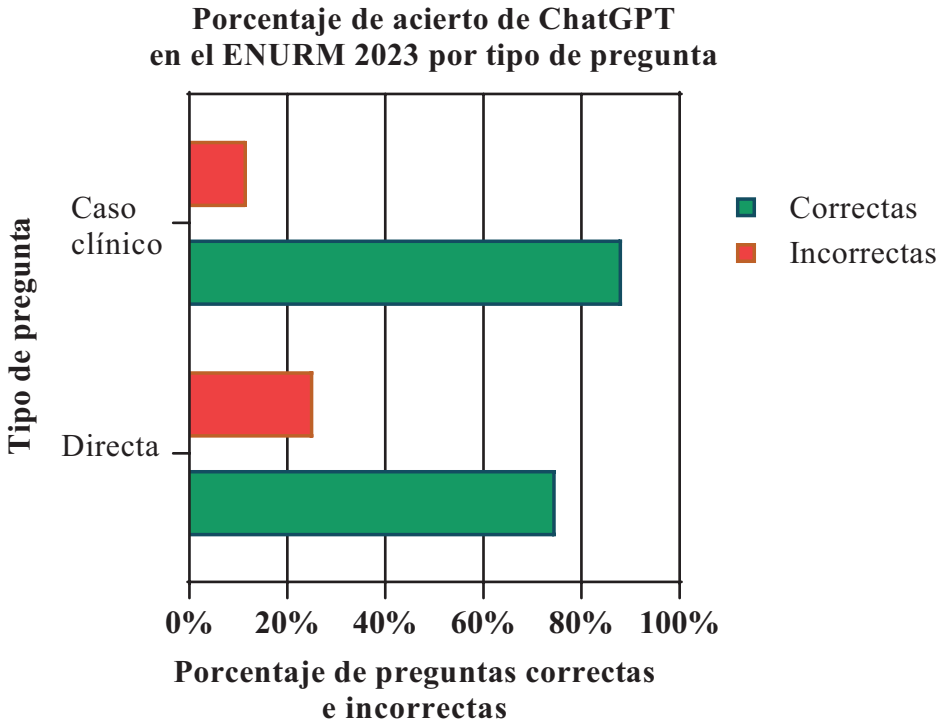
El examen ENURM constaba de 100 preguntas, de las cuales el ChatGPT contesto 77% correctamente y 23% incorrectamente, según las plantillas ofrecida por organizadores.

El 83% de las preguntas fueron de tipo directas y 17% en formato de casos clínicos.

El 74.6% (62/83) de las preguntas directas y el 88.2% (15/17) de los casos clínicos fueron respondidos correctamente, según la plantilla publicada (Grafica 1).

Entre 21 categorías diferentes de especialidades, 11 categorías presentaron todas las respuestas correctas (Anestesiología, Neumología, Fisiología, Dermatología, Endocrinología, Salud Pública, Embriología, Traumatología, farmacología, Neurología, Nefrología). Las especialidades que registraron un mayor porcentaje de respuestas incorrectas, en orden descendente fueron: Hematología 100% (4/4), gastroenterología 57.1% (4/7), cardiología 50% (2/4), anatomía 25% (1/5), genética 25% (1/5), cirugía 20% (2/10), pediatría

Gráfica 1. Acierto de ChatGPT según tipo de preguntas. El total de preguntas en el examen fue de 100



20% (3/15), ginecología 17.6% (3/17), infectología 14.2% (1/7) (Grafica 2).

En relación con 5 agrupaciones de especialidades, las respuestas incorrectas tienen las siguientes distribuciones: Clínico y comunitario 36.3% (12/33), pediatría 20% (3/15), ginecología 17.6% (3/17), quirúrgicas, 17.6% (3/17) y ciencias básicas 11.1% (2/18) (Grafica 3).

Discusión

El objetivo de este estudio fue evaluar la capacidad de ChatGPT para responder correctamente las respuestas del examen ENURM.

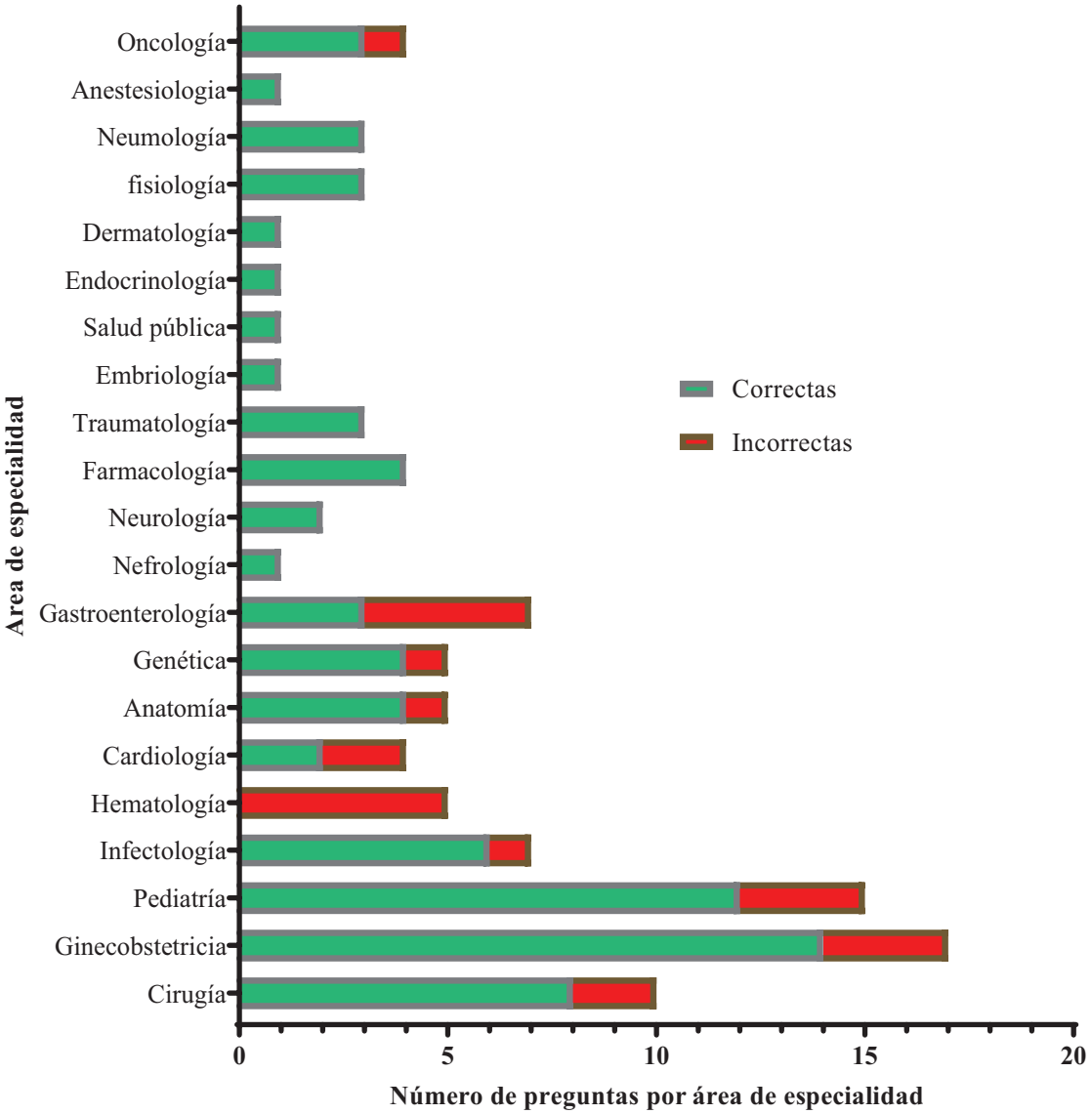
Un total de 5,031 aspirantes a cursar especialidades participaron en el Examen Nacional Único para Aspirantes a Residencias Médicas (ENURM 2023), de los cuales 3,588 correspondieron al sexo femenino, para un 71.31%; mientras que, los de sexo masculino alcanzaron la cifra de 1,443, representando el 28.69 % de los participantes.

La media de notas fue 53.4 puntos para ambos sexos, con un máximo de 85 y mínimo de 19. La media de puntuación por sexo fueron 54.20 para los hombres y 53.12 para las mujeres.

Los resultados muestran que ChatGPT acertó el 77% de las preguntas, a pesar de no haber sido entrenado previamente. Estos resultados superan la media de la nota por los aspirantes a plazas de residencias médicas, que fue un 53.4%.

Este modelo de IA ha sido capaz de demostrar un desempeño aceptable en exámenes de medicina más complejos, como los del *United States Medical Licensing Exam* (USMLE) donde fue capaz de aprobar con aproximadamente el 60%, algo menor que el examen de ENURM con un 77%[13,14].

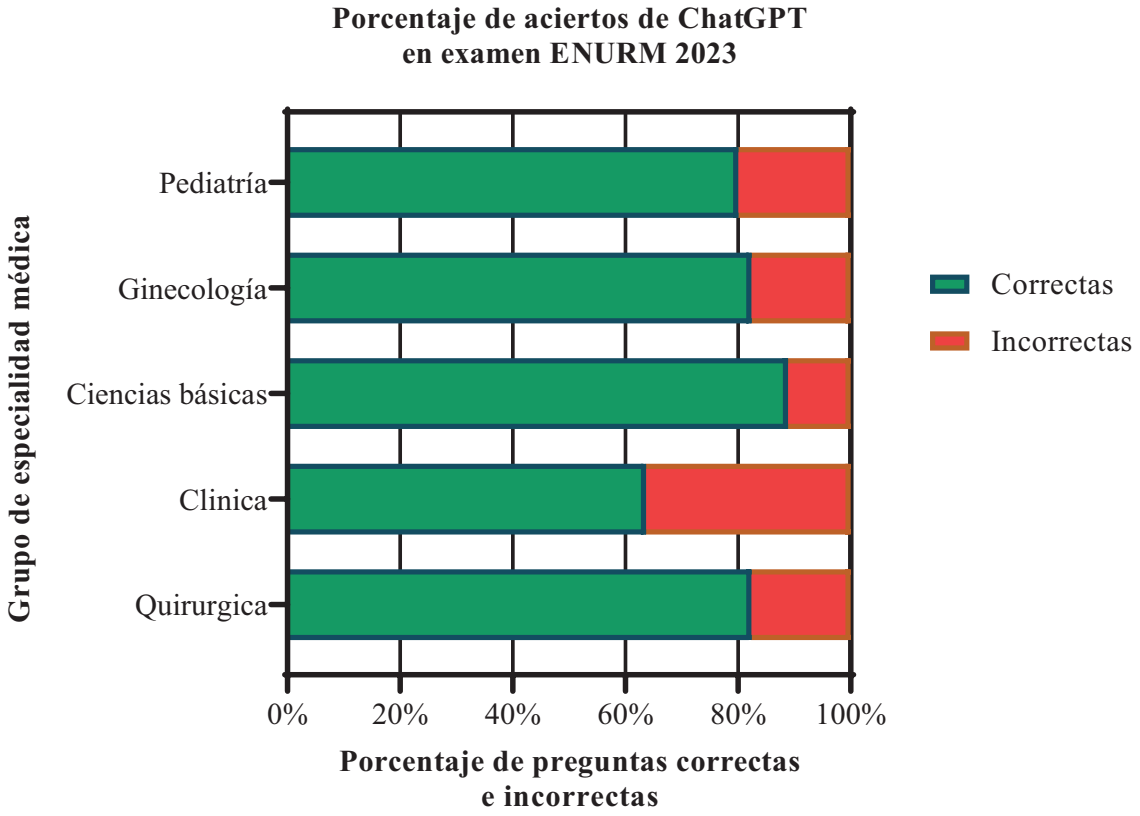
Gráfica 2. Distribución del total de preguntas por área de especialidad y acierto de ChatGPT



Es importante enfatizar que el examen USMLE se caracteriza por su elevada complejidad, la cual no se limita únicamente al volumen de preguntas, sino que se extiende a su formato particular. Dicho formato abarca preguntas con gráficos y figuras, introduciendo así un nivel de dificultad adicional tanto para los candidatos como para sistemas basados en inteligencia artificial, como ChatGPT.

En otro estudio realizado en Corea sobre el conocimiento y la capacidad de interpretación de ChatGPT comparado con la de estudiantes de medicina mediante la administración de un examen de parasitología con 79 preguntas[11], la tasa de respuestas correctas de ChatGPT fue un 60,8% y no necesariamente fue un indicador de un mal desempeño, ya que las preguntas no eran fáciles de responder correctamente para los estudiantes de medicina.

Gráfica 3. Porcentaje de acierto del ChatGPT según especialidad médica



El puntaje promedio considerablemente más alto (89,6%) de los estudiantes de medicina puede haberse debido a su aprendizaje previo en parasitología y al hecho de que el examen se administró 4 días después de la clase[11]. En comparación con nuestro estudio (77% Vs 60.8%) con una mayor tasa de respuesta correctas pude ser que en el examen de ENURM no se incluyó figuras ni tablas, lo que limita la capacidad de análisis de ChatGP en este ámbito.

En otro estudio realizado en Canadá, se examinó la precisión de ChatGPT en el área de oftalmología utilizando dos bancos de preguntas de opción múltiple populares que se emplean en el examen de Evaluación del Conocimiento Oftálmico (OKAP) de alto riesgo. Los resultados obtenidos por ChatGPT marcaron una precisión del 55,8% y del 42,7% en los dos exámenes simulados de 260 preguntas[14].

Estos resultados son significativamente inferiores a los obtenidos en nuestro estudio. Sin embargo, es posible que esta diferencia se deba al grado de dificultad y subespecialización de las preguntas en el área de oftalmología.

En la prueba del ENURM la mayoría de las preguntas eran directas, con respuestas de selección múltiples. A pesar de esto el Chat GPT acertó en un menor porcentaje que en las preguntas tipo caso clínicos que requieren un nivel de razonamiento mayor para ser respondidas correctamente, y es que esta IA ha sido creada como un sistema lingüístico y tiene la capacidad de responder preguntas complejas[15].

Limitaciones

Cabe destacar que la veracidad de las respuestas ofrecidas tanto por los organizadores como por ChatGPT no ha sido verificada. Los resultados presentados reflejan exclusivamente el desempeño de ChatGPT hasta el 8 de marzo de 2023.Además, es crucial reconocer la posibilidad de que ChatGPT no esté actualizado con información específica, debido a que se basa en datos que no incluyen los del año en curso (2023). Los resultados de este estudio no son necesariamente aplicables a contextos distintos, tales como otros países, idiomas, escuelas de medicina o exámenes.

Conflicto de interés

Ninguno, según los autores.

Financiamiento

Esta investigación no ha recibido financiación.

Contribución de autoría

Vahid Nouri Kandany: Idea de investigación, recolección de datos, Introducción, Resultados, recomendaciones.

Alexander Valdez Disla: Gráficas, Discusión y conclusiones.

Pascual Valdez: Revisiones, Metodología, conclusiones.

Conclusiones

La evolución constante de los modelos de inteligencia artificial, como ChatGPT, gracias a la integración de nuevos datos, destaca su potencial para alcanzar una precisión casi perfecta. Este avance subraya la viabilidad de incorporar dichas tecnologías en la creación de preguntas para exámenes médicos futuros, lo que podría enriquecer su diversidad y complejidad y ofrecer retroalimentación más eficaz sobre el desempeño de los candidatos. El notable rendimiento de ChatGPT en el examen ENURM, superando el promedio de los postulantes a residencias médicas, ilustra su valor como herramienta complementaria en la educación y práctica clínica. Aunque ChatGPT se presenta como un recurso útil en contextos clínicos complejos, no reemplaza la importancia de una educación médica especializada y la experiencia clínica.

Referencias

1. Masters K. Artificial intelligence in medical education. *Med Teach*. 2019; 41(9): 976–980.
2. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. Pławiak P, editor. *PLOS ONE*. 2020; 15(2): e0229596.
3. Kirmani AR. Artificial Intelligence-Enabled Science Poetry. *ACS Energy Lett*. 2023; 8(1): 574–576.
4. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J Assoc Inf Sci Technol*. 2023; 74(5): 570–581.
5. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Libr Hi Tech News*. 2023; 40(3): 26–29.
6. Liu X, Zheng Y, Du Z, Ding M, Qian Y, Yang Z, et al. GPT Understands, Too. arXiv; 2023 [cited 2024 Feb 27]. Available from: http://arxiv.org/abs/2103.10385
7. Dale R. NLP in a post-truth world. *Nat Lang Eng*. 2017; 23(2): 319–324.

8. Jp C. ¿Es capaz "ChatGPT" de aprobar el examen MIR de 2022? Implicaciones de la inteligencia artificial en la educación médica en España. Is "ChatGPT" capable of passing the 2022 MIR exam? Implications of artificial intelligence in medical education in Spain. *Revista Española de Educación Médica*. 2023; 4(1): 12-18.
9. WMA - The World Medical Association-WMA Statement on Augmented Intelligence in Medical Care. [cited 2024 Feb 27]. Available from: https://www.wma.net/policies-post/wma-statement-on-augmented-intelligence-in-medical-care/
10. Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *Medical Education*. 2022 [cited 2023 Mar 9]. Available from: http://medrxiv.org/lookup/doi/10.1101/2022.12.23.22283901
11. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023; 20: 1.
12. Residencias Medicas. Facultad de Ciencias de la Salud. [cited 2024 Feb 27]. Available from: https://www.fcsuasd.net/examen.php
13. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023; 2(2): e0000198.
14. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings. *Ophthalmology*. 2023 [cited 2023 Mar 9]. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.01.22.23284882
15. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation*. 2023; 185: 109732.