Data and Metadata. 2024; 3:.433
doi: 10.56294/dm2024.433

ORIGINAL

Evaluation of the efficacy of ChatGPT versus medical students in clinical case resolution

Evaluación de la eficacia de ChatGPT frente a estudiantes de medicina en la resolución de casos clínicos

Fernanda Marizande[1], Andrea Cevallos[1], Diana Bustillos[2], Cristina Arteaga[2], Fabricio Vásquez de la Bandera[3], Alberto Bustillos[1]

[1]Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Medicina. Ambato, Ecuador.
[2]Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Nutrición y Dietética. Ambato, Ecuador.
[3] Universidad Técnica de Ambato, Facultad de Ciencias de la Salud, Carrera de Psicología Clínica. Ambato, Ecuador.

Citar como: Marizande F, Cevallos A, Bustillos D, Arteaga C, Vásquez de la Bandera F, Bustillos A. Evaluation of the efficacy of ChatGPT versus medical students in clinical case resolution. Data and Metadata. 2024; 3:.433. https://doi.org/10.56294/dm2024.433

Enviado: 10-02-2024 **Revisado:** 02-06-2024 **Aceptado:** 01-10-2024 **Publicado:** 02-10-2024

Editor: Adrián Alejandro Vitón Castillo

Autor para la correspondencia: Alberto Bustillos

ABSTRACT

Introduction: the use of artificial intelligence in medical education has gained relevance, and tools like ChatGPT offer support in solving clinical cases. This study compared the average performance of ChatGPT against medical students to evaluate its potential as an educational tool.
Methods: a cross-sectional quantitative study was conducted with 110 sixth-semester medical students from the Technical University of Ambato. Four clinical cases were designed, covering cardiology, endocrinology, gastroenterology, and neurology scenarios. Multiple-choice questions were used to assess both the participants and ChatGPT. Data were analyzed using the Student's t-test for independent samples.
Results: chatGPT outperformed the students in all cases, with an average score of 8,25 compared to 7,35 for the students. A statistically significant difference was found between the two groups (p = 0,0293).
Conclusions: chatGPT demonstrated superior performance in solving clinical cases compared to medical students. However, limitations such as potential inaccuracies in information highlight the need for further studies and supervision when integrating AI into medical education.

Keywords: ChatGPT; Medical Education; Clinical Case Resolution; Artificial Intelligence; Student Performance.

RESUMEN

Introducción: el uso de la inteligencia artificial en la educación médica ha cobrado relevancia, y herramientas como ChatGPT ofrecen apoyo en la resolución de casos clínicos. Este estudio comparó el promedio que obtiene ChatGPT frente a estudiantes de medicina, para evaluar su potencial como herramienta educativa.
Métodos: se realizó un estudio cuantitativo transversal con 110 estudiantes de sexto semestre de medicina de la Universidad Técnica de Ambato. Se diseñaron cuatro casos clínicos que incluyen escenarios de cardiología, endocrinología, gastroenterología y neurología. Se utilizaron preguntas de opción múltiple para evaluar a los participantes y a ChatGPT. Los datos fueron analizados con la prueba t de Student para muestras independientes.
Resultados: chatGPT superó a los estudiantes en todos los casos, con una puntuación promedio de 8,25, frente a 7,35 para los estudiantes. Se encontró una diferencia estadísticamente significativa entre los dos grupos (p = 0,0293)

Conclusiones: chatGPT demostró un rendimiento superior en la resolución de casos clínicos en comparación con los estudiantes de medicina. No obstante, limitaciones como la posible inexactitud de la información destacan la necesidad de realizar más estudios y supervisión al integrar la IA en la educación médica.

Palabras clave: ChatGPT; Educación Médica; Resolución de Casos Clínicos; Inteligencia Artificial; Rendimiento Estudiantil.

INTRODUCCIÓN

La integración de la inteligencia artificial (IA) en el campo de la educación médica ha marcado un hito significativo en cómo se enseñan y aprenden tanto el conocimiento teórico como las habilidades clínicas en el siglo XXI. ChatGPT, un modelo avanzado de procesamiento de lenguaje natural desarrollado por OpenAI, es uno de los exponentes más notables de esta tecnología, ofreciendo posibilidades antes impensables para la formación médica. La capacidad de ChatGPT para simular conversaciones y proporcionar respuestas coherentes lo convierte en una herramienta potencialmente útil para la enseñanza de la medicina, especialmente en el contexto de la resolución de casos clínicos.[1]

El aprendizaje basado en problemas (ABP), una metodología educativa centrada en el estudiante, ha sido ampliamente adoptado en las facultades de medicina en todo el mundo. Esta técnica pone a los estudiantes en el centro del proceso de aprendizaje y les facilita construir su conocimiento a través de la resolución de casos clínicos complejos que imitan situaciones reales.[2] La eficacia del ABP ha sido bien documentada, mostrando mejoras significativas en la retención del conocimiento y en las habilidades de pensamiento crítico de los estudiantes.[3] Sin embargo, el ABP también enfrenta desafíos, como la demanda de recursos significativos y la necesidad de facilitadores altamente capacitados que guíen el proceso de aprendizaje.[4]

En este contexto, herramientas como ChatGPT podrían desempeñar un papel crucial al proporcionar una plataforma adicional de aprendizaje autodirigido y a demanda. Los modelos de IA como ChatGPT tienen el potencial de complementar la educación médica tradicional, ofreciendo respuestas y retroalimentación instantáneas, lo cual es esencial para el desarrollo de habilidades diagnósticas y de toma de decisiones en un entorno controlado y sin riesgos.[5] Más aún, la capacidad de ChatGPT para procesar y generar información basada en enormes volúmenes de datos puede ayudar a exponer a los estudiantes a una variedad más amplia de casos clínicos, aumentando así su experiencia y su capacidad de aplicar el conocimiento médico en prácticas reales.[6]

A pesar de sus promesas, la implementación de IA en la educación médica no está exenta de críticas y preocupaciones. Uno de los temas más debatidos es la calidad de las interacciones que ofrece ChatGPT. Mientras que algunos estudios indican que las interacciones con sistemas de IA pueden ser menos ricas comparadas con las interacciones humanas, otros sugieren que la continua mejora de algoritmos podría eventualmente minimizar o eliminar esta brecha.[7] Además, la ética de usar IA, especialmente en relación con la privacidad de los datos y la autonomía del paciente, sigue siendo un campo de intensa discusión y regulación.

Otro desafío importante es la exactitud de la información proporcionada por ChatGPT. Aunque ChatGPT está entrenado y tiene varios textos médicos, no está exento de errores, y su uso como herramienta educativa requiere una supervisión cuidadosa y una evaluación constante para asegurar que la información que proporciona es actual y médicamente precisa.[8] La formación en habilidades críticas de evaluación debe ser una parte integral de cualquier currículo que integre IA, para que los estudiantes no solo aprendan a utilizar estas herramientas, sino también a cuestionarlas y verificarlas críticamente.[9]

Esta investigación busca evaluar la eficacia de ChatGPT frente a estudiantes de medicina en la resolución de casos clínicos. Comparar la precisión y la utilidad de las respuestas proporcionadas por ChatGPT con las soluciones generadas por estudiantes puede ofrecer perspectivas valiosas sobre el rol potencial de la IA en la educación médica futura. Además, explorar cómo los estudiantes perciben y confían en la tecnología para proporcionar información clínica relevante y precisa es fundamental para entender y optimizar su integración en los currículos médicos.

MÉTODOS

Diseño del estudio y población evaluada

Se llevó a cabo un estudio cuantitativo transversal para evaluar y comparar la eficacia de ChatGPT frente a estudiantes de sexto semestre de medicina de la Universidad Técnica de Ambato en la resolución de casos clínicos. El estudio se diseñó para medir la precisión y calidad de las respuestas dadas tanto por los estudiantes como por ChatGPT, basándose en un conjunto estándar de preguntas de opción múltiple diseñadas específicamente para este propósito.

La población de estudio consistió en todos los estudiantes de sexto semestre de medicina de la Universidad

Técnica de Ambato durante el ciclo académico 2024. Se utilizó un muestreo no probabilístico por conveniencia para seleccionar a 110 estudiantes que voluntariamente acordaron participar en la investigación. Se aseguró que los estudiantes comprendieran el propósito del estudio y se obtuvo su consentimiento informado antes de la participación.

Selección y Diseño de los Casos Clínicos

Se seleccionaron cuatro casos clínicos que cubrían diversas áreas de la medicina, tales como cardiología, endocrinología, gastroenterología y neurología. Cada caso clínico fue diseñado para simular situaciones reales que un médico podría encontrar en su práctica diaria. Los casos clínicos se pueden revisar en la siguiente base de datos: https://doi.org/10.7910/DVN/V4TACO

Desarrollo de Instrumentos de Evaluación

Para cada uno de los cuatro casos clínicos, se redactaron 10 preguntas de opción múltiple enfocadas en el diagnóstico y la solución de los casos. Las preguntas fueron diseñadas para evaluar la comprensión del estudiante sobre el caso, su capacidad para aplicar el conocimiento médico relevante y su habilidad para elegir el plan de manejo más adecuado. Cada pregunta tenía cuatro opciones de respuesta, de las cuales solo una era correcta. La guía de preguntas y la clave de calificación se encuentra en la base de datos de los casos clínicos.

Los estudiantes respondieron las preguntas en un ambiente controlado y supervisado para evitar el uso de ayudas externas. Paralelamente, las mismas preguntas fueron ingresadas a ChatGPT, y las respuestas obtenidas fueron registradas. La evaluación se llevó a cabo en una única sesión.

Análisis de Datos

Las respuestas fueron calificadas automáticamente utilizando un sistema de puntuación, donde cada respuesta correcta otorgaba un punto. Las puntuaciones de los estudiantes y de ChatGPT fueron registradas para cada pregunta de cada caso clínico.

Se verificó la normalidad de las distribuciones de las notas usando la prueba de Shapiro-Wilk debido al tamaño relativamente pequeño de la muestra.

Dado que las notas siguieron una distribución normal, se empleó la prueba t de Student para muestras independientes para comparar las medias de las puntuaciones obtenidas por los estudiantes y por ChatGPT en cada caso clínico. Se utilizó el software estadístico SPSS para realizar todos los análisis. Se consideraron significativos estadísticamente aquellos resultados con valores de $p<0,05$.

Consideraciones Éticas

Este estudio fue revisado y aprobado por el Comité ético de Investigación en seres humanos de la Universidad UTE, con el código de registro CEISH-UTE-003455-2422. Todos los datos fueron tratados de manera confidencial, y se preservó el anonimato de los estudiantes participantes. Se informó a los estudiantes que podían retirarse del estudio en cualquier momento sin ninguna consecuencia académica.

RESULTADOS

En la Tabla 1 se presentan los promedios de las notas obtenidas por los estudiantes en cada caso clínico. Los datos revelan que los estudiantes tuvieron un desempeño consistente a través de los diferentes casos, con promedios que oscilan en torno a 7,35 en un rango de 0 a 10 puntos posibles. La prueba estadística en la tabla 1, muestra un valor p de 0,0293, sugiriendo una diferencia estadísticamente significativa en las notas promedio entre los dos grupos evaluados.

Tabla 1. Comparación del rendimiento en los cuestionarios de los casos clínicos, entre estudiantes y ChatGPT

Casos clínicos (Enfoque)	Promedio Estudiantes	ChatGPT	t student
Cardiología	7,2	8	p valor: 0,0293
Endocrinología	7	8	
Gastroenterología	7,3	9	
Neurología	7,9	8	
Promedio General	7,35	8,25	

En la figura 1 se ilustra las diferencias en las notas promedio entre los estudiantes y ChatGPT, destacando visualmente la superioridad estadística de las notas de ChatGPT sobre las de los estudiantes.

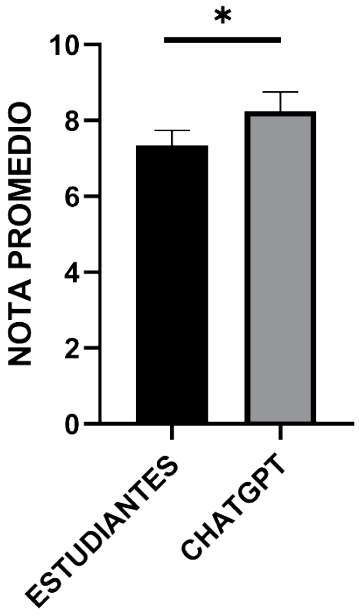

Figura 1. Diferencias estadísticas en las notas promedio entre estudiantes y ChatGPT

DISCUSIÓN

Los resultados del estudio indican que ChatGPT obtiene un mejor rendimiento que los estudiantes de medicina en la resolución de casos clínicos en términos de precisión de las respuestas. Estos hallazgos son consistentes con estudios previos que exploran el uso de herramientas de inteligencia artificial en entornos educativos, donde se ha demostrado que la IA puede complementar y a veces superar el rendimiento humano en tareas específicas de aprendizaje y resolución de problemas.[10,11]

El rendimiento superior de ChatGPT en las especialidades evaluadas, como cardiología, gastroenterología y endocrinología, se alinea con estudios recientes sobre el uso de IA en la educación médica. Estos trabajos sugieren que modelos como ChatGPT tienen un gran potencial en la enseñanza de temas complejos, ya que pueden generar respuestas coherentes y comprensibles a partir de grandes volúmenes de datos entrenados.[12,13]

La capacidad de sintetizar información y proporcionar explicaciones detalladas es especialmente útil en la formación médica, donde la comprensión profunda de los casos clínicos es fundamental. Sin embargo, estudios previos también señalan que la IA aún enfrenta limitaciones importantes, particularmente en la toma de decisiones clínicas reales.

En ocasiones, ChatGPT ha mostrado una tendencia a generar información inexacta o "alucinaciones artificiales", donde los datos presentados, aunque aparentemente correctos, son inventados. Esto plantea un riesgo considerable cuando se utiliza para generar documentos científicos sin la revisión adecuada por parte de profesionales humanos.[8,10]

Además, la falta de transparencia en las fuentes de información utilizadas por el modelo plantea desafíos en cuanto a la fiabilidad de los datos presentados.

Si bien ChatGPT ha mostrado un rendimiento aceptable en la generación de diagnósticos diferenciales y la evaluación de opciones terapéuticas, existen limitaciones significativas que impiden su aplicación directa en la práctica clínica sin supervisión humana. En particular, su incapacidad para interactuar dinámicamente con los pacientes y obtener información contextual adicional lo convierte en una herramienta subóptima para la toma de decisiones clínicas complejas. Además, su desempeño en el diagnóstico de enfermedades raras ha sido desigual, con mejores resultados en condiciones más comunes.[11]

Una de las principales limitaciones de nuestro estudio es el tamaño y la diversidad de la muestra, que se limitó a estudiantes de un solo semestre de una universidad. Futuros estudios podrían beneficiarse de una muestra más grande y más diversa para generalizar los resultados. Además, el diseño del estudio no permitió evaluar cómo los estudiantes utilizan el razonamiento clínico para llegar a sus respuestas, algo que ChatGPT podría no replicar completamente. Por lo tanto, mientras que las respuestas de ChatGPT pueden ser técnicamente precisas, no necesariamente reflejan un proceso de pensamiento clínico que es crucial en la formación médica.

El papel de ChatGPT en la medicina podría expandirse significativamente con mejoras en la precisión de los datos y la reducción de sesgos. Sin embargo, es esencial que los profesionales médicos mantengan un control

riguroso sobre la información generada por estos modelos para garantizar que las decisiones clínicas se basen en datos confiables y bien fundamentados. Las investigaciones futuras podrían centrarse en la capacitación específica de modelos como ChatGPT en áreas más especializadas, como las enfermedades raras, lo que podría aumentar su utilidad en contextos clínicos más complejos. Con avances en la formación específica y el refinamiento de los modelos de lenguaje, herramientas como ChatGPT podrían convertirse en aliados valiosos en el ámbito médico, mejorando la eficiencia y la precisión en diversas tareas clínicas y educativas.

AGRADECIMIENTO

Los autores agradecen a la Dirección de Investigación y Desarrollo DIDE de la Universidad Técnica de Ambato.

REFERENCIAS BIBLIOGRÁFICAS

1. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg2024;110(6):3701–6.

2. Hernández R, Moreno SM. El aprendizaje basado en problemas: una propuesta de cualificación docente. Praxis & Saber 2021;12(31):e11174.

3. Rosa N, Palomino J, Cesar U, Piura V, Diaz Espinoza M, Giovanna L, et al. Evaluación del Impacto del aprendizaje basado en proyectos frente a la clase invertida en el desarrollo de habilidades de investigación Comparison of the development of investigative skills between Project Based Learning and the Flipped Class. Ciencia y Tecnología, 2024;28:40. Available from: https://doi.org/10.47460/uct.v28i124.800

4. Valverde-Gutiérrez KV, Esteves-Fajardo ZI. Aprendizaje Basado en Problemas para el Desarrollo del Pensamiento Crítico desde Tempranas Edades. Revista Arbitrada Interdisciplinaria Koinonía 2023;8(1):150–71.

5. Javaid M, Haleem A, Singh RP, Khan S, Khan IH. Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system. BenchCouncil Transactions on Benchmarks, Standards and Evaluations 2023;3(2).

6. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in Medical Research: Current Status and Future Directions. J Multidiscip Healthc2023;16:1513–20.

7. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. J Pediatr Urol (Internet) 2023;19(5):598–604. Available from: https://www.sciencedirect.com/science/article/pii/S1477513123002243

8. Baumgartner C. The opportunities and pitfalls of ChatGPT in clinical and translational medicine. Clin Transl Med 2023;13(3).

9. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst 2023;47(1).

10. Goodman RS, Patrinely JR, Stone CA, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open 2023;E2336483.

11. Zampatti S, Peconi C, Megalizzi D, Calvino G, Trastulli G, Cascella R, et al. Innovations in Medicine: Exploring ChatGPT's Impact on Rare Disorder Management. Genes (Basel)2024;15(4).

12. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023; 6.

13. Chen TC, Multala E, Kearns P, Delashaw J, Dumont A, Maraganore D, et al. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023;5(2).

FINANCIACIÓN

Se recibió financiación de la Dirección de Investigación y Desarrollo DIDE, de la Universidad Técnica de Ambato, bajo el proyecto de investigación aprobado por Resolución Nro. UTA-CONIN-2024-0025-R

CONFLICTO DE INTERESES

Los autores declaran que no existe conflicto de intereses.

CONTRIBUCIÓN DE AUTORÍA

Conceptualización: Alberto Bustillos, Fernanda Marizande.

Curación de datos: Alberto Bustillos, Fernanda Marizande.

Análisis formal: Fernanda Marizande, Alberto Bustillos, Cristina Arteaga, Fabricio Vásquez de la Bandera.

Investigación: Alberto Bustillos, Fernanda Marizande, Andrea Cevallos, Cristina Arteaga.

Metodología: Fernanda Marizande, Fabricio Vásquez de la Bandera, Diana Bustillos, Cristina Arteaga, Alberto Bustillos.

Administración del proyecto: Fernanda Marizande, Alberto Bustillos.

Recursos: Alberto Bustillos, Fernanda Marizande, Diana Bustillos.

Supervisión: Alberto Bustillos.

Redacción – borrador original: Alberto Bustillos, Fernanda Marizande, Fabricio Vásquez de la Bandera, Diana Bustillos, Cristina Arteaga.

Redacción – revisión y edición: Alberto Bustillos, Fernanda Marizande, Fabricio Vásquez de la Bandera, Diana Bustillos, Cristina Arteaga.

https://doi.org/10.56294/dm2024.433