Endocrinología, Diabetes y Nutrición

418 - EVALUACIÓN COMPARATIVA DEL RENDIMIENTO DE PLATAFORMAS DE INTELIGENCIA ARTIFICIAL GENERATIVA MÁS COMUNES (CHATGPT, GEMINI Y COPILOT) EN 1.140 PREGUNTAS DE ENDOCRINOLOGÍA Y NUTRICIÓN

C. Lozano Aida[1], R. Gómez Almendros[2], P. Pérez Castro[3], R. Fernández García-Salazar[4], A. Gutiérrez Hurtado[4], J. Napky Rajo[2], D. Rivas Otero[1], I. Masid Sánchez[1], E. Redondo[5] y M. García Villarino[6]

[1]Hospital Universitario Central de Asturias, Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo.[2]Hospital Universitario Torrecárdenas, Almería.[3]Complejo Hospitalario Universitario de Vigo.[4]Hospital Universitario Central de Asturias, Oviedo.[5]Hospital Universitario Clínico San Cecilio, Granada.[6]Instituto de Investigación Sanitaria del Principado de Asturias.

Resumen

Introducción: Las plataformas de inteligencia artificial generativa (IAG) han irrumpido como herramientas potenciales en educación médica. Su rendimiento en contextos específicos, como la resolución de preguntas clínicas tipo test en el ámbito médico, y más concretamente en el área de la endocrinología y la nutrición, aún no ha sido evaluado de forma sistemática.

Objetivos: Comparar el rendimiento de ChatGPT, Copilot y Gemini –en sus versiones gratuitas y de pago– al resolver preguntas tipo test de endocrinología y nutrición extraídas de oposiciones oficiales del sistema sanitario español.

Métodos: Se incluyeron un total de 1.140 preguntas tipo test procedentes de exámenes de oposición en endocrinología y nutrición celebrados entre los años 2022 y 2024 en diversas comunidades autónomas españolas, eliminándose las impugnadas. Cada pregunta fue introducida simultáneamente en seis modelos de IAG (ChatGPT 4mini, ChatGPT4o, Gemini, Gemini Advanced, Copilot y Copilot Pro). Se evaluó el porcentaje de aciertos y la concordancia intermodelo mediante el índice Kappa.

Resultados: El análisis de rendimiento mostró que la plataforma con mayor tasa de aciertos fue ChatGPT-4o (versión de pago), con un 81,4% de respuestas correctas. Le siguieron Gemini Advanced (74,6%) y Copilot Pro (70,7%). Entre las versiones gratuitas, Copilot (69,2%) y ChatGPT 4mini (68,3%) presentaron resultados similares, mientras que Gemini gratuito obtuvo el porcentaje más bajo (61,8%). La concordancia entre modelos fue moderada (? = 0,45 entre ChatGPT 4mini y Copilot), siendo superior en las versiones de pago.

Conclusiones: Todas las versiones de pago mostraron un rendimiento superior a un 70%, siendo superior ChatGPT-4o con más de un 80%, por lo que es evidente que las IAG podrían constituir una herramienta de apoyo útil en formación médica, especialmente si se utilizan de forma complementaria. Estos datos no son equiparables a los resultados obtenidos en el examen MIR, lo que puede ser debido a una menor dificultad del mismo.