Endocrinología, Diabetes y Nutrición

416 - ¿QUÉ HERRAMIENTA DE INTELIGENCIA ARTIFICIAL ACIERTA MÁS PREGUNTAS DE ENDOCRINO EN EL MIR?

C. Lozano Aida[1], I. Masid Sánchez[1], R.P. Fernández García-Salazar[1], A. Gutiérrez Hurtado[1], M. García Villarino[2], A.V. García Gómez[3], E. Villa Fernández[3], P. Pérez Castro[4], E. Delgado Álvarez[5] y E.L. Menéndez Torre[5]

[1]Servicio de Endocrinología y Nutrición, Hospital Universitario Central de Asturias, Oviedo. [2]Instituto de Investigación Sanitaria del Principado de Asturias. Universidad de Oviedo. [3]Instituto de Investigación Sanitaria del Principado de Asturias, Oviedo. [4]Servicio de Endocrinología y Nutrición, Complejo Hospitalario Universitario de Vigo. [5]Servicio de Endocrinología y Nutrición, Hospital Universitario Central de Asturias. Instituto de Investigación Sanitaria del Principado de Asturias, Universidad de Oviedo.

Resumen

Introducción: Nos planteamos determinar qué asistente de inteligencia artificial acierta más preguntas de Endocrinología y Nutrición del examen de acceso a la formación especializada MIR en España en los últimos cinco años. Como objetivos secundarios, valorar si existe diferencia en la tasa de aciertos en función de si la pregunta es caso clínico o no, comparando las versiones estándar y avanzadas y valorar su concordancia.

Métodos: Se realizó un análisis transversal y descriptivo, usando las versiones estándar de tres sistemas de inteligencia artificial (ChatGPT 3,5, Gemini y Copilot) y sus ediciones avanzadas (ChatGPT 4, Gemini Advanced y Copilot Pro) para responder a las 62 preguntas (32 casos clínicos) de los últimos cinco exámenes MIR (2020-2024).

Resultados: ChatGPT 4 es el asistente que tiene el mayor porcentaje de aciertos con un 91,4%, mientras que la versión estándar de Copilot presenta el menor con un 56,45%. En las versiones avanzadas de los tres asistentes el porcentaje de acierto es similar en las preguntas que incluyen casos clínicos y en las que no; presentando mayor disparidad en las básicas. Si bien los porcentajes de aciertos individuales son elevados, la concordancia general es solo entre débil y moderada, con mejores resultados en las versiones básicas.

Porcentaje de aciertos.						
	ChatGPT 3.5	Copilot	Gemini	ChatGPT4	Copilot Pro	Gemini Advanced
Total	66,13	56,45	69,35	91,94	90,32	83,87
Caso clínico	56,25	50	71,88	90,63	90,63	81,25
No caso clínico	76,67	63,33	66,67	93,33	90	86,67

Conclusiones: Las versiones avanzadas de los asistentes de inteligencia artificial presentan porcentajes de aciertos superiores y con una mayor similitud entre distintos tipos de preguntas, en

comparación con las versiones estándar, siendo ChatGPT4 el que más preguntas acertó. Sin embargo, la concordancia entre ellas es inferior. La inteligencia artificial ha supuesto un avance en el ámbito de la educación médica, pero se debe usar con un enfoque crítico y razonado.