

ORIGINAL

# Explorando el potencial de la inteligencia artificial en traumatología: respuestas conversacionales a preguntas específicas



F. Canillas del Rey<sup>a,b,\*</sup> y M. Canillas Arias<sup>b</sup>

<sup>a</sup> Servicio de Cirugía Ortopédica y Traumatología, Hospital Universitario Cruz Roja, Madrid, España

<sup>b</sup> Facultad de Medicina, Universidad Alfonso X el Sabio, Madrid, España

Recibido el 19 de marzo de 2024; aceptado el 15 de mayo de 2024

Disponible en Internet el 21 de mayo de 2024

## PALABRAS CLAVE

Inteligencia artificial generativa;  
Robot conversacional;  
Respuesta a preguntas de opción múltiple;  
ChatGPT;  
Bard;  
Perplexity

## Resumen

**Antecedentes y objetivo:** La inteligencia artificial generativa es una tecnología que ofrece su mayor conectividad con las personas gracias a los bots conversacionales («chatbot»). Estos pueden mantener un diálogo con un lenguaje natural indistinguible del humano y son una fuente potencial de información para los pacientes. El objetivo de este trabajo es estudiar el rendimiento de estos bots en la resolución de cuestiones específicas de cirugía ortopédica y traumatología empleando las preguntas del examen MIR español entre 2008 y 2023.

**Material y métodos:** Se analizaron 3 modelos de «chatbots» (ChatGPT, Bard y Perplexity) respondiendo a 114 preguntas del MIR. Se compararon aciertos, se valoró la legibilidad de las respuestas y se examinó su dependencia con el razonamiento lógico y la información interna y externa. En los fallos también se evaluó el tipo de error.

**Resultados:** ChatGPT obtuvo un 72,81% de aciertos, seguido por Perplexity (67,54%) y Bard (60,53%). Las respuestas más legibles y completas las ofrece Bard. Las respuestas demostraron un razonamiento lógico y el uso de información interna de los enunciados de preguntas. En 16 preguntas (14%) las 3 aplicaciones fallaron simultáneamente. Se identificaron errores, que incluían fallos lógicos y de información.

**Conclusiones:** Aunque los bots conversacionales pueden ser útiles en la resolución de preguntas médicas, se señala la necesidad de precaución debido a la posibilidad de errores. Actualmente deben considerarse como una herramienta en desarrollo y la opinión humana debe prevalecer sobre la inteligencia artificial generativa.

© 2024 SECOT. Publicado por Elsevier España, S.L.U. Este es un artículo Open Access bajo la CC BY-NC-ND licencia (<http://creativecommons.org/licencias/by-nc-nd/4.0/>).

\* Autor para correspondencia.

Correo electrónico: [fercanillas@yahoo.es](mailto:fercanillas@yahoo.es) (F. Canillas del Rey).

**KEYWORDS**

Generative Artificial Intelligence;  
Chatbot;  
Multi-choice question answering;  
ChatGPT;  
Bard;  
Perplexity

**Exploring the potential of Artificial Intelligence in Traumatology: Conversational answers to specific questions****Abstract**

**Introduction:** Generative Artificial Intelligence is a technology that provides greater connectivity with people through conversational bots («chatbots»). These bots can engage in dialogue using natural language indistinguishable from humans and are a potential source of information for patients. The aim of this study is to examine the performance of these bots in solving specific issues related to orthopedic surgery and traumatology using questions from the Spanish MIR exam between 2008 and 2023.

**Material and methods:** Three «chatbot» models (ChatGPT, Bard and Perplexity) were analyzed by answering 114 questions from the MIR. Their accuracy was compared, the readability of their responses was evaluated, and their dependence on logical reasoning and internal and external information was examined. The type of error was also evaluated in the failures.

**Results:** ChatGPT obtained 72.81% correct answers, followed by Perplexity (67.54%) and Bard (60.53%). Bard provides the most readable and comprehensive responses. The responses demonstrated logical reasoning and the use of internal information from the question prompts. In 16 questions (14%), all 3 applications failed simultaneously. Errors were identified, including logical and information failures.

**Conclusions:** While conversational bots can be useful in resolving medical questions, caution is advised due to the possibility of errors. Currently, they should be considered as a developing tool, and human opinion should prevail over Generative Artificial Intelligence.

© 2024 SECOT. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Introducción**

La inteligencia artificial (IA) engloba desarrollos tecnológicos que emulan las capacidades cognitivas de los humanos. En cirugía ortopédica y traumatología, las aplicaciones incluyen reconocimiento y diagnóstico de imágenes, registros de textos médicos, rehabilitación y cuidados postoperatorios, formación quirúrgica o algoritmos predictivos<sup>1</sup>. En los últimos años, la IA dio un paso más al convertirse en una IA generativa. Es decir, ya no solo analiza problemas y los resuelve, sino que con los datos que se le aporta mejora su aprendizaje y genera contenidos originales (texto, imágenes, videos, presentaciones, moléculas, etc.)<sup>2</sup>.

Desde hace poco tiempo, se han popularizado los bots de charla o conversacionales que son IA generativas que tienen la capacidad de mantener una conversación dando respuestas coherentes y similares a un humano. Incluso se puede modular su comportamiento a la hora de responder, permitiendo cierta «personalidad» en las respuestas.

En noviembre de 2022, se lanza el bot conversacional («chatbot») gratuito denominado ChatGPT (Generative Pre-trained Transformer) de la empresa OpenAI (OpenAI, LLC, San Francisco, California, EE. UU.) desarrollada en lenguaje Python. Su principal limitación era que solo tenía acceso a Internet hasta el año 2021, pero la nueva versión (de pago) ya ha actualizado el acceso a la red. Además, es multimodal permitiendo entrada de texto e imagen para generar la respuesta.

En contestación a esta aplicación, distintas empresas tecnológicas lanzan al mercado otros «chatbot» como Bing de Microsoft (Microsoft Corporation, Redmond, Washington, EE. UU.) o Bard de Google (Google LLC, Mountain View, Cali-

fornia, EEUU) o Perplexity de los diseñadores Denis Yarats, Aravind Srinivas, Johnny Ho y Andy Konwinski. Esta última con la particularidad de aportar citas para respaldar su información.

Estos desarrollos tecnológicos disruptivos abren un abanico de opciones aún por explorar y tienen enormes implicaciones en el campo de la medicina y docencia<sup>3</sup>. No solo porque ofrecen respuestas a cuestiones que se le plantean, sino que pueden generar información con múltiples utilidades. Cada día se amplían sus capacidades y ya empiezan incluso a reconocer voz, imágenes o videos de modo que sus potencialidades crecen exponencialmente.

Los pacientes lo emplean para resolver preguntas sobre medicina y salud<sup>4</sup> y los profesionales para resolución de casos<sup>5</sup> o creación de textos científicos<sup>6,7</sup>. Incluso generan respuestas con más información, de más calidad y con más empatía que la humana<sup>8</sup>, aunque algunos pacientes muestran poca aceptación para su uso como sustituto de profesionales<sup>9</sup>. Evidentemente, esto tiene importantes connotaciones legales y éticas, que involucran la responsabilidad en toma de decisiones o en autoría de producción científica<sup>10</sup>. Máxime cuando se han detectado la presencia de errores, conocidos como «alucinaciones»<sup>11</sup>, que son más comunes de lo que aparentemente debería ser una «inteligencia» que genera datos sensibles. No debemos olvidar que, en el año 2022, el 40% de los usuarios de Internet recurrieron a Internet en busca de información sobre temas de salud<sup>12</sup>.

La IA conversacional se basa en 3 pilares para su desarrollo que son: aprendizaje automático («machine learning»), gran volumen de datos y procesamiento de lenguaje natural<sup>2</sup>. El aprendizaje automático, o «machine

learning», permite a los ordenadores aprender y mejorar automáticamente a partir de la experiencia sin necesidad de haber sido programada específicamente. Los datos se analizan mediante algoritmos para identificar patrones y tomar decisiones con mayor o menor supervisión humana. La arquitectura de la máquina emplea redes neuronales que imita el funcionamiento del cerebro humano. Esto permite el aprendizaje profundo, o «deep learning», al procesar una inmensa cantidad de datos, o «big data», que incluye datos no estructurados o sin etiquetar, como imágenes, audio y texto, para realizar tareas como reconocimiento de voz o imagen. La versión 3 de ChatGPT se entrenó con 175.000 millones de parámetros. Esto significa que la IA no busca la información en la red, sino que genera respuestas siguiendo modelos predictivos desde la información que ha recogido y procesado en unidades más pequeñas (conocidas como «token»). En este sentido, hay autores que abogan por hablar de «aprendizaje estadístico computacional» en vez de «inteligencia artificial»<sup>13</sup>. Esto explica por qué las IA tienen fallos, o «alucinaciones» en el argot informático, debidos en gran medida a su habilidad o no para manejar los datos de los que se nutre y generar los resultados siguiendo patrones estocásticos. Si a esto se le añade el efecto conocido como GIGO (Garbage In Garbage Out, Entra basura, Sale basura)<sup>7</sup> nos encontramos con un sistema que requiere un proceso de aprendizaje y supervisión.

El procesamiento del lenguaje natural es otra área de la IA que se encarga de la interacción entre máquina y el lenguaje humano. Gracias a esta aplicación, se puede comprender, interpretar y generar texto de manera eficiente. Todo ello conduce a procesar la orden o «prompt», «comprenderla» y responder de forma natural y aparentemente acertada. Además añade la capacidad de recordar conversaciones previas.

En el campo médico<sup>14</sup>, permite el tratamiento de enormes cantidades de registros médicos electrónicos, permitiendo analizarlos y organizarlos para obtener información eficiente y precisa. En el ámbito quirúrgico<sup>1</sup>, la IA permite la simulación de procedimientos quirúrgicos complejos en entornos virtuales. Además se pueden identificar patrones de éxito y áreas de mejora en la formación quirúrgica, mejorando la calidad y seguridad de la atención médica.

El objetivo de este trabajo es estudiar y comparar la capacidad de las IA conversacionales de resolver cuestionarios de cirugía ortopédica y traumatología que se emplean en el examen nacional para la obtención de plaza de médico residente.

## Material y método

Se han revisado todas las preguntas del examen de acceso a la formación médica especializada española (examen MIR) desde los años 2008 al 2023. Los criterios de inclusión son: preguntas referidas a traumatología y cirugía ortopédica. Los criterios de exclusión son: preguntas anuladas por el tribunal o que contengan imágenes que son necesarias para poder contestar la pregunta.

Las preguntas son de tipo test con 4 o 5 opciones de respuestas y una única correcta. Algunas de las preguntas incluyen imágenes. Dado que no todas las aplicaciones



**Figura 1** Logos de los «chatbot» conversaciones ChatGPT, Bard y Perplexity.

tienen reconocimiento de imagen, se han excluido las que exigen ver la imagen para contestar. Las preguntas se han agrupado según la temática en traumatología, ortopedia del adulto, ortopedia infantil y columna, y según el año. Además, se han clasificado en 2 tipos según la necesidad de conocimientos para dar la respuesta<sup>15</sup>: tipo 1, solo requieren un conocimiento, y tipo 2, necesitan varios pasos para conseguir la respuesta.

Las respuestas se han analizado valorando la coherencia narrativa<sup>16</sup>. Los datos son binarios e incluyen 3 apartados: razonamiento lógico (si la respuesta es seleccionada según la información presentada), información interna (la respuesta incluye información facilitada en la pregunta) e información externa (la respuesta aporta información externa a la facilitada). También hemos analizado la legibilidad de Flesch-Kincaid<sup>17,18</sup>, adaptada al español por Fernández Huerta<sup>19</sup> y corregida por Gwillim Law<sup>20</sup>. El resultado muestra la facilidad de comprensión de un texto y lo correlaciona con el nivel de estudios como resultado de esta fórmula:

$$L = 206,84 - 0,60P - 1,02F$$

$L$  es la «lecturabilidad»;  $P$ , el promedio de sílabas por palabra;  $F$ , la media de palabras por frase.

También se ha empleado la escala INFLESZ<sup>21</sup> que mide la facilidad para leer un texto, así como el número de palabras de cada respuesta.

En caso de haber errores, se etiquetaron atendiendo a ser un error lógico (encuentra información correcta pero no la traslada a la respuesta), error de información (no identifica un dato clave aportado en la pregunta o en información externa) y error estadístico (basados en errores aritméticos p. ej. estimación errónea de la frecuencia de una enfermedad)<sup>22</sup>.

Los motores de IA evaluados son ChatGPT (versión 3.5), Bard y Perplexity. (fig. 1). Todas las respuestas se obtuvieron en un periodo de 48 h. En la redacción de este trabajo, la aplicación Bard cambió el nombre a Gemini, pero dado que los resultados se obtuvieron con la primera denominación se ha mantenido en los resultados y conclusiones.

Los datos recogidos son del tipo cuantitativo, cualitativo y descriptivo. Las variables cualitativas se presentan con su frecuencia absoluta y su porcentaje relativo. Para comparar valores cualitativos de carácter dicotómico se ha empleado la prueba de Q de Cochran, y para los valores cuantitativos se ha empleado la prueba de ANOVA si cumple normalidad y la prueba de Kruskal-Wallis si no la cumple. El riesgo alfa aceptado para todos los contrastes de hipótesis es de 0,05. Si la prueba es significativa, se realiza una prueba *post-hoc* para comparación por pares de variables. Los datos fueron analizados empleando el programa MedCalc versión 22.016 (MedCalc Software Ltd, Ostend, Bélgica; <https://www.medcalc.org>; 2023) para procesamiento

**Tabla 1** Frecuencias de temática/año

Año	Tema	Frecuencias	% del total	Año	Tema	Frecuencias	% del total
2009	Trauma	6	5,3%	2017	Trauma	4	3,5%
	OrtoAdulto	2	1,8%		OrtoAdulto	3	2,6%
	OrtoInfantil	2	1,8%		OrtoInfantil	0	0,0%
	Columna	1	0,9%		Columna	1	0,9%
2010	Trauma	2	1,8%	2018	Trauma	6	5,3%
	OrtoAdulto	2	1,8%		OrtoAdulto	3	2,6%
	OrtoInfantil	2	1,8%		OrtoInfantil	1	0,9%
	Columna	1	0,9%		Columna	0	0,0%
2011	Trauma	4	3,5%	2019	Trauma	1	0,9%
	OrtoAdulto	4	3,5%		OrtoAdulto	3	2,6%
	OrtoInfantil	2	1,8%		OrtoInfantil	1	0,9%
	Columna	0	0,0%		Columna	0	0,0%
2012	Trauma	0	0,0%	2020	Trauma	5	4,4%
	OrtoAdulto	2	1,8%		OrtoAdulto	3	2,6%
	OrtoInfantil	1	0,9%		OrtoInfantil	1	0,9%
	Columna	0	0,0%		Columna	1	0,9%
2013	Trauma	2	1,8%	2021	Trauma	1	0,9%
	OrtoAdulto	3	2,6%		OrtoAdulto	7	6,1%
	OrtoInfantil	0	0,0%		OrtoInfantil	1	0,9%
	Columna	1	0,9%		Columna	2	1,8%
2014	Trauma	1	0,9%	2022	Trauma	2	1,8%
	OrtoAdulto	1	0,9%		OrtoAdulto	2	1,8%
	OrtoInfantil	1	0,9%		OrtoInfantil	1	0,9%
	Columna	0	0,0%		Columna	2	1,8%
2015	Trauma	4	3,5%	2023	Trauma	3	2,6%
	OrtoAdulto	2	1,8%		OrtoAdulto	3	2,6%
	OrtoInfantil	0	0,0%		OrtoInfantil	2	1,8%
	Columna	0	0,0%		Columna	2	1,8%
2016	Trauma	3	2,6%				
	OrtoAdulto	2	1,8%				
	OrtoInfantil	1	0,9%				
	Columna	1	0,9%				

de datos y estudio estadístico. La hipótesis nula ( $H_0$ ) que asumimos es que la capacidad de aciertos de los diferentes programas evaluados es la misma.

## Resultados

Inicialmente se recogieron 129 preguntas en los exámenes del MIR referidas a cirugía ortopédica y traumatología. Se excluyeron 15 debido a que presentaban imágenes (14 preguntas) y una por haber sido anulada. En la [tabla 1](#) se muestra la distribución de las preguntas por año y temática.

Según el tipo de pregunta, hay 49 (43%) del tipo I (requieren un conocimiento) y 65 (57%) del tipo II (necesitan varios para obtener la respuesta).

La aplicación ChatGPT ha obtenido 83 aciertos (72,81%), frente a los 69 (60,53%) de Bard y los 77 (67,54%) de Perplexity. Encontramos significación estadística en la prueba Q de Cochran (0,049) respecto a la distribución de frecuencias entre los grupos, y en cuando se realiza el análisis *post-hoc* solo objetiva diferencia significativa en ChatGPT frente a Bard, pero no entre el resto de pares. Si agrupamos las respuestas según el tipo de pregunta, observamos que en el tipo I ChatGPT acierta 36 (73,47%), Bard obtiene 27 (55,10%)

y Perplexity 36 (73,47%) que muestra una diferencia significativa (0,030) pero en las comparaciones múltiples no encuentra diferencias entre pares debido al tamaño muestral del subgrupo. En el tipo II de preguntas no encontramos diferencias significativas entre grupos, ChatGPT ha obtenido 47 (72,31%) aciertos, frente a los 42 (64,62%) de Bard y a los 41 (63,08%) de Perplexity.

En la [tabla 2](#) se muestra el acúmulo de aciertos de cada una de las preguntas al contestar los 3 «chatbot».

En el análisis de respuestas de cada programa, ChatGPT usa el razonamiento lógico en 92 (80,7%), información interna en 108 (94,7%) e información externa en 50 (43,8%). En Bard, se emplea el razonamiento lógico en 114 (100%) respuestas, información interna en 114 (100%) e información externa en 111 (97,3%). En Perplexity, se utiliza el razonamiento lógico en 109 (95,6%) respuestas, información interna en 113 (99,1%) e información externa en 60 (52,6%). En este último apartado, señalamos que esta aplicación incluye citas que enlazan a páginas web pero no se han considerado información externa ya que exigiría una evaluación diferente al propósito de este trabajo. Al analizar el razonamiento lógico, encontramos significación estadística ( $p < 0,001$ ) a favor de Bard frente al resto. Esta diferencia

**Tabla 2** Número de aciertos de «Chatbots»

Tres aciertos	Dos aciertos	Un acierto	Ningún acierto	Preguntas
49 43%	33 28,90%	16 14%	16 14%	114
Dos aciertos			Un acierto	
ChatGPT y Bard		9	ChatGPT	7
ChatGPT y Perplexity		19	Bard	6
Bard y Perplexity		5	Perplexity	3

**Tabla 3** Análisis de respuestas

	ChatGPT		Bard		Perplexity	
	Aciertos (n = 83)	Fallos (n = 31)	Aciertos (n = 69)	Fallos (n = 45)	Aciertos (n = 77)	Fallos (n = 37)
<b>RAZ. LOG.</b>						
Sí	69 (83,1%)	23 (74,4%)	69 (100%)	45 (100%)	72 (93,5%)	37 (100%)
No	14 (16,9%)	8 (25,8%)	0 (0%)	0 (0%)	5 (6,5%)	0 (0%)
<b>INF. INT.</b>						
Sí	79 (95,2%)	29 (93,5%)	69 (100%)	45 (100%)	76 (98,7%)	37 (100%)
No	4 (8%)	2 (6,5%)	0 (0%)	0 (0%)	1 (1,3%)	0 (0%)
<b>INF. EXT.</b>						
Sí	41 (49,4%)	9 (29,1%)	66 (95,7%)	45 (100%)	44 (57,1%)	16 (43,3%)
No	42 (50,6%)	22 (70,9%)	3 (4,3%)	0 (0%)	33 (42,9%)	21 (56,7%)

( $p < 0,001$ ) también está presente en la información externa aportada favoreciendo a Bard respecto a sus competidores. La información interna muestra una diferencia significativa ( $p = 0,012$ ) entre Bard y ChatGPT a favor del primero. En la [tabla 3](#) se analizan las respuestas atendiendo a los aciertos y fallos.

El análisis estadístico de los subgrupos «Aciertos» y «Fallos» presenta diferencias significativas ( $p < 0,001$ ) en la información externa y en el razonamiento lógico a favor de Bard. No hallamos diferencias significativas en el análisis de la información interna en los subgrupos que aciertan o fallan la pregunta.

Al comparar el número de palabras de cada respuesta, se objetiva que hay una diferencia significativa. El análisis de las diferencias por pares se muestra en la [tabla 4](#).

Esta diferencia significativa también se ha obtenido en el análisis del índice modificado de «lecturabilidad» de Fernández Huerta y en la escala INFLESH. Las comparaciones entre grupos se exponen en la [tabla 5](#). Para mostrar la relación entre las puntuaciones y los niveles de «lecturabilidad», se facilita en la tabla los diferentes niveles de cada escala.

Las respuestas erradas son 31 en ChatGPT. La razón es por error lógico en 10 (32,2%), error de información en 13 (41,9%) ([fig. 2](#)), error estadístico en una (3,2%) y combinación de error lógico y de información 7 (22,5%).

Bard obtuvo 45 respuestas fallidas. Por error lógico en 3 (6,6%), error de información en 3 (6,6%), error estadístico en 1 (2,2%) ([fig. 3](#)) y combinación de error lógico y de información 38 (84,4%).

Perplexity tuvo 37 respuestas inexactas, la causa fue por error lógico en 9 (24,3%) ([fig. 4](#)), error de información en 14

(37,8%), error estadístico en 2 (5,4%) y combinación de error lógico y de información 12 (32,4%).

## Discusión

Nuestros resultados indican que los programas de IA conversacional analizados (ChatGPT, Bard y Perplexity) aprueban el examen de preguntas de cirugía ortopédica y traumatología extraídos de la prueba MIR del periodo 2008-2023. La aplicación ChatGPT es mejor que Bard pero parecida en resultados a Perplexity. Nuestros datos mejoran los previos de JIN et al.<sup>15</sup> en exámenes norteamericanos similares y los de Carrasco et al.<sup>23</sup> que analizaron el examen MIR español de 2022 en donde se acertó un 54,8% de preguntas globales sin imágenes y que ascendió al 62,5% en el subgrupo de preguntas de traumatología. Esta mejoría hay que enmarcarla dentro del proceso de mejora continuo de estos sistemas. El análisis de aciertos atendiendo al tipo de pregunta no es concluyente, lo cual sugiere que responde de manera similar a preguntas con un razonamiento o varios, al menos con este tamaño muestral.

Nuestros resultados revelan que las 3 aplicaciones usan razonamiento lógico e información externa en gran número de respuestas correctas, destacando Bard. Sin embargo, esta ventaja no redundo en obtener los mejores resultados.

Las alucinaciones de la IA incluyen sesgos (pueden dar respuestas xenófobas p. ej.), equivocaciones u omisiones<sup>24</sup>. Este problema, inherente al diseño y a la estructura operativa del modelo, socava la reputación de la IA, afecta negativamente en la toma de decisiones y puede dar lugar a



**Tabla 4** Estadística palabras/respuesta

IA	Palabras	Tukey Post-Hoc Test – Palabras			
			ChatGPT	Bard	Perplexity
ChatGPT	98,4 ± 56,4	Diferencia de medias	—	–148	13,1
		Valor p	—	< 0,001*	0,143
Bard	246 ± 63,7	Diferencia de medias		—	160,8
		Valor p		—	< 0,001*
Perplexity	85,3 ± 30,8	Diferencia de medias			—
		Valor p			—

\* Significación estadística.

**Tabla 5** Legibilidad de las respuestas

	ChatGPT	Bard	Perplexity	FH	INFLESZ	ChatGPT	Bard	Perplexity
		FH					INFLESZ	
100				Muy fácil	Muy fácil			
90				Fácil				
80				Algo fácil	Bastante fácil			
70				Normal				
60		62,7 ± 8,83			Normal		57,9 ± 9,1	
50	58,2 ± 11,1		56,3 ± 10	Algo difícil	Algo difícil	53,2 ± 11,3		51,4 ± 10,3
40				Difícil				
30					Muy difícil			
20				Muy difícil				
10								
0								
Comparación entre parejas - FH				Comparación entre parejas - INFLESZ				
			p					p
ChatGPT	Bard		0,004*			ChatGPT	Bard	0,002*
ChatGPT	Perplexity		0,422			ChatGPT	Perplexity	0,507
Bard	Perplexity		< 0,001*			Bard	Perplexity	< 0,001*

FH: índice de Fernández Huerta.

\* Significación estadística.

conflictos éticos y legales<sup>25</sup>. Para evitar este problema, se ha recomendado emplear varias IA para reforzar la calidad de la respuesta. Nuestros resultados indican que casi la mitad de las preguntas han sido acertadas por las 3 IA simultáneamente, sin embargo, un 14% de preguntas fueron falladas por todas ellas. De modo que emplear varias IA no despeja completamente la presencia de alucinaciones.

Resulta interesante señalar que incluso las respuestas incorrectas incluyen razonamiento lógico y uso de información interna. En ChatGPT, era conocido que los errores tienen menos respaldo de estos 2 factores<sup>22</sup>. En el caso de Bard y Perplexity, aparecen en el 100% de los fallos, lo que

indicaría que es un modelo que apuesta por justificar su información a base de generar respuestas más completas y avaladas por la propia información de la pregunta. Respecto del uso de la información externa, destaca Bard aportando tanto en aciertos como en fallos. En las 3, la información externa aparece más en aciertos que en los fallos, especialmente en ChatGPT<sup>22,23</sup>. Como se señaló previamente, Perplexity incluye citas a otras páginas, pero el propósito de este trabajo es conocer la respuesta que da la aplicación y no analizar sus fuentes. La presencia de citas puede considerarse de gran valor en el respaldo de una respuesta, sin embargo, está fuera del objetivo de este trabajo.

MIR 2019

198. Hombre de 45 años que acude a consulta por dolor en el hombro izquierdo de 6 meses de evolución que ha aumentado de intensidad en las dos últimas semanas. No relaciona el dolor con traumatismo previo. El dolor es de predominio nocturno y empeora con el movimiento. En la exploración presenta un rango de movilidad pasiva completa con disminución de los últimos grados de rotación interna y elevación activa. No se objetiva deformidad, crepitación ni aumento de temperatura. El diagnóstico más probable es:

1. Inestabilidad de hombro asociada a laxitud.
2. Rotura del manguito de los rotadores del hombro.
3. Capsulitis adhesiva de hombro.
4. Tendinitis calcificante del tendón supraespinoso del hombro.

ChatGPT

Dado el cuadro clínico descrito, el diagnóstico más probable es **Capsulitis adhesiva de hombro**. Este trastorno se caracteriza por dolor en el hombro, especialmente en la noche, y pérdida progresiva de la movilidad articular. La limitación de la rotación interna y la elevación activa son hallazgos comunes en la capsulitis adhesiva. Sin embargo, la evaluación clínica completa y, en algunos casos, estudios de imagen adicionales pueden ser necesarios para confirmar el diagnóstico.

Figura 2 Ejemplo de error de información.

MIR 2009

91. De todas estas fracturas de etiología osteoporótica, señale cuál es la más frecuente:

1. Fractura de cadera.
2. Fractura de extremidad distal de radio.
3. Fractura de tobillo.
4. Fractura de extremidad proximal de hombro.
5. Fractura vertebral.

BARD

La respuesta correcta es (1), fractura de cadera.

Las fracturas osteoporóticas más frecuentes son las fracturas de cadera, seguidas de las fracturas vertebrales. Las fracturas de cadera son las más graves, ya que pueden provocar discapacidad permanente.

Las fracturas de extremidad distal de radio, fractura de tobillo, fractura de extremidad proximal de hombro y fractura vertebral también son frecuentes en pacientes con osteoporosis, pero menos que las fracturas de cadera.

En concreto, las fracturas de cadera representan el 60% de todas las fracturas osteoporóticas. Las fracturas vertebrales representan el 25% de todas las fracturas osteoporóticas. Las otras fracturas osteoporóticas representan el 15% restante.

Por lo tanto, la respuesta correcta es (1), fractura de cadera.

Figura 3 Ejemplo de error estadístico.

Mujer de 70 años, diabética e hipertensa que sufre una caída en su domicilio, presentando una herida de 9 cm que comunica con un foco de fractura de la tibia derecha. Radiográficamente se observa una fractura oblicua corta de tercio medio-distal de tibia. Se opera de urgencias mediante limpieza (Friedrich) y colocación de un clavo endomedular acerrojado. A los 11 meses presenta una pseudoartrosis atrófica de tibia con supuración en la zona de la herida. ¿Cuál será su mejor opción terapéutica inmediata?

1. Triple antibioterapia (grampositivos, gramnegativos y anaerobios) y limpieza de la herida quirúrgica, retirando los cerrojos distales para favorecer la consolidación ósea.
2. Actitud expectante y tratamiento antibiótico con quinolonas.
3. Aporte de injerto autólogo y de factores de crecimiento (BMP 2 y 7) para estimular el proceso de consolidación ósea, que está retardado.
4. Retirada del clavo, desbridamiento, colocación de fijador externo y antibioterapia ajustada a los resultados de los cultivos.

#### Respuesta

La mejor opción terapéutica inmediata para la paciente de 70 años con pseudoartrosis atrófica de tibia y supuración en la zona de la herida, tras sufrir una caída en su domicilio y ser operada de urgencias mediante limpieza y colocación de un clavo endomedular acerrojado, es la siguiente:

1. Triple antibioterapia (grampositivos, gramnegativos y anaerobios) y limpieza de la herida quirúrgica, retirando los cerrojos distales para favorecer la consolidación ósea <sup>1</sup>.

Esta opción terapéutica se basa en el tratamiento con antibióticos de amplio espectro para cubrir diferentes tipos de bacterias, junto con la limpieza de la herida quirúrgica y la retirada de los cerrojos distales para favorecer la consolidación ósea. Dado el cuadro de pseudoartrosis atrófica de tibia con supuración, es crucial abordar la infección y favorecer la consolidación ósea para el manejo efectivo de la complicación.

Las otras opciones no abordan de manera integral la infección y la pseudoartrosis, por lo que no serían las más adecuadas en este caso.

 Compartir  Reescribir

  ...

Figura 4 Ejemplo de error lógico.

La legibilidad del texto es fundamental en los «chatbot» generativos que ofrecen información médica. Sabemos que mejora con frases con menos palabras y más cortas<sup>21</sup>, aunque también se relaciona con la complejidad de las palabras

o presencia de representaciones visuales<sup>26</sup>. Es interesante señalar que Bard ofrece mejor capacidad de comprensión que las otras aplicaciones analizadas. Además, añade en muchas ocasiones imágenes para mejorar la información.

Las aplicaciones ChatGPT y Perplexity producen respuestas «algo difíciles» lo cual puede afectar a la interacción dialógica con usuarios no formados.

Este trabajo tiene limitaciones. En primer lugar, las preguntas del examen MIR no son comparables a las preguntas que podría hacer un usuario de IA. El objetivo de este trabajo no es tanto saber si aprobaría un examen, sino cómo responde a la pregunta y cómo ofrece mayor o menor información. En segundo lugar, sabemos que los «chatbot» permiten modular la respuesta según las indicaciones que le demos, mediante órdenes o «prompt», lo cual permite mejorarla en términos de cantidad o calidad, pero hemos evitado emplear esta modulación para que la respuesta sea lo más «espontánea» posible. En tercer lugar, se han empleado 2 sistemas de evaluación de tipo de respuesta frente a otros existentes. La decisión se ha basado en que habían sido contrastados para el idioma español y para textos médicos sin estudios comparativos entre ellos. También se puede aducir como limitación que el sistema ChatGPT 3.5 no tiene acceso a información de Internet a partir del 2021 frente a los otros 2 sistemas, pero entendemos que la mayoría de la población accede a las aplicaciones que no son de pago por lo que la comparación es objetiva y además las preguntas analizadas no incluyen datos que requieran información a partir del 2021. Es interesante señalar que sistemas con más tamaño de respuestas, legibilidad e información externa no han conseguido mejores resultados.

Este trabajo da pie a seguir investigando este tema tan novedoso e interesante. Sugerimos investigar cómo es la interacción dialógica de pacientes con sus cuestiones médicas o cómo un chat puede explicar la información médica que ofrecemos a nuestros pacientes o cómo la modulación de las preguntas mediante los «prompts» puede mejorar la calidad de las respuestas. Asimismo, se podría comparar la fiabilidad de estos programas frente a estudiantes o personal médico, tanto en formación como en su actividad.

En conclusión, los «chatbot» conversacionales pueden ser un instrumento muy interesante para resolver cuestiones médicas, pero no están exentos de cometer errores, o «alucinaciones», que pueden tener importantes implicaciones para los pacientes y médicos. Se debe advertir de su uso en la población general no formada y recordar que la información médica producida por profesionales médicos debe prevalecer jerárquicamente sobre la elaboración de una IA generativa.

## Financiación

No ha habido apoyo financiero para este artículo.

## Consideraciones éticas

El trabajo no se ha realizado sobre humanos o animales y no requiere consentimiento informado. Ni precisa conformidad del Comité de Ética.

## Conflicto de intereses

Los autores declaran no tener conflicto de intereses.

## Nivel de evidencia

Nivel de evidencia IV.

## Bibliografía

1. Lisacek-Kiosoglous AB, Powling AS, Fontalis A, Gabr A, Mazomenos E, Haddad FS. Artificial intelligence in orthopaedic surgery: Exploring its applications, limitations, and future direction. *J Bone Joint Res.* 2023;12:447–54, <http://dx.doi.org/10.1302/2046-3758.127>.
2. Benavent Núñez D, Colomer Mascaró J, Quecedo Gutiérrez L, Gol-Montserrat J, del Llano Señaris JE. *Inteligencia artificial y decisiones clínicas: Cómo está cambiando el comportamiento del médico.* Madrid: Fundación Gaspar Casal; 2020.
3. Mayol J. Inteligencia artificial generativa y educación médica. *Educación Médica.* 2023;4:1–3, <http://dx.doi.org/10.1016/j.edumed.2023.100851>.
4. Chung K, Park RC. Chatbot-based healthcare service with a knowledge base for cloud computing. *Cluster Comput.* 2019;22 Suppl 1:1925–37, <http://dx.doi.org/10.1007/s10586-018-2334-5>.
5. García-Vidal C, Sanjuan G, Puerta-Alcalde P, Moreno-García E, Soriano A. Artificial intelligence to support clinical decision-making processes. *EBioMedicine.* 2019;46:27–9, <http://dx.doi.org/10.1016/j.ebiom.2019.07.019>.
6. Marlow R, Wood D. Ghost in the machine or monkey with a typewriter-generating titles for Christmas research articles in The BMJ using artificial intelligence: observational study. *BMJ.* 2021;375:e067732, <http://dx.doi.org/10.1136/bmj-2021-067732>.
7. Cárdenas J. Inteligencia artificial, investigación y revisión por pares: escenarios futuros y estrategias de acción. *RES.* 2023;32:1–99, <http://dx.doi.org/10.22325/fes/res.2023.184>.
8. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183:589–96, <http://dx.doi.org/10.1001/jamainternmed.2023.1838>.
9. Miles O, West R, Nadarzynski T. Health chatbots acceptability moderated by perceived stigma and severity: A cross-sectional survey. *Digit Health.* 2021;7, <http://dx.doi.org/10.1177/20552076211063012>, 20552076211063012.
10. Aedo-Martín D. Inteligencia artificial: futuro y desafíos en la medicina moderna. *Rev Esp Cir Ortop Traumatol.* 2024, <http://dx.doi.org/10.1016/j.recot.2024.01.020>. S1888-4415(24)00047-X.
11. 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;6:1169595, <http://dx.doi.org/10.3389/frai.2023.1169595>.
12. Encuesta sobre equipamiento y uso de tecnologías de la información y comunicación en los hogares. Instituto Nacional de Estadística (INE) [actualizada 16 Dic 2022; consultada 15 Mar 2023]. Disponible en: [www.ine.es](http://www.ine.es)
13. Pasquinelli M. Cómo una máquina aprende y falla – Una gramática del error para la Inteligencia Artificial. *Hipertextos.* 2022;10:13–29, <http://dx.doi.org/10.24215/23143924e054>.
14. Clement ND, Simpson AHRW. Artificial intelligence in orthopaedics: What level of evidence does it represent and how is it validated? *Bone Joint Res.* 2023. p. 494–6.
15. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *App Sci.* 2021;11, <http://dx.doi.org/10.3390/app11146421>.



16. Trabasso T. The development of coherence in narratives by understanding intentional action. *Advances in Psychology*. 1991;79:297–314, [http://dx.doi.org/10.1016/S0166-4115\(08\)61559-9](http://dx.doi.org/10.1016/S0166-4115(08)61559-9).
17. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32:221–33, <http://dx.doi.org/10.1037/h0057532>.
18. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Institute for Simulation and Training. 1975;56:8-75.
19. Fernández Huerta J. Medidas sencillas de lecturabilidad. *Consigna*. 1959;214:29–32.
20. Law G. Error in the Fernandez Huerta Readability Formula [consultada 15 Mar 2023]. Disponible en: <https://linguistlist.org/issues/22/22-2332/>
21. Barrio-Cantalejo IM, Simón-Lorda P, Melguizo M, Escalona I, Marijuán MI, Hernando P. Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. *An Sist Sanit Navar*. 2008;31:135–52.
22. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312, <http://dx.doi.org/10.2196/45312>.
23. 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 Educ Méd*. 2023;4:55–9, <http://dx.doi.org/10.6018/edumed.556511>.
24. Van Deemter K, Reiter E. Lying and computational linguistics. En: Meibauer Jörg, editor. *The Oxford Handbook of Lying Oxford Handbooks*; 2018. p. 420–35, <http://dx.doi.org/10.1093/oxfordhb/9780198736578.013.32>.
25. Athaluri SA, Manthena SV, Kesapragada VKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*. 2023;15:e37432, <http://dx.doi.org/10.7759/cureus.37432>.
26. Blanco Pérez A, Gutierrez Couto U. Legibilidad de las páginas web sobre salud dirigidas a pacientes y lectores de la población general. *Rev Esp Salud Pública*. [online]. 2002;76:321–31. Disponible en: <https://scielo.isciii.es/pdf/resp/v76n4/original5.pdf>