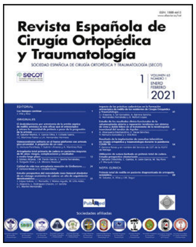
Revista Española de Cirugía Ortopédica y Traumatología 69 (2025) T38–T46

Sociedad Española de Cirugía Ortopédica y Traumatología

Revista Española de Cirugía Ortopédica y Traumatología

www.elsevier.es/rot



ORIGINAL PAPER

[Translated article] Exploring the potential of artificial intelligence in traumatology: Conversational answers to specific questions

F. Canillas del Rey [a,b,*], **M. Canillas Arias** [b]

[a] *Servicio de Cirugía Ortopédica y Traumatología, Hospital Universitario Cruz Roja, Madrid, Spain*
[b] *Facultad de Medicina, Universidad Alfonso X el Sabio, Madrid, Spain*

Received 19 March 2024; accepted 15 May 2024
Available online 7 November 2024

KEYWORDS
Generative artificial intelligence;
Chatbot;
Multi-choice question answering;
ChatGPT;
Bard;
Perplexity

Abstract

Background and objective: 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 three 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/).

DOI of original article: https://doi.org/10.1016/j.recot.2024.05.004

[*] Corresponding author.
E-mail address: fercanillas@yahoo.es (F. Canillas del Rey).

https://doi.org/10.1016/j.recot.2024.11.005
1888-4415/© 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/).

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

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

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/).

Introduction

Artificial intelligence (AI) encompasses technological developments that emulate the cognitive abilities of humans. In orthopaedic surgery and traumatology, applications include image recognition and diagnosis, medical text records, rehabilitation and postoperative care, surgical training, and predictive algorithms.[1] In recent years, AI has gone a step further by becoming generative AI. That is to say, it no longer simply analyses problems and solves them, but with the data it is provided with, it improves its learning and generates original content (text, images, videos, presentations, molecules, etc.).[2]

Recently, chatbots or conversational bots have become popular. These are generative AIs that have the ability to maintain a conversation, giving coherent and human-like responses. Their behaviour can even be modulated when responding, lending a certain amount of ''personality'' to the responses.

In November 2022, the free conversational bot (or ''chatbot'') called ChatGPT (generative pre-trained transformer) by the company OpenAI (OpenAI, LLC, San Francisco, California, USA) developed in Python language was launched. Its main limitation was that it only had access to the Internet until 2021, but the new (paid) version has already updated access to the network. In addition, it is multimodal, allowing text and image input to generate response.

In response to this application, several technology companies launched other ''chatbots'' onto the market, such as Bing by Microsoft (Microsoft Corporation, Redmond, Washington, USA) or Bard by Google (Google LLC, Mountain View, California, USA) or Perplexity by designers Denis Yarats, Aravind Srinivas, Johnny Ho and Andy Konwinsk. The latter has the particularity of providing citations to support its information.

These ground-breaking technological developments are creating a range of options yet to be explored and have enormous implications in the fields of medicine and teaching.[3] This is not only because they offer answers to questions that are raised, but because they can generate information with multiple uses. Their capabilities continuously grow and they are even beginning to recognise voices, images and videos, resulting in exponential potential.

Patients use this technology to answer questions about medicine and health[4] and professionals to solve cases[5] or create scientific texts.[6,7] They even generate answers which contain more, better quality information, and with greater empathy than humans are capable of,[8] although some patients rarely accept its use as a substitute for professionals.[9] Obviously, this has important legal and ethical connotations, which involve responsibility in decision-making or in authorship of scientific production.[10] All the more so, when the presence of errors, known as ''hallucinations'',[11] is detected, and these are more common than what a sensitive-data-generating ''intelligence'' should supposedly produce. We should remember that, in 2022, 40% of Internet users turned to the Internet in search of information on health issues.[12]

The development of conversational AI is based on three pillars: machine learning, big data and natural language processing.[2] Machine learning enables computers to automatically learn and improve from experience without having been specifically programmed to do so. Data is analysed using algorithms to identify patterns and make decisions with varying degrees of human supervision. The machine architecture uses neural networks that mimic the functioning of the human brain. This enables deep learning by processing vast amounts of data, or ''big data'', which includes unstructured or unlabelled data such as images, audio and text, to perform tasks such as voice or image recognition. Version 3 of ChatGPT was trained with 175 billion parameters. This means that the AI does not search the network for information, but rather generates responses by following predictive models from the information it has collected and processed into smaller units (known as ''tokens''). In this sense, there are authors who advocate talking about ''computational statistical learning'' instead of ''artificial intelligence''.[13] This explains why AI has failures, or ''hallucinations'' in computer jargon, largely due to its ability or lack thereof to handle the data it feeds on and generate results following stochastic patterns. If we add to this the effect known as GIGO (Garbage In Garbage Out)[7] we are looking at a system that requires a learning and supervision process.

Natural language processing is another area of AI that deals with the interaction between machine and human language. Thanks to this application, text can be understood, interpreted and generated efficiently. All of this leads to processing the order or ''prompt'', ''understanding'' it and responding in a natural and suposedly correct way. It also adds the ability to remember previous conversations.

In the field of medicine,[14] huge amounts of electronic medical records can be processed, enabling them to be analysed and organised to obtain efficient and accurate information. In the surgical field,[1] AI can provide the simulation of complex surgical procedures in virtual environments. Moreover, patterns of success and areas of improvement in surgical training can be identified, improving the quality and safety of medical care.

The objective of this study was to study and compare the capacity of conversational AI to solve orthopaedic surgery and traumatology questionnaires used in the national exam to obtain a position as a resident physician.

Material and method

All questions from the Spanish medical training entrance exam (MIR exam) from 2008 to 2023 were reviewed. The inclusion criteria were: questions relating to traumatology and orthopaedic surgery. The exclusion criteria were: questions cancelled by the governing body or those containing images that were necessary for answering the question.

The questions were multiple choice with four or five answer options and only one correct answer. Some of the questions included images. Since not all applications have image recognition, those that required perusal of the image to answer were excluded. The questions were grouped according to the subject in traumatology, adult orthopaedics, paediatric orthopaedics and spine, and according to the year. In addition, they were classified into two types according to the knowledge needed to give the answer[15]: type 1, only one piece of knowledge was required and type 2, several steps were needed to reach the answer.

Figure 1 Logos of the ''chatbot'' conversations ChatGPT, Bard and Perplexity.

The answers were analysed by assessing narrative coherence.[16] The data were binary and included three sections: logical reasoning (if the answer was selected based on the information presented), internal information (the answer included information provided in the question) and external information (the answer provided information external to that provided). We also analysed the readability of Flesch–Kincaid,[17,18] adapted to Spanish by Fernández Huerta[19] and corrected by Law.[20] The result shows the ease of understanding a text and correlates it with the level of education as a result of this formula:

$$L = 206.84 - 0.60\,P - 1.02\,F$$

L is ''readability''; P, the average number of syllables per word; F, the average number of words per sentence.

The INFLESZ[21] scale was also used, which measures the ease of reading a text, as well as the number of words in each response.

When errors occurred, they were labelled as logical errors (correct information is found but not transferred to the response), information errors (a key piece of information provided in the question or in external information is not identified) and statistical errors (based on arithmetic errors, e.g. incorrect estimation of the frequency of a disease).[22]

The AI engines evaluated were ChatGPT (version 3.5), Bard and Perplexity (Fig. 1). All responses were obtained within a period of 48 h. When writing this work, the Bard application was renamed Gemini, but since the results were obtained with the first name, it was kept in the results and conclusions.

The data collected were quantitative, qualitative and descriptive. Qualitative variables are presented with their absolute frequency and relative percentage. To compare qualitative values of a dichotomous nature, Cochran's Q test was used, and for quantitative values, the ANOVA test was used if it met normality and the Kruskal–Wallis test if it did not. The accepted alpha risk for all hypothesis contrasts was .05. If the test was significant, a post-hoc test was performed for pairwise comparison of variables. The data were analysed using the MedCalc program version 22.016 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2023) for data processing and statistical study. The null hypothesis (H0) that we assumed was that the success rate of the different programmes evaluated was the same.

Table 1 Frequency of topic/year.

Year	Topic	Frequencies	% of the total
2009	Trauma	6	5.3%
	OrthoAdult	2	1.8%
	OrthoChild	2	1.8%
	Spine	1	.9%
2010	Trauma	2	1.8%
	OrthoAdult	2	1.8%
	OrthoChild	2	1.8%
	Spine	1	.9%
2011	Trauma	4	3.5%
	OrthoAdult	4	3.5%
	OrthoChild	2	1.8%
	Spine	0	.0%
2012	Trauma	0	.0%
	OrthoAdult	2	1.8%
	OrthoChild	1	.9%
	Spine	0	.0%
2013	Trauma	2	1.8%
	OrthoAdult	3	2.6%
	OrthoChild	0	.0%
	Spine	1	.9%
2014	Trauma	1	.9%
	OrthoAdult	1	.9%
	OrthoChild	1	.9%
	Spine	0	.0%
2015	Trauma	4	3.5%
	OrthoAdult	2	1.8%
	OrthoChild	0	.0%
	Spine	0	.0%
2016	Trauma	3	2.6%
	OrthoAdult	2	1.8%
	OrthoChild	1	.9%
	Spine	1	.9%
2017	Trauma	4	3.5%
	OrthoAdult	3	2.6%
	OrthoChild	0	.0%
	Spine	1	.9%
2018	Trauma	6	5.3%
	OrthoAdult	3	2.6%
	OrthoChild	1	.9%
	Spine	0	.0%
2019	Trauma	1	.9%
	OrthoAdult	3	2.6%
	OrthoChild	1	.9%
	Spine	0	.0%
2020	Trauma	5	4.4%
	OrthoAdult	3	2.6%
	OrthoChild	1	.9%
	Spine	1	.9%
2021	Trauma	1	.9%
	OrthoAdult	7	6.1%
	OrthoChild	1	.9%
	Spine	2	1.8%
2022	Trauma	2	1.8%
	OrthoAdult	2	1.8%
	OrthoChild	1	.9%
	Spine	2	1.8%
2023	Trauma	3	2.6%
	OrthoAdult	3	2.6%
	OrthoChild	2	1.8%
	Spine	2	1.8%

Results

Initially, 129 questions were collected from the MIR exams related to orthopaedic surgery and traumatology. Fifteen were excluded because they included images (14 questions) and one because it had been cancelled. Table 1 shows the distribution of the questions by year and topic.

According to the type of question, there were 49 (43%) of type I (requiring one piece of knowledge) and 65 (57%) of type II (requiring several to obtain the answer).

The ChatGPT application obtained 83 correct answers (72.81%), compared to 69 (60.53%) for Bard and 77 (67.54%) for Perplexity. We found statistical significance in the Cochran *Q* test (0.049) regarding the frequency distribution between the groups, and when the post-hoc analysis was performed, a significant difference was only found in ChatGPT compared to Bard, but not between the other pairs. If we group the answers according to the type of question, we observe that in type I ChatGPT gets 36 (73.47%) correct, Bard gets 27 (55.10%) and Perplexity 36 (73.47%) which shows a significant difference (.030) but in the multiple comparisons it does not find differences between pairs due to the sample size of the subgroup. In type II questions we do not find significant differences between groups, ChatGPT obtained 47 (72.31%) correct answers, compared to 42 (64.62%) for Bard and 41 (63.08%) for Perplexity.

Table 2 shows the accumulation of correct answers for each of the questions when answering the three ''chatbots''.

In the analysis of each programme's responses, ChatGPT used logical reasoning in 92 (80.7%), internal information in 108 (94.7%) and external information in 50 (43.8%). In Bard, logical reasoning was used in 114 (100%) responses, internal information in 114 (100%) and external information in 111 (97.3%). In Perplexity, logical reasoning was used in 109 (95.6%) responses, internal information in 113 (99.1%) and external information in 60 (52.6%). In this last section, we noted that this application included citations that were linked to web pages but external information was not considered since it would require an evaluation that differed from the purpose of this study. When analysing logical reasoning, we found statistical significance ($p < .001$) in favour of Bard compared to the rest. This difference ($p < .001$) is also present in the external information provided, favouring Bard over its competitors. The internal information shows a significant difference ($p = .012$) between Bard and Chat-

Table 2 Number of chatbot correct answers.

Three	Two	One	None	Questions
49	33	16	16	114
43%	28.90%	14%	14%	

Two correct answers		One correct answer	
ChatGPT and Bard	9	ChatGPT	7
ChatGPT and Perplexity	19	Bard	6
Bard and Perplexity	5	Perplexity	3

Table 3 Analysis of answers.

	ChatGPT		Bard		Perplexity	
	Correct ($n = 83$)	Incorrect ($n = 31$)	Correct ($n = 69$)	Incorrect ($n = 45$)	Correct ($n = 77$)	Incorrect ($n = 37$)
Logical reasoning						
Yes	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%)
Internal information						
Yes	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%)
External information						
Yes	41 (49.45)	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%)

GPT in favour of the former. Table 3 analyses the responses according to the correct and incorrect answers.

Statistical analysis of the subgroups ''Correct'' and ''Failure'' flagged up significant differences ($p < .001$) in the external information and in the logical reasoning in favour of Bard. We found no significant differences in the analysis of the internal information in the subgroups that get the question right or wrong.

Comparison between the number of words in each answer clarified a significant difference. The analysis of the pairwise differences is shown in Table 4.

This significant difference was also obtained in the analysis of the modified index of ''readability'' of Fernández Huerta and in the INFLESZ scale. The comparisons between groups are shown in Table 5. To show the relationship between the scores and the levels of ''readability'', the different levels of each scale are provided in the table.

There are 31 wrong answers in ChatGPT. The reason is logical error in 10 (32.2%), information error in 13 (41.9%) (Fig. 2), statistical error in one (3.2%) and combination of logical and information error in 7 (22.5%).

Bard had 45 incorrect answers. Logical error in 3 (6.6%), information error in 3 (6.6%), statistical error in 1 (2.2%) (Fig. 3) and combination of logical and information error in 38 (84.4%).

Perplexity had 37 inaccurate answers, the cause was logical error in 9 (24.3%) (Fig. 4), information error in 14 (37.8%), statistical error in 2 (5.4%) and combination of logical and information error in 2 (32.4%).

Discussion

Our results indicate that the conversational AI programmes analysed (ChatGPT, Bard and Perplexity) pass the exam with orthopaedic surgery and traumatology questions taken from the MIR exam for the period 2008–2023. The ChatGPT application is better than Bard but similar in results to Perplexity. Our data improves on the previous data of Jin et al.[15] in similar North American exams and those of Carrasco et al.[23] who analysed the Spanish MIR exam of 2022 where 54.8% of global questions without images were answered correctly and which rose to 62.5% in the subgroup of traumatology questions. This improvement relies on the process of continuous improvement these systems are prone to. The analysis of correct answers according to the type of question is inconclusive, which suggests that the app responds similarly to questions with one, or several, reasons, at least with this sample size.

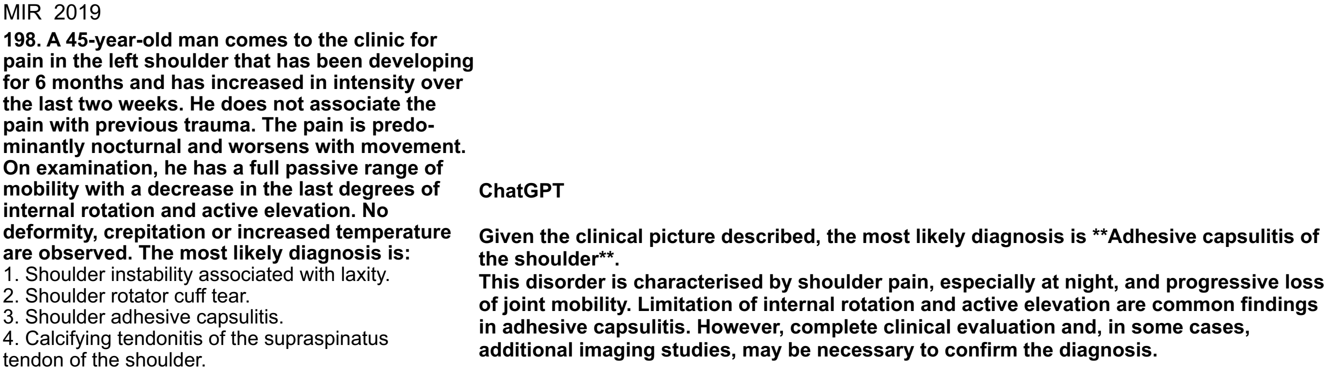
Our results reveal that all three apps use logical reasoning and external information in a large number of their correct answers, with Bard ranking first here. However, this advantage does not translate into obtaining the best results.

AI hallucinations include biases (they can give xenophobic answers, for example), mistakes or omissions.[24] This problem, inherent to the design and operational structure of the model, undermines the reputation of the AI, negatively affects decision-making and can lead to ethical and legal conflicts.[25] To avoid this problem, it has been recommended to use several AIs to reinforce the quality of the

Table 4 Word/answer statistics.

AI	Words	Tukey post-hoc test – words			
			ChatGPT	Bard	Perplexity
ChatGPT	98.4 ± 56.4	Mean differences	–	−148	13.1
		p value	–	<.001*	.143
Bard	246 ± 63.7	Mean differences		–	160.8
		p value		–	<.001*
Perplexity	85.3 ± 30.8	Mean differences			–
		p value			–

* Statistical significance.


MIR 2019

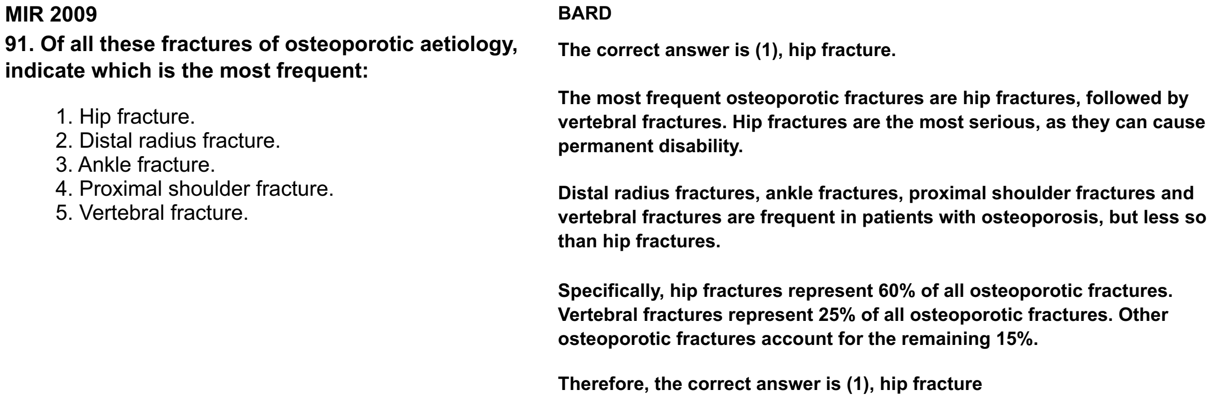
198. A 45-year-old man comes to the clinic for pain in the left shoulder that has been developing for 6 months and has increased in intensity over the last two weeks. He does not associate the pain with previous trauma. The pain is predominantly nocturnal and worsens with movement. On examination, he has a full passive range of mobility with a decrease in the last degrees of internal rotation and active elevation. No deformity, crepitation or increased temperature are observed. The most likely diagnosis is:

1. Shoulder instability associated with laxity.
2. Shoulder rotator cuff tear.
3. Shoulder adhesive capsulitis.
4. Calcifying tendonitis of the supraspinatus tendon of the shoulder.

ChatGPT

Given the clinical picture described, the most likely diagnosis is **Adhesive capsulitis of the shoulder.**
This disorder is characterised by shoulder pain, especially at night, and progressive loss of joint mobility. Limitation of internal rotation and active elevation are common findings in adhesive capsulitis. However, complete clinical evaluation and, in some cases, additional imaging studies, may be necessary to confirm the diagnosis.

Figure 2 Example of information error.


MIR 2009

91. Of all these fractures of osteoporotic aetiology, indicate which is the most frequent:

1. Hip fracture.
2. Distal radius fracture.
3. Ankle fracture.
4. Proximal shoulder fracture.
5. Vertebral fracture.

BARD

The correct answer is (1), hip fracture.

The most frequent osteoporotic fractures are hip fractures, followed by vertebral fractures. Hip fractures are the most serious, as they can cause permanent disability.

Distal radius fractures, ankle fractures, proximal shoulder fractures and vertebral fractures are frequent in patients with osteoporosis, but less so than hip fractures.

Specifically, hip fractures represent 60% of all osteoporotic fractures. Vertebral fractures represent 25% of all osteoporotic fractures. Other osteoporotic fractures account for the remaining 15%.

Therefore, the correct answer is (1), hip fracture

Figure 3 Example of statistical error.

answer. Our results indicate that almost half of the questions were answered correctly by the three AIs simultaneously. However, 14% of the questions were incorrectly answered by all of them. The use of several AIs does not therefore completely eliminate the presence of hallucinations.

It is interesting to note that incorrect answers include logical reasoning and use of internal information. In ChatGPT it was known that the errors had less support from these two factors.[22] In the case of Bard and Perplexity, they appear in 100% of the errors, which would indicate that this model justifies its information based on providing more complete answers backed up by the actual information from the question. Regarding the use of external information, Bard stands out for providing as many correct answers as erroneous ones. In all three, external information appears more in correct answers than in incorrect ones, particularly in ChatGPT.[22,23] As previously stated, Perplexity includes citations to other sites, but the objective of this study was to determine the response the app provided, not to analyse its sources. The provision of citations could be considered of great value in supporting a response, but it outside the realm of this study.

Text legibility is essential in generative ''chatbots'' that offer medical information. We know that the chatbot improves with shorter sentences containing fewer words,[21] although it is also related to the complexity of the words or the presence of visuals.[26] It is interesting to note that Bard offers a better comprehension capacity than the other analysed apps. Furthermore, it often adds images to

Table 5 Legibility of answers.

	ChatGPT	Bard	Perplexity	FH	INFLESZ	ChatGPT	Bard	Perplexity
100		FH					INFLESZ	
				Muy fácil	Muy fácil			
90								
				Fácil				
80								
				Algo fácil	Bastante fácil			
70								
		62.7 ± 8.83		Normal				
60					Normal		57.9 ± 9.1	
	58.2 ± 11.1		56.3 ± 10	Algo difícil				
50					Algo difícil	53.2 ± 11.3		51.4 ± 10.3
				Difícil				
40								
					Muy difícil			
30								
				Muy difícil				
20								
10								
0								

Comparación entre parejas - FH

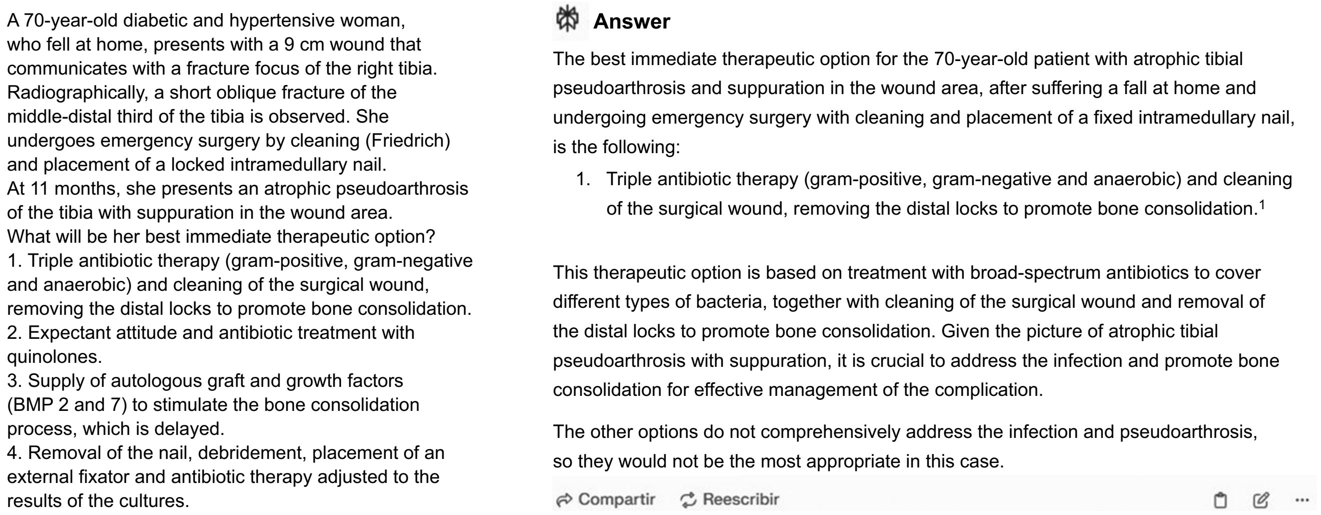
		p
ChatGPT	Bard	.004*
ChatGPT	Perplexity	.422
Bard	Perplexity	<.001*

Comparación entre parejas - Inflesz

		p
ChatGPT	Bard	.002*
ChatGPT	Perplexity	.507
Bard	Perplexity	<.001*

FH: Fernández Huerta index.

[*] Statistical significance.



A 70-year-old diabetic and hypertensive woman, who fell at home, presents with a 9 cm wound that communicates with a fracture focus of the right tibia. Radiographically, a short oblique fracture of the middle-distal third of the tibia is observed. She undergoes emergency surgery by cleaning (Friedrich) and placement of a locked intramedullary nail.
At 11 months, she presents an atrophic pseudoarthrosis of the tibia with suppuration in the wound area.
What will be her best immediate therapeutic option?
1. Triple antibiotic therapy (gram-positive, gram-negative and anaerobic) and cleaning of the surgical wound, removing the distal locks to promote bone consolidation.
2. Expectant attitude and antibiotic treatment with quinolones.
3. Supply of autologous graft and growth factors (BMP 2 and 7) to stimulate the bone consolidation process, which is delayed.
4. Removal of the nail, debridement, placement of an external fixator and antibiotic therapy adjusted to the results of the cultures.

Answer

The best immediate therapeutic option for the 70-year-old patient with atrophic tibial pseudoarthrosis and suppuration in the wound area, after suffering a fall at home and undergoing emergency surgery with cleaning and placement of a fixed intramedullary nail, is the following:

1. Triple antibiotic therapy (gram-positive, gram-negative and anaerobic) and cleaning of the surgical wound, removing the distal locks to promote bone consolidation.[1]

This therapeutic option is based on treatment with broad-spectrum antibiotics to cover different types of bacteria, together with cleaning of the surgical wound and removal of the distal locks to promote bone consolidation. Given the picture of atrophic tibial pseudoarthrosis with suppuration, it is crucial to address the infection and promote bone consolidation for effective management of the complication.

The other options do not comprehensively address the infection and pseudoarthrosis, so they would not be the most appropriate in this case.

Figure 4 Example of logical error.

improve the information. The ChatGPT and Perplexity produce ''somewhat difficult'' responses, which may affect the dialogic interaction with untrained users.

This study has its limitations. Firstly, the MIR exam questions are incomparable with the questions an AI user could ask. The purpose of this study was not so much to know if they would pass an exam but how they would respond to the questions and how they would offer greater or lesser information. Secondly, we know that the ''chatbot'' allows the response to be modulated according to the indications we give it, using orders or prompts which improve it in terms of quantity or quality, but we have avoided using this modulation so that the response was as ''spontaneous'' as possible. Thirdly, two assessment systems out of other existing ones were used for the type of responses. The decision was based on the fact that they had been contrasted for the Spanish language and for medical texts without comparative studies between them. Another possible limitation is that ChatGPT 3.5 had no access to Internet information from 2021 onwards compared with the other two systems, but we understand that the majority of the population uses free apps and the comparison was therefore objective and also the questions analysed did not include data that required information from 2021 onwards. It is interesting to note that systems with a greater volume of answers, legibility and external information did not achieve better results.

This study provides the basis for further research into this novel and interesting topic. We suggest investigating the dialogic interaction of patients with their medical questions, or how a chat can explain the medical information we offer to our patients, or how the modulation of questions through prompts can improve the quality of responses. The reliability of these programmes could also be compared with students or medical personnel, both in training and in practice.

To conclude, conversational chatbots can be a very interesting tool for resolving medical issues, but they are not exempt from making mistakes, or ''hallucinations'', which can have important implications for patients and doctors. We should warn against their use in the general untrained population and remember that medical information produced by medical professionals must prevail hierarchically over the development of generative AI.

Level of evidence

Level of evidence IV.

Ethical considerations

The study was not conducted on humans or animals and did not require informed consent. Nor did it require approval from the Ethics Committee.

Funding

No funding was received for this article.

Conflict of interests

The authors have no conflict of interests to declare.

References

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 Señarís 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 heathcare 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. Garcia-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;68:428–9, http://dx.doi.org/10.1016/j.recot.2024.01.020.
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). Available from: www.ine.es [updated 16.12.22; consulted 15.03.23].
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;12:494–6.
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. Appl Sci. 2021;11, http://dx.doi.org/10.3390/app11146421.
16. Trabasso T. The development of coherence in narratives by understanding intentional action. Adv Psychol. 1991;79:297–314, 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, vol. 56. Institute for Simulation and Training; 1975. p. 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. Available from: https://linguistlist.org/issues/22/22-2332/ [consulted 15.03.23].
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. In: 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. Available from: https://scielo.isciii.es/pdf/resp/v76n4/original5.pdf