DOI: https://doi.org/10.30795/jfootankle.2024.v18.1757

Original Article

Artificial intelligence in foot and ankle pathology: Can large language models replace us?

Florencio Pablo Segura[1,2,3], Facundo Manuel Segura[1,2,3], Julieta Porta[4], Natalia Heredia[3], Ignacio Masquijo[3], Federico Anain[5], Leandro Casola[6], Agustina Trevisson[6], Virginia Cafruni[7], María Paz Lucero Zudaire[3], Ignacio Toledo[4], Florencio Vicente Segura[1,2]

1. Segura, Centro Privado de Ortopedia y Traumatología, Córdoba, Argentina.
2. Universidad Nacional de Córdoba, Córdoba, Argentina.
3. Instituto Modelo de Cardiología, Córdoba, Argentina.
4. Sanatorio Allende, Córdoba, Argentina.
5. Unidad de Pierna y Pie, Ciudad Autónoma de Buenos Aires, Argentina.
6. Sanatorio Dupuytren, Ciudad Autónoma de Buenos Aires, Argentina.
7. Hospital Italiano de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina.

Abstract

Objective: Determine if large language models (LLMs) provide better or similar information compared to an expert trained in foot and ankle pathology in various aspects of daily practice (definition and treatment of pathology, general questions).

Methods: Three experts and two artificial intelligent (AI) models, ChatGPT (GPT-4) and Google Bard, answered 15 specialty-related questions, divided equally among definitions, treatments, and general queries. After coding, responses were redistributed and evaluated by five additional experts, assessing aspects like clarity, factual accuracy, and patient usefulness. The Likert scale was used to score each question, enabling experts to gauge their agreement with the provided information.

Results: Using the Likert scale, each question could score between 5 and 25 points, totaling 375 or 75 points for evaluations. Expert 2 led with 69.86%, followed by Expert 1 at 68.53%, ChatGPT at 64.80%, Expert 3 at 58.40%, and Google Bard at 54.93%. Comparing experts, significant differences emerged, especially with Google Bard. The rankings varied in specific sections like definitions and treatments, highlighting GPT-4's variability across sections. The results emphasize the differences in performance among experts and AI models.

Conclusion: Our findings indicate that GPT-4 often performed comparably to or even better than experts, particularly in definition and general question sections. However, both LLMs lagged notably in the treatment section. These results underscore the potential of LLMs as valuable tools in orthopedics but highlight their limitations, emphasizing the irreplaceable role of expert expertise in intricate medical contexts.

Evidence Level: III, observational, analytics.

Keywords: Large language models; Artificial intelligence; Foot and ankle; ChatGPT; Google Bard; Generative AI.

Introduction

The Turing test, proposed by Alan Turing in 1950, evaluates a machine's intelligence by observing whether it can generate responses indistinguishable from those provided by a human[1]. With the emergence of large language models (LLMs), such as ChatGPT (GPT-4) or Google Bard, demonstrating a significant ability to understand or coherently generate medical responses, Turing's work has gained paramount importance. These models have become powerful tools in various fields, including medicine[2].

The term artificial intelligence (AI) was first described by McCarthy et al. in 1955 when they referred to AI as "the science and engineering of making intelligent machines." They believed these machines would be capable of performing

Study performed at the Hospital Italiano de Buenos Aires. Ciudad Autónoma de Buenos Aires, Argentina.

Correspondence: Facundo Manuel Segura. **Email:** facusegura@gmail.com. **Conflicts of interest**: none. **Source of funding:** none. **Date received:** March 1, 2024. **Date accepted:** January 17, 2024. **Online:** April 30, 2024.

How to cite this article: Segura FP, Segura, FM, J, Heredia N, Masquijo I, Anain, Casola L. et al. Artificial intelligence in foot and ankle pathology: Can large language models replace us? J Foot Ankle. 2024;18(1):52-8.

tasks traditionally considered exclusive to humans, with the primary limitation being the speed and memory of programs. Jerrold S. Maxmen, a psychiatry professor at Columbia University, predicted that AI would usher in the "post-medical era" for the 21st century, describing the shift as "possible, inevitable, and desirable"[3].

Artificial intelligence is now regarded as the primary potential catalyst of the fourth industrial revolution, following the steam engines of the 1760s, the electricity and oil revolution of the 1870s, and computers of the 1970s[4].

GPT-4 and Google Bard represent the latest introductions in AI and have quickly found their place in healthcare services. Both models employ a hybrid language format that includes supervised learning and unsupervised or reinforcement learning with human feedback. They can provide an overview of existing literature on a specific topic within our specialty[5-6].

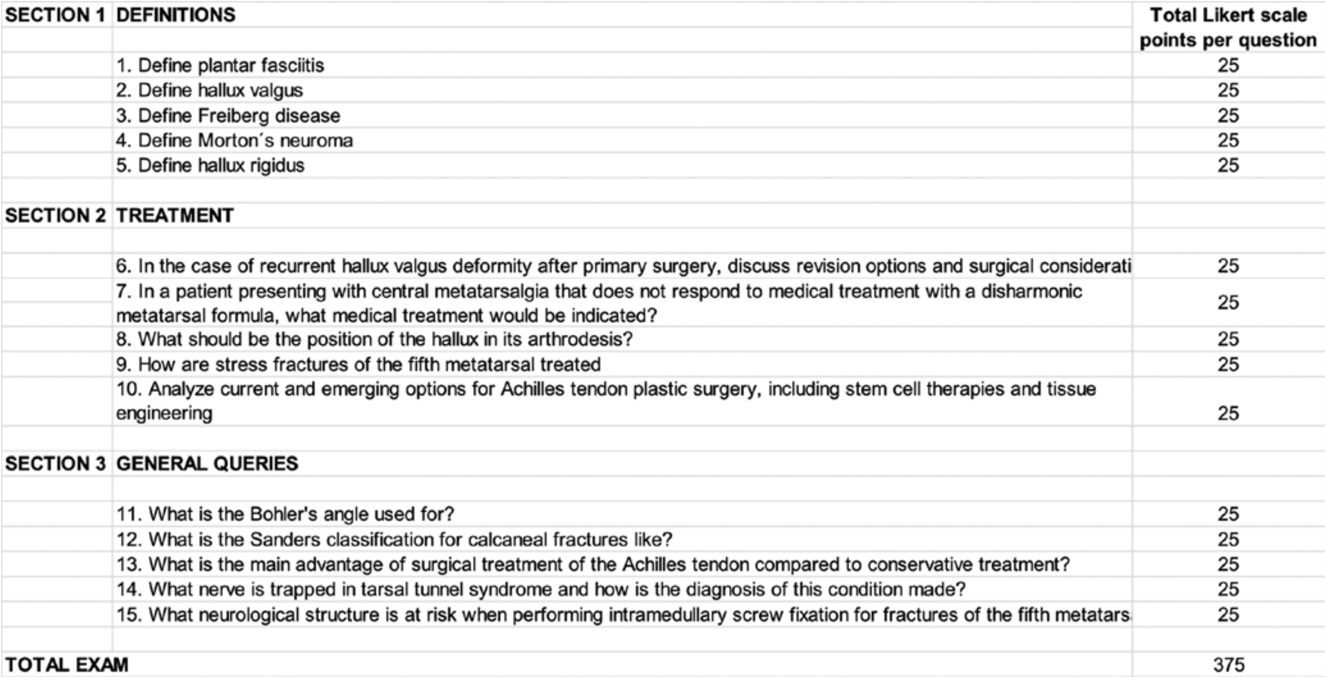
GPT-4 has been tested to pass high-level exams such as the United States Medical Licensing Examination (USMLE)[7] and the American Board of Orthopedic Surgery (ABOS)[8] exam. While there is no doubt about the high potential and capabilities of various AI tools like GPT-4 or Google Bard, there are several concerns regarding their application in medicine, particularly in orthopedics, and even more specifically in a subspecialty foot and ankle pathology.

The objective of this study is to examine the ability of LLMs to respond to medical queries related to foot and ankle pathology. The information provided by GPT-4/Google Bard will be compared and evaluated by various experts from the foot, ankle, and leg society. The accuracy, timeliness, and relevance of the information provided by GPT-4/Google Bard will be assessed.

Methods

In November 2023, three experts from Argentina Society of Medicine and Foot and Leg Surgery (SAMeCiPP) and two LLMs (GPT-4 and Google Bard) responded to 15 specialty-related questions. Five questions were definitions related to the specialty, five were treatments that should be performed based on a specific pathology, and the remaining five were general information queries. The questions are detailed in Figure 1. Five tests were conducted (3 experts/2 LLMs). The 15 responses were coded and redistributed into five new tests for evaluation. Five more experts from SAMeCiPP evaluated the responses from the experts and the LLMs without knowing to whom each question belonged (Table 1a). The instructions given to the experts and the LLMs are detailed in Table 1b.

For each response, five aspects were evaluated: 1) the provided information is comprehensive; 2) the provided response is confusing; 3) there are factual errors in the provided information; 4) the information is up-to-date; and 5) the response is a good source of information for the patient. The maximum possible score for each question was 25 points, and the minimum was 5. Each question was scored according to the Likert scale (Table 2)[9]. This scale allows the evaluator's experts to express their agreement or disagreement with provided statements or assertions, assigning a numerical value indicating their degree of agreement or disagreement.



SECTION 1	DEFINITIONS	Total Likert scale points per question
	1. Define plantar fasciitis	25
	2. Define hallux valgus	25
	3. Define Freiberg disease	25
	4. Define Morton´s neuroma	25
	5. Define hallux rigidus	25
SECTION 2	TREATMENT	
	6. In the case of recurrent hallux valgus deformity after primary surgery, discuss revision options and surgical considerati	25
	7. In a patient presenting with central metatarsalgia that does not respond to medical treatment with a disharmonic metatarsal formula, what medical treatment would be indicated?	25
	8. What should be the position of the hallux in its arthrodesis?	25
	9. How are stress fractures of the fifth metatarsal treated	25
	10. Analyze current and emerging options for Achilles tendon plastic surgery, including stem cell therapies and tissue engineering	25
SECTION 3	GENERAL QUERIES	
	11. What is the Bohler's angle used for?	25
	12. What is the Sanders classification for calcaneal fractures like?	25
	13. What is the main advantage of surgical treatment of the Achilles tendon compared to conservative treatment?	25
	14. What nerve is trapped in tarsal tunnel syndrome and how is the diagnosis of this condition made?	25
	15. What neurological structure is at risk when performing intramedullary screw fixation for fractures of the fifth metatars	25
TOTAL EXAM		375

Figure 1. Questions for experts and LLMs with the Likert scale score for each question and the total exam.
LLMs: Large language models.

Determine if the LLMs provide better or similar information compared to an expert trained in foot and ankle pathology in various aspects of daily practice (definition and treatment of pathology, general questions).

Results

The maximum possible score for each question was 25 points, and the minimum was 5, according to the Likert scale. Thus, the highest possible total score for each evaluation was 375 points, and the minimum was 75. The score obtained by each expert is described in Table 3, along with the percentage of the total possible points.

Expert 2 (E2) achieved the highest value from external evaluators, scoring 69.86% of the total score (269/375). Expert 1 (E1) scored 68.53%, representing 257/375, GPT-4 scored 64.80%, representing 243/375, Expert 3 (E3) scored 58.40%, representing 219/375, and Google Bard scored 54.93%, representing 206/375 (Figure 2). The results were compared among the evaluator's experts and between the experts and LLMs, as detailed in Table 3.

In the comparative analysis among the experts, no significant differences were observed between E1 and E2 (z-stat: -0.395, p-value: 0.692). However, notable differences were identified in the overall exam scores when comparing E1 with E3 (z-stat: 2.882, p-value: 0.004) and E2 with E3 (z-stat: -3.274, p-value: 0.001). No significant differences were found in the overall exam score between E1 and GPT-4 (z-stat: 1.084, p-value: 0.278) nor between E2 and GPT-4 (z-stat: 1.479, p-value: 0.139) when comparing the results between the experts and the LLMs. No significant differences were observed when comparing E3 with GPT-4 (z-stat: -1.802, p-value: 0.072). On the other hand, significant differences were found in all cases (E1 vs Bard: z-stat: 3.832, p-value: 0.0001; E2 vs Bard: z-stat: 4.222, p-value: 0.0001; E3 vs Bard: z-stat: 0.958, p-value: 0.338; GPT-4 vs Bard: z-stat: 2.756, p-value: 0.006) when comparing the results of all experts with Google Bard. These results suggest notable discrepancies in performance between the experts and Google Bard, while no significant differences were evident between the experts and GPT-4 in the overall exam.

If each set of questions (definition, treatment, and general) were considered separately, the maximum score (5 questions per set) would be 125 points, and the minimum score would be 25 points.

In the definitions section (Table 4), the same analysis was conducted for the entire exam. The total scores were compared (Figure 3), and the percentage of each exam between the experts and LLMs.

Table 1. Distribution of questions among evaluators to avoid biases in the exam evaluation

	Expert 1 Exam	Expert 2 Exam	Expert 3 Exam	ChatGPT Exam	Google Bard Exam
A	15 questions	15 questions	15 questions	15 questions	15 questions
	15 answers	15 answers	15 answers	15 answers	15 answers
	5 Q&A definitions	5 Q&A definitions	5 Q&A definitions	5 Q&A definitions	5 Q&A definitions
	5 Q&A treatment	5 Q&A treatment	5 Q&A treatment	5 Q&A treatment	5 Q&A treatment
	5 Q&A generalities	5 Q&A generalities	5 Q&A generalities	5 Q&A generalities	5 Q&A generalities
B	**Exam evaluator 1**	**Exam evaluator 2**	**Exam evaluator 3**	**Exam evaluator 4**	**Exam evaluator 5**
	15 answers	15 answers	15 answers	15 answers	15 answers
	3 definitions answers	3 definitions answers	3 definitions answers	3 definitions answers	3 definitions answers
	2 definitions answers	2 definitions answers	2 definitions answers	2 definitions answers	2 definitions answers
	3 treatment answers	3 treatment answers	3 treatment answers	3 treatment answers	3 treatment answers
	2 treatment answers	2 treatment answers	2 treatment answers	2 treatment answers	2 treatment answers
	3 generalities answers	3 generalities answers	3 generalities answers	3 generalities answers	3 generalities answers
	2 generalities answers	2 generalities answers	2 generalities answers	2 generalities answers	2 generalities answers

Table 2. Likert scale for the evaluation of 15 questions. Each question had a maximum score of 25 points and a minimum of 5 points.

1. The provided information is comprehensive	1. Strongly disagree	2. Disagree	3. Neither agree nor disagree	4. Agree	5. Strongly agree
2. The provided response is confusing	1. Strongly disagree	2. Disagree	3. Neither agree nor disagree	4. Agree	5. Strongly agree
3. There are factual errors in the provided information	1. Strongly disagree	2. Disagree	3. Neither agree nor disagree	4. Agree	5. Strongly agree
4. The information is up-to-date	1. Strongly disagree	2. Disagree	3. Neither agree nor disagree	4. Agree	5. Strongly agree
5. The response is a good source of information for the patient	1. Strongly disagree	2. Disagree	3. Neither agree nor disagree	4. Agree	5. Strongly agree

Table 3. Comparative table between the experts and LLMs (total exam). The comparative p-values where significant differences exist are highlighted.

	Total points	Percentage of exam	p-value < 0.05
Expert 1 (E1)	257 / 375	68.53 %	vs E2 = 0.692 **vs E3 = 0.004** vs CG = 0.578 **vs GB = 0.001**
Expert 2 (E2)	269 / 375	69.86 %	**vs E3 = 0.001** vs CG = 0.139 **vs GB = 0.001**
Expert 3 (E3)	219 / 375	58.40 %	vs CG = 0.578 **vs GB = 0.038**
ChatGPT (CG)	243 / 375	64.80 %	**vs GB = 0.006**
Google Bard GB)	206 / 375	54.93 %	

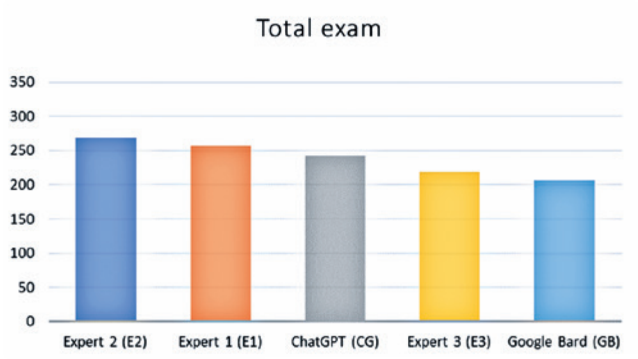


Figure 2. Likert scale scores for each expert and LLMs on the total exam. The maximum score is 375.

Table 4. Comparative table between the experts and LLMs (definitions section). The comparative p-values where significant differences exist are highlighted.

	Total points	Percentage of exam	p-value < 0.05
Expert 1 (E1)	87 / 125	69.6 %	vs E2 = 0.265 **vs E3 = 0.036** vs CG = 0.891 vs GB = 0.347
Expert 2 (E2)	95 / 125	76 %	**vs E3 = 0.001** vs CG = 0.203 **vs GB = 0.038**
Expert 3 (E3)	71 / 125	56.8 %	vs CG = 0.578 vs GB = 0.244
ChatGPT (CG)	86 / 125	68.8 %	vs GB = 0.422
Google Bard (GB)	80 / 125	64 %	

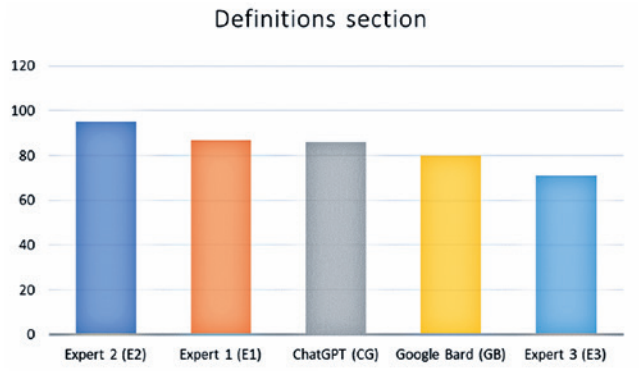


Figure 3. Likert scale scores for each expert and LLMs in the definitions section. The maximum score is 125.

No significant differences were observed between E1 and E2 (z-stat: -1.137, p-value: 0.256) when comparing the results in the definitions section among the experts. However, significant differences were evident between experts and E3, who received the lowest evaluation (E1 vs E3: z-stat: 2.098, p-value: 0.036; E2 vs E3: z-stat: 3.214, p-value: 0.001). No significant differences were found in any of the cases during the definitions section (E1 vs GPT-4: z-stat: 0.137, p-value: 0.891; E1 vs Bard: z-stat: 0.940, p-value: 0.347; E2 vs GPT-4: z-stat: 1.273, p-value: 0.203; E2 vs Bard: z-stat: 2.070, p-value: 0.038; E3 vs GPT-4: z-stat: 1.963, p-value: 0.050; E3 vs Bard: z-stat: 1.164, p-value: 0.244) when evaluating the experts against the LLMs. These results indicate consistency between the experts and LLMs in the definitions section without observing statistically different performances in any of the cases.

In the questions related to treating various pathologies in the specialty, the results were evaluated similarly to the previous section (Figure 4 and Table 5). When comparing the results among the experts, there were no differences between E1 and E2, who remain the experts with the best assessment (z-stat: 0.592, p-value: 0.554). However, there were differences between both in relation to E3 (E1 vs E3: z-stat: 2.622, p-value: 0.009; E2 vs E3: z-stat: 2.041, p-value: 0.041) and to both LLMs (E1 vs GPT-4: z-stat: 4.116, p-value: 0.0001; E1 vs Bard: z-stat: 5.800, p-value: 0.0001; E2 vs GPT-4: z-stat: 1.774, p-value: 0.076). Lastly, there were no differences between E3 and GPT-4 (E3 vs GPT-4: z-stat: 1.536, p-value: 0.125) but there were with Google Bard (z-stat: 3.291, p-value: 0.001). In this section, when comparing both LLMs, they behaved similarly with no differences in their results (z-stat: 1.774, p-value: 0.076).

The final analysis focuses on general queries. Figure 5 presents the results obtained for each evaluator. Contrary to previous sections, in this case, GPT-4 achieved the best performance, while Google Bard, which had received a less favorable evaluation in the two previous, managed an improved score. This shift in performance underscores the

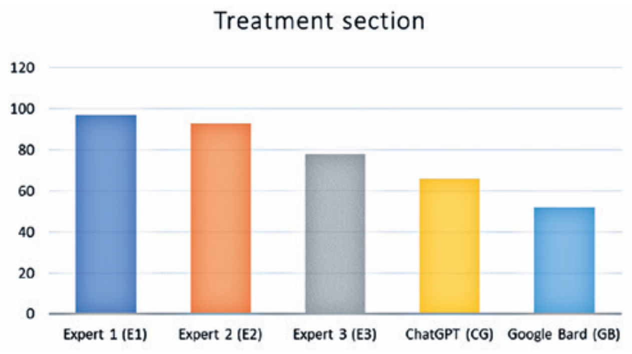


Figure 4. Likert scale scores for each expert and LLMs on the treatment section. The maximum score is 125.

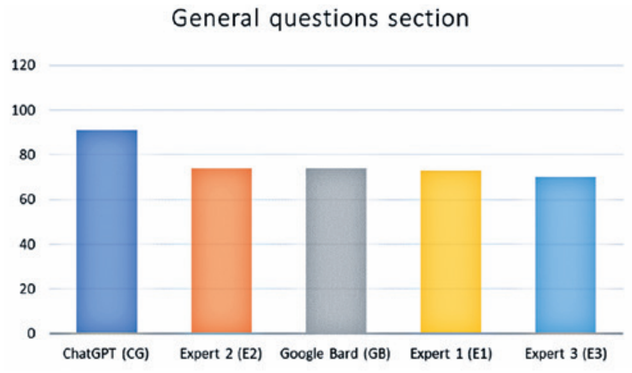


Figure 5. Likert scale scores for each expert and LLMs on the general questions section. The maximum score is 125.

Table 5. Comparative table between the experts and LLMs (treatment section). The comparative p-values where significant differences exist are highlighted.

	Total points	Percentage of exam	p-value < 0.05
Expert 1 (E1)	97 / 125	77.6 %	vs E2 = 0.554 **vs E3 = 0.009** **vs CG = 0.001** **vs GB = 0.001**
Expert 2 (E2)	93 / 125	74.4 %	**vs E3 = 0.041** **vs CG = 0.001** **vs GB = 0.001**
Expert 3 (E3)	78 / 125	62.4 %	vs CG = 0.125 **vs GB = 0.001**
ChatGPT (CG)	66 / 125	52.8 %	vs GB = 0.076
Google Bard (GB)	52 / 125	41.6 %	

Table 6. Comparative table between the experts and LLMs (general questions section). The comparative p-values where significant differences exist are highlighted.

	Total points	Percentage of exam	p-value < 0.05
Expert 1 (E1)	73 / 125	58.4 %	vs E2 = 0.898 vs E3 = 0.701 **vs CG = 0.017** vs GB = 0.898
Expert 2 (E2)	74 / 125	59.2 %	vs E3 = 0.609 **vs CG = 0.023** vs GB = 1
Expert 3 (E3)	70 / 125	56%	**vs CG = 0.006** vs GB = 0.609
ChatGPT (CG)	91 / 125	72.8 %	vs GB = 0.023
Google Bard (GB)	74 / 125	59.2 %	

variability of results between thematic sections and suggests that the performance of the evaluators, whether experts or LLMs, can vary depending on the specific nature of the questions and topics addressed. As seen in Table 6, significant differences were observed in favor of GPT-4 against the other four (GPT-4 vs E1: z-stat: -2.396, p-value: 0.017; GPT-4 vs E2: z-stat: -2.270, p-value: 0.023; GPT-4 vs E3: z-stat: -2.774, p-value: 0.006; GPT-4 vs Bard: z-stat: -2.270, p-value: 0.023) when comparing the results. When comparing the results among the experts themselves in this case, there were no differences between them, nor were there differences between the experts and Google Bard.

Discussion

The objective of the study was to evaluate the performance of GPT-4 and Google Bard in providing answers to complex questions about foot and ankle pathology and compare the results with the experts, all of whom are full members of the association.

Our study reveals that faced with 15 specific questions regarding leg, ankle, and foot pathology, a subspecialty of orthopedics and traumatology, GPT-4 demonstrates comparable behavior and, in some cases, even superior to experts with experience in this area. In contrast, Google Bard does not exhibit such outstanding performance, which aligns with previous findings in the literature that have compared both LLMs[(10)].

The ability of GPT-4 and Google Bard to answer questions in this subspecialty is evident in their capacity to understand and generate coherent responses. These LLMs have demonstrated relative competence in interpreting medical queries. However, it is crucial to note that their knowledge stems from previous data, not from practical experience or direct clinical interaction[(11)].

This assertion is supported by analyzing the different thematic sections of the questionnaire. In the definitions section, GPT-4 and Google Bard exhibited behaviors comparable to those of experts. Similarly, in the general questions section, it was observed that GPT-4's performance even surpassed that of the experts, while Google Bard's performance improved compared to the other two sections. However, in the treatment section, the performance of both LLMs was notably inferior to that of the experts, highlighting that, for the definition of surgical treatments, LLMs might not be the most suitable source of information[12-13].

Although these LLMs possess extensive knowledge, their lack of clinical context and direct experience could limit their ability to handle specific cases or complex clinical situations. As noted by Lopes et al., the accurate interpretation of medical data often requires a deep understanding of the clinical context, something these LLms might lack [14].

The efficacy of LLMs in this field is also influenced by the quality and quantity of available training data[15]. Moreover, constant updates in medical research can affect their ability to stay updated.

Despite these limitations, it is evident that AI, represented by GPT-4 and Google Bard, can be a valuable tool in the orthopedics and traumatology field, providing general information and initial support. However, consultation with an expert remains essential for more detailed evaluations and informed surgical decisions.

Both LLMs have the potential to improve access to healthcare for patients. However, it is important to remember that these technologies are not flawless and cannot replace specialized medical personnel. Medical consultations with LLMs may be prone to errors and may not always provide accurate or updated information[7]. It is crucial for the medical community to use these medical applications as a complementary tool to the healthcare provided by expert doctors.

Conclusion

A detailed analysis of the thematic sections of the questionnaire reveals that both GPT-4 and Google Bard demonstrated notable skills in definitions and general questions, even equating and surpassing, in some cases, the performance of experts. However, this efficacy significantly decreased in the treatment section, where both LLMs exhibited considerably inferior performance compared to experts.

This finding underscores the importance of considering the limitations of the LLMs, especially in clinical contexts where the precise definition of surgical treatments requires deeper knowledge and practical experience that these LLMs currently lack. It is crucial to recognize that, despite their capabilities, LLMs cannot completely replace the expertise and clinical judgment of healthcare professionals in specific and complex situations.

These results emphasize the need for effective collaboration between AI and clinical experts to achieve a more comprehensive and accurate approach to medical decision-making.

Authors' contributions: Each author contributed individually and significantly to the development of this article: AT *(https://orcid.org/0009-0006-5634-0823), and FPS *(https://orcid.org/0000-0002-2376-4834) Conceived and planned the activity that led to the study, wrote the article, participated in the review process; JP *(https://orcid.org/0000-0001-9662-0367) Data collection, bibliographic review; NH *(https://orcid.org/0009-0002-7215-2137) Formatting of the article, bibliographic review; IM *(https://orcid.org/0000-0002-6284-6410) Interpreted the results of the study, participated in the review process; FA *(https://orcid.org/0000-0001-6577-8911), and LC *(https://orcid.org/0000-0003-1187-0864), and VC *(https://orcid.org/0000-0002-8115-6300), and MPLZ *(https://orcid.org/0009-0009-8632-480X), and IT *(https://orcid.org/0000-0003-4033-8818) FVS *(https://orcid.org/0009-0004-0424-8334), and FMS *(https://orcid.org/0009-0000-7101-9145). Performed the surgeries; data collection, statistical analysis. All authors read and approved the final manuscript. *ORCID (Open Researcher and Contributor ID).

References

1. Turing AM. Computing machinery and intelligence. Mind. 1950; 59(236):433-60.
2. Sasanelli F, Le KDR, Tay SBP, Tran P, Verjans JW. Applications of Natural Language Processing Tools in Orthopaedic Surgery: A Scoping Review. Appl Sci. 2023;13(20):11586.
3. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence. AI Magazine. 1955;*27*(4):12-4.
4. Microsoft [Internet]. La inteligencia artificial es "la cuarta revolución industrial", según Microsoft. Europa Press. [March 30th, 2017]. Available from: https://www.europapress.es/portaltic/sector/noticia-inteligencia-artificial-cuarta-revolucion-industrial-microsoft-20161115143434.html
5. ChatGPT. Information provided by the GPT-3.5 language model. Retrieved from ChatGPT, a platform developed by OpenAI. November, 2023.
6. BERT [Internet]. Information provided by the BERT language model. [November 2023]. Available from: https://bard.google.com/
7. Lum ZC. Can Artificial Intelligence Pass the American Board of Orthopaedic Surgery Examination? Orthopaedic Residents Versus ChatGPT. Clin Orthop Relat Res. 2023;481(8):1623-30.
8. Hofmann HL, Guerra GA, Le JL, Wong AM, Hofmann GH, Mayfield CK, et al. The Rapid Development of Artificial Intelligence: GPT-4's Performance on Orthopedic Surgery Board Questions. Orthopedics. 2023:1-5.

9. Likert, R. A technique for the measurement of attitudes. Arch Psycholo. 1932;140:1-55.
10. Thibaut G, Dabbagh A, Liverneaux P. Does Google's Bard Chatbot perform better than ChatGPT on the European hand surgery exam? Int Orthop. 2024;48(1):151-8.
11. Agharia S, Szatkowski J, Fraval A, Stevens J, Zhou Y. The ability of artificial intelligence tools to formulate orthopaedic clinical decisions in comparison to human clinicians: An analysis of ChatGPT 3.5, ChatGPT 4, and Bard. J Orthop. 2023;50:1-7.
12. Cuthbert R, Simpson AI. Artificial intelligence in orthopaedics: can Chat Generative Pre-trained Transformer (ChatGPT) pass Section 1 of the Fellowship of the Royal College of Surgeons (Trauma & Orthopaedics) examination? Postgrad Med J. 2023;99(1176): 1110-4.
13. Chatterjee S, Bhattacharya M, Pal S, Lee SS, Chakraborty C. ChatGPT and large language models in orthopedics: from education and surgery to research. J Exp Orthop. 2023;10(1):128.
14. Lopes FM, Catarino JD, Lima JS, Neves JF. Challenges in the adoption of Natural Language Processing in Clinical Decision Support Systems. In: Advances in Human Factors and System Interactions. Springer; 2020. pp. 428-41.
15. Smith, LN, Yoon K, Jiménez PM. (2019). Improving language understanding by generative pretraining. J. open source softw. 2019;4(43):1851.