Journal of Orthopaedic Surgery and Research

RESEARCH

Open Access

Assessment of ChatGPT's adherence to evidence-based clinical practice guidelines for plantar fasciitis management

Le Zhang[1†], Tianyi Wang[1†], Yinfeng Zheng[1], Xiaochuan Kong[1], Gang Hong[1] and Lei Zang[1*]

Abstract

Purpose This study aimed to test the multidimensional performance of Chat-Generative Pre-trained Transformer (ChatGPT) in generating recommendations for the management of plantar fasciitis (PF) that adhere to well-established clinical practice guidelines.

Materials and methods 21 queries were raised from the 2023 APTA guideline recommendations for PF and prompted into ChatGPT-4o and ChatGPT-4 Turbo. Two experienced orthopaedic physicians evaluated the responses for accuracy, consistency, self-awareness, and fabrication and falsification using five-point Likert scales. The group-wise comparisons were conducted between the two models and subgroups.

Results The interrater agreement between evaluators was moderate to good (intraclass correlation coefficients of 0.573–0.757). Both versions of ChatGPT were outperformed and comparable across all dimensions, including accuracy ([4.1 ± 0.8] vs. [4.1 ± 0.7], $P = 0.959$), consistency ([4.6 ± 0.5] vs. [4.6 ± 0.6], $P = 0.890$), self-awareness ([4.3 ± 0.6] vs. [4.5 ± 0.5], $P = 0.407$), and fabrication and falsification ([4.6 ± 0.6] vs. [4.5 ± 0.4], $P = 0.681$). In the subgroup comparisons, better performance was identified in closed-ended questions and for positive rather than negative recommendations ($P < 0.05$). No significant differences were found between recommendation strength subgroups, except in fabrication and falsification ([4.4 ± 0.6] vs. [5.0 ± 0], $P = 0.001$).

Conclusions The two mainstream versions of ChatGPT showed comparable and superior performance in generating recommendations concordant with clinical guidelines for PF management. However, notable specific issues included performance variations between different prompt strategies, recommendation grades, and recommendation type, and the models should still be utilized with caution.

Keywords Plantar fasciitis, Artificial intelligence, ChatGPT, Large Language models, Clinical guidelines, Physical therapy

[†]Le Zhang and Tianyi Wang contributed equally to this work.

*Correspondence:
Lei Zang
zanglei@ccmu.edu.cn
[1]Department of Orthopedics, Beijing Chaoyang Hospital, Capital Medical University, 5 JingYuan Road, Shijingshan District, Beijing 100043, China

Introduction

Plantar fasciitis (PF) is the primary cause of plantar heel pain, affecting 4–7% of the population and accounting for > 1 million patient visits in the United States per year [1–3]. PF greatly diminishes mobility and foot functionality, and reduces the ability to work, leading to a decline in health-related quality of life [3]. Although PF is typically self-limiting, a substantial proportion of patients continue to experience residual symptoms for anywhere from six months to over 15 years [4]. Therefore, effective management is essential for satisfactory outcomes. Several guidelines have been developed for PF to standardize the clinical practice diagnosis and treatment, and recommended physical therapy, such as plantar fascia stretching and low dye taping, as the primary core management for PF [1, 5–7]. Besides, approximately 62% of PF patient visits are to primary care providers [1]. The high demand for medical visits coupled with the importance of physical therapy underscores the critical role that primary care physicians play in the management of PF. However, it can be challenging for primary care physicians to find a concise and reliable source of comprehensive and up-to-date standards for diagnosis and intervention, thereby making accurate and in-time clinical decision in routine workflow.

The introduction and advancement of artificial intelligence (AI) large language models (LLMs) offers tremendous potential for bridging that gap [8, 9]. Chat-Generative Pre-trained Transformer (ChatGPT) is among the most widely used LLMs. It was released for public use in November 2022 by OpenAI (San Francisco, CA, USA) and is continuously updated [8, 10]. Based on the vast open-source corpus of online training data, it is able to effectively conduct conversations with human users. The versatility of ChatGPT makes it a promising source of healthcare information, potentially broadening the spectrum of its acceptable uses [9]. Thus, there has been a considerable body of research assessing its performance in clinical decision support, physician education and training, patient education, medical question answering, and more [11]. Investigating ChatGPT's ability to generate specialized medical information that adheres to well-established clinical guidelines is among these areas of research interest, while previous studies focusing on various medical issues showed discrepant results [12–15]. However, to the best of our knowledge, no previous research to date has evaluated the performance of current LLMs in the provision of specialized medical guidance for the management of PF.

Therefore, this study aimed to test the multidimensional performance of ChatGPT in the generation of recommendations for the management of PF the adheres to the American Physical Therapy Association (APTA) clinical practice guidelines.

Materials and methods

The institutional ethics committee exempted this study from the ethical approval requirement because no human data were gathered or used. As an observational cross-sectional study for LLM evaluation, this study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline [16] and the QUEST (Quality of information, Understanding and reasoning, Expression style and persona, Safety and harm, and Trust and confidence) framework for human evaluation for LLMs [11].

Data source and recommendation selection

The APTA drafted the initial version of clinical practice guideline for PF in 2008, and revised in 2014 and 2023 respectively [5–7]. The APTA guidelines provide six grades of recommendations based on the strength of the supporting evidence: A (strong evidence), B (moderate evidence), C (weak evidence), D (conflicting evidence), E (theoretical/foundational evidence), and F (expert opinion). As the most up-to-date and comprehensive, the 2023 APTA guidelines were used as the primary source, with the 2008 and 2014 APTA guidelines accessed where necessary for further reference. A total of 19 recommendations were identified and selected as standardized benchmarks. These covered six topics, including risk factors, diagnosis, differential diagnosis, examination, physical impairment measures, and interventions.

ChatGPT selection and prompt design

Two updated versions of ChatGPT-4: ChatGPT-4o (OpenAI) and ChatGPT-4 Turbo (OpenAI), were selected for investigation. ChatGPT-4 Turbo was the more efficient version than ChatGPT-4, trained with data up to April 2023 [17]. Whereas, ChatGPT-4o is a newer, flagship model with performance improvement and more natural human-computer interaction, which is able to accept multimedia input (text, audio, image, and video) [18]. The training data cutoff for ChatGPT-4o was October 2023.

Based on the suitability, the 19 recommendations were rephrased as either closed-ended or open-ended questions. Recommendations comprised of two independent items were divided into two queries. If the query did not contain a reference to plantar fasciitis or heel pain, it was adjusted to contain the words *plantar fasciitis/heel pain*. Each unique query was entered into ChatGPT three times, with no additional prompting. A new window was used for each repetition of the prompt to avoid interference. All prompts and answers were recorded verbatim (Supplementary File 1). The query process was performed by one physician (TW, with 5 years of experience) on January 7th, 2025.

Performance evaluation

To evaluate the performance of the AI models, we compared the answers from ChatGPT with the standardized benchmarks. Four performance dimensions were evaluated [11]: (1) Accuracy (concordance between LLM and the benchmarks); (2) Consistency (stability and uniformity of the LLM responses to three repetitions of each query); (3) Self-awareness (the ability of the LLM to recognize its limitations and avoid overconfidence); and (4) Fabrication and falsification (the presence of made-up or distorted information in the LLM responses). Each dimension was assessed using five-point Likert scales. Aside from consistency, the dimensions were assessed based on the first response of ChatGPT to each query.

Two experienced orthopaedic physicians (LZ, with 20 years of experience, and XK, with 37 years of experience) independently reviewed and scored all the responses. The physicians were blinded to the version of LLM model that produced each response during the evaluation process. The final score for each dimension was calculated as the average score of the two physicians. Interrater agreement between the physicians was also assessed.

Statistical analysis

As five-point Likert scales have equidistance, scores were presented as both mean ± standard deviation and median (interquartile range [IQR]). Comparisons of ranked data were performed using the Mann-Whitney U test. Interrater agreement was evaluated using the intraclass correlation coefficient (ICC), with a two-way mixed-effects model for absolute agreement. Excellent, good, moderate, and poor agreements were quantified as ICC ≥ 0.90, 0.75–0.90, 0.50–0.75, and < 0.50, respectively. Statistical analysis was conducted using SPSS 24.0 (IBM, Armonk, NY, USA), and a P value < 0.05 was deemed statistically significant.

Results

A total of 21 prompts were rephased from 19 recommendations, and 63 pairs of responses (containing 21 pairs of first-time responses) were generated by two versions of ChatGPT (Supplementary File 1). In first-time responses, one was on risk factors, diagnosis, differential diagnosis, and physical impairment measures, respectively. Another two were on examination, and the rest 15 were about interventions. The interrater agreements for the four dimensions ranged from moderate to good. The ICC were 0.757 (0.592–0.861) for accuracy, 0.573 (0.327–0.746) for consistency, 0.580 (0.335–0.750) for self-awareness, and 0.671 (0.463–0.809) for fabrication and falsification (Table 1).

Both ChatGPT-4o and ChatGPT-4 Turbo performed well overall (average score for each dimension > 4). The mean accuracy score was 4.1 ± 0.8 (median [IQR], 4.5 [3.5–5.0]) for ChatGPT-4o, and 4.1 ± 0.7 (median [IQR], 4.0 [3.5–5.0]) for ChatGPT-4 Turbo. For consistency, the mean score was 4.6 ± 0.5 (median [IQR], 5.0 [4.5–5.0]) for ChatGPT-4o, and 4.6 ± 0.6 (median [IQR], 5.0 [4.3–5.0]) for ChatGPT-4 Turbo. The mean self-awareness scores were 4.3 ± 0.6 (median [IQR], 4.5 [3.5–5.0]) and 4.5 ± 0.5 (median [IQR], 4.5 [4.0–5.0]) for ChatGPT-4o and ChatGPT-4 Turbo, respectively. The mean fabrication and falsification scores for ChatGPT-4o and ChatGPT-4 Turbo were 4.6 ± 0.6 (median [IQR], 5.0 [4.0–5.0]) and 4.5 ± 0.4 (median [IQR], 4.5 [4.0–5.0]), respectively. There were no statistical differences between the two models on any of the dimensions (Fig. 1). Only one unsatisfactory response (either of dimension < 3 points) was generated by ChatGPT-4o (query 21 with an accuracy of 2.5) and one by ChatGPT-4 Turbo (query 17 with a consistency of 2.5). Both queries were on the topic of PF interventions.

Three subgroup analyses were conducted. The subgrouping was by prompts strategy (open/closed-ended prompts), recommendation type (positive/negative recommendation), and recommendation strength (recommendation grades A–C or D–F). The answers to closed-ended queries were better than those for open-ended queries on the dimensions of consistency ([4.8 ± 0.5] vs. [4.2 ± 0.3], $p < 0.001$), self-awareness ([4.6 ± 0.5] vs. [3.8 ± 0.6], $p < 0.001$), and fabrication and falsification ([4.7 ± 0.5] vs. [4.1 ± 0.5], $p = 0.002$), but not for accuracy ([4.2 ± 0.9] vs. [3.9 ± 0.3], $p = 0.108$). In closed-ended queries, positive recommendations outperformed negative recommendations on all four dimensions ($p < 0.05$). No significant differences were found between the recommendation strength subgroups (grade A–C vs. D–F), except for fabrication and falsification ([4.4 ± 0.6] vs. [5.0 ± 0], $P = 0.001$; Table 2).

Discussion

PF is a global public health issue that has considerable deleterious effects on both athletic and non-athletic populations [19, 20]. As the standardized management guidelines for PF are regularly updated, it is important for physicians to remain abreast of the latest guidelines [1, 5–7]. The present study investigated the ability of ChatGPT

Table 1 Interrater reliability between physicians for ChatGPT performance evaluation

Dimension	Accuracy	Consistency	Self-awareness	Fabrication and falsification
ICC (95%CI)	0.757 (0.592–0.861)	0.573 (0.327–0.746)	0.580 (0.335–0.750)	0.671 (0.463–0.809)

LLMs large language models, *ICC* intraclass correlation coefficient

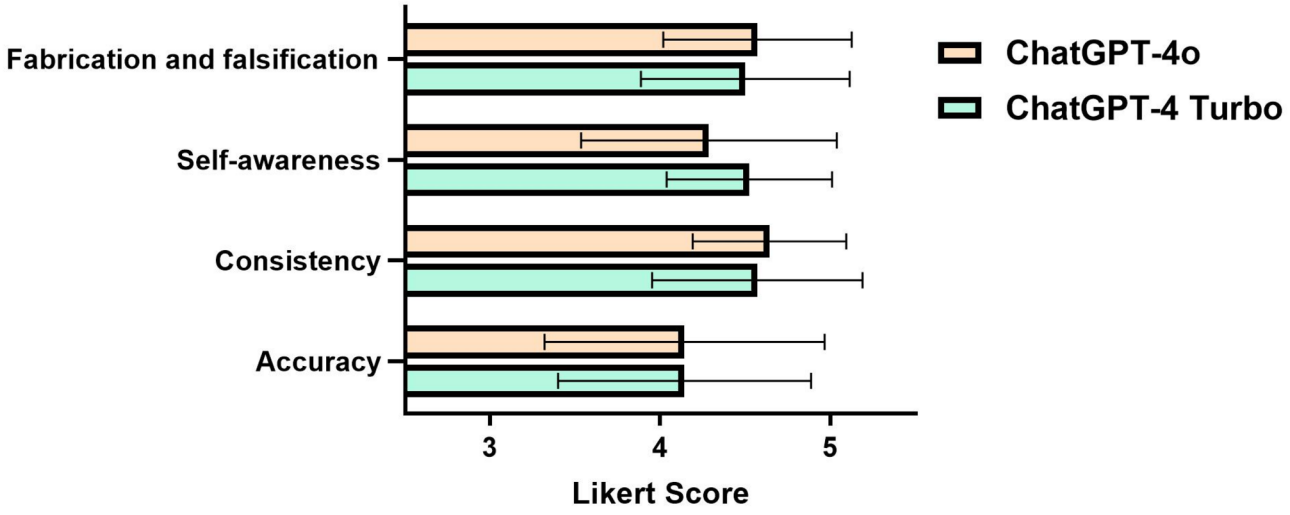

Fig. 1 Performance of ChatGPT in providing recommendations for plantar fasciitis

Table 2 Subgroup analysis based on characteristics of prompts or recommendations

	N	Accuracy		Consistency		Self-awareness		Fabrication and falsification	
		Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Open/Close-ended queries									
Open	10	3.9 (0.3)	4.0 (3.5–4.0)	4.2 (0.3)	4.0 (4.0–4.5)	3.8 (0.6)	3.8 (3.0–4.5)	4.1 (0.5)	4.0 (3.9–4.5)
Close	32	4.2 (0.9)	4.5 (3.5–5.0)	4.8 (0.5)	5.0 (4.5–5.0)	4.6 (0.5)	5.0 (4.1–5.0)	4.7 (0.5)	5.0 (4.5–5.0)
P value		0.108		**< 0.001**		**< 0.001**		**0.002**	
Positive/Negative recommendation*									
Positive	28	4.4 (0.8)	4.8 (4.0–5.0)	4.8 (0.3)	5.0 (4.6–5.0)	4.7 (0.5)	5.0 (4.5–5.0)	4.8 (0.5)	5.0 (4.6–5.0)
Negative	4	3.1 (0.3)	3.0 (3.0–3.4)	4.1 (1.1)	4.5 (3.0–4.9)	4.1 (0.5)	4.3 (3.6–4.5)	4.0 (0.4)	4.0 (3.6–4.4)
P value		**0.009**		**0.039**		**0.029**		**0.007**	
Recommendation grade									
A-C	32	4.0 (0.7)	4.0 (3.5–4.9)	4.5 (0.6)	4.5 (4.1–5.0)	4.4 (0.7)	4.5 (4.0–5.0)	4.4 (0.6)	4.5 (4.0–5.0)
D-F	10	4.5 (0.8)	5.0 (4.1–5.0)	4.8 (0.3)	5.0 (4.5–5.0)	4.5 (0.5)	4.5 (4.0–5.0)	5.0 (0)	5.0 (5.0–5.0)
P value		0.078		0.180		0.805		**0.001**	

* Comparison between answers based on positive and negative recommendations was conducted only in close-ended queries

Bold values indicate statistic difference ($P < 0.05$)

SD standard deviation, *IQR* interquartile range

to provide PF recommendations that adhere to the latest clinical practice guideline. The two versions of ChatGPT were evaluated on multiple dimensions, on which they showed remarkable and comparable performance. However, the chatbots are prone to generate more satisfactory responses to closed-ended questions and for positive rather than negative recommendations. The results may provide novel insights for levering the current LLMs to bridge the gap between routine clinical practice and comprehensive and up-to-date PF standard.

Several previous studies have conducted binary qualitative evaluations of LLM performance on dimensions such as response accuracy [12, 13, 21–24]. However, given that the chatbots are prone to generate vague responses that not entirely precise or incorrect [12, 25], the advantages of semiquantitative or quantitative score system have been highlighted and leveraged in precise investigation [14, 15, 26, 27]. In this vein, by using 5-point Likert scale, this study found a satisfactory accuracy (> 4) in both versions of ChatGPT-4 in line with the latest guidelines

for PF. Besides, the reliability of medical information is essential in clinical decision-making because even occasionally erroneous information can result in catastrophic health and economic consequences [10, 28, 29]. Fortunately, both ChatGPT-4o and ChatGPT-4 Turbo demonstrated high consistency (mean scores of 4.6) to repeated prompts. Furthermore, only one unsatisfactory response (4.8%) each was generated by the models, indicating a trustworthy level of stability in current ChatGPT advice for the management of PF.

Another crucial dimension in the evaluation of LLMs' capacity to guide clinical management is safety. Self-awareness, reflecting the recognition of its limitations in data source, processing patterns, and responses, is one of the specific aspects of safety [11]. Some chatbots, such as the Claude models, definitely declare itself as an AI assistant [30]. This may help prevent blind trust or excessive expectations in users. However, the assessment for the adherence of chatbots to well-established guideline raised higher demand for the professionalism essentially. Promisingly, we obtained high self-awareness scores (4.3–4.5) for both versions of ChatGPT, and most responses included disclaimers and recommendations to consult professional institutions or physicians. However, self-awareness may be a double-edged sword, in that undue recognition of inherent limitations without a premise of high accuracy may undermine users' confidence in the model [14]. Thus, it is imperative to weigh the pros and cons in future model update and fine-tuning. Fabrication and falsification are also a threat to the safety of chatbot responses, which are defined as responses containing plausible but made-up or distorted information or data despite plausible [11, 31]. Both models performed well on this dimension, with little evidence of fabrication or falsification (mean score of 4.5–4.6).

However, some specific issues are still needed to be considered. First, there is ongoing debate about whether the performance of LLMs differs in their responses to closed- and open-ended questions. Goodman et al. conducted a comparison between chatbot responses to descriptive and binary physician-developed medical questions and found no significant difference [10]. Conversely, Zaidat et al. investigated ChatGPT's responses to questions about for antibiotic prophylaxis in spinal surgery. Although there was no statistical analysis of the data, ChatGPT-3.5 and 4.0 demonstrated accuracy levels of 83.3% and 100%, respectively, in the responses to closed questions, and of 40.0% and 80.0% in the responses to open questions [21]. Similarly, this study also identified a better performance of ChatGPT in response of closed-ended questions about PF. This is unsurprising as closed-ended queries do not reflect the nuances of medical decision-making and open-ended queries are inherently more complex, while LLMs are generally believed under-performing in complicated queries [12, 26, 28]. The poor performance may be attributed to "hallucinations", referring to generating plausible but inaccurate responses [32], which was also confirmed by the differences of fabrication and falsification between the subgroups shown in our results. Furthermore, the answers from open-ended queries tend to be equivocal and not entirely in line with the reference standards [25]. Of note, we interestingly observed no unacceptable responses to open-ended queries. A plausible explanation is the open-ended queries for PF in this study were mainly about risk factors, diagnosis, differential diagnosis, and examination, which are relatively simple and easier to access from open-source corpus compared to interventions.

Furthermore, recommendation grades based on evidence levels may potentially impacts the quality of chatbot responses. A previous study found that, compared to issues with insufficient or conflicting evidence, ChatGPT-3.5 performed better in providing information in line with high-grade recommendations in clinical guidelines for low back pain [13]. However, instead of recognizing a superiority of ChatGPT-4 in responding prompts of high-grade recommendations, this study even found less fabrication and falsification among low-grade recommendations. We believe there may be several reasons. Among these, an unclear source of LLM training data may be the prime culprit. Typically, ChatGPT training begins with unsupervised learning from a vast open-source corpus, followed by supervised fine-tuning and reinforcement learning [25]. Thus, it was originally trained by easily accessed online data without quality control, rather than comprehensive authoritative evidence-based data. This likely leads to arbitrary or opportunistic output with regard to specific issues. Besides, the "black box" nature of LLM analysis makes it hard to grasp the logic of generating outputs based on inputs [33]. Therefore, the impact of recommendation grades to ChatGPT responses may be indirect and inconclusive, and further well-designed studies are warranted to shed light on this issue.

Moreover, we interestingly found that ChatGPT underperformed in generating negative recommendations on all four dimensions. For instance, the APTA guidelines recommend that "Clinicians should not use orthoses, either prefabricated or custom fabricated/fitted, as an isolated treatment for short-term pain relief in individuals with plantar fasciitis", while two ChatGPT models were unlikely to "refuse" despite binary prompts. Instead, they gave uncertain responses such as "The use of orthoses, whether prefabricated or custom-fabricated, can be an effective component of treatment for short-term pain relief in individuals with PF…using orthoses as an isolated treatment might not address all underlying factors contributing to PF." Unfortunately, few previous studies have

mentioned this issue. Echoing our findings, Gianola et al. evaluated the performance of ChatGPT-3.5 responses to questions about lumbosacral radicular pain in line with clinical practice guidelines [12]. Surprisingly, evidently wrong answers were found in 60% (3/5) responses on "do not do" recommendations, compared to none (0/4) on "should/could do" recommendations. It may be that providing an assertive negative response requires more complex logic or reverse thinking when the AI analyzes the input. It is also possible that this result could be due to selection bias caused by the small sample size in the present study, and further modified study are warranted to calibrate the findings.

Certain limitations should be acknowledged in this study. First, despite the relatively rigorous study design, the small number of models, raters, and questions may have introduced biases. Therefore, the results of our statistical analyses should be interpreted with caution. Second, the APTA clinical guidelines for PF was developed only for physical therapist practice, thus lacking recommendations for pharmacological or surgical interventions [5]. However, this source is the most authoritative and recent set of standardized organizational guidelines for PF, and we recommend that future work should make a more comprehensive evaluation in premise of newly well-established clinical guidelines. Furthermore, prompt engineering is an emerging discipline that affects the performance of LLMs in responding to different prompts [34]. Hence, research using more well-designed prompts may be conducted to further test the performance of these models. Finally, as we mainly focused on assessing ChatGPT's potential in providing specialized information and supporting decision-making for physicians, determining the performance of the chatbots in responding to the questions of patients and scenarios were beyond the scope of this study. Nevertheless, investigations into this will improve our understanding of the capability and applicability of chatbots in healthcare.

Conclusion

On the dimensions of accuracy, consistency, self-awareness, and fabrication and falsification, the two mainstream versions of ChatGPT showed equivalent superior performance in the generation of recommendations concordant with clinical guidelines for the management of PF. However, specific issues including performance variations between different prompt strategy, recommendation grade, and recommendation type should be noted, and the models should still be utilized with caution.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13018-025-05831-y.

Supplementary Material 1

Acknowledgements
Not applicable.

Author contributions
ZL and WT contributed equally to this work. Conceptualization: ZL, WT, ZY and ZL; methodology: ZL, WT, and DP; formal analysis and investigation: WT and ZY; writing—original draft preparation: ZL and WT; writing—review and editing: KX, HG, FN, DP, and CR; resources: ZL; supervision: ZL.

Funding
None.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval
Ethical approval for this study was not required as all chatbots used were publicly resources, and no human or animal data were involved.

Consent to publish
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 February 2025 / Accepted: 20 April 2025
Published online: 30 April 2025

References

1. Morrissey D, Cotchett M, Said J, Bari A, et al. Management of plantar heel pain: a best practice guide informed by a systematic review, expert clinical reasoning and patient values. Br J Sports Med. 2021;55(19):1106–18. https://doi.org/10.1136/bjsports-2019-101970.
2. Cooper MT. Common painful foot and ankle conditions: a review. JAMA. 2023;330(23):2285–94. https://doi.org/10.1001/jama.2023.23906.
3. Babatunde OO, Legha A, Littlewood C, et al. Comparative effectiveness of treatment options for plantar heel pain: a systematic review with network meta-analysis. Br J Sports Med. 2019;53(3):182–94. https://doi.org/10.1136/bjsports-2017-098998.
4. Hansen L, Krogh TP, Ellingsen T, Bolvig L, Fredberg U. Long-term prognosis of plantar fasciitis: a 5- to 15-year follow-up study of 174 patients with ultrasound examination. Orthop J Sports Med. 2018;6(3):2325967118757983. https://doi.org/10.1177/2325967118757983.
5. Koc TA Jr, Bise CG, Neville C, Carreira D, Martin RL, McDonough CM. Heel pain - plantar fasciitis: revision 2023. J Orthop Sports Phys Ther. 2023;53(12):CPG1–39. https://doi.org/10.2519/jospt.2023.0303.
6. Martin RL, Davenport TE, Reischl SF, McPoil TG, Matheson JW, Wukich DK, McDonough CM, American Physical Therapy Association. Heel pain-plantar fasciitis: revision 2014. J Orthop Sports Phys Ther. 2014;44(11):A1–33. https://doi.org/10.2519/jospt.2014.0303.
7. McPoil TG, Martin RL, Cornwall MW, Wukich DK, Irrgang JJ, Godges JJ. Heel pain–plantar fasciitis: clinical practice Guildelines linked to the international classification of function, disability, and health from the orthopaedic section of the American physical therapy association. J Orthop Sports Phys Ther. 2008;38(4):A1–18. https://doi.org/10.2519/jospt.2008.0302.
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233–9. https://doi.org/10.1056/NEJMsr2214184.
9. Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. J Med Internet Res. 2024;26:e56930. https://doi.org/10.2196/56930.

10. Goodman RS, Patrinely JR, Stone CA Jr, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483. https://doi.org/10.1001/jamanetworkopen.2023.36483.
11. Tam TYC, Sivarajkumar S, Kapoor S, et al. A framework for human evaluation of large Language models in healthcare derived from literature review. NPJ Digit Med. 2024;7(1):258. https://doi.org/10.1038/s41746-024-01258-7.
12. Gianola S, Bargeri S, Castellini G, et al. Performance of ChatGPT compared to clinical practice guidelines in making informed decisions for lumbosacral radicular pain: a cross-sectional study. J Orthop Sports Phys Ther. 2024;54(3):222–8. https://doi.org/10.2519/jospt.2024.12151.
13. Shrestha N, Shen Z, Zaidat B, et al. Performance of ChatGPT on NASS clinical guidelines for the diagnosis and treatment of low back pain: a comparison study. Spine (Phila Pa 1976). 2024;49(9):640–51. https://doi.org/10.1097/BRS.0000000000004915.
14. Sciberras M, Farrugia Y, Gordon H, et al. Accuracy of information given by ChatGPT for patients with inflammatory bowel disease in relation to ECCO guidelines. J Crohns Colitis. 2024;18(8):1215–21. https://doi.org/10.1093/ecco-jcc/jjae040.
15. Nwachukwu BU, Varady NH, Allen AA et al. (2024) Currently available large Language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. Arthroscopy S0749–8063(24)00575-9. https://doi.org/10.1016/j.arthro.2024.07.040
16. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7. https://doi.org/10.1016/S0140-6736(07)61602-X.
17. Zhou S, Luo X, Chen C, et al. The performance of large Language model-powered chatbots compared to oncology physicians on colorectal cancer queries. Int J Surg. 2024;110(10):6509–17. https://doi.org/10.1097/JS9.0000000000001850.
18. Miyazaki Y, Hata M, Omori H, et al. Performance of ChatGPT-4o on the Japanese medical licensing examination: evalution of accuracy in text-only and image-based questions. JMIR Med Educ. 2024;10:e63129. https://doi.org/10.2196/63129.
19. Llurda-Almuzara L, Labata-Lezaun N, Meca-Rivera T, et al. Is dry needling effective for the management of plantar heel pain or plantar fasciitis? An updated systematic review and meta-analysis. Pain Med. 2021;22(7):1630–41. https://doi.org/10.1093/pm/pnab114.
20. Zheng Y, Wang T, Zang L, et al. A novel combination strategy of ultrasound-guided percutaneous radiofrequency ablation and corticosteroid injection for treating recalcitrant plantar fasciitis: a retrospective comparison study. Pain Ther. 2024;13(5):1137–49. https://doi.org/10.1007/s40122-024-00629-y.
21. Zaidat B, Shrestha N, Rosenberg AM, et al. Performance of a large Language model in the generation of clinical guidelines for antibiotic prophylaxis in spine surgery. Neurospine. 2024;21(1):128–46. https://doi.org/10.14245/ns.2347310.655.
22. Mejia MR, Arroyave JS, Saturno M, et al. Use of ChatGPT for determining clinical and surgical treatment of lumbar disc herniation with radiculopathy: a North American spine society guideline comparison. Neurospine. 2024;21(1):149–58. https://doi.org/10.14245/ns.2347052.526.
23. Hoang T, Liou L, Rosenberg AM, et al. An analysis of ChatGPT recommendations for the diagnosis and treatment of cervical radiculopathy. J Neurosurg Spine. 2024;41(3):385–95. https://doi.org/10.3171/2024.4.SPINE231148.
24. Duey AH, Nietsch KS, Zaidat B, et al. Thromboembolic prophylaxis in spine surgery: an analysis of ChatGPT recommendations. Spine J. 2023;23(11):1684–91. https://doi.org/10.1016/j.spinee.2023.07.015.
25. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. J Med Internet Res. 2023;25:e47479. https://doi.org/10.2196/47479.
26. Tsai CY, Cheng PY, Deng JH, Jaw FS, Yii SC. ChatGPT v4 outperforming v3.5 on cancer treatment recommendations in quality, clinical guideline, and expert opinion concordance. Digit Health. 2024;10:20552076241269538. https://doi.org/10.1177/20552076241269538.
27. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep. 2024;14(1):243. https://doi.org/10.1038/s41598-023-50884-w.
28. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection Doctor?? Lancet Infect Dis. 2023;23(4):405–6. https://doi.org/10.1016/S1473-3099(23)00113-5.
29. Manohar N, Prasad SS. Use of ChatGPT in academic publishing: a rare case of seronegative systemic lupus erythematosus in a patient with HIV infection. Cureus. 2023;15(2):e34616. https://doi.org/10.7759/cureus.34616.
30. Rewthamrongsris P, Burapacheep J, Trachoo V, Porntaveetus T. Accuracy of large Language models for infective endocarditis prophylaxis in dental procedures. Int Dent J. 2025;75(1):206–12. https://doi.org/10.1016/j.identj.2024.09.033.
31. OpenAI. (2024) Introducing ChatGPT. Accessed December 5, 2024. https://openai.com/blog/chatgpt
32. Antaki F, Milad D, Chia MA, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol. 2024;108(10):1371–8. https://doi.org/10.1136/bjo-2023-324438.
33. Scaff SPS, Reis FJJ, Ferreira GE, Jacob MF, Saragiotto BT. (2024) Assessing the performance of AI chatbots in answering patients' common questions about low back pain. Ann Rheum Dis ard-2024-226202. https://doi.org/10.1136/ard-2024-226202
34. Wang L, Chen X, Deng X. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. 2024;7(1):41. https://doi.org/10.1038/s41746-024-01029-4.

Publisher's note