Are Large Language Model–Based Chatbots Effective in Providing Reliable Medical Advice for Achilles Tendinopathy?

An International Multispecialist Evaluation

Zuru Liang, PhD, Ming Wang, PhD, Nasef Mohamed Nasef Abdelatif, MD, Marut Arunakul, MD, Keen Wai Chong, MD, Yinghui Hua, MD, David Oji, MD, Ken Jin Tan, MD, Yasuhito Tanaka, MD, Akira Taniguchi, MD, Patrick Shu-Hang Yung, MBChB,
and Samuel Ka-Kin Ling,* MBChB(CUHK), ChM(Edin), FHKAM
Investigation performed at The Chinese University of Hong Kong, Hong Kong

Background: Large language model (LLM)–based chatbots have shown potential in providing health information and patient education. However, the reliability of these chatbots in offering medical advice for specific conditions like Achilles tendinopathy remains uncertain. Mixed outcomes in the field of orthopaedics highlight the need for further examination of these chatbots' reliability.

Hypothesis: Three leading LLM-based chatbots can provide accurate and complete responses to inquiries related to Achilles tendinopathy.

Study Design: Cross-sectional study.

Methods: Eighteen questions derived from the Dutch clinical guideline on Achilles tendinopathy were posed to 3 leading LLM-based chatbots: ChatGPT 4.0, Claude 2, and Gemini. The responses were incorporated into an online survey assessed by orthopaedic surgeons specializing in Achilles tendinopathy. Responses were evaluated using a 4-point scoring system, where 1 indicates unsatisfactory and 4 indicates excellent. The total scores for the 18 responses were aggregated for each rater and compared across the chatbots. The intraclass correlation coefficient was calculated to assess consistency among the raters' evaluations.

Results: Thirteen specialists from 9 diverse countries and regions participated. Analysis showed no significant difference in the mean total scores among the chatbots: ChatGPT (59.7 $\pm$ 5.5), Claude 2 (53.4 $\pm$ 9.7), and Gemini (53.6 $\pm$ 8.4). The proportions of unsatisfactory responses (score 1) were low and comparable across chatbots: 0.9% for ChatGPT 4.0, 3.4% for Claude 2, and 3.4% for Gemini. In terms of excellent responses (score 4), ChatGPT 4.0 outperformed the others, with 43.6% of the responses rated as excellent, significantly higher than Claude 2 at 27.4% and Gemini at 25.2% ($P < .001$ for both comparisons). Intraclass correlation coefficients indicated poor reliability for ChatGPT 4.0 (0.420) and moderate reliability for Claude 2 (0.522) and Gemini (0.575).

Conclusion: While LLM-based chatbots such as ChatGPT 4.0 can deliver high-quality responses to queries regarding Achilles tendinopathy, the inconsistency among specialist evaluations and the absence of standardized assessment criteria significantly challenge our ability to draw definitive conclusions. These issues underscore the need for a cautious and standardized approach when considering the integration of LLM-based chatbots into clinical settings.

Keywords: large language model; AI; chatbot; Achilles tendinopathy

Achilles tendinopathy is prevalent among physically active individuals and the general population, with an annual incidence rate of 5.2 per 1000 people.[13] In specific active populations like runners, its lifetime cumulative incidence

The Orthopaedic Journal of Sports Medicine, 13(4), 23259671251332596
DOI: 10.1177/23259671251332596

can exceed 50%.[20] Characterized by localized pain at the distal insertion and midportion of the Achilles tendon, along with impaired function, Achilles tendinopathy often progresses to chronic symptoms and disability.[4,18] Research underscores the importance of managing Achilles tendinopathy effectively, highlighting that it requires not only physical interventions but also a strong emphasis on patient education.[3,21] Given the long-term nature of the disease and its tendency to recur, enhancing patient education and providing accurate medical advice could effectively improve treatment outcomes.

In recent years, large language model (LLM)–based chatbots, such as OpenAI's ChatGPT, has gained significant interest for their potential role as supplements in clinical practice. Initial studies assessing ChatGPT's capabilities in medical licensing examinations[14] and specialty board examinations[2,19] have shown that it performs close to the passing threshold. Furthermore, comparisons between chatbot-generated content and that produced by clinicians, for instance, in drafting informed consent documents, have demonstrated that chatbots often generate more readable and complete content.[6] This development holds particular relevance as an increasing number of patients turn to the internet for health-related advice.[7] Consequently, it is plausible that patients may begin to utilize LLM-based chatbots more frequently as a source of medical information.

Building on these findings, numerous studies across various medical specialties have begun to explore the capability of chatbots to handle common patient inquiries effectively.[1,8,16] Within orthopaedics specifically, a study by Mika et al[17] assessed ChatGPT's performance in answering common questions about total hip replacement. The study demonstrated that the chatbot could effectively provide evidence-based responses. Additionally, 2 studies focusing on anterior cruciate ligament (ACL) reconstruction reached different conclusions, one suggesting that ChatGPT frequently provided satisfactory responses,[15] while the other revealed less reliable performance.[11] These mixed findings in orthopaedics highlight potential limitations in study design, such as a narrow pool of questions, evaluations conducted by a small number of specialists from single institutions, and assessments focused on only 1 LLM model. While there are scenarios in which inaccurate information may be tolerated as a mere nuisance, the margin of error regarding medical information is minimal. Thus, more rigorous study designs are needed to evaluate the performance of leading LLM-based chatbots in providing health care information, particularly in orthopaedics.

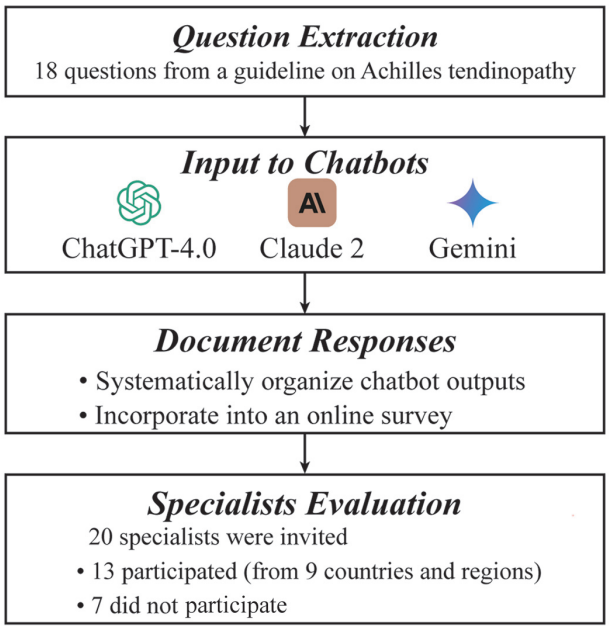

Figure 1. Flowchart of the study design.

Our study aimed to assess and compare the responses provided on Achilles tendinopathy by 3 leading LLMs: OpenAI's ChatGPT 4.0, Anthropic's Claude 2, and Google's Gemini. We hypothesized that these LLM-based chatbots could deliver accurate and complete responses, potentially providing high-quality medical advice in the field of Achilles tendinopathy.

METHODS

Study Design

This multicenter observational study from an international group of experts was conducted between February 25, 2024, and April 4, 2024, at The Chinese University of Hong Kong. The objective was to evaluate the performance of 3 leading LLM-based chatbots in answering 18 questions derived from a recent clinical guideline on Achilles tendinopathy.[5] The flow of the study is illustrated in Figure 1. Given the nature of the study, ethics approval was deemed unnecessary by the institutional review board.

*Address correspondence to Samuel Ka-Kin Ling, MBChB(CUHK), ChM(Edin), FHKAM, Department of Orthopaedics & Traumatology, Faculty of Medicine, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, Hong Kong, SAR, PR China (email: samuel.ling@cuhk.edu.hk).
Z.L. and M.W. contributed equally to this work.
All authors are listed in the Authors section at the end of this article.
Final revision submitted October 27, 2024; accepted December 2, 2024.
The authors declared that they have no conflicts of interest in the authorship and publication of this contribution. AOSSM checks author disclosures against the Open Payments Database (OPD). AOSSM has not conducted an independent investigation on the OPD and disclaims any liability or responsibility relating thereto.
Ethical approval was not sought for the present study.

Question Source

A thorough review of the last 5 years of clinical guidelines on Achilles tendinopathy was conducted. The Dutch guideline[5] was selected for its comprehensive and evidence-based approach, covering 6 critical aspects: risk factors, diagnosis, imaging, treatment, prognosis, and prevention. From this guideline, 18 scoping questions were extracted and validated by our research team for relevance and accuracy, ensuring they appropriately reflected current clinical understandings and practices.

Chatbot Interaction and Data Collection

Between February 25, 2024, and February 26, 2024, a set of 18 questions was posed to 3 LLM-based chatbots: ChatGPT 4.0, Claude 2, and Gemini (Table 1). Each question was posed as an isolated query to eliminate the possibility of influence from previous conversations. The specific prompts used for each chatbot are detailed in Appendix Table A1. Responses were then standardized into plaintext format, with any generic suggestions to seek professional health care advice removed for brevity.

Survey Design and Data Analysis

An online survey incorporating 18 questions and corresponding answers was created using the Qualtrics platform (Qualtrics). Twenty orthopaedic surgeons specializing in Achilles tendinopathy were invited via email to evaluate the responses. Thirteen surgeons from 9 different countries and regions (Chile, Egypt, Hong Kong, Japan, Mainland China, Philippines, Singapore, Thailand, and the United States) completed the survey between February 28 and April 4, 2024. Raters were instructed to assess the accuracy and completeness of each response based on their clinical expertise, knowledge of current best practices, and familiarity with recent literature in Achilles tendinopathy management. Accuracy was defined as alignment with current clinical guidelines and evidence-based practice, while completeness was based on how comprehensively the response addressed all aspects of the question. Importantly, the Dutch consensus was not provided as a specific reference to the evaluators. This decision was made to allow for assessments based on broader clinical expertise and to avoid biasing individual expertise-based evaluations.

Responses were scored using a modified system[17]:

- Score 4: Excellent response not requiring clarification
- Score 3: Satisfactory requiring minimal clarification
- Score 2: Satisfactory requiring moderate clarification
- Score 1: Unsatisfactory requiring substantial clarification

To minimize bias, the presentation order of the responses was randomized in the survey. The complete set of questions and corresponding answers is available in their entirety in Appendix Table A2. For each rater, the total score for a set of 18 responses was summed up. For example, if a rater evaluated all 18 responses from ChatGPT 4.0, the individual scores for these responses were added to get a total score, which could range from 18 to 72. In addition, to evaluate the reliability of the ratings among different specialists, the intraclass correlation coefficient (ICC) was calculated. This metric helps determine the extent of consistency and consensus among the raters across the evaluations.

TABLE 1
Summary of 18 Questions Regarding Achilles Tendinopathy

Module 1: Risk Factors
1. Which modifiable and nonmodifiable factors increase the risk of Achilles tendinopathy?
2. Which primary prevention strategy is most effective for Achilles tendinopathy?
Module 2: Diagnosis
3. What are the criteria for diagnosing Achilles tendinopathy?
4. Which differential diagnoses of posterior ankle pain should be considered and which underlying pathology might be related to Achilles tendinopathy?
Module 3: Imaging
5. Which imaging techniques can be used for assessing Achilles tendinopathy in clinical practice?
6. Which qualifications are required to perform imaging?
7. Which imaging findings are characteristic for Achilles tendinopathy?
8. Which imaging findings have prognostic value in Achilles tendinopathy?
Module 4: Treatment
9. Which measurement instruments are best suited for monitoring a treatment effect?
10. What is the effect of a wait-and-see policy in Achilles tendinopathy?
11. Which nonsurgical treatment is most effective for Achilles tendinopathy?
12. Is surgery more effective than nonsurgical treatment for Achilles tendinopathy?
13. Which factors influence treatment effects in Achilles tendinopathy?
14. What advice (self-management and patient education) should be given to patients with Achilles tendinopathy regarding lifestyle, work, and sports loading?
Module 5: Long-term Prognosis
15. What percentage of patients with Achilles tendinopathy have persistent symptoms for >1 year?
16. What percentage of patients with Achilles tendinopathy return to their original level of sport over a period of >1 year?
17. What factors affect the long-term prognosis (>1 year) in patients with Achilles tendinopathy?
Module 6: Preventing Recurrence
18. Which prevention strategies are effective in patients who have recovered from Achilles tendinopathy?

Statistical Analysis

Statistical analyses were performed using the SPSS statistical package (Version 22.0; IBM Corp). The mean word

count of 18 responses provided by 3 chatbots was compared using a 1-way analysis of variance. Dunnett T3 post hoc tests were conducted to adjust for multiple comparisons. The Kruskal-Wallis test was utilized to compare the mean total scores assigned to the responses of the 3 LLM-based chatbots. Proportions of scores were analyzed using a chi-square test, with a Bonferroni correction applied for multiple comparisons. The consistency among specialists' ratings was evaluated using an ICC. A P value $<.05$ was considered statistically significant.

RESULTS

Heatmap Presentation of Scores

The scores of each response provided by 3 LLM-based chatbots are visually represented in Figure 2 as a heatmap, where:

- Each row corresponds to a specialist, identified by an alphanumeric identifier.
- Each column represents one of the 18 questions.
- Each cell shows the score for a response, with color coding indicating the quality of the answer (eg, blue for unsatisfactory answers).

Preliminary observations suggest that ChatGPT 4.0's responses generally received fewer unsatisfactory scores (score 1, cells in blue) and moderate clarification ratings (score 2, cells in white) compared with those from Claude 2 and Gemini. The results of the mean word count of responses are shown in Appendix Table A4.

Comparison of Scores

The total scores for the 18 responses from each specialist were calculated and compared across the 3 LLM-based chatbots, as depicted in Figure 3. The analysis showed no significant difference in the mean total scores among ChatGPT 4.0 (59.7 ± 5.5), Claude 2 (53.4 ± 9.7), and Gemini (53.6 ± 8.4). In addition, Table 2 provides a detailed proportion of the 4 assigned scores for responses from ChatGPT 4.0, Claude 2, and Gemini. For score 1 (unsatisfactory), the proportions of unsatisfactory responses were 0.9% for ChatGPT 4.0, 3.4% for Claude 2, and 3.4% for Gemini, with statistical analysis finding no significant differences among the chatbots. For score 2 (requiring moderate clarification), ChatGPT 4.0 had significantly fewer responses at 10.3% compared with 23.9% for Claude 2 and 20.5% for Gemini ($P < .001$ for both comparisons). For score 3 (satisfactory requiring minimal clarification), there were no significant differences observed between ChatGPT 4.0 at 45.3%, Claude 2 at 45.3%, and Gemini at 50.9%. For score 4 (excellent), ChatGPT 4.0 outperformed the other chatbots with 43.6% of its responses rated as excellent, significantly higher than Claude 2 at 27.4% and Gemini at 25.2% ($P < .001$ for both comparisons).

Intraclass Correlation Coefficients

Table 3 presents the ICCs for the grading consistency among specialists evaluating the responses from 3 LLM-based chatbots: ChatGPT, Claude 2, and Gemini. The ICCs were 0.420 for ChatGPT, 0.522 for Claude 2, and 0.575 for Gemini. The ICC for ChatGPT indicates poor reliability (ICC < 0.5), and the ICCs for Claude 2 and Gemini fall within the range of 0.5 to 0.75, suggesting moderate reliability.

DISCUSSION

We hypothesized that 3 LLM-based chatbots could deliver accurate and complete responses in the context of Achilles tendinopathy. The results largely support this hypothesis, with 88.9% of ChatGPT's responses, 72.7% of Claude 2's responses, and 70.5% of Gemini's responses rated as 3 or 4, indicating either excellent or satisfactory performances with minimal clarification needed. ChatGPT 4.0 demonstrated a significantly higher proportion of excellent responses (score 4) compared with the other chatbots, aligning with previous studies that have underscored its efficiency in handling medical inquiries.[16] Despite these promising results, the wide range of scores observed in the heatmap (Figure 2) and low ICCs (Table 3) reveal substantial variability in how specialists assessed the responses. These findings introduce a critical nuance to our initial hypothesis, suggesting that while LLM-based chatbots can frequently provide high-quality responses, achieving consistent and consensual evaluations among different specialists poses a significant challenge. This inconsistency complicates any definitive conclusions the authors can draw about the reliability of these chatbots as tools for medical advice or their superiority over one another. In addition, it is crucial to recognize that chatbots synthesize information available on the internet. The variability in specialists' assessments likely reflects areas in which clear consensus or definitive knowledge in Achilles tendinopathy management is lacking. In fields with strong professional consensus and clear guidelines, chatbots are more likely to provide consistent, accurate responses. This relationship between available medical information consistency and chatbot performance highlights both the potential and limitations of these tools in medical contexts, emphasizing the importance of establishing clear clinical guidelines.

The exploration of LLM-based chatbots in providing medical advice has gained increasing attention across various medical specialties. Pioneering this investigation within the field of orthopaedics, Mika et al[17] demonstrated that ChatGPT could provide appropriate and evidence-based answers to common questions about total hip replacement, prompting further evaluations in other orthopaedic subspecialties. However, subsequent assessments of ChatGPT 3.5 in ACL reconstructions presented inconsistent results. While Li et al. reported accurate information from ChatGPT,[15] Johns et al[11] criticized the responses as outdated and superficial. Apart from the mixed results,

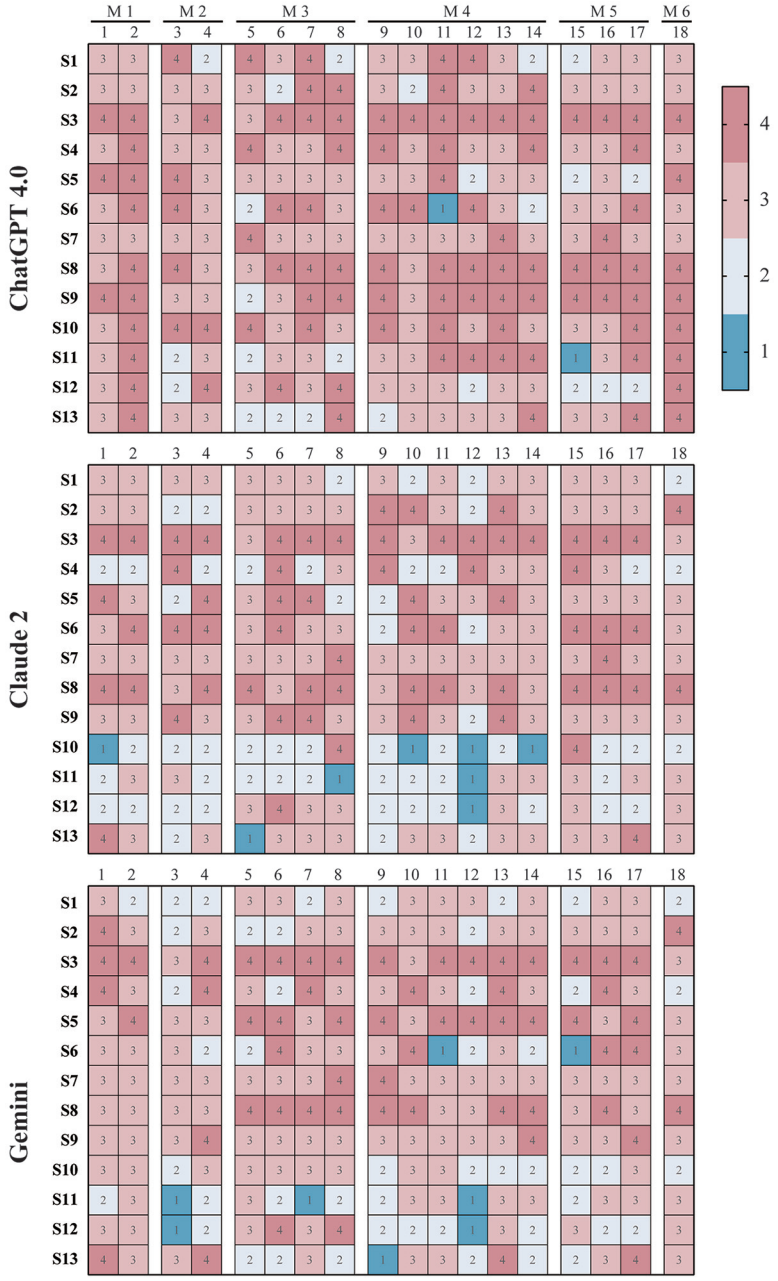

Figure 2. Scoring heatmap of responses from 3 LLM-based chatbots. The heatmap illustrates the scores assigned by specialists to the responses generated by 3 LLM-based chatbots, specifically ChatGPT 4.0, Claude 2, and Gemini. Each row within the heatmap corresponds to a specialist, identified by an alphanumeric identifier (eg, S1, S2, and S3), who assessed the chatbot responses. Each column corresponds to one of the 18 questions answered by the chatbots. Every cell in the heatmap denotes a specific score assigned to a response, with higher scores indicating greater accuracy and completeness. The questions are categorized into 6 modules (M1-M6), addressing topics such as risk factors (M1), diagnosis (M2), imaging (M3), treatment (M4), prognosis (M5), and preventing recurrence (M6).

these studies exhibited several methodological limitations that could affect their reliability. For example, the question sets were derived from various medical institution websites and were subsequently reduced to 10 questions by the authors without clear selection criteria. In addition, evaluations were performed by a limited number of raters from a single medical center, and crucially, there was no thorough assessment of the consistency and consensus in the grading by the raters. Another significant limitation of these studies was their exclusive focus on ChatGPT 3.5, which overlooks the variety of available LLM-based chatbots.

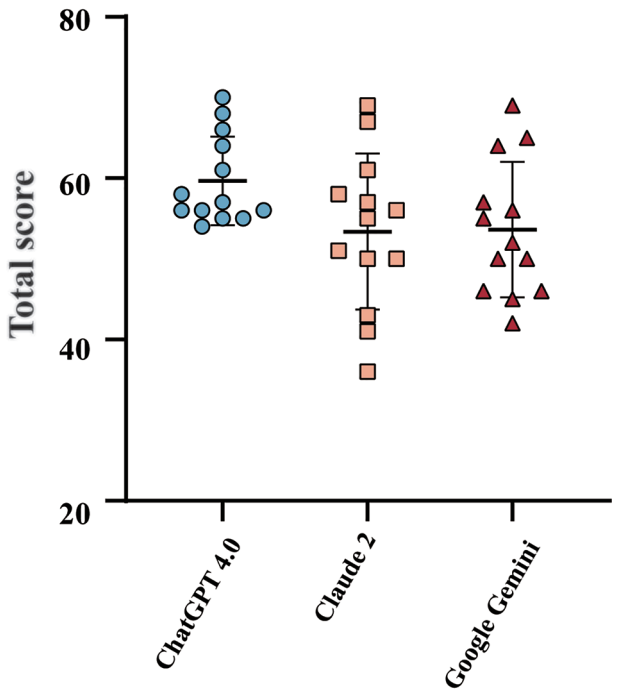

Figure 3. Comparative evaluation of total scores. The total scores for responses to the 18 Achilles tendinopathy-related questions by ChatGPT 4.0, Claude 2, and Gemini were assessed by 13 specialists. Scores are presented as means with standard deviations. Individual specialist assessments are represented by symbols (circles for ChatGPT 4.0, squares for Claude 2, and triangles for Gemini).

TABLE 2
Proportion of Response Scores Assigned by Specialists[a]

Score	ChatGPT 4.0	Claude 2	Gemini
4	102 (43.6)[b,c]	64 (27.4)	59 (25.2)
3	106 (45.3)	106 (45.3)	119 (50.9)
2	24 (10.3)[b,c]	56 (23.9)	48 (20.5)
1	2 (0.9)	8 (3.4)	8 (3.4)

[a]Data are presented as n (%). Higher scores indicate responses with greater accuracy and completeness. The row colors correspond to the scoring in the heatmap (eg, blue for score 1 and dark red for score 4).
[b]Significant differences between ChatGPT 4.0 and Claude 2.
[c]Significant differences between ChatGPT 4.0 and Gemini.

To address these shortcomings, we implemented several methodological improvements. First, we utilized 18 questions derived from a Dutch guideline on Achilles tendinopathy, developed through an evidence-based process involving a patient panel and a multidisciplinary working group.[5] This approach ensured that our questions were both clinically relevant and structured effectively for evaluation. Second, we expanded our pool of raters to specialists from different regions and countries, thereby increasing the diversity and expertise of our raters. Lastly,

TABLE 3
ICCs and 95% CIs[a]

LMM	ICC	95% CI Lower Bound	95% CI Upper Bound
ChatGPT 4.0	0.420	−0.068	0.747
Claude 2	0.522	0.119	0.791
Google Gemini	0.575	0.216	0.814

[a]Intraclass correlation coefficients (ICCs) measure the consistency of evaluations among specialists, with values closer to 1 indicating higher agreement. LMM, large language model.

we broadened our evaluation by including 2 additional leading LLM-based chatbots, Claude 2 and Gemini. This comparative approach allowed us to assess the current landscape of advanced LLM technologies and mirrors real-world scenarios in which patients and health care providers might interact with various LLM-based chatbots.

In our improved study design, we evaluated the performance of 3 chatbots. Our evaluation particularly focused on responses categorized as unsatisfactory, as these could provide deeper insights compared with the high-scoring responses. From the heatmap (Figure 2), we identified that 3 responses were scored as "unsatisfactory" by multiple raters: 1 response from Gemini to question 3 and 2 responses to question 12 by both Claude 2 and Gemini. Specifically, Gemini's response to question 3 (What are the criteria for diagnosing Achilles tendinopathy?) provided a generalized list of diagnostic criteria but was not comprehensive, failing to distinguish between the diagnostic criteria for insertional and midportion Achilles tendinopathy. This issue was also prevalent in a large portion of responses from all 3 chatbots. For question 12 (Is surgery more effective than nonsurgical treatment for Achilles tendinopathy?), both chatbots Claude 2 and Gemini delivered overly generalized responses that failed to directly address the question. Notably, Claude 2's response has contradictory statements. Initially, it suggested that surgery should be considered only after a minimum of 3 to 6 months of nonoperative therapy, but later it stated that surgery is typically reserved for those who do not see improvement after 6 to 12 months of nonsurgical management. This inconsistency not only undermines the reliability of the response but also highlights a critical gap in the chatbots' ability to provide logical and precise medical advice. Similarly, Gemini's response misclassified Achilles tendinopathy into acute rupture and chronic tendinopathy, further highlighting the unreliability of chatbot responses with misleading information. ChatGPT 4.0 was not without its own issues. Although fewer of its responses were rated as unsatisfactory, some provided inaccurate information. For example, in response to question 5 (Which imaging techniques can be used for assessing Achilles tendinopathy in clinical practice?), which required moderate clarification by 4 raters, ChatGPT 4.0 included an exhaustive list of imaging techniques including ultrasound, magnetic resonance imaging, computed tomography, dual-energy x-ray absorptiometry, and even bone

scans, indiscriminately grouping common orthopaedic imaging methods without distinguishing their specific relevance to Achilles tendinopathy, thus demonstrating a lack of critical filtering of relevant information. These findings highlight significant limitations in providing updated, accurate, and appropriately filtered health care information in all 3 chatbots. Such shortcomings are particularly concerning in the field of health care, where even minor inaccuracies can complicate communication between clinicians and patients and potentially influence treatment compliance and outcomes.

In addition to the misinformation and lack of depth in the responses, we also observed inconsistencies and a lack of consensus among the raters. The heatmap (Figure 2) provides a visual representation of this issue. For instance, in the column corresponding to question 15 (What percentage of patients with Achilles tendinopathy have persistent symptoms for >1 year?) in the ChatGPT 4.0 block, the cells displayed a wide range of scores from 1 to 4, highlighting different opinions. A similar pattern of scoring variation was evident in responses from the other 2 chatbots. Furthermore, certain raters, such as S3 and S8, consistently awarded high scores across all models, while others, like S11 and S12, exhibited significant scoring discrepancies among the different chatbots' responses. Moreover, the ICC for ChatGPT 4.0 was 0.420, indicating poor reliability, whereas Claude 2 and Gemini demonstrated moderate reliability with ICCs of 0.522 and 0.575, respectively. These low ICCs underscore the inconsistency in rater opinions. As a result, it remains challenging to conclusively determine the reliability of LLM-based chatbots as a supplement to patient education or to effectively compare their performance.

As discussed in an editorial commentary,[10] the complexity and contentious nature of the topics under discussion likely contribute to the varying opinions among raters. More importantly, the lack of standardized assessment and reporting methods may also contribute to the mixed results observed. Our study adopts a 4-point scale for evaluating responses based on accuracy and completeness, building on the framework used by 3 previous studies.[11,15,17] Additionally, other different evaluation methods are also used, such as a 9-point scale segmented into correctness, completeness, and adaptability,[12] or a combined approach using a 5-point accuracy scale alongside a 5-point relevance scale.[22] Although these methods differ, they fundamentally assess the responses on similar dimensions, accuracy, and completeness. However, it is essential to acknowledge that evaluating responses based simply on these 2 criteria might be insufficient. Additional factors such as adaptability, evidence-based information, currency of medical evidence, safety considerations, and ethical implications are also critical. Supporting this perspective, Huo et al[9] emphasized that the lack of consistent and scientific reporting and assessment standards in this kind of research prevents the interpretation of results and hinders readers' ability to critically evaluate study methodologies. As research in this domain expands, there is a clear and pressing need for standardized assessment methods and structured reporting protocols. Such guidelines would standardize the evaluation of LLM-based chatbots across studies and ensure that assessments are comprehensive and reflective of all relevant dimensions.

Limitations

Several limitations of this study warrant consideration. First, the inherent variability in expert assessments poses a challenge to establishing a true gold standard for evaluating LLM-generated medical information. While experienced clinicians are often considered the best available standard, their judgments can vary significantly based on individual experiences, regional practices, and interpretations of current evidence. Although our panel included experts from North America, South America, and Africa, the majority were primarily based in Asia, potentially limiting the global diversity of clinical perspectives on Achilles tendinopathy management. Nevertheless, the inclusion of specialists from diverse centers improves this study's generalizability compared with previous research. Second, the rapid evolution of LLM-based chatbots, exemplified by Anthropic's release of its Claude 3 model during our data analysis phase, indicates that our findings may not capture the capabilities of the latest models. Third, our use of a previously reported scoring system, which assesses responses based solely on accuracy and completeness, may not encompass all relevant criteria. This limitation underscores the critical need for developing comprehensive, standardized evaluation and reporting guidelines in future research. To address these limitations, future studies could benefit from a larger, more geographically diverse panel of evaluators, the incorporation of standardized guidelines in the evaluation process, and the development of more comprehensive assessment criteria that keep pace with rapidly advancing LLM technologies.

CONCLUSION

Overall, our findings reveal that while LLM-based chatbots like ChatGPT 4.0 can deliver high-quality responses to questions related to Achilles tendinopathy, the substantial variability in the assessments by specialists, evidenced by a wide range of scores and low ICCs, highlights the difficulty in achieving consistent and consensual evaluations. Although these LLM-based technologies hold promise for enhancing access to medical information, their use should be carefully managed with expert oversight and should not replace direct consultations with appropriately experienced health care professionals.

Authors

Zuru Liang, PhD (Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, SAR, China); Ming Wang, PhD (Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, SAR, China); Nasef Mohamed Nasef Abdelatif, MD (DrNasef OrthoClinic for Foot and Ankle Orthopedic Disorders, Cairo, Egypt); Marut Arunakul, MD (Department of Orthopedic

Surgery, Faculty of Medicine, Thammasat University, Pathumthani, Thailand); Carlo Angelo V Borbon, MD (Makati Medical Center, Makati, Philippines); Keen Wai Chong, MD (Duke-NUS Graduate Medical School, Singapore); Man Wai Chow, MD (Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, SAR, China); Yinghui Hua, MD (Department of Sports Medicine, Huashan Hospital, Fudan University, Shanghai, China); David Oji, MD (Foot and Ankle Surgery, Department of Orthopaedic Surgery, Stanford University School of Medicine, Redwood City, California, USA); Ximena Ahumada, MD (Hospital Clínico San Borja Arriarán, Santiago, Chile); Kwai Ming Siu, MD (Department of Orthopaedics and Traumatology, Princess Margaret Hospital, Hong Kong, China); Ken Jin Tan, MD (OrthoSports Clinic for Orthopedic Surgery and Sports Medicine, Mt. Elizabeth Novena Specialist Centre, Singapore); Yasuhito Tanaka, MD (Department of Orthopaedic Surgery, Nara Medical University, Kashihara, Nara, Japan); Akira Taniguchi, MD (Department of Orthopaedic Surgery, Nara Medical University, Kashihara, Nara, Japan); Patrick Shu-Hang Yung, MRCS(Edin) (Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, SAR, China); and Samuel Ka-Kin Ling, MRCS(Edin) (Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, Hong Kong, SAR, China).

Supplemental Material for this article is available at https://journals.sagepub.com/doi/full/10.1177/23259671251332596#supplementary-materials.

REFERENCES

1. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model Chatbot responses to online patient eye care questions. *JAMA Netw Open*. 2023;6(8):e2330320.
2. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):e230582.
3. Chimenti RL, Post AA, Rio EK, et al. The effects of pain science education plus exercise on pain and function in chronic Achilles tendinopathy: a blinded, placebo-controlled, explanatory, randomized trial. *Pain*. 2023;164(1):e47-e65.
4. Cooper MT. Common painful foot and ankle conditions: a review. *JAMA*. 2023;330(23):2285-2294.
5. de Vos RJ, van der Vlist AC, Zwerver J, et al. Dutch multidisciplinary guideline on Achilles tendinopathy. *Br J Sports Med*. 2021;55(20):1125-1134.
6. Decker H, Trang K, Ramirez J, et al. Large language model-based Chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw Open*. 2023;6(10):e2336997.
7. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, et al. Online health information seeking among US adults: measuring progress toward a Healthy People 2020 objective. *Public Health Rep*. 2019;134(6):617-625.
8. Garcia P, Ma SP, Shah S, et al. Artificial intelligence-generated draft replies to patient inbox messages. *JAMA Netw Open*. 2024;7(3):e243201.
9. Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med*. 2023;29(12):2988.
10. Hurley ET, Crook BS, Dickens JF. Editorial commentary: At present, ChatGPT cannot be relied upon to answer patient questions and requires physician expertise to interpret answers for patients. *Arthroscopy*. 2024;40(7):2080-2082.
11. Johns WL, Martinazzi BJ, Miltenberg B, Nam HH, Hammoud S. ChatGPT provides unsatisfactory responses to frequently asked questions regarding anterior cruciate ligament reconstruction. *Arthroscopy*. 2024;40(7):2067-2079.e1.
12. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(11):5190-5198.
13. Kearney RS, Ji C, Warwick J, et al. Effect of platelet-rich plasma injection vs sham injection on tendon dysfunction in patients with chronic midportion Achilles tendinopathy: a randomized clinical trial. *JAMA*. 2021;326(2):137-144.
14. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health*. 2023;2(2):e0000198.
15. Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. *Arthroscopy*. 2024;40(7):2058-2066.
16. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBioMedicine*. 2023;95:104770.
17. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. 2023;105(19):1519-1526.
18. Reider B. Achilles' heel. *Am J Sports Med*. 2021;49(7):1707-1710.
19. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. *JAMA Netw Open*. 2023;6(12):e2346721.
20. van der Vlist AC, Winters M, Weir A, et al. Which treatment is most effective for patients with Achilles tendinopathy? A living systematic review with network meta-analysis of 29 randomised controlled trials. *Br J Sports Med*. 2021;55(5):249-256.
21. Verges J, Martínez N, Pascual A, Bibas M, Santiña M, Rodas G. Psychosocial and individual factors affecting quality of life (QoL) in patients suffering from Achilles tendinopathy: a systematic review. *BMC Musculoskelet Disord*. 2022;23(1):1114.
22. Zhang S, Liau ZQG, Tan KLM, Chua WL. Evaluating the accuracy and relevance of ChatGPT responses to frequently asked questions regarding total knee replacement. *Knee Surg Relat Res*. 2024;36(1):15.