

ORIGINAL RESEARCH

ChatGPT-o1 Preview Outperforms ChatGPT-4 as a Diagnostic Support Tool for Ankle Pain Triage in Emergency Settings

Pooya Hosseini-Monfared¹, Shayan Amiri², Alireza Mirahmadi³, Amirhossein Shahbazi⁴, Aliasghar Alamian¹, Mohammad Azizi¹, Seyed Morteza Kazemi^{1*}

1. Bone Joint and Related Tissues Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran

2. Bone and Joint Reconstruction Research Center, Department of Orthopedics, School of Medicine, Iran University of Medical Sciences, Tehran, Iran

3. Musculoskeletal Translational Innovation Initiative, Carl J. Shapiro Department of Orthopaedic Surgery, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA

4. Student Research Committee, School of Medicine, Ilam University of Medical Sciences, Ilam, Iran

Received: February 2025; Accepted: March 2025; Published online: 5 April 2025

Abstract: **Introduction:** ChatGPT, a general-purpose language model, is not specifically optimized for medical applications. This study aimed to assess the performance of ChatGPT-4 and o1-preview in generating differential diagnoses for common cases of ankle pain in emergency settings. **Methods:** Common presentations of ankle pain were identified through consultations with an experienced orthopedic surgeon and a review of relevant hospital and social media sources. To replicate typical patient inquiries, questions were crafted in simple, non-technical language, requesting three possible differential diagnoses for each scenario. The second phase involved designing case vignettes reflecting scenarios typical for triage nurses or physicians. Responses from ChatGPT were evaluated against a benchmark established by two experienced orthopedic surgeons, with a scoring system assessing the accuracy, clarity, and relevance of the differential diagnoses based on their order. **Results:** In 21 ankle pain presentations, ChatGPT-o1 preview outperformed ChatGPT-4 in both accuracy and clarity, with only the clarity score reaching statistical significance ($p < 0.001$). ChatGPT-o1 preview also had a significantly higher total score ($p = 0.004$). In 15 case vignettes, ChatGPT-o1 preview scored better on diagnostic and management clarity, though differences in diagnostic accuracy were not statistically significant. Among 51 questions, ChatGPT-4 and ChatGPT-o1 preview produced incorrect responses for 5 (9.8%) and 4 (7.8%) questions, respectively. Inter-rater reliability analysis demonstrated excellent reliability of the scoring system with interclass coefficients of 0.99 (95% CI, 0.998–0.999) for accuracy scores and 0.99 (95% CI, 0.990–0.995) for clarity scores. **Conclusion:** Our findings demonstrated that both ChatGPT-4 and ChatGPT-o1 preview provide acceptable performance in the triage of ankle pain cases in emergency settings. ChatGPT-o1 preview outperformed ChatGPT-4, offering clearer and more precise responses. While both models show potential as supportive tools, their role should remain supervised and strictly supplementary to clinical expertise.

Keywords: Artificial intelligence; Ankle; Emergency medicine; Triage

Cite this article as: Hosseini-Monfared P, Amiri S, Mirahmadi A, et al. ChatGPT-o1 Preview Outperforms ChatGPT-4 as a Diagnostic Support Tool for Ankle Pain Triage in Emergency Settings. Arch Acad Emerg Med. 2025; 13(1): e42. <https://doi.org/10.22037/aaemj.v13i1.2580>.

1. Introduction

Large Language Models (LLMs), such as ChatGPT, are a specific type of artificial intelligence (AI) that uses natural language processing (NLP) and generative AI to provide humanized text output (1). ChatGPT is a general-purpose LLM that is trained on the wide source of data with the purpose of wide application uses (2, 3). Since the ChatGPT models are not designed specifically for medical applications, there is doubt

about the accuracy and performance of these models in the field of medicine (4, 5). Different uses have been suggested for LLMs like ChatGPT in medicine and surgery, including assistance for diagnosis, prognosis, management, and patient education (6-9). Despite their potential, LLMs present challenges, including the generation of inaccurate information and hallucinations, which contribute to physicians' hesitancy to rely on these models in medical practice (2, 10).

Ankle injury is one of the most common causes for people to visit the emergency departments accounting for around 20% of all injury-related visits to emergency departments (11, 12). These injuries are mostly caused by sport-related activities following landing and direction changes (13-15). Ankle pain can be due to ligament injuries, mostly lateral ligaments or ankle fractures (16-18). Although not life-threatening, ac-

* **Corresponding Author:** Seyed Morteza Kazemi; Bone Joint and Related Tissues Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: kazemimort@gmail.com, Tel: +989121141341, ORCID: <https://orcid.org/0000-0003-3631-1305>.

curate diagnosis and proper management are necessary to prevent chronic ankle instability, residual ankle symptoms, or recurrent injuries (14, 19, 20). The application of LLMs in emergency departments as diagnostic and triage assistants can improve performance accuracy and save time for healthcare providers in busy emergency departments (21, 22). There have been a few studies evaluating the accuracy of LLMs in assisting medical diagnoses (23-25). These studies mainly focused on complaints related to internal medicine and have largely assessed the performance of earlier ChatGPT models. This study aimed to evaluate the performance and accuracy of ChatGPT-4 and o1-preview models in providing diagnoses as a potential assistant for both patients and triage staff in the emergency departments.

2. Methods

2.1. Study design and setting

We conducted an observational study to evaluate the accuracy of different LLM models for the diagnosis of ankle pain complaints in September 2024 in Iran. Frequent clinical presentations of patients with ankle pain were determined by consulting an experienced orthopedic surgeon and observing frequently asked questions on the hospital websites. Questions were phrased in non-technical and simple language to mimic typical patient inquiries. For patient inquiries, each scenario asked for 3 differential diagnoses. For the second section of the study, case vignettes were developed from the viewpoint of a triage nurse or physician according to common scenarios and the experience of the authors. For these cases, we asked the ChatGPT models and physicians to provide the diagnosis and the next management step.

We used ChatGPT-4 and ChatGPT-o1 preview as two large language models to evaluate their accuracy in answering the questions about ankle pain. ChatGPT-4 is the advanced model of ChatGPT developed by OpenAI. The o1-preview model by OpenAI is their latest model designed to use "the chain of thought" to process queries and engineering to perform better in reasoning and solving problems. All searches were done on a single date (September 25, 2024). No custom instructions were used. The memory option of ChatGPT was turned off. ChatGPT responses were recorded after the initial query, and no follow-up or repeat queries were conducted.

Two expert orthopedic surgeons with more than 10 years of experience (SM.K and A.A) also separately answered these questions. After the initial responses were recorded a consensus was reached for cases of disagreement. A third independent orthopedic surgeon (Sh.A) graded the answers of the LLMs in comparison to the experts' answers.

We used a modified version of a rating system suggested by Mika et al. for scoring the accuracy and correctness of differential diagnoses considering the order of the diagnoses (26). Table 1 summarizes the scoring system for differential diagnoses. For case vignettes, clarity was scored according to the

scoring system presented in Table 1, and accuracy was rated 1 for correct and 0 for incorrect answers.

In evaluating the next steps of management for the case vignettes, common initial treatments, such as rest, ice, compression, and elevation (RICE), were considered standard for all cases and not included in the scoring. The focus of our evaluation was on the accuracy of the next steps recommended by the LLMs beyond RICE, which is the well-established initial management for ankle pain or sprain. Two independent raters rated the answers of ChatGPT models based on the physician's answers and the average score of the two raters was used for analysis. Inter-rater reliability was calculated using an intraclass correlation coefficient (ICC) based on absolute agreement, a two-way mixed effect model.

2.2. Statistical analysis

Scores were reported as mean \pm standard deviation (SD). The normality of the scores was evaluated by the Kolmogorov-Smirnov test. Independent t-test or Mann Whitney U-test were employed to compare the scores between the two ChatGPT models. Inter-rater reliability was evaluated via ICC based on the absolute agreement, two-way mixed effect model. P-values less than 0.05 were considered statistically significant. All the analyses were performed using IBM SPSS v.26.0 software (IBM, Armonk, New York).

2.3. Ethical considerations

This study did not involve direct interaction with human or animal subjects. The research evaluated the accuracy of LLMs using synthetic, predefined clinical scenarios for common ankle pain complaints and scenarios. No real patient data were utilized, and no individuals were recruited. Since this work did not involve human participants, identifiable data, or interventions, institutional review board (IRB) or ethics committee approval was not required.

3. Results

We asked ChatGPT and orthopedic surgeons about 21 common presentations of patients with ankle pain. The questions and differential diagnoses provided by the physicians and ChatGPT models are summarized in Table 2. ChatGPT-o1 preview's accuracy score was higher than ChatGPT-4, but the difference was not significant (Table 3). The clarity score of ChatGPT-o1 preview was significantly higher than ChatGPT-4. Also, the total score of ChatGPT-o1 preview was higher than ChatGPT-4 (Figure 1).

3.1. Case vignettes

We asked the ChatGPT models about the most probable diagnosis and the next step of the management regarding 15 case vignettes. A summary of the case vignettes, along with the physicians' and ChatGPT models' responses, is presented in Table 4. ChatGPT-o1 preview had a higher total score for the diagnosis questions compared to ChatGPT-4. However, the difference was not statistically significant. Also, the clarity

and accuracy scores were higher for ChatGPT-o1 preview. For the management questions, ChatGPT-o1 preview had higher clarity and total scores (Figure 1, Table 5).

Out of the total of 51 questions asked from ChatGPT models, including both presentations and case vignettes, ChatGPT-4 and ChatGPT-o1 preview gave incorrect answers to 5 (9.8%) and 4 (7.8%) questions, respectively. The agreement between the two evaluators of the ChatGPT model's performance demonstrated an inter-rater reliability of 0.99 (95% CI: 0.998–0.999) for accuracy scores and 0.98 (95% CI: 0.979–0.990) for clarity scores.

4. Discussion

We evaluated the accuracy of the two most advanced models of ChatGPT for the triage of patients with ankle pain by comparing the answers of the common models for common scenarios with the expert's answers. ChatGPT-o1 preview outperformed ChatGPT-4 in providing differential diagnoses for common ankle pain complaints. ChatGPT-o1 preview described the diagnoses more clearly and used more specific medical terminology. However, both models had similar accuracy and clarity for case vignette diagnosis and management steps. The excellent inter-rater reliability observed among the evaluators of ChatGPT responses demonstrated that the scoring system proposed in this study is consistently interpretable by different evaluators.

Ankle injuries are among the most common extremity complaints among patients presenting to emergency departments (13, 27). Although most of these injuries are benign, accurate diagnosis and management of these injuries are necessary to prevent prolonged morbidity (27).

In a study by Hoppe et al., the performance of ChatGPT-3.5 and ChatGPT-4 were compared with emergency department physicians for patients with internal medicine complaints based on the discharge diagnosis as ground truth (25). They found that the ChatGPT-4 outperformed its previous version and also the emergency department resident physicians (25). Kunze et al. evaluated the performance of ChatGPT-4 for its triage capabilities for patients presenting with knee pain (28). They evaluated 20 knee complaints and extended case vignettes and found that ChatGPT provided reasonable diagnoses to triage patients. Their study highlighted the importance of the prompting strategy and found that refinements are necessary to get the best performance of the ChatGPT-4 model (28, 29).

The diagnostic accuracy of LLMs is dependent on the quality of information and physical examination provided. Therefore, if the patient or the healthcare provider is unable to accurately identify the primary source of the patient's pain, these models often provide incorrect diagnoses (30). Also, the phrasing of the prompt input plays a crucial role in guiding the diagnostic outcome, which may account for incorrect diagnoses generated by the ChatGPT models in the second case vignette of our study (29). In this case, ChatGPT-4 diagnosed knee injury involving the posterior cruciate liga-

ment (PCL), and ChatGPT-o1-preview provided the diagnosis of posterior hip dislocation, neither of which aligned with the patient's ankle pain, which was probably due to ankle fracture that required x-ray evaluation for confirmation.

This study is the first study to compare the performance and accuracy of ChatGPT-o1 preview with ChatGPT-4 for medical assistance in the emergency department. We found that the newer reasoning model of ChatGPT had improved in clarity and accuracy and may be useful as a clinical support tool. It is important to ensure these models demonstrate equity, fairness, and a low error rate. We observed that the ChatGPT models provided incorrect answers for 7.8% to 9.8% of the answers, indicating that these models have not reached the ideal accuracy to be used in clinical settings without supervision and should only be used as a support tool for healthcare providers. We also observed higher accuracy in cases that healthcare professionals described compared to patient inquiries. However, we should note that ChatGPT advised patients to visit a healthcare professional at the end of the responses to the patient's questions.

5. Limitations

Our study had some limitations that should be considered. The first ChatGPT answers were considered for evaluation, and the performance of LLMs might have improved by providing multiple inputs and prompts. We did not use prompt engineering methods to guide the ChatGPT models to act as a healthcare expert to evaluate how a patient may use this tool as a guide, although for case vignettes, we used more medical terms to mimic the condition where a triage nurse would use the ChatGPT models. Grading of the ChatGPT responses was somehow subjective, even by the use of a defined scoring system. Also, it is possible that the wording of the case scenarios and questions influence the generated outputs, known as prompt engineering bias, though we tried to minimize this bias by asking three different people to develop the case scenarios.

6. Conclusions

Our study demonstrates that both ChatGPT-4 and ChatGPT-o1 preview provide acceptable performance in triage and assessment of patients with ankle pain presenting to emergency rooms. ChatGPT-o1 preview outperformed ChatGPT-4 in most scenarios, offering answers with more clearly described conditions and more precise medical terminology. However, while both models show promise as supportive tools, their use should be supervised and regarded solely as an aid for triage nurses and staff rather than as a replacement for clinical expertise.

7. Declarations

7.1. Acknowledgments

We would like to thank those who contributed to the performance of this study.

7.2. Author Contribution

PHM, ShA, AM, AS and SMK conceptualized the study. PHM, ShA, AS, AA, and MA Performed the investigation. PHM Carried out the analysis. PHM, ShA, AM, AS, and MA drafted and AA and SMK revised the paper. All authors read and approved the final manuscript.

7.3. Funding

The authors received no financial or material support for the research, authorship, and/or publication of this article.

7.4. Conflicts of interest

The authors have no relevant financial or non-financial interests to disclose.

7.5. Data Availability

The data used and analyzed during the current study are available from the corresponding author upon reasonable request.

7.6. Using artificial intelligence chatbots

AI chatbots, specifically ChatGPT models, were utilized solely for the purpose of generating responses to clinical scenarios in order to analyze their accuracy for triaging ankle pain. However, no AI chatbots or large language models were used in drafting, editing, or preparing the manuscript text. All writing, data analysis, and interpretation were conducted independently by the authors.

References

- Hadi MU, Qureshi R, Shah A, Irfan M, Zafar A, Shaikh MB, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Author Preprints*. 2023.
- Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Taylor & Francis*; 2023. p. 277-304.
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*. 2023;388(13):1233-9.
- Yao JJ, Aggarwal M, Lopez RD, Namdari S. Current Concepts Review: Large Language Models in Orthopaedics: Definitions, Uses, and Limitations. *JBJS*. 2021;10.2106.
- Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. *JAMA*. 2023;330(9):866-9.
- Merrell LA, Fisher ND, Egol KA. Large language models in orthopaedic trauma: a cutting-edge technology to enhance the field. *JBJS*. 2023;105(17):1383-7.
- Kaboudi N, Firouzbakht S, Shahir Eftekhari M, Fayazbakhsh F, Joharivarnoosfaderani N, Ghaderi S, et al. Diagnostic Accuracy of ChatGPT for Patients' Triage; a Systematic Review and Meta-Analysis. *Arch Acad Emerg Med*. 2024;12(1):e60.
- Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. *Journal of Medical Internet Research*. 2023;25:e47479.
- Chlorogiannis DD, Apostolos A, Chlorogiannis A, Palaodimos L, Giannakoulas G, Pargaonkar S, et al. The Role of ChatGPT in the Advancement of Diagnosis, Management, and Prognosis of Cardiovascular and Cerebrovascular Disease. *Healthcare (Basel)*. 2023;11(21).
- Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH, et al. Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. *Obesity surgery*. 2023;33(6):1790-6.
- Shah S, Thomas AC, Noone JM, Blanchette CM, Wikstrom EA. Incidence and cost of ankle sprains in United States emergency departments. *Sports health*. 2016;8(6):547-52.
- Shah S, Thomas A, Noone J, Blanchette C, Wikstrom E. Emergency department visits associated with ankle sprains in the United States. *Value in Health*. 2016;19(3):A12.
- Doherty C, Delahunt E, Caulfield B, Hertel J, Ryan J, Bleakley C. The incidence and prevalence of ankle sprain injury: a systematic review and meta-analysis of prospective epidemiological studies. *Sports medicine*. 2014;44:123-40.
- Halabchi F, Hassabi M. Acute ankle sprain in athletes: Clinical aspects and algorithmic approach. *World journal of orthopedics*. 2020;11(12):534.
- Kerkhoffs GM, van den Bekerom M, Elders LA, van Beek PA, Hullegie WA, Bloemers GM, et al. Diagnosis, treatment and prevention of ankle sprains: an evidence-based clinical guideline. *British journal of sports medicine*. 2012;46(12):854-60.
- Court-Brown CM, McBurnie J, Wilson G. Adult ankle fractures—an increasing problem? *Acta Orthopaedica Scandinavica*. 1998;69(1):43-7.
- Renström PA, Lynch SA. Ankle ligament injuries. *Revista Brasileira de Medicina do Esporte*. 1998;4:71-80.
- Polzer H, Kanz KG, Prall WC, Haasters F, Ockert B, Mutschler W, Grote S. Diagnosis and treatment of acute ankle injuries: development of an evidence-based algorithm. *Orthop Rev (Pavia)*. 2012;4(1):e5.
- Manske RC, Wolfe C, Page P, Voight M. Enhancing the Diagnosis of Lateral Ankle Sprains: The Role of MSK Diagnostic Ultrasound in Evaluating ATFL and CFL. *International Journal of Sports Physical Therapy*. 2024;19(2):245.
- Cavazos Jr GJ, Harkless LB. The epidemiology, evaluation, and assessment of lateral ankle sprains in athletes.

- Journal of Sports Medicine and Therapy. 2021.
21. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*. 2023;23(1):689.
 22. HaghighiKian SM, Shirinzadeh-Dastgiri A, Vakili-Ojarood M, Naseri A, Barahman M, Saberi A, et al. A Holistic Approach to Implementing Artificial Intelligence in Lung Cancer. *Indian Journal of Surgical Oncology*. 2024;1-22.
 23. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada health and WebMD symptom checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. *JMIR mHealth and uHealth*. 2023;11(1):e49995.
 24. Le K, Chen J, Mai D, Le KDR. An Evaluation on the Potential of Large Language Models for Use in Trauma Triage. *Emergency Care and Medicine*. 2024;1(4):350-67.
 25. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT with GPT-4 outperforms emergency department physicians in diagnostic accuracy: retrospective analysis. *Journal of Medical Internet Research*. 2024;26:e56110.
 26. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *JBJS*. 2023;105(19):1519-26.
 27. Wedmore I, Young S, Franklin J. Emergency department evaluation and management of foot and ankle pain. *Emergency Medicine Clinics*. 2015;33(2):363-96.
 28. Kunze KN, Varady NH, Mazzucco M, Lu AZ, Chahla J, Martin RK, et al. The large language model ChatGPT-4 exhibits excellent triage capabilities and diagnostic performance for patients presenting with various causes of knee pain. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*. 2024.
 29. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*. 2023;25:e50638.
 30. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-80.

Table 1: Scoring for accuracy and clarity of ChatGPT responses

Score	Description
Accuracy	
0	The diagnosis was not among the top 3 differential diagnoses
1	The diagnosis was among the top 3 differential diagnoses but was ranked 2 levels apart from its true position
2	The diagnosis was among the top 3 differential diagnoses but was ranked 1 level apart from its true position
3	The diagnosis was among the top 3 differential diagnoses, and its ranking was right
Clarity	
0	Unclear: the response is unclear and fails to convey the intended message, making it difficult to understand or requiring significant clarification, or is incorrect.
1	Partially Clear: the response is clear but may need moderate clarification in certain areas to ensure complete understanding.
2	Moderately Clear: the response is mostly clear but may require minimal clarification for full understanding.
3	Completely Clear: the response is entirely clear and requires no additional clarification.

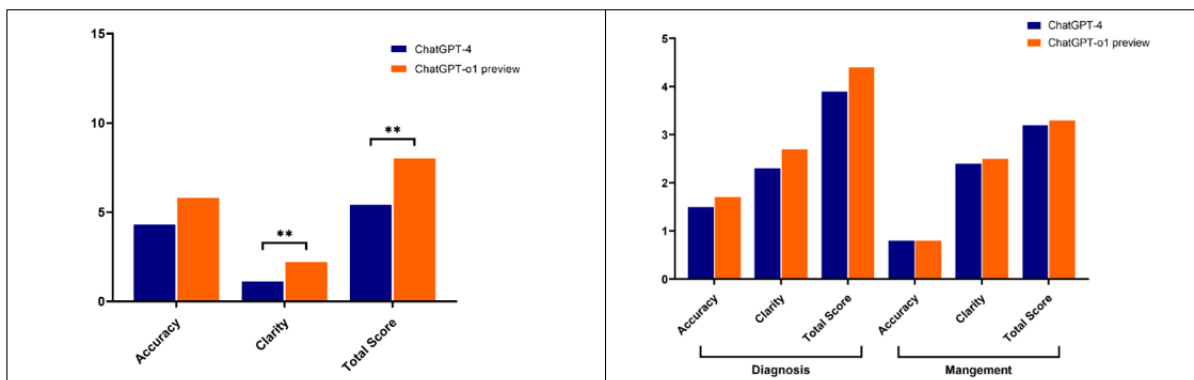


Figure 1: Comparison of ChatGPT models in diagnosing ankle pain presentations (Left) as well as in diagnosing and managing clinical case vignettes (right).

Table 2: The questions and differential diagnoses provided by the physicians and ChatGPT models (continue)

No.	Questions	Physicians' Response	ChatGPT-4		ChatGPT-o1 preview	
			Response	Score	Response	Score
1	My left foot rotated outward during playing soccer and felt pain with a little bit of swelling but I can bear my weight and walk on it. What are the 3 most likely diagnoses?	1. Deltoid Ligament Sprain 2. Medial Malleol fracture (tip of medial Malleol) 3. Syndesmotom Injury	1. Ankle Sprain 2. Tendon Strain or Tendinitis 3. Subtle Fracture	Accuracy: 5/9 Clarity: 1/3	1. High Ankle Sprain (Syndesmotom Injury) 2. Medial Ankle Sprain (Deltoid Ligament Sprain) 3. Lateral Ankle Sprain (Anterior Talofibular and Calcaneofibular Ligament Injury)	Accuracy: 3/9 Clarity: 2.5/3
2	In a football game my right foot rotated inward and I wasn't able to continue the game because I couldn't bear my weight and was unable to ambulate. What are the 3 most likely diagnoses?	1. Anterior talofibular ligament (ATFL) injury 2. Lat. Malleolar fracture 3. Syndesmotom Injury	1. Ankle Sprain (Inversion Injury) 2. Fracture (Ankle or Foot) 3. Tendon Injury (e.g., Peroneal Tendon Dislocation or Tear)	Accuracy: 6/9 Clarity: 1.5/3	1. Posterior hip dislocation 2. Femoral neck fracture 3. Intertrochanteric femur fracture	Accuracy: 0/9 Clarity: 0/3
3	I twisted my ankle outward and have a swelling, bruises, and pain in my left ankle and I feel joint instability in my ankle. What are the 3 most likely diagnoses?	1. Deltoid ligament tear 2. Medial malleolar fracture 3. Syndesmotom tears	1. Ankle Sprain (Lateral Ligament Injury) 2. Peroneal Tendon Injury 3. Fracture (Avulsion Fracture or Fibular Fracture)	Accuracy: 2.5/9 Clarity: 2/3	1. Deltoid Ligament Sprain or Tear (Medial Ankle Sprain) 2. Medial Malleolus 3. Fracture (Ankle Fracture) Syndesmotom Injury (High Ankle Sprain)	Accuracy: 9/9 Clarity: 3/3
4	I just twisted my ankle during running and I can barely walk because of it. What are the 3 most likely diagnoses?	1. Ligament injury 2. Malleolar fractures 3. Syndesmotom Injury	1. Ankle Sprain 2. Ankle Strain 3. Fracture	Accuracy: 5/9 Clarity: 1/3	1. Ankle Sprain 2. Ankle Fracture 3. Tendon Injury or Rupture	Accuracy: 6/9 Clarity: 1.5/3
5	My right foot twisted outward during hiking, now I have a swelling and bruises and can't even touch my ankle due to severe pain. What are the 3 most likely diagnoses?	1. Medial malleolar fracture 2. Syndesmotom injuries 3. Ligament injury	1. Ankle Sprain 2. Fracture 3. Tendon Injury	Accuracy: 4/9 Clarity: 1/3	1. Ankle Sprain 2. Ankle Fracture 3. Peroneal Tendon Injury	Accuracy: 4/9 Clarity: 1/3
6	During a basketball game suddenly felt a sharp pain on the outside of my left ankle and I was not able to continue the game because I couldn't bear my weight and it immediately started to swell up. What are the 3 most likely diagnoses?	1. ATFL + FC tear 2. Lateral malleolar fracture 3. Peroneus tendon subluxation	1. Ankle Sprain 2. Ankle Fracture 3. Tendon Injury	Accuracy: 9/9 Clarity: 1/3	1. Lateral Ankle Sprain 2. Fracture of the Lateral Malleolus 3. Peroneal Tendon Injury	Accuracy: 9/9 Clarity: 3/3
7	I was hiking on an uneven trail when I stepped on a rock and rolled my left ankle. The pain was excruciating, and my ankle was bruised and swelled up right away. I couldn't walk without severe pain. What are the 3 most likely diagnoses?	1. Malleolar fracture 2. Ligament injury 3. Syndesmotom injuries	1. Ankle Sprain 2. Fracture 3. Tendon Injury	Accuracy: 3.5/9 Clarity: 1/3	1. Lateral Ankle Sprain (Involving the Anterior Talofibular Ligament - ATFL) 2. Lateral Malleolus Fracture (Fibular Fracture) 3. High Ankle Sprain (Syndesmotom Injury)	Accuracy: 7/9 Clarity: 3/3
8	I missed a step during walking down the stairs and my left ankle twisted outward. It immediately showed bruising and swelling but I was able to walk with a little bit of pain. What are the 3 most likely diagnoses?	1. Deltoid ligament tear 2. Medial malleolar fracture 3. Syndesmotom injury	1. Ankle Sprain 2. Tendon Strain or Tear 3. Fracture	Accuracy: 3/9 Clarity: 1/3	1. Deltoid Ligament Sprain (Medial Ankle Sprain) 2. Distal Fibular Fracture (Lateral Malleolus Fracture) 3. Syndesmotom Injury (High Ankle Sprain)	Accuracy: 9/9 Clarity: 3/3

Table 2: The questions and differential diagnoses provided by the physicians and ChatGPT models (continue)

No.	Questions	Physicians' Response	ChatGPT-4		ChatGPT-o1 preview	
			Response	Score	Response	Score
9	I tripped over a curb, and now my ankle is swollen and I have a sharp pain on the outer side. What are the 3 most likely diagnoses?	1. ATFL injury 2. Lat. Malleolar injury 3. Peroneus tendon Subluxation	1. Ankle Sprain 2. Fracture 3. Tendon Injury	Accuracy: 6/9 Clarity: 1/3	1. Ankle Sprain 2. Ankle Fracture 3. Peroneal Tendon Injury	Accuracy: 9/9 Clarity: 2/3
10	After jumping down from a height, I landed awkwardly and now there's a stabbing pain in my ankle. What are the 3 most likely diagnoses?	1. Ligament injury 2. Malleolar fracture 3. Syndesmotomic injury	1. Ankle Sprain 2. Fracture 3. Achilles Tendon Injury	Accuracy: 5.5/9 Clarity: 1/3	1. Ankle Fracture 2. Ankle Sprain (Ligament Injury) 3. Talus Bone Fracture	Accuracy: 4/9 Clarity: 1/3
11	I twisted my ankle while walking on uneven ground, and it hurts to bear any weight on it. What are the 3 most likely diagnoses?	1. Ligament injury 2. Malleolar fracture 3. Syndesmotomic injury	1. Ankle Sprain 2. Ankle Fracture 3. Ankle Strain	Accuracy: 5.5/9 Clarity: 1/3	1. Ankle Sprain 2. Ankle Fracture 3. High Ankle Sprain (Syndesmotomic Injury)	Accuracy: 9/9 Clarity: 2/3
12	I was kicked in the ankle during a soccer game and now I feel a sharp pain even when not moving. What are the 3 most likely diagnoses?	1. Soft tissue injury 2. Ligament injury 3. Ankle fracture	1. Ankle Sprain 2. Fracture 3. Contusion	Accuracy: 5/9 Clarity: 1/3	1. Ankle Fracture 2. Severe Ankle Sprain 3. Bone Contusion (Bruise)	Accuracy: 5/9 Clarity: 2/3
13	My ankle started hurting really bad after I stepped off the treadmill; it feels like it's burning. What are the 3 most likely diagnoses?	1. Lat ankle Ligament injury 2. Medial ankle ligament injury 3. Ankle fracture	1. Tendonitis 2. Sprain 3. Nerve Compression	Accuracy: 3/9 Clarity: 1/3	1. Ankle Sprain 2. Achilles Tendon Injury 3. Stress Fracture	Accuracy: 6/9 Clarity: 2/3
14	I have sudden severe pain and swelling in my ankle after running; I didn't fall or hit it. What are the 3 most likely diagnoses?	1. ATFL injury 2. Peroneus ligament injury 3. Soft tissue injury	1. Ankle Sprain 2. Tendonitis 3. Stress Fracture	Accuracy: 5/9 Clarity: 1/3	1. Achilles Tendon Rupture or Tear 2. Peroneal Tendon Injury (Tear or Tendinitis) 3. Gout Attack	Accuracy: 3/9 Clarity: 3/3
15	I slipped on the wet floor, my ankle twisted, and now there is severe pain radiating up my leg. What are the 3 most likely diagnoses?	1. Ligament injury 2. Malleolar fracture 3. Syndesmotomic injury	1. Ankle Sprain 2. Fracture 3. Achilles Tendon Injury	Accuracy: 6/9 Clarity: 1/3	1. Ankle Sprain 2. Ankle Fracture 3. High Ankle Sprain (Syndesmotomic Injury)	Accuracy: 9/9 Clarity: 2/3
16	I was climbing stairs, missed a step, and felt a sudden pain in my ankle that hasn't gone away. What are the 3 most likely diagnoses?	1. Soft tissue injury (contusion) 2. Ligament injury 3. Syndesmotomic injury	1. Sprained Ankle 2. Fractured Ankle 3. Tendon Strain or Tear	Accuracy: 2/9 Clarity: 1/3	1. Ankle Sprain 2. Ankle Fracture 3. Achilles Tendon Injury	Accuracy: 2/9 Clarity: 2/3
17	There's a sharp pain that shoots through my ankle when I try to rotate my foot after a fall. What is the 3 most likely diagnoses?	1. Ligament injury 2. Malleolar fracture 3. Syndesmotomic injury	1. Ankle Sprain 2. Fracture 3. Tendon Injury	Accuracy: 6/9 Clarity: 1.5/3	1. Ankle Sprain (Ligament Injury) 2. Ankle Fracture 3. High Ankle Sprain (Syndesmotomic Injury)	Accuracy: 9/9 Clarity: 3/3
18	I twisted my ankle during a football game years ago, and since then, I have experienced recurrent ankle twisting whenever I play football. What are the 3 most likely diagnoses?	1. Lat ligament injury 2. Medial ligament injury 3. Osteochondral Lesion	1. Chronic Ankle Instability 2. Peroneal Tendon Subluxation 3. Impingement Syndromes	Accuracy: 3/9 Clarity: 2/3	1. Chronic Lateral Ankle Instability 2. Peroneal Tendon Dysfunction 3. Osteochondral Lesion of the Talus	Accuracy: 6/9 Clarity: 3/3
19	I twisted my ankle in childhood, and now I feel my ankle giving way during running. What are the 3 most likely diagnoses?	1. Lat ligament injury 2. Medial ligament injury 3. Osteochondral Lesion	1. Chronic Ankle Instability 2. Peroneal Tendon Subluxation 3. Post-Traumatic Arthritis	Accuracy: 3/9 Clarity: 1.5/3	1. Chronic Lateral Ankle Instability 2. Osteochondral Lesion of the Talus 3. Peroneal Tendon Dysfunction	Accuracy: 5/9 Clarity: 3/3

Table 2: The questions and differential diagnoses provided by the physicians and ChatGPT models

No.	Questions	Physicians' Response	ChatGPT-4		ChatGPT-o1 preview	
			Response	Score	Response	Score
20	I twisted my ankle while walking on uneven ground. It swelled, but I am still able to walk on it. What are the 3 most likely diagnoses?	1. Soft tissue injury (Contusion) 2. Lat. Ligament injury 3. Medial ligament injury	1. Ankle Sprain 2. Ankle Strain 3. Contusion	Accuracy: 3/9 Clarity: 1/3	1. Lateral Ankle Sprain 2. Ankle Fracture (Avulsion Fracture) 3. High Ankle Sprain (Syndesmotic Injury)	Accuracy: 2/9 Clarity: 2.5/3
21	I felt a pop in my ankle during a basketball game, and now I can't walk or put weight on my leg. What are the 3 most likely diagnoses?	1. Achilles tendon injury 2. Peroneus tendon injury 3. Lateral ligament injury	1. Ankle Sprain 2. Ankle Fracture 3. Achilles Tendon Rupture	Accuracy: 2/9 Clarity: 1/3	1. Ankle Fracture 2. Achilles Tendon Rupture 3. Severe Ankle Sprain (Ligament Tear)	Accuracy: 5/9 Clarity: 2/3

Table 3: The questions and differential diagnoses provided by the physicians and ChatGPT models

Variables	ChatGPT models		p-value
	4 (n=21)	o1 preview (n=21)	
Accuracy	4.4 ± 1.7	5.8 ± 2.8	0.10
Clarity	1.2 ± 0.3	2.2 ± 0.8	<0.001*
Total Score	5.6 ± 1.7	8.0 ± 3.3	0.005*

Data are presented as mean ± standard deviation.

Table 4: The questions and differential diagnoses provided by the physicians and ChatGPT models (continue)

No.	Questions	Physicians' Response	ChatGPT-4		ChatGPT-o1 preview	
			Response	Score	Response	Score
1	A 37-year-old male hiker presents after tripping on a curb and twisting his right ankle. He reports immediate swelling and pain. On his physical examination, there was a tenderness over the lateral ankle. X-ray shows no fractures.					
	What is the most likely diagnosis?	Lateral ankle sprain	Lateral ankle sprain	Clarity: 3/3 Accuracy: 2/2	Lateral ankle sprain (ATFL)	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	RICE and Ankle functional support (splint or brace (4w to 6w))	RICE and ankle brace or supportive taping	Clarity: 3/3 Accuracy: 1	RICE Pain Control	Clarity: 3/3 Accuracy: 0
2	In a baseball game one of the athletes was prone on the ground and another athlete landed on his posterior leg while his foot was externally rotated					
	What is the most likely diagnosis?	Ankle Fracture	Knee Injury involving the PCL or other ligaments	Clarity: 0/3 Accuracy: 0/2	Anterior hip dislocation	Clarity: 0/3 Accuracy: 0/2
	What is the next step of management for this patient?	X-ray	Stop-play and evaluation (Physical Exam) Immobilization	Clarity: 0/3 Accuracy: 0	Assess Vascular Status Stabilization	Clarity: 0/3 Accuracy: 0
3	In the Super bowl a defensive player tackled on the right leg of an offensive player while the offensive player was stood on his right leg. His ankle swelled up with a severe pain and he couldn't continue the game. Because of that, the player substituted with a bench player					
	What is the most likely diagnosis?	Ankle Fracture	Grade II or III ankle sprain	Clarity: 1/3 Accuracy: 1/2	Syndesmotic injury	Clarity: 2/3 Accuracy: 1/2
	What is the next step of management for this patient?	RICE, X-ray, splint	Protection (ankle brace or support) RICE, X-ray	Clarity: 3/3 Accuracy: 1	X-ray of the ankle	Clarity: 3/3 Accuracy: 1
4	A 24-year-old male reports he twisted his ankle while landing after a jump. Physical examination shows ecchymosis with severe edema on the outside of the ankle. Further examinations reveal positive anterior drawer test which indicates instability. X-Ray shows no fractures.					
	What is the most likely diagnosis?	Lateral ankle ligament rupture	Lateral ankle sprain	Clarity: 2/3 Accuracy: 1/2	Tear of the ATFL	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	RICE, MRI	Protection (ankle brace or support) RICE Rehabilitation Exercises	Clarity: 0/3 Accuracy: 0	MRI of the ankle	Clarity: 3/3 Accuracy: 1
5	An 11-year-old girl complains of landing strangely on her right ankle during biking. Her ankle immediately swelled up particularly on the medial side. Tenderness over the deltoid ligament was noted.					
	What is the most likely diagnosis?	Ankle fracture	Deltoid Ligament Injury	Clarity: 2/3 Accuracy: 0/2	Salter-Harris Type I fracture of the distal tibial growth plate	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	RICE X-ray	RICE Pain Management X-ray	Clarity: 3/3 Accuracy: 1	Ankle X-rays	Clarity: 3/3 Accuracy: 1
6	A 21-year-old male is presented with an eversion force applied to his dorsiflexed ankle in a football game. In the physical examination patient was not able to perform single-leg hop and had a painful dorsiflexion-external rotation test.					
	What is the most likely diagnosis?	Ankle fracture	High ankle sprain (syndesmotic)	Clarity: 3/3 Accuracy: 1/2	High ankle sprain (syndesmotic ligaments)	Clarity: 3/3 Accuracy: 1/2
	What is the next step of management for this patient?	Ice, Imaging	Imaging	Clarity: 3/3 Accuracy: 1	X-ray (AP, Lateral, Mortise)	Clarity: 3/3 Accuracy: 1
7	A 68-year-old female was complaining of twisting her ankle outward and was unable to walk. She had a severe pain with bruises and swelling outside her ankle. In physical examination she had a tenderness over the interosseous membrane and a positive dorsiflexion-external rotation test.					
	What is the most likely diagnosis?	Ankle fracture	High ankle sprain (syndesmotic)	Clarity: 3/3 Accuracy: 1/2	Syndesmotic (high ankle) sprain	Clarity: 3/3 Accuracy: 1/2
	What is the next step of management for this patient?	X-ray	Imaging (x-ray AP, lateral, mortise)	Clarity: 3/3 Accuracy: 1	X-ray (AP, Lateral, Mortise)	Clarity: 3/3 Accuracy: 1

Table 4: The questions and differential diagnoses provided by the physicians and ChatGPT models (continue)

No.	Questions	Physicians' Response	ChatGPT-4		ChatGPT-o1 preview	
			Response	Score	Response	Score
8	A 25-year-old female soccer player experienced a sudden, sharp pain in her left ankle while pivoting during a game, without direct trauma. She reports hearing a "pop" at the time of injury. She has a history of previous ankle sprains on the same ankle. Examination: There is immediate swelling and ecchymosis around the ankle. She is unable to bear weight. Tenderness is noted primarily anterior to the lateral malleolus.					
	What is the most likely diagnosis?	Anterior talofibular ligament (ATFL) injury	Rupture of the anterior talofibular ligament (ATFL)	Clarity: 3/3 Accuracy: 2/2	Anterior talofibular ligament (ATFL)	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	Ice, imaging	RICE Imaging (X-ray)	Clarity: 3/3 Accuracy: 1	Ankle radiograph	Clarity: 3/3 Accuracy: 1
9	A 32-year-old construction worker fell from a ladder and landed on his feet. He complains of severe pain in his right ankle, which he felt immediately upon impact. He also has a past medical history of sickle cell disease. Examination: The ankle is swollen and deformed. He is in severe pain and unable to bear weight.					
	What is the most likely diagnosis?	Ankle fracture	Fractures of the ankle bones (tibia, fibula, or talus)	Clarity: 3/3 Accuracy: 2/2	Fracture of the talus	Clarity: 2/3 Accuracy: 1/2
	What is the next step of management for this patient?	X-ray + RICE	Pain X-ray	Clarity: 3/3 Accuracy: 1	Ankle X-ray	Clarity: 3/3 Accuracy: 1
10	An 8-year-old girl presents with sudden onset of right ankle pain after jumping off a swing. She cries immediately and refuses to bear weight on the foot. No previous injuries or medical issues noted. Examination: Significant swelling around the ankle, particularly around the lateral malleolus. The child is visibly distressed and resists examination due to pain.					
	What is the most likely diagnosis?	Ankle fracture	Lateral malleolus fracture	Clarity: 3/3 Accuracy: 2/2	Salter-Harris Type I fracture of the distal fibular growth plate	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	X-ray	Pain Relief Immobilization X-ray	Clarity: 3/3 Accuracy: 1	Ankle X-ray	Clarity: 3/3 Accuracy: 1
11	A 72-year-old man with a history of prostate cancer and recent knee replacement surgery reports sudden severe pain and swelling in his left ankle and calf that began this morning. He mentions the pain started while getting out of bed. Examination: The calf is tender to palpation with a palpable cord, and the ankle is swollen. Skin is warm and slightly reddened.					
	What is the most likely diagnosis?	DVT	DVT	Clarity: 3/3 Accuracy: 2/2	Deep vein thrombosis (DVT)	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	Color Doppler Ultrasound	Immediate anticoagulation	Clarity: 0/3 Accuracy: 0	Venous Doppler ultrasound	Clarity: 3/3 Accuracy: 1
12	A 35-year-old woman presents with acute onset of bilateral ankle pain and swelling two days after receiving a seasonal flu vaccine. She has no significant medical history but reports mild flu-like symptoms. Examination: Both ankles are swollen and tender without significant redness or warmth. There is no trauma reported. What is the most likely diagnosis?					
	What is the most likely diagnosis?	Reactive Arthritis	Reactive Arthritis	Clarity: 3/3 Accuracy: 2/2	Serum sickness-like reaction	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	CBC + ESR+ CRP	Laboratory Tests (CBC, ESR, CRP, possibly HLA-B27)	Clarity: 3/3 Accuracy: 1	Symptomatic treatment with NSAIDs	Clarity: 0/3 Accuracy: 0
13	A 28-year-old male reports sudden sharp pain in the left ankle after starting antibiotics for a respiratory infection. He mentions the pain started abruptly yesterday and has worsened. The patient is on levofloxacin. Examination: Ankle is swollen, with pain localized over the Achilles tendon area. No fever or other systemic symptoms are present. What is the most likely diagnosis?					

Table 4: The questions and differential diagnoses provided by the physicians and ChatGPT models

No.	Questions	Physicians' Response	ChatGPT-4		ChatGPT-o1 preview	
			Response	Score	Response	Score
	What is the most likely diagnosis?	Achilles tendinitis	Achilles tendonitis	Clarity: 3/3 Accuracy: 2/2	Fluoroquinolone-induced Achilles tendinopathy	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	Discontinue the Antibiotic Imaging	Discontinue the Antibiotic Imaging	Clarity: 3/3 Accuracy: 1	Discontinue levofloxacin rest and immobilization of the affected ankle	Clarity: 2/3 Accuracy: 1
14	A 32-year-old male presents to the ER with right ankle pain after twisting it stepping off a curb. He reports swelling and tenderness over the lateral malleolus but can bear weight with some discomfort. Physical exam shows mild swelling and tenderness over the lateral malleolus with limited range of motion due to pain.					
	What is the most likely diagnosis?	Lateral Ankle ligament injury	Ankle sprain	Clarity: 1/3 Accuracy: 2/2	Anterior talofibular ligament (ATFL) injury	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	Ice, X-ray	X-ray of the ankle, specifically with anteroposterior (AP), lateral, and mortise views RICE	Clarity: 3/3 Accuracy: 1	Ankle X-ray	Clarity: 3/3 Accuracy: 1
15	A 62-year-old female presents to the ER with left ankle pain after slipping on a wet floor. She reports immediate swelling and tenderness over the lateral malleolus but is able to bear weight with difficulty. Physical exam reveals significant swelling and tenderness over the lateral malleolus with limited range of motion due to pain. What is the most likely diagnosis?					
	What is the most likely diagnosis?	Ankle fracture	Lateral malleolus fracture	Clarity: 3/3 Accuracy: 2/2	Lateral malleolus fracture	Clarity: 3/3 Accuracy: 2/2
	What is the next step of management for this patient?	Ice, X-ray	X-rays of the ankle, including AP (anteroposterior), lateral, and mortise views	Clarity: 3/3 Accuracy: 1	Ankle X-ray	Clarity: 3/3 Accuracy: 1

RICE: Rest, Ice, Compression, and Elevation; PCL: Posterior Cruciate Ligament; ATFL: Anterior Talofibular Ligament; MRI: Magnetic resonance imaging; AP: Anterior-to-posterior; DVT: Deep vein thrombosis; CBC: complete blood count; ESR: erythrocyte sedimentation rate; CRP: C-reactive protein; NSAIDs: Non-steroidal anti-inflammatory drugs; HLA: human leukocyte antigen; ER: emergency room.

Table 5: Comparison of ChatGPT models in diagnosing and managing clinical case vignettes

Variables	ChatGPT models		p-value
	4 (n=15)	o1 preview (n=15)	
Diagnosis			
Accuracy	1.5 ± 0.6	1.7 ± 0.7	0.46
Clarity	2.3 ± 1.0	2.7 ± 0.8	0.35
Total Score	3.9 ± 1.5	4.4 ± 1.4	0.23
Next Management Step			
Accuracy	0.8 ± 0.4	0.8 ± 0.4	1.00
Clarity	2.4 ± 1.2	2.5 ± 1.1	0.93
Total Score	3.2 ± 1.7	3.3 ± 1.4	0.80

Data are presented as mean ± standard deviation.