

AI vs AI: clinical reasoning performance of language models in orthopedic rehabilitation

 Ertuğrul Safran¹,  Yusuf Yaşasın²

¹Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Bezmialem Vakıf University, İstanbul, Türkiye

²Department of Physiotherapy and Rehabilitation, Institute of Graduate Studies, Bahçeşehir University, İstanbul, Türkiye

Cite this article as: Safran E, Yaşasın Y. AI vs AI: clinical reasoning performance of language models in orthopedic rehabilitation. *J Health Sci Med.* 2025;8(5):825-831.

Received: 15.07.2025

Accepted: 06.08.2025

Published: 16.09.2025

ABSTRACT

Aims: This study aimed to compare the clinical reasoning and treatment planning performance of three advanced large language models (LLMs)-ChatGPT-4o, Gemini 2.5 Pro, and DeepSeek-V3-in orthopedic rehabilitation. Their responses to standardized clinical scenarios were evaluated to determine alignment with evidence-based physiotherapy practices, focusing on relevance, accuracy, completeness, applicability, and safety awareness.

Methods: Three fictional but clinically realistic scenarios involving rotator cuff tendinopathy, lumbar disc herniation with radiculopathy, and anterior cruciate ligament (ACL) reconstruction were developed by an experienced physiotherapist. These scenarios were independently queried on the same day by three AI models using identical prompts. A blinded expert physiotherapist evaluated each model's detailed responses using a 5-point Likert Scale across five domains: clinical accuracy, relevance, completeness, applicability, and safety awareness. Mean scores and descriptive statistics were calculated.

Results: DeepSeek-V3 was consistently rated highest (5/5) across all domains and scenarios, demonstrating comprehensive and clinically rigorous plans. ChatGPT-4o showed strong performance overall, with total scores ranging from 19 to 20 out of 25, though it exhibited lower completeness scores due to less specific milestones. Gemini 2.5 Pro scored lower overall (average total score 18/25), with particular weaknesses in applicability and clinical relevance in complex cases such as lumbar disc herniation. All models provided evidence-based treatment approaches emphasizing pain management, postural correction, gradual strengthening, and return-to-activity progression. Differences arose in emphasis on lifestyle modification, patient education depth, and integration of psychosocial factors, with Gemini uniquely addressing psychological readiness in ACL rehabilitation.

Conclusion: AI-generated rehabilitation plans show substantial concordance with current physiotherapy guidelines but vary in detail and clinical practicality. DeepSeek-V3 outperformed the other models in consistency and safety considerations, while ChatGPT-4o balanced clinical accuracy with moderate completeness. Gemini 2.5 Pro's inclusion of biopsychosocial components offers valuable insights but may require further refinement for clinical applicability. These findings highlight the potential and current limitations of AI tools in orthopedic rehabilitation, suggesting careful model selection based on clinical context and user needs.

Keywords: Artificial Intelligence, clinical reasoning, musculoskeletal diseases, rehabilitation, physical therapy modalities

INTRODUCTION

In recent years, the integration of Artificial Intelligence (AI) into healthcare has accelerated, driven largely by the development of large language models (LLMs) such as OpenAI's ChatGPT, Google's Gemini, and DeepSeek.^{1,2} These models are capable of generating human-like responses to diverse queries, including those related to clinical reasoning, patient education, and therapeutic guidance.³⁻⁶ As their accessibility increases, physiotherapy students, clinicians, and even patients are beginning to consult AI tools to gain rapid information on musculoskeletal conditions and rehabilitation strategies.^{7,8} However, concerns regarding the reliability and clinical relevance of AI-generated responses particularly

in complex decision-making contexts persist and warrant systematic investigation.

Orthopedic rehabilitation represents a central and foundational pillar of physiotherapy practice. It involves the assessment and management of musculoskeletal injuries and disorders such as low back pain, rotator cuff tears, ligament injuries, and post-operative recovery following orthopedic surgeries.⁹ Effective orthopedic rehabilitation demands accurate clinical reasoning, individualized treatment planning, and ongoing adaptation based on the patient's response and risk profile.^{10,11} For these reasons, it serves as an ideal context

Corresponding Author: Ertuğrul Safran, esafran@bezmialem.edu.tr



This work is licensed under a Creative Commons Attribution 4.0 International License.

in which to evaluate the potential and limitations of AI-generated clinical content.

Despite the growing use of AI in health education and preliminary clinical decision support, comparative analyses of different language models are scarce. Existing studies have largely focused on general medical information delivery¹² or single-model evaluations¹³ while no published research has systematically compared multiple AI models in orthopedic physiotherapy. Given their differing architectures and data sources, it is plausible that these models may vary in the accuracy, depth, and clinical applicability of their responses.^{14,15}

This pilot study was conducted to compare the clinical reasoning performance of ChatGPT-4o, Gemini 2.5 Pro, and DeepSeek-V3 in orthopedic rehabilitation. Their responses to standardized clinical scenarios were analyzed for relevance, accuracy, completeness, applicability, and safety awareness. By identifying strengths and limitations of each model, the study aims to inform the potential role of AI in physiotherapy education and practice while clarifying its current boundaries.

METHODS

Study Design

This study employed a structured comparative design to evaluate the clinical reasoning performance of three large language models—OpenAI's ChatGPT, Google's Gemini, and DeepSeek-V3—in the context of orthopedic rehabilitation. The models' responses to standardized fictional clinical scenarios were independently rated by expert physiotherapists using a multi-domain scoring rubric (assessing accuracy, clinical reasoning, clarity, and safety) that was developed for this study. This study did not involve human or animal participants. Therefore, ethical approval was not required according to national guidelines. All procedures were carried out in accordance with the ethical rules and the principles.

Clinical Scenario Development and AI Response Collection

Three structured clinical scenarios were developed to reflect common musculoskeletal conditions encountered in orthopedic physiotherapy practice, deliberately covering one spinal condition (lumbar disc herniation), one upper extremity condition (rotator cuff tendinopathy), and one lower extremity post-surgical case (anterior cruciate ligament reconstruction). These scenarios were entirely fictional but constructed by an experienced physiotherapist to ensure clinical realism without involving any identifiable patient data.

Importantly, the process of querying AI models was conducted by a different researcher, independent of the scenario author. This separation ensured neutrality and minimized the risk of bias in prompt formulation or interpretation. Each case was submitted to the AI systems using the following standardized prompt:

“Based on the case details below, outline an appropriate physiotherapy treatment plan, including clinical priorities, precautions, and justification of therapeutic choices.”

ChatGPT (GPT-4, accessed via OpenAI), Gemini (accessed via Google AI), and DeepSeek-V3 (accessed via DeepSeek AI platform) were queried with each scenario under identical conditions and on the same day to ensure consistency. No prior context or model-specific optimization was applied.

The three case scenarios, designed to reflect real-life clinical situations and reviewed by the artificial intelligence, are presented below.

Scenario 1: A 52-year-old right-handed female presents with a 6-month history of progressive right shoulder pain. The pain began insidiously without any specific trauma and is localized over the anterolateral aspect of the shoulder. She describes the pain as dull and aching, worsened by overhead activities, reaching behind her back (such as fastening her bra), and sleeping on the affected side. Morning stiffness is minimal, and there is no reported numbness, tingling, or neck pain. She works as an office administrator but is also an amateur tennis player, playing twice a week. Over the past few months, she has reduced her tennis activities due to increasing discomfort. NSAIDs provide partial relief. She reports occasional night pain, particularly after increased shoulder use during the day. Past medical history includes mild hypertension, well controlled with medication. No history of diabetes or rheumatoid disease. No prior shoulder injuries or surgeries. On physical examination, her posture shows mild protraction of the right shoulder and forward head posture. Active range of motion reveals painful arc between 70-120° of abduction and flexion. Passive range of motion is mostly preserved, with mild discomfort at end ranges. Strength testing shows mild weakness (4+/5) in right shoulder abduction and external rotation. Special tests: Positive Neer impingement test, Positive Hawkins-Kennedy test, Jobe (empty can) test positive for pain, minimal weakness, Negative external rotation lag sign, Negative drop arm test. Palpation reveals tenderness over the supraspinatus tendon and subacromial region. No signs of gross instability or labral pathology. Ultrasound report shows hypoechoic areas and mild thickening of the supraspinatus tendon without full-thickness tear; mild subacromial-subdeltoid bursitis noted. Functional impact includes difficulty with overhead reaching, carrying moderate loads, and participating in tennis. Patient expresses concern about long-term shoulder function and returning to recreational sports.

Scenario 2: A 38-year-old right-handed male presents with a 3-month history of low back pain radiating to the right leg. The pain is described as a dull ache in the lumbar region, with burning/shooting radiation to the right gluteal area and posterior thigh. Prolonged sitting and forward bending exacerbate the symptoms, while walking provides partial relief. There is no morning stiffness, and pain decreases with rest. The patient reports a pain intensity of 7 out of 10 on the Visual Analog Scale (VAS 7/10). He works as an office employee, spending approximately 8-10 hours per day at a computer. There is no prior history of lumbar disc herniation or low back pain. His medical history is unremarkable for systemic diseases. He leads a sedentary lifestyle and does not engage in regular exercise. He has been smoking about 10 cigarettes per day for the past 10 years. He occasionally uses

nonsteroidal anti-inflammatory drugs (NSAIDs) for symptom relief. Physical examination reveals mild lumbar hypolordosis and mild atrophy of the right gluteal muscles. Lumbar flexion is restricted and painful; right lateral flexion is painful, left lateral flexion is limited; extension is slightly painful. There is tenderness over the right L4-L5 paravertebral area and right gluteal region with noted myofascial tightness. On orthopedic and neurological testing, the right straight leg raise (SLR) is painful at 45°, and the right Lasègue test is positive. Patellar and Achilles reflexes are normal. Muscle strength testing shows 4/5 right ankle dorsiflexion and 5/5 plantarflexion. There is mild hypoesthesia in the right L5 dermatome. Gait analysis reveals an antalgic pattern with right-side unloading. Lumbar magnetic resonance imaging (MRI) shows a right posterolateral disc herniation at the L4-L5 level, consistent with nerve root compression. Clinically, the patient is diagnosed with L4-L5 disc herniation associated with right L5 radiculopathy. There is pain, mild motor weakness, and sensory alteration. Chronicity risk is high due to the 3-month duration, sedentary lifestyle, and smoking habit.

Scenario 3: A 25-year-old, male, six weeks post-right knee anterior cruciate ligament (ACL) reconstruction, presents for physiotherapy. The injury occurred during an amateur football game when his right knee twisted, resulting in immediate sharp pain, a giving-way sensation, and swelling. Surgery was performed using a hamstring graft, and postoperative instructions included two weeks of bracing and crutch use. He now presents at week 6 post-surgery for rehabilitation. The patient reports occasional pain and a sense of swelling in the right knee, with feelings of instability when going up and down stairs. He experiences fatigue during prolonged walking. He has partially resumed daily activities but is anxious about returning to sports. His rehabilitation goals include achieving full knee range of motion, regaining muscle strength, and ensuring a safe return to sports.

Past medical history: No systemic diseases, non-smoker, normal BMI, previously active lifestyle.

Physical examination findings: Mild right knee swelling, well-healed surgical scar, knee flexion approximately 0-110°, near full extension, quadriceps and hamstring muscle atrophy, notable weakness of the right vastus medialis obliquus (VMO), slightly antalgic gait, full weight-bearing, stability tests; negative Lachman and pivot shift (postoperative), negative valgus/varus stress tests, single-leg stance shows

poor balance, minimal pain on palpation, especially around the patellofemoral region.

Functional status: Experiences fatigue during long walks, difficulty climbing stairs, avoids squatting and jumping activities. He is highly motivated with the goal of returning to sports.

Clinical summary: Right ACL reconstruction, postoperative week 6, early-to-mid rehabilitation phase.

Evaluation Criteria and Expert Scoring

To minimize bias and ensure blinding during the evaluation process, the study employed a two-phase approach. One licensed physiotherapist with seven years of clinical experience was responsible for designing the clinical scenarios, inputting them into the AI systems, and collecting the generated responses. A second physiotherapist, with 12 years of experience in the same field and blinded to the identity of the AI model that produced each response, independently scored the answers using a predefined evaluation rubric. This procedure was implemented to reduce potential evaluator bias and maintain objectivity in the scoring process.

A 5-point Likert scale (1 = Very poor, 5 = Excellent) was used across five domains, with explicit scoring anchors provided for each rating level (Table 1). This rubric was adapted from prior frameworks used in studies evaluating AI-generated health content^{2,16} and refined for orthopedic rehabilitation by physiotherapy experts.^{2,16}

Clinical accuracy: Consistency with evidence-based physiotherapy approaches.

Applicability: Practical relevance and feasibility of the proposed treatment.

Safety awareness: Consideration of contraindications, red flags, and patient-specific risks.

Relevance: The extent to which the model’s response directly addresses the clinical scenario and aligns with the presented case context.

Completeness: The degree to which the response provides all essential components of assessment, reasoning, and management without omitting key information.

Domain-specific and total average scores were calculated separately for ChatGPT, Gemini, and Deepseek (Table 1).

Table 1. Multi-domain evaluation rubric for AI-generated responses					
Domain	1-very poor	2-poor	3-fair	4-good	5-excellent
Clinical accuracy	Provides incorrect or misleading information; contradicts evidence-based practice.	Mostly inaccurate; several critical errors present.	Mixed accuracy; includes some correct elements but notable gaps.	Largely accurate with only minor errors or omissions.	Fully accurate and aligned with current evidence-based physiotherapy guidelines.
Applicability	No practical application; advice is unusable in clinical context.	Limited relevance; many suggestions unrealistic or impractical.	Moderately relevant; some feasible elements but lacks clarity on implementation.	Mostly practical and relevant with minor feasibility issues.	Fully relevant and readily applicable to clinical physiotherapy practice.
Safety awareness	Ignores major red flags, contraindications, or patient risks.	Acknowledges some risks but omits important safety considerations.	Basic safety elements present but lacks depth or misses subtler risks.	Covers most safety considerations; only minor omissions.	Thoroughly integrates safety, contraindications, and patient-specific precautions.
Relevance	Does not address the case; largely off-topic or generic.	Partially relevant; addresses only some aspects of the case.	Moderately relevant; case-related but missing key contextual links.	Mostly relevant; minor digressions but generally on point.	Fully relevant, directly addressing all aspects of the presented scenario.
Completeness	Major components (assessment, reasoning, management) missing.	Many important aspects missing; very superficial response.	Partially complete; includes key components but lacks depth.	Almost complete with only minor omissions.	Fully complete, covering assessment, reasoning, and management comprehensively.
AI: Artificial Intelligence					

Statistical Analysis

Mean scores were calculated for each evaluation domain and case based on the single rater's assessments. Descriptive statistics (mean±standard deviation) are reported. All analyses were performed using IBM SPSS Statistics version 25.

RESULTS

The responses of the three different AI models to the three clinical scenarios are summarized below. Therapist ratings were calculated based on the detailed, full-length responses.

Quantitative assessment revealed significant performance differences across AI models ([Table 1](#)). DeepSeek-V3 achieved perfect scores (5/5) in all parameters across all cases, totaling 25/25 for each clinical scenario. ChatGPT-4o demonstrated consistent performance with total scores ranging 19-20/25, though it showed relative weakness in Completeness (average: 3.0/5). Gemini 2.5 Pro scored lowest in Applicability (average: 3.0/5) and total scores (18/25 in all cases), with notable vulnerability in Lumbar Disc Herniation (Relevance: 3/5).

Scenario 1: Shoulder Pathology (Rotator Cuff Tendinopathy & Postural Correction)

ChatGPT-4o provided a comprehensive, phased physiotherapy treatment plan focusing on reducing pain and inflammation associated with supraspinatus tendinopathy and subacromial bursitis. The plan emphasized restoring pain-free shoulder range of motion, correcting rotator cuff and scapular muscle imbalances, and addressing postural abnormalities such as forward head posture and shoulder protraction. Strengthening progressed from pain-free isometrics to functional and sports-specific drills with a controlled return to tennis.

DeepSeek-V3 outlined a pain-guided approach with gradual progression of exercise intensity based on patient tolerance. The plan included manual therapy interventions, neuromuscular control exercises, ergonomic adjustments, and patient education emphasizing self-management. Early management targeted inflammation and pain with careful monitoring to avoid aggravation.

Gemini 2.5 Pro proposed a detailed multi-phase rehabilitation program with a strong focus on symptom control, postural correction, progressive strengthening, and tennis-specific retraining. The approach incorporated manual therapy, neuromuscular education, and technique analysis to optimize biomechanical function and prevent recurrence.

Scenario 2: Lumbar Disc Herniation with Radiculopathy

ChatGPT-4o recommended a conservative treatment plan centered on alleviating radicular pain and decompressing the L5 nerve root. Key elements included restoring lumbar mobility, correcting posture, strengthening core and gluteal muscles, and modifying lifestyle factors such as reducing sedentary behavior and smoking cessation. Neural mobilization techniques and gradual progression of exercises were emphasized.

DeepSeek-V3 presented a comprehensive rehabilitation strategy addressing symptom centralization, neurological function restoration, and lifestyle modification. The plan

incorporated ergonomics education, postural retraining, core stabilization, and gradual functional activity retraining, with detailed progression criteria and patient education.

Gemini 2.5 Pro outlined a phase-based approach involving acute symptom management, motor control enhancement, and advanced functional strengthening. Lifestyle interventions, including smoking cessation and home exercise compliance, were integral components to ensure long-term success and recurrence prevention.

Scenario 3: ACL Reconstruction Rehabilitation

ChatGPT-4o emphasized restoring full knee extension and flexion, managing swelling and mild pain, and progressively strengthening the quadriceps and hamstrings. Proprioceptive and neuromuscular control exercises were integrated to improve gait and prevent compensatory patterns. Return to sport was carefully phased with milestones and precautions to protect the graft.

DeepSeek-V3 focused on symmetrical range of motion recovery, dynamic stability, and graft protection through progressive closed-chain strengthening and neuromuscular retraining. Balance and functional mobility exercises were advanced gradually, with consideration for sport-specific demands.

Gemini 2.5 Pro-proposed a criteria-based rehabilitation plan incorporating pain and swelling control, progressive strength training, plyometric and agility drills, and psychological readiness assessment. The program included detailed milestones for return to football, emphasizing safe progression through non-contact to full-contact activities ([Table 2](#)).

DISCUSSION

This study compared the clinical reasoning performance of three large language models (LLMs)-ChatGPT-4o, DeepSeek-V3, and Gemini 2.5 Pro-using three scenarios that reflected key areas of orthopedic rehabilitation; one spinal, one upper extremity, and one lower extremity post-surgical case. Across all scenarios, the models produced rehabilitation plans largely consistent with physiotherapy guidelines, but important differences emerged in how each model approached clinical reasoning, integrated psychosocial elements, and managed safety considerations. These findings highlight both the potential and the current boundaries of AI in supporting rehabilitation practice. Although DeepSeek-V3 achieved the highest possible scores across domains, this reflects its alignment with guideline-based benchmarks rather than flawless real-world applicability. In clinical practice, factors such as patient variability, contextual judgment, and therapist-patient interaction remain essential and cannot be fully replicated by an AI-generated plan. Recent advances in LLM have sparked growing interest in their application across various healthcare domains, including physiotherapy.¹⁷⁻¹⁹ As these models, such as ChatGPT, become increasingly integrated into clinical and educational contexts, questions arise regarding their reliability, consistency, and alignment with professional guidelines. In the field of musculoskeletal rehabilitation, where clinical reasoning and individualized decision-making are critical, evaluating the performance

Table 2. AI model performance scores across clinical case

Case	AI model	Relevance	Accuracy	Completeness	Applicability	Safety awareness	Total score
1 (Rotator cuff tendinopathy)	ChatGPT-4o	5	4	3	4	4	20
	Deepseek V3	5	5	5	5	5	25
	Gemini 2.5 Pro	4	3	4	3	4	18
2 (Lumbar disc herniation)	ChatGPT-4o	4	4	3	4	4	19
	Deepseek V3	5	5	5	5	5	25
	Gemini 2.5 Pro	3	4	4	3	4	18
3 (ACL reconstruction)	ChatGPT-4o	4	4	3	4	4	19
	Deepseek V3	5	5	5	5	5	25
	Gemini 2.5 Pro	4	4	3	3	4	18
Overall average	ChatGPT-4o	4.3	4	3	4	4	19.3
	Deepseek V3	5	5	5	5	5	25
	Gemini 2.5 Pro	3.7	3.7	3.7	3	4	18

AI: Artificial Intelligence, ACL: Anterior cruciate ligament

of AI-based tools has become both timely and essential. Hao et al.²⁰ investigated the alignment of ChatGPT-4's responses with clinical practice guidelines (CPGs) across 30 physical therapy-related questions and found that 80% of its outputs were consistent with established recommendations. The model performed best in upper extremity conditions (100% accuracy) and showed lower performance in spinal cases (60%). Inter-rater reliability was very good ($\kappa=0.847$), supporting the model's consistency in clinical interpretation. Bilika et al.²¹ evaluated the clinical reasoning capacity of ChatGPT-3.5 in both simple and complex physiotherapy scenarios. While the model demonstrated a reasonable ability to generate evidence-based reasoning in straightforward cases, key assessment components-such as re-assessment and subjective evaluation-were missing in up to 40% of responses. The authors emphasized the variability in output quality and the strong influence of prompt phrasing on model performance.

As interest in the clinical applications of LLMs grows, recent research has shifted from examining single models like ChatGPT to conducting comparative analyses involving multiple platforms-including ChatGPT, Google's Gemini, and DeepSeek-to evaluate their reasoning accuracy, consistency, and alignment with evidence-based guidelines.^{14,15,22-25} These comparisons offer a broader perspective on the capabilities and limitations of different language models in healthcare contexts.^{12,26} Marcaccini et al.¹⁵ evaluated the diagnostic and management capabilities of ChatGPT-4o, DeepSeek-V3, and Gemini 1.5 in 58 real-world hand fracture cases. While ChatGPT-4o achieved high diagnostic accuracy (98.28%) and recall, DeepSeek-V3 showed moderate performance, and Gemini 1.5 failed to deliver clinically reliable outputs. These findings underscore significant variability in model reliability across complex orthopedic scenarios. Wu et al.²⁷ investigated the alignment of ChatGPT-4.0 and DeepSeek-R1 with Chinese clinical practice guidelines for knee osteoarthritis (KOA). Among 17 therapeutic recommendations, DeepSeek-R1 demonstrated slightly higher concordance (71%) compared to ChatGPT-4.0 (59%). However, both models showed inconsistent outputs for treatments such as ozone therapy and arthroscopy. Despite high inter-rater agreement, the study

emphasized the need for critical scrutiny of AI-generated medical advice. In contrast to the above study focused on KOA, the present research explores AI performance in broader orthopedic rehabilitation contexts with structured scenarios and multi-criteria scoring. Gültekin et al.²⁸ compared ChatGPT-4o and DeepSeek R1-each equipped with real-time internet access features (Deep Research and DeepThink, respectively)-in generating educational responses to frequently asked questions about ACL surgery. While both models demonstrated high accuracy and consistency, ChatGPT-4o provided more comprehensive answers, whereas DeepSeek R1 produced significantly clearer and more readable content, as supported by Flesch-Kincaid metrics. These findings emphasize model-specific strengths relevant to evidence-based patient education. Saglam et al.²⁹ compared the clinical decision-making performance of GPT-4 and GPT-3.5 in the fields of sports surgery and physiotherapy using standardized scenarios evaluated by clinicians. GPT-4 outperformed GPT-3.5 across all parameters, including diagnostic accuracy, treatment planning, and rehabilitation protocols. While the findings support the potential of LLMs in medical decision support, the study focused solely on GPT-based models, limiting generalizability to the broader AI landscape. In contrast, the present study expands the comparative framework to include diverse AI architectures-ChatGPT-4o, Gemini 2.5 Pro, and DeepSeek-V3-thus providing a more comprehensive assessment of language model performance in orthopedic rehabilitation contexts. Recent work by Karaagac and Carkit³⁰ evaluated the performance of four AI-based chatbots-GPT, DeepSeek, Gemini, and Copilot-in providing public health information on liver cancer. Using established quality assessment tools such as the DISCERN instrument and guideline concordance checks (AASLD, NCCN, ESMO), they found that while all models generally offered relevant content, there were notable differences in actionability, clarity, and adherence to guidelines. GPT and DeepSeek were more action-oriented, whereas Gemini and Copilot excelled in user understandability. However, none fully met the standards of consistent, guideline-referenced medical communication. Although their focus was on oncology-related public education, this line of research reinforces the importance

of evaluating AI performance across diverse domains. From a practical perspective, model selection may depend on the user profile. Clinicians might prefer DeepSeek-V3 for its strong emphasis on safety and detailed planning, while students and trainees may benefit from ChatGPT-4o's structured, phased rehabilitation frameworks that support learning. Gemini's inclusion of psychosocial elements and accessible explanations could be particularly valuable for patient education or wellness-oriented applications. This differentiation underscores the importance of matching the model's strengths to the specific needs of its intended user. To our knowledge, the present study is the first to systematically compare ChatGPT-4o, Gemini 2.5 Pro, and DeepSeek-V3 in the specialized context of orthopedic rehabilitation, where clinical reasoning and treatment planning accuracy are critical.

Only one expert physiotherapist rated the AI responses, which prevented assessment of inter-rater reliability. Future studies should involve multiple raters to enhance scoring robustness and allow reliability testing. In addition, there is a possibility of prior familiarity with the models, which, although minimized by separating the scenario author from the evaluator, cannot be completely ruled out. The scenarios themselves were fictional and, despite being designed for clinical realism, inherently involve a degree of subjectivity that may have influenced how responses were interpreted and scored. Finally, because only three standardized scenarios were included for each model, the dataset was too small for meaningful inferential statistical testing. Future studies with larger and more diverse datasets should incorporate appropriate statistical analyses to allow robust comparisons between models.

CONCLUSION

Overall, the AI-generated plans demonstrated substantial alignment with evidence-based physiotherapy principles but varied in the integration of psychosocial components, lifestyle advice, and patient education depth. These differences may inform the choice of AI tool depending on clinical context and user expertise. These findings underscore the growing potential of AI-generated rehabilitation plans to support clinical decision-making; however, as their integration into healthcare settings increases, it is crucial to critically assess their consistency and adherence to established professional guidelines.

ETHICAL DECLARATIONS

Ethics Committee Approval

This study did not require ethical approval as it did not involve any human subjects or animal experiments.

Informed Consent

Because the study did not involve human participants, no written consent was obtained.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

Data availability

Data are available upon request.

REFERENCES

1. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. *Cureus*. 2023;15(4):e37589. doi:10.7759/cureus.37589
2. Onan D, Arıkan H, Can İ, Güven Ş, Işııkay L, Ozge A. Examining the ability of artificial intelligence with ChatGPT-4.0 to create an exercise program: case scenario examples" lumbar disc herniation, chronic migraine, and urge urinary incontinence". *Turk J Kinesiol*. 2025;11(1):28-44.
3. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: a taxonomy and systematic review. *Comput Method Progr Biomed*. 2024;245:108013. doi:10.1016/j.cmpb.2024.108013
4. Tustumi F, Andreollo NA, Aguilar-Nascimento JEd. Future of the language models in healthcare: the role of ChatGPT. *Arq Bras Cir Dig*. 2023;36:e1727. doi:10.1590/0102-672020230002e1727
5. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? *Diabetes Metabol Syndr*. 2023;17(4):102744. doi:10.1016/j.dsx.2023.102744
6. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical considerations of using ChatGPT in health care. *J Med Internet Res*. 2023;25:e48009. doi:10.2196/48009
7. Giorgino R, Alessandri-Bonetti M, Del Re M, Verdoni F, Peretti GM, Mangiavini L. Google Bard and ChatGPT in orthopedics: which is the better doctor in sports medicine and pediatric orthopedics? The role of AI in patient education. *Diagnostics*. 2024;14(12):1253. doi:10.3390/diagnostics14121253
8. Artioli E, Veronesi F, Mazzotti A, et al. Assessing ChatGPT responses to common patient questions regarding total ankle arthroplasty. *J Exp Orthop*. 2025;12(1):e70138. doi:10.1002/jeo2.70138
9. Giangarra CE, Manske RC. Clinical orthopaedic rehabilitation: a team approach E-book: Elsevier Health Sciences; 2017.
10. Sueki D, Brechter J. Orthopedic rehabilitation clinical advisor: Elsevier Health Sciences; 2009.
11. Powden CJ, Hoch MC, Hoch JM. Examination of response shift after rehabilitation for orthopedic conditions: a systematic review. *J Sport Rehabil*. 2018;27(5):469-479. doi:10.1123/jsr.2017-0104
12. Sahin Ozdemir M, Ozdemir YE. Comparison of the performances between ChatGPT and Gemini in answering questions on viral hepatitis. *Sci Rep*. 2025;15(1):1712. doi:10.1038/s41598-024-83575-1
13. Erol E, Arıkan H. Does ChatGPT provide comprehensive and accurate information regarding the effects, types and programming of core exercises? *Turk J Kinesiol*. 2024;10(3):178-182. doi:10.31459/turkjkin.1516614
14. Rossetini G, Barger S, Cook C, et al. Accuracy of ChatGPT-3.5, ChatGPT-4o, Copilot, Gemini, Claude, and Perplexity in advising on lumbosacral radicular pain against clinical practice guidelines: crosssectional study. *Front Digit Health*. 2025;7:1574287. doi:10.3389/fgdth.2025.1574287
15. Marcaccini G, Seth I, Xie Y, et al. Breaking bones, breaking barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. *J Clin Med*. 2025;14(6):1983. doi:10.3390/jcm14061983
16. Boone Jr HN, Boone DA. Analyzing likert data. *J Extension*. 2012;50(2):48. doi:10.34068/joe.50.02.48
17. Sidiq M, Sharma J. ChatGPT in physiotherapy research and education: a boon or bane?-an overview. *Eur J Physiother*. 2025;1-3. doi:10.1080/12679169.2025.2513354

18. Safran E, Yildirim S. A cross-sectional study on ChatGPT's alignment with clinical practice guidelines in musculoskeletal rehabilitation. *BMC Musculoskelet Disord*. 2025;26(1):411. doi:10.1186/s12891-025-08650-8
19. Naqvi WM, Shaikh SZ, Mishra GV. Large language models in physical therapy: time to adapt and adept. *Front Public Health*. 2024;12:1364660. doi:10.3389/fpubh.2024.1364660
20. Hao J, Yao Z, Tang Y, Remis A, Wu K, Yu X. Artificial intelligence in physical therapy: evaluating ChatGPT's role in clinical decision support for musculoskeletal care. *Ann Biomed Eng*. 2025;53(1):9-13. doi:10.1007/s10439-025-03676-4
21. Bilika P, Stefanouli V, Strimpakos N, Kapreli EV. Clinical reasoning using ChatGPT: is it beyond credibility for physiotherapists use? *Physiother Theory Pract*. 2024;40(12):2943-2962. doi:10.1080/09593985.2023.2291656
22. Arbel Y, Gimmon Y, Shmueli L. Evaluating the potential of large language models for vestibular rehabilitation education: a comparison of ChatGPT, Google Gemini, and Clinicians. *Phys Ther*. 2025;105(4):pzaf010. doi:10.1093/ptj/pzaf010
23. Kleebayoon A, Wiwanitkit V. On "evaluating the potential of large language models for vestibular rehabilitation education: a comparison of ChatGPT, Google Gemini, and Clinicians." *Phys Ther*. 2025;105(4):pzaf010. doi:10.1093/ptj/pzaf010
24. Gürses ÖA, Özudoğru A, Tuncay F, Kararti C. The role of artificial intelligence large language models in personalized rehabilitation programs for knee osteoarthritis: an observational study. *J Med Syst*. 2025;49(1):73. doi:10.1007/s10916-025-02207-x
25. Puce L, Bragazzi NL, Curra A, Trompetto C. Harnessing generative artificial intelligence for exercise and training prescription: applications and implications in sports and physical activity-a systematic literature review. *Appl Sci (2076-3417)*. 2025;15:7.
26. Usen A, Kuculmez O. Evaluation of the performance of large language models in the management of axial spondyloarthritis: analysis of EULAR 2022 recommendations. *Diagnostics*. 2025;15(12):1455. doi:10.3390/diagnostics15121455
27. Wu H, Yao S, Bao H, Guo Y, Xu C, Ma J. ChatGPT-4.0 and DeepSeek-R1 does not yet provide clinically supported answers for knee osteoarthritis. *Knee*. 2025;56:386-396. doi:10.1016/j.knee.2025.06.007
28. Gültekin O, Inoue J, Yilmaz B, et al. Evaluating DeepResearch and DeepThink in anterior cruciate ligament surgery patient education: ChatGPT-4o excels in comprehensiveness, DeepSeek R1 leads in clarity and readability of orthopaedic information. *Knee Surg Sports Traumatol Arthrosc*. 2025. doi:10.1002/ksa.12711
29. Saglam S, Uludag V, Karaduman ZO, Arıcan M, Yücel MO, Dalaslan RE. Comparative evaluation of artificial intelligence models GPT-4 and GPT-3.5 in clinical decision-making in sports surgery and physiotherapy: a cross-sectional study. *BMC Med Inform Decis Mak*. 2025;25(1):163. doi:10.1186/s12911-025-02996-8
30. Karaagac M, Carkit S. Evaluation of AI-based chatbots in liver cancer information dissemination: a comparative analysis of GPT, DeepSeek, Copilot, and Gemini. *Oncology*. 2025. doi:10.1159/000546726