Research Report

JMST

Comparing ChatGPT and DeepSeek for Generating Clinically Relevant Responses related to Physical Therapy

Jun-hee Kim, PT, Ph.D

Department of Physical Therapy, College of Software and Digital Healthcare Convergence, Yonsei University, Wonju, South Korea

Background Integrating large language models, such as ChatGPT, into healthcare has introduced new opportunities in medical education and clinical decision support. Recently, DeepSeek—an alternative artificial intelligence (AI) model optimized for computational efficiency—has emerged as a potential competitor to ChatGPT. However, the clinical accuracy and relevance of these models in physical therapy remain unclear.

Purpose This study aimed to compare ChatGPT and DeepSeek in generating responses relevant to musculoskeletal sciences and rehabilitation.

Study design A technical evaluation study

Methods A comparative analysis was conducted to evaluate ChatGPT and DeepSeek using six standardized questions related to musculoskeletal rehabilitation. Both models' responses were evaluated by clinical expert using a 5-point scale based on six criteria including accuracy, coherence, fluency, reason-ing ability, justification, and medical suitability.

Results ChatGPT provided comprehensive and structured explanations with strong clinical reasoning and justification, rendering it suitable for healthcare professionals. Meanwhile, DeepSeek generated concise, accessible responses optimized for quick understanding but lacked depth and justification. Although both models demonstrated good accuracy, ChatGPT's responses were more suitable for professional use, whereas DeepSeek's responses were more user-friendly for nonspecialists.

Conclusions ChatGPT exhibited superior clinical depth and justification, rendering it more appropriate for medical professionals and educators. DeepSeek's computational efficiency and concise responses suggested its potential utility in patient education and telemedicine. Overall, a combined AI approach integrating depth and computational efficiency can enhance AI-driven healthcare applications. However, further validation in this regard is needed to optimize AI deployment in rehabilitation.

Key words ChatGPT; DeepSeek; Large language models; Physical therapy; Rehabilitation.

J Musculoskelet Sci Technol
2025; 9(1): 9-18
Published Online Jun 30, 2025
pISSN 2635-8573
eISSN 2635-8581

Article History
Received 19 Feb 2025
Revised 09 Mar 2025 (1st)
Revised 19 Mar 2025 (2nd)
Accepted 20 Mar 2025

CONTACT
move@yonsei.ac.kr
Jun-hee Kim,
Department of Physical Therapy, College of Software and Digital Healthcare Convergence, Yonsei University, Wonju, South Korea

INTRODUCTION

The rapid advancement of large language models (LLM) has notably impacted various fields, including education, business, scientific research, and healthcare. LLM-based generative artificial intelligence (AI) models, such as OpenAI's ChatGPT, Google's Gemini, and Meta's Llama, utilize vast datasets and deep learning–based algorithms to generate human-like responses.[1–3] These models have demonstrated remarkable capabilities in natural-language understanding, content generation, data analysis, and decision support, making them increasingly valuable in academic research, professional communication, customer service, and knowledge management.[4–7] As AI technology continues to

evolve, the potential of LLMs to enhance efficiency, improve access to information, and support complex problem solving across multiple domains is being increasingly explored.

Among the available generative AI models, ChatGPT has garnered considerable attention with regard to medical and rehabilitative applications. It has been widely employed for summarizing research papers, answering medical queries, and even providing preliminary diagnostic insights.[8–10] In physical therapy, ChatGPT has proven its ability to deliver information regarding musculoskeletal disorders, rehabilitation techniques, and clinical guidelines, making it a valuable resource for clinicians, students, and educators. For example, a study analyzing ChatGPT's responses for the shoulder impingement syndrome found that ChatGPT could provide definitions, risk factors, symptoms, and treatment options, including rehabilitation exercises.[11] However, the same study also highlighted ChatGPT's tendency to present biased or potentially inaccurate medical information, reinforcing the need for human oversight. Similarly, in orthopedic education, ChatGPT has been employed to simplify patient education materials associated with rotator cuff injuries, improving accessibility while maintaining medical accuracy.[12] Further, in sports rehabilitation, ChatGPT has been integrated into patient support systems, allowing individuals to ask questions regarding treatment plans, exercise modifications, and recovery strategies while receiving real-time and personalized feedback on their rehabilitation progress.[13] Furthermore, academic physical-therapy programs have initiated leveraging ChatGPT to assist in curriculum development, streamline research documentation, and create case-based learning materials, highlighting the growing role of ChatGPT in education and professional training.[14] These applications underscore ChatGPT's expanding role in physical therapy and rehabilitation and highlight the need for further validation to ensure clinical accuracy and alignment with evidence-based practices.

Despite the potential benefits of generative AI in healthcare, notable challenges remain. AI-generated responses can contain inaccurate information, hallucinated facts, and biases, raising concerns in terms of clinical accuracy and trustworthiness.[15] Although ChatGPT and similar LLMs process vast datasets, they do not always provide responses consistent with evidence-based medical literature and may struggle with specialized terminology and clinical reasoning, which are essential for effective patient care.[16] Moreover, the high computational costs associated with LLMs pose sustainability concerns.[17] Training and deploying these models require energy-intensive processes, leading to high operational costs and a negative environmental impact.[18] The need for continuous real-time processing during the operation of medical applications further amplifies these challenges, making it crucial to develop highly efficient and sustainable AI solutions before their widespread adoption in healthcare.[19]

As AI-driven medical applications have become more prevalent, the demand for cost-effective and computationally efficient alternatives to proprietary models such as ChatGPT has increased. One such emerging model is DeepSeek, a mixture-of-experts (MoE) LLM designed to offer high-performance language processing at reduced computational costs.[20,21] DeepSeek employs the multihead latent attention (MLA) and DeepSeekMoE architectures, which improve inference efficiency and reduce overall training expenses. Unlike ChatGPT, which requires extensive computational resources, DeepSeek achieves competitive performance with only 2.788 million GPU hours of training, making it a more affordable option for AI-driven medical applications.[21] DeepSeek has undergone supervised fine tuning and reinforcement learning to enhance its consistency with human-like responses, thereby improving its applicability across diverse domains.[21] However, despite these advancements, DeepSeek has not been rigorously validated for clinical use in the fields of physical therapy and rehabilitation sciences. Unlike those of specialized models explicitly designed for healthcare applications, the responses of DeepSeek with regard to musculoskeletal diagnosis and treatment have not been systematically assessed for clinical accuracy, relevance, and reliability. The lack of formal validation raises concerns about its suitability for providing evidence-based medical guidance in physical therapy. Given these uncertainties, a direct performance comparison with ChatGPT is necessary to determine whether DeepSeek can serve as a viable tool for medical and rehabilitative applications.

The current study aims to compare the performances of ChatGPT and DeepSeek in generating responses related to physical therapy and rehabilitation sciences. The research focuses on evaluating accuracy, clinical relevance, and readability to determine whether low-cost LLMs such as DeepSeek can be viable alternatives to more established AI models in medical education and clinical practice. By conducting a structured comparative analysis, this study addresses the strengths and limitations of these AI models in providing evidence-based physical-therapy knowledge. The findings will provide valuable insights to AI developers, healthcare professionals, and educators, guiding the integration of generative AI into medical training and patient care. Ultimately, this research will contribute to the growing discourse on the role of AI in healthcare and help shape future advancements in medical AI applications.

METHODS

Study design

This study was designed as a technical evaluation employing a comparative qualitative analysis of AI-generated responses from two generative language models—ChatGPT and DeepSeek. Figure 1 outlines the study process. This study employed a comparative qualitative research design in line with established frameworks for systematically and context-sensitively appraising new technologies in healthcare.[22,23] Specifically, this study compared AI-generated responses from ChatGPT and DeepSeek in musculoskeletal rehabilitation, integrating domain-expert evaluations to assess each model's depth, accuracy, and clinical relevance.

AI model selection

On January 31, 2025, AI-generated responses were collected from ChatGPT o1 and DeepSeek with the R1 functionality activated (Table 1). Both models were tested under identical conditions, without external fine-tuning or prompt modifications. This approach ensured comparison of the models' baseline responses to medical questions related to musculoskeletal sciences and rehabilitation.

Question selection

Six questions were selected based on their relevance to musculoskeletal sciences and rehabilitation. These questions, which covered musculoskeletal functions, movement impairments, clinical assessments, and postoperative management (Table 2), were derived from clinical scenarios and key biomechanical principles frequently encountered in physical-therapy practice.[24] Each question was input into each model, after which the generated responses were collected without modification.

Evaluation criteria

Responses were analyzed based on the following six predefined criteria (Table 3): (1) accuracy, the extent to which the involved response aligned with established medical and biomechanical knowledge; (2) coherence, the logical structure and flow of information within the

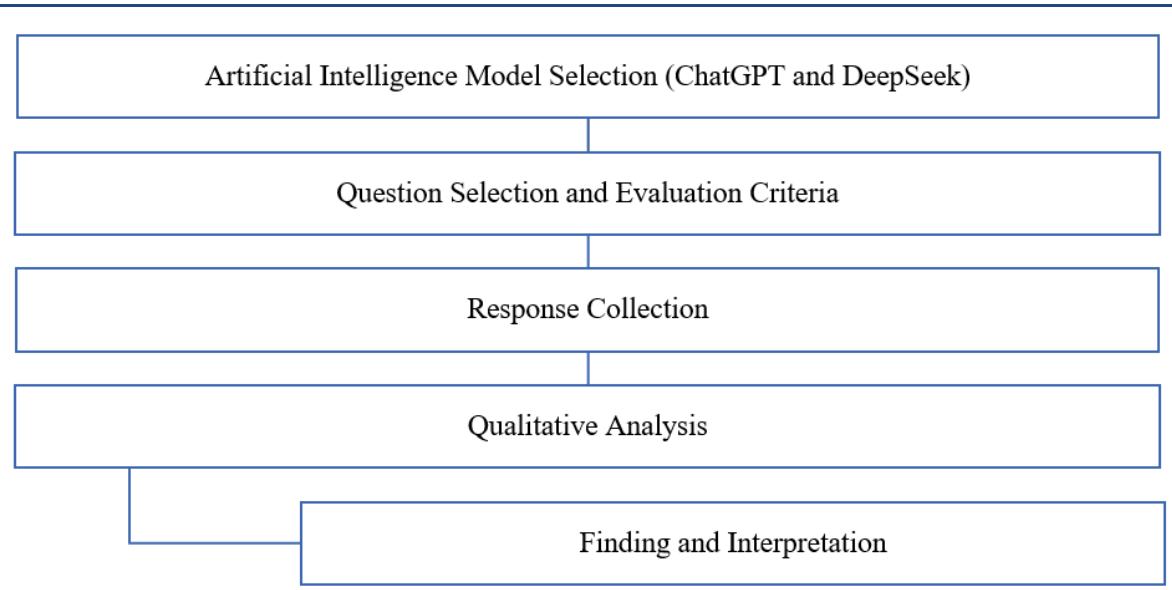

Figure 1. Flowchart of this study.

Table 1. Description of artificial intelligence models used in the study

Model	Description
ChatGPT	A large-scale generative language model by OpenAI, trained on diverse datasets including medical literature and scientific texts. The o1 model was selected for its advanced reasoning and improved domain-specific response generation.
DeepSeek	A competing generative AI model, optimized for concise and structured responses, with R1 functionality activated to enhance response accuracy and contextual understanding in specialized topics.

Table 2. Questions related diagnosis and treatment of musculoskeletal system

List of questions
1) Tell me about the action of the serratus anterior muscle.
2) Tell me about the scapulo-humeral rhythm.
3) Tell me about the scapular downward rotation movement impairment syndrome.
4) Tell me about the upper crossed syndrome.
5) Tell me about the kinetic medial rotation test.
6) Tell me about the management for the glenohumeral joint after surgery.

response; (3) fluency, the clarity and readability of the language used; (4) reasoning ability, the depth of biomechanical analysis and logical explanation; (5) justification, the presence of supporting details, evidence, or rationale for the response; and (6) medical suitability, the relevance of the response for clinical and educational use in rehabilitation and physical therapy. These evaluation criteria were established by integrating methodological and conceptual frameworks drawn from recent investigations into AI-driven medical assessment and content analysis.[25–27]

Qualitative analysis

Qualitative analysis according to evaluation criteria was performed by a single evaluator with more than 5 years of clinical experience in musculoskeletal therapy and more than 10 years of research experience in that field. Each response generated by ChatGPT and DeepSeek was evaluated using a structured five-point rating scale across six predefined criteria. The analysis process involved identifying notable differences in the scope and detail of content coverage, assessing the depth of biomechanical reasoning and clinical relevance, comparing the clarity and logical flow of each explanation, and recording observations on each model's strengths and limitations in addressing domain-specific inquiries.

RESULTS

Responses generated by ChatGPT and DeepSeek to each question were collected, which are comparatively summarized in Table 4. And evaluation results for each criterion, rated on the five-point scale, are presented in Table 5. The full responses are presented in Supplementary File 1.

Accuracy

ChatGPT generated highly detailed and precise explanations, incorporating anatomical terminology, physiological principles, and clinical implications. It elaborated on complex biomechanical processes and included phase transitions, specific muscle activations, and pathological implications. Meanwhile, DeepSeek presented accurate and concise responses that summarized key concepts without comprehensive analysis. Although both models delivered factually correct information, the explanations of ChatGPT were more comprehensive than those of DeepSeek, whereas DeepSeek's responses were optimized for quick understanding.

Coherence

ChatGPT structured its responses in a hierarchical manner, progressing logically from basic definitions to clinical applications. Each section followed a clear sequence, enabling smooth information flow. Meanwhile, DeepSeek presented its responses in a bulleted list, which although enhanced readability but occasionally resulted in fragmented information that lacked connection between related concepts.

Fluency

Both models exhibited high fluency in generating natural-sounding responses. However, the language used in ChatGPT's responses resembled that used in academic or medical literature, rendering them more suitable for healthcare professionals and researchers. Meanwhile, DeepSeek used simpler language with a more direct communication style, making its content more accessible to nonspecialists, such as fitness professionals, patients, and general readers.

Reasoning ability

ChatGPT demonstrated strong reasoning ability, particularly in biomechanical explanations and clinical applications. It frequently expounded on cause–effect relations, compensatory mechanisms, and assessment methodologies, providing comprehensive responses regarding diagnostic

Table 3. Evaluation criteria for assessing generated responses

Evaluation criteria	Rating	Description
Accuracy	1	Contains multiple major factual errors or the information significantly deviates from established facts
	2	Contains noticeable factual inaccuracies or omissions that reduce reliability
	3	Mostly accurate but minor imprecisions or missing details may appear
	4	Generally factually sound with only minor oversights and main points are reliable and consistent with known facts
	5	Highly accurate and factually sound with no or negligible errors and aligns well with established knowledge
Coherence	1	Extremely disjointed or unclear making it very hard to follow the argument or narrative
	2	Some sections flow awkwardly or contain logical gaps that disrupt readability
	3	Overall coherence is acceptable though occasional abrupt transitions or mild logical gaps may occur
	4	The writing is mostly well-organized with sections and paragraphs linking smoothly and minimal logical gaps
	5	Very clear and logically consistent throughout and paragraphs and sentences link seamlessly for a highly readable text
Fluency	1	Language use is awkward with frequent grammatical or spelling errors and comprehension is significantly hindered
	2	Style or grammar issues occasionally impede reading and some expressions feel unnatural
	3	Basic clarity is maintained though some minor awkwardness or errors can appear but do not severely impair understanding
	4	The text reads smoothly with few grammatical errors and language style is appropriate and content is easy to understand
	5	Demonstrates excellent command of language with near-perfect grammar, style, and fluidity making it effortless to read
Reasoning ability	1	Lacks clear explanations or causal links and conclusions seem unfounded or are drawn abruptly
	2	Some rationales are given but key steps in reasoning are missing or not well explained
	3	Provides reasonable explanations and causal links but may omit deeper details or skip certain logical steps
	4	Offers solid rationales and logical links that explain causes, processes, and outcomes in a coherent manner
	5	Thorough structured reasoning with detailed cause-effect analysis and robust argumentation suitable for expert review
Justification	1	Does not offer supporting evidence or references and claims and recommendations appear unsubstantiated
	2	References or examples are mentioned but insufficiently support the main arguments
	3	Includes general supporting details or references though some may be vague or incomplete
	4	Claims and recommendations are consistently backed by relevant evidence, examples, or explanations and overall persuasive
	5	Provides robust specific evidence, references, or data that thoroughly validate and strengthen the claims offering high credibility
Medical suitability	1	Contains information that is largely inapplicable or potentially harmful if applied in clinical or educational settings
	2	Some parts could be applied but major content does not align well with medical knowledge or requires significant correction
	3	Generally usable medical information but some sections require verification or expert supervision for practical use
	4	Suitable for clinical and educational contexts with minimal adjustments and overall aligned with professional standards
	5	Highly aligned with professional practice and thoroughly appropriate for direct application in clinical or educational settings with little to no modification needed

Table 4. Summary of models' responses

Question	ChatGPT	DeepSeek
Action of serratus anterior muscle	Detailed explanation covering protraction, upward rotation, stabilization, accessory breathing role, innervation, clinical implications, and sports relevance.	Concise summary focusing on protraction, upward rotation, stabilization, innervation, and functional significance. Less detail on biomechanics and clinical applications.
Scapulohumeral rhythm	Thorough breakdown of the 2:1 movement ratio, phases of movement, muscles involved, clinical relevance, and common deviations. Well-structured with cause-effect explanations.	Summarized key points on the 2:1 ratio, movement phases, involved muscles, and clinical importance, but with fewer details on biomechanics and reasoning.
Scapular downward rotation movement impairment syndrome	Comprehensive discussion on pathomechanics, muscle imbalances, assessment, contributing biomechanics, treatment approaches, and prognosis. Includes rehabilitation exercises.	Summarized key aspects, highlighting muscle imbalances, symptoms, diagnostic tests, and general rehabilitation principles but without in-depth assessment techniques.
Upper crossed syndrome	Detailed explanation of postural imbalances, affected muscles, symptoms, common causes, assessment techniques, and specific corrective exercises.	Covers major features, emphasizing tight/weak muscles, causes, symptoms, and broad rehabilitation approaches but lacks detailed assessment methods.
Kinetic medial rotation test	Describes functional assessment of lower limb control during movement, related to knee valgus and ACL injury risk. Explains biomechanics, test procedure, abnormal findings, and clinical applications.	Focuses on shoulder medial rotation assessment, particularly in overhead athletes. Covers procedure, ROM measurement, and clinical relevance.
Management of glenohumeral joint after surgery	Structured post-operative management plan with detailed phase-by-phase rehabilitation, including immobilization, ROM progression, strengthening, neuromuscular control, and return to function.	Summarizes key phases of rehabilitation with emphasis on pain control, ROM, strengthening, and functional recovery but lacks in-depth justifications for specific interventions.

Table 5. Evaluation results

Criterion	ChatGPT	DeepSeek	Explanation
Accuracy	5	4	ChatGPT offers highly detailed information with minimal error, whereas DeepSeek remains broadly accurate but omits some finer details.
Coherence	5	4	ChatGPT presents information step by step from definitions to clinical application, while DeepSeek's bullet points occasionally disrupt the overall flow.
Fluency	5	5	Both models write clearly and grammatically, but ChatGPT's language is more academic compared to DeepSeek's simpler style.
Reasoning ability	5	4	ChatGPT thoroughly explains cause–effect relations and clinical reasoning, whereas DeepSeek focuses on key points without extensive depth.
Justification	5	3	ChatGPT frequently cites assessments and guidelines to support its claims, while DeepSeek delivers concise responses with fewer references.
Medical suitability	5	4	ChatGPT provides in-depth discussions suitable for professional healthcare settings, whereas DeepSeek offers quick, easily digestible content for general users.

and therapeutic approaches. Although DeepSeek was able to identify key elements of movement dysfunctions and rehabilitation approaches, it failed to expound as much as ChatGPT on clinical reasoning or underlying biomechanical principles.

Justification

ChatGPT consistently included strong justifications, ref-

erencing clinical assessments, rehabilitation protocols, and evidence-based treatment guidelines. It also provided rationale for each intervention and explained the biomechanical mechanisms behind specific conditions. Meanwhile, DeepSeek delivered short and practical answers but often lacked justification or supporting details for its statements; although it covered relevant information, it did not provide as much rationale as ChatGPT for clinical assessments or treatments.

Medical suitability

ChatGPT's responses, which contained detailed discussions, clinical assessments, and treatment strategies, were highly suitable for medical and rehabilitation professionals, catering to medical practitioners, physical therapists, and sports scientists who require detailed and research-backed explanations. However, DeepSeek's responses were more suitable for general audiences, fitness trainers, and individuals seeking practical takeaways without excessive technical details; its focus on clarity and conciseness made it more suitable for nonmedical professionals.

DISCUSSION

The current study comparatively evaluated the performances of ChatGPT and DeepSeek in generating responses related to musculoskeletal sciences and rehabilitation. Overall, our findings showed that ChatGPT demonstrated superior ability to provide detailed, clinically relevant responses that incorporated anatomical terminology, physiological principles, and evidence-based rehabilitation strategies. Its responses were structured and offered hierarchical explanations that aligned with established medical education frameworks. In contrast, DeepSeek produced responses that were notably more concise and computationally efficient. This brevity allowed for quick retrieval of essential facts but often lacked depth in biomechanical explanations and justification of its recommendations based on underlying anatomical or physiological principles.

Prior research has extensively explored the application of AI in medical education, diagnostics, and rehabilitation, highlighting both the advantages and challenges of integrating LLM into healthcare and professional training.[28–30] Previous studies have demonstrated that LLM can enhance clinical decision-making by rapidly synthesizing vast amounts of medical knowledge and providing structured, evidence-based responses.[31,32] Compared to traditional clinical decision-support systems, which primarily function as rule-based algorithms, LLMs such as ChatGPT offer a more dynamic and context-sensitive approach by integrating multimodal data and providing explanations that go beyond rigid protocol adherence.[33,34] However, despite these advantages, concerns regarding AI-generated hallucinations remain a notable limitation.[35,36] Prior studies investigating AI reliability in clinical decision support have identified instances wherein models fabricate non-existent conditions, misinterpret medical guidelines, or provide inaccurate citations.[35,36] Compared to rule-based expert systems, which strictly adhere to predefined medical guidelines, LLMs can occasionally generate plausible-sounding but incorrect information due to their probabilistic nature. This limitation underscores the need for human oversight in AI-assisted medical decision-making.

ChatGPT's detailed explanations and biomechanical reasoning suggest its potential use in medical education and professional training. AI-driven educational tools enhance student engagement and understanding of complex physiological processes, particularly in musculoskeletal sciences.[37,38] Unlike traditional educational tools, which often rely on static content delivery, ChatGPT enables interactive and adaptive learning experiences, allowing students to engage in case-based reasoning and receive context-sensitive explanations.[39] This capability renders it particularly valuable in clinical training, where real-time feedback and exposure to diverse patient scenarios are crucial for skill development. Furthermore, as AI-driven models continue to evolve, their integration into medical curricula can complement existing pedagogical approaches, bridging the gap between theoretical knowledge and practical application.

From a resource efficiency perspective, however, the computational demands of ChatGPT raise concerns about its sustainability and scalability in healthcare settings. The energy demands of LLMs are quite substantial, with recent studies indicating that inference now surpasses training in energy consumption, which could have notable environmental impacts.[40] For instance, serving a single ChatGPT prompt generates >4 g of CO_2 equivalent emissions, corresponding to >20 times the carbon footprint of a typical web search.[40] Moreover, depending on the GPU platform and batch size, LLM inference can exhibit notable trade-offs between latency, energy efficiency, and total carbon emissions, necessitating optimized AI infrastructure to mitigate environmental impact.[40] In contrast, DeepSeek's optimized computational efficiency could potentially be more energy-efficient, particularly for applications wherein concise, fact-based responses may be prioritized, such as telemedicine consultations, patient education, and preliminary assessments.[20] DeepSeek is designed to leverage the MoE architecture and MLA strategies, which can considerably reduce inference costs while maintaining high per-

formance across reasoning tasks.[20] Moreover, DeepSeek's reinforcement learning-driven models, such as DeepSeek-R1, have the potential to exhibit faster response times and lower hardware requirements, making them possible alternatives for resource-constrained environments, including rural healthcare settings and mobile health applications.[21] Considering the characteristics of these language models, we believe that a combined approach integrating DeepSeek's computational efficiency with ChatGPT's advanced clinical reasoning capabilities may be necessary. Such a model could enable dynamic switching between highly detailed reasoning and rapid, low-power responses, potentially enhancing real-world applicability while mitigating the environmental impact of AI-driven healthcare solutions.

This research has some limitations. First, this study was conducted by a single evaluator without the participation of multiple experts, and thus, an objective reliability assessment such as inter-rater agreement could not be performed. In addition, a blind procedure to reduce bias was not implemented. The involved evaluation was qualitative, relying on expert analysis rather than quantitative metrics, which may introduce subjectivity. Future research should involve multiple evaluators and incorporate objective reliability measures to strengthen the validity of the findings. Additionally, this study assessed AI models based on relatively simple, information-based questions. To further explore AI's potential in clinical decision-making, future studies should incorporate complex patient scenarios that reflect real-world clinical reasoning processes. Furthermore, this study primarily focused on physical therapy; hence, expanding the dataset to include broader rehabilitation topics could provide a more comprehensive evaluation of AI performance across medical disciplines. Further studies should also explore the integration of AI feedback mechanisms to enhance response accuracy, allowing models to learn from clinician interactions and continuously refine their outputs. Addressing these limitations will enhance the reliability and real-world applicability of AI in rehabilitation sciences, ultimately improving patient care and medical education.

CONCLUSIONS

The current study compared the ability of ChatGPT and DeepSeek in providing response related to musculoskeletal sciences and rehabilitation. Overall, our results showed that ChatGPT provided detailed, structured, and clinically relevant explanations, emphasizing its utility for medical professionals and educators. In contrast, DeepSeek, although concise and computationally efficient, offered quick, easy-to-understand responses that lacked depth in biomechanical reasoning, highlighting its suitability for nonspecialists. ChatGPT excels in clinical reasoning, whereas DeepSeek is more efficient at rapid information retrieval. A hybrid approach combining ChatGPT's depth with DeepSeek's efficiency could balance accuracy with resource sustainability and optimize AI applications in healthcare.

Key Points

Question How do ChatGPT and DeepSeek compare in providing clinically relevant responses regarding musculoskeletal sciences and rehabilitation?

Findings ChatGPT offers detailed, well-justified responses suitable for professionals, whereas DeepSeek provides concise, computationally efficient answers with less depth.

Meaning Although artificial intelligence–based language models can assist in physical-therapy education and patient support, their implementation must balance accuracy, computational efficiency, and clinical applicability.

Article information

Conflict of Interest Disclosures: The author certifies that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria, educational grants, participation in speakers' bureaus, membership, employment, consultancies, stock ownership, or other equity interest, expert testimony, or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge, or beliefs) in the subject matter or materials discussed in this manuscript. The author reports the following details of affiliation or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript.

Funding/Support: I would like to disclose that this research was conducted without the receipt of any external funding. No specific grants, financial support, or sponsorships were utilized in the design, data collection, analysis, interpretation, or composition of this study. The author acknowledges that the study is entirely self-funded, and the absence of external funding has not influenced the independence or objectivity of our research findings.

Acknowledgment: None.

Ethic Approval: This manuscript does not require IRB approval because there are no human and animal participants.

Author contributions

Conceptualization: JH Kim.
Data acquisition: JH Kim.
Design of the work: JH Kim.
Data analysis: JH Kim.
Project administration: JH Kim.
Interpretation of data: JH Kim.
Writing – original draft: JH Kim.
Writing–review&editing: JH Kim.

Supplementary materials

Supplementary materials are only available online from: https://doi.org/10.29273/jmst.2025.9.1.9

REFERENCES

1. Team G, Georgiev P, Lei VI, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:240305530*. Published online 2024.
2. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. *Minds Mach (Dordr)*. 2020;30:681-694.
3. Touvron H, Lavril T, Izacard G, et al. Llama: open and efficient foundation language models. *arXiv preprint arXiv:230213971*. Published online 2023.
4. Mosch L, Fürstenau D, Brandt J, et al. The medical profession transformed by artificial intelligence: qualitative study. *Digit Health*. 2022;8:20552076221143904.
5. Lund BD, Wang T, Mannuru NR, et al. ChatGPT and a new academic reality: artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J Assoc Inf Sci Technol*. 2023;74(5):570-581.
6. Chaturvedi R, Verma S. Opportunities and challenges of AI-driven customer service. *Artif Intell Customer Serv*. Published online 2023:33-71.
7. Pokhrel S, Banjade SR. AI Content generation technology based on Open AI language model. *J Artif Intell Capsule Netw*. 2023; 5(4):534-548.
8. Caruccio L, Cirillo S, Polese G, et al. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. *Expert Syst Appl*. 2024; 235:121186.
9. Oeding JF, Lu AZ, Mazzucco M, et al. ChatGPT-4 performs clinical information retrieval tasks using consistently more trustworthy resources than does Google search for queries concerning the latarjet procedure. *Arthroscopy*. Published online 2024.
10. Wei Q, Yao Z, Cui Y, et al. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. Published online 2024: 104620.
11. Gwak GT, Kim JH, Hwang UJ, Jung SH. Search for medical information and treatment options for musculoskeletal disorders through an artificial intelligence chatbot: focusing on shoulder impingement syndrome. *J Musculoskelet Sci Technol*. 2023;7(1):8-16.
12. Miskiewicz MJ, Perez M, Capotosto S, et al. Enhancing access to orthopedic education: exploring the potential of generative artificial intelligence (AI) in improving health literacy on rotator cuff injuries. *Cureus*. Published online November 1, 2024. doi:10.7759/cureus.72833
13. Ahsan M. Chatbot generative pre-trained transformer and artificial intelligence in sports physical therapy and rehabilitation. *Saudi J Sports Med*. 2023; 23(2):61-62. doi:10.4103/sjsm.sjsm_16_23
14. Severin R, Gagnon K. An early snapshot of attitudes toward generative artificial intelligence in physical therapy education. *J Phys Ther Educ*. Published online 2024. doi:10.1097/JTE.000000000000 0381
15. Gangavarapu A. Enhancing guardrails for safe and secure healthcare AI. *arXiv preprint arXiv:240917190*. Published online 2024.
16. Templin T, Perez MW, Sylvia S, et al. Addressing 6 challenges in generative AI for digital health: a scoping review. *PLOS Digital Health*. 2024;3(5):e0000503.
17. Lorenz P, Perset K, Berryhill J. *Initial policy considerations for generative artificial intelligence*. Paris, France: OECD Publishing; 2023.
18. Al-kfairy M, Mustafa D, Kshetri N, et al. Ethical challenges and solutions of generative AI: an interdisciplinary perspective. In: *Informatics*. Vol 11. Multidisciplinary Digital Publishing Institute; 2024:58.
19. Biswas A, Talukdar W. Intelligent clinical documentation: harnessing generative AI for patient-centric clinical note generation. *arXiv preprint arXiv:240518346*. Published online 2024.
20. DeepSeek-AI, Liu A, Feng B, et al. DeepSeek-V3 technical report. Published online December 26, 2024. http://arxiv.org/abs/2412.19437
21. Guo D, Yang D, Zhang H, et al. Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:250112948*. Published online 2025.
22. Creswell JW, Poth CN. *Qualitative inquiry and research*

design: choosing among five approaches. Thousand Oaks, CA: Sage Publications; 2016.
23. Patton MQ. *Qualitative research & evaluation methods: integrating theory and practice*. Thousand Oaks, CA: Sage publications; 2014.
24. Kim J hee. Fine-tuning the llama2 large language model using books on the diagnosis and treatment of musculoskeletal system in physical therapy. *Journal of Musculoskeletal Science and Technology*. 2024;8(2):65-73. doi:10.29273/jmst.2024.8.2.65
25. Wang Z, Xiao C, Sun J. AutoTrial: prompting language models for clinical trial design. *arXiv preprint arXiv: 230511366*. Published online 2023.
26. Genovese A, Borna S, Gomez-Cabello CA, et al. Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation. *Ann Transl Med*. 2024;12(6):117.
27. Sarraf S. Evaluating Generative AI-enhanced content: a conceptual framework using qualitative, quantitative, and mixed-methods approaches. *arXiv preprint arXiv: 241117943*. Published online 2024.
28. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Med Educ*. 2023; 9(1):e48291.
29. Chow JCL, Wong V, Li K. Generative pre-trained transformer-empowered healthcare conversations: current trends, challenges, and future directions in large language model-enabled medical chatbots. *BioMedInformatics*. 2024;4(1):837-852.
30. Zhou H, Liu F, Gu B, et al. A survey of large language models in medicine: progress, application, and challenge. *arXiv preprint arXiv:231105112*. Published online 2023.
31. Porter R, Diehl A, Pastel B, et al. LLMD: A large language model for interpreting longitudinal medical records. *arXiv preprint arXiv:241012860*. Published online 2024.
32. Vaid A, Lampert J, Lee J, et al. Natural language programming in medicine: administering evidence based clinical workflows with autonomous agents powered by generative large language models. *arXiv preprint arXiv:240102851*. Published online 2024.
33. Umerenkov D, Zubkova G, Nesterov A. Deciphering diagnoses: how large language models explanations influence clinical decision making. *arXiv preprint arXiv:231001708*. Published online 2023.
34. Delourme S, Redjdal A, Bouaud J, et al. Measured performance and healthcare professional perception of large language models used as clinical decision support systems: a scoping review. *Stud Health Technol Inform*. 2024;316:841-845.
35. Giuffrè M, You K, Shung DL. Evaluating ChatGPT in medical contexts: the imperative to guard against hallucinations and partial accuracies. *Clin Gastroenterol Hepatol*. 2024;22(5):1145-1146.
36. Bélisle-Pipon JC. Why we need to be careful with LLMs in medicine. *Front Med (Lausanne)*. 2024;11: 1495582.
37. Guo AA, Li J. Harnessing the power of ChatGPT in medical education. *Med Teach*. 2023;45(9):1063.
38. Kundakcı YE. ChatGPT's Capabilities for use in anatomy education and anatomy research. *Eur J Ther*. 2024;30(2):200-202.
39. Huang Z, Mao Y, Zhang J. The influence of artificial intelligence technology on college students' learning effectiveness from the perspective of constructivism—taking ChatGPT as an example. *J Educ Humanit Soc Sci FMESS*. 2024;30(1):40-46.
40. Nguyen S, Zhou B, Ding Y, et al. Towards sustainable large language model serving. Published online December 30, 2024. http://arxiv.org/abs/2501.01990