Article

Not peer-reviewed version

AI-Powered Physiotherapy: Evaluating LLMs Against Students in Clinical Rehabilitation Scenarios

Ioanna Michou , Athanasios Fouras , Dionysia Chrysanthakopoulou , Marina Theodoritsi , Sotiria Stellatou , Savvina Mariettou , Constantinos Koutsojannis *

Posted Date: 30 July 2025

doi: 10.20944/preprints202507.2498.v1

Keywords: generative AI; large language models; physiotherapy; rehabilitation; clinical decision-making; medical education; ChatGPT; DeepSeek; low back pain; multiple sclerosis; frozen shoulder; knee osteoarthritis; AI voice assistants; AI characters

 doi:10.20944/preprints202507.2498.v1

Article

AI-Powered Physiotherapy: Evaluating LLMs Against Students in Clinical Rehabilitation Scenarios

Ioanna Michou, Athanasios Fouras, Dionysia Chrysanthakopoulou, Marina Theodoritsi, Sotiria Stellatou, Savvina Mariettou and Constantinos Koutsojannis *

Health Physics & Computational Intelligence lab, Department of Physiotherapy, School of Rehabilitation Sciences, University of Patras, Patras, Greece

* Correspondence: ckoutsog@upatras.gr (C.K.)

Abstract

Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs) like ChatGPT and DeepSeek, is transforming healthcare by enhancing clinical decision-making, education, and patient interaction. This study compares ChatGPT (GPT-4) and DeepSeek against 60 final-year physiotherapy students in Greece answering 60 clinical questions across four rehabilitation domains: low back pain, multiple sclerosis, frozen shoulder, and knee osteoarthritis (15 questions per domain). Questions spanned basic knowledge, diagnosis, alternative treatments, and rehabilitation practices. Responses were evaluated for relevance, accuracy, clarity, completeness, and consistency with clinical practice guidelines (CPGs), emphasizing conceptual understanding. Results indicate LLMs outperformed students in most domains, particularly in global response quality and conceptual depth, raising questions about AI's role in physiotherapy. This manuscript explores these findings, compares them with related work, and discusses whether GenAI could transform or threaten physiotherapy. Ethical considerations, limitations, and future directions, including AI voice assistants and AI characters, are addressed.

Keywords: generative AI; large language models; physiotherapy; rehabilitation; clinical decision-making; medical education; ChatGPT; DeepSeek; low back pain; multiple sclerosis; frozen shoulder; knee osteoarthritis; AI voice assistants; AI characters

Introduction

The integration of Artificial Intelligence (AI), particularly Generative AI (GenAI) and Large Language Models (LLMs), into healthcare has ushered in a new era of innovation, transforming clinical practice, education, and research [1]. LLMs, such as OpenAI's ChatGPT (GPT-4) and DeepSeek AI, leverage advanced natural language processing (NLP) to generate human-like text, offering applications in clinical decision support, patient education, and professional training [2,3]. In physiotherapy, a discipline that blends scientific knowledge with hands-on clinical skills, AI technologies have shown significant promise in areas such as motion analysis, wearable technologies, and predictive modeling for patient outcomes [4]. However, the application of LLMs in physiotherapy remains relatively underexplored, particularly in their ability to address complex clinical queries compared to human expertise. This study is one of the first to systematically evaluate LLMs against physiotherapy students in clinical question-answering, shedding light on their potential to augment or challenge traditional physiotherapy education and practice.

Beyond text-based LLMs, emerging AI technologies such as AI voice digital assistants and AI characters are gaining traction in healthcare [5]. AI voice assistants, like Amazon's Alexa or Google Assistant, enable hands-free interaction, delivering real-time clinical information, guiding patients through exercises, or assisting clinicians with documentation [6]. For instance, voice-activated systems can provide verbal instructions for home-based rehabilitation programs, improving patient adherence and accessibility [7]. AI characters, or virtual avatars powered by GenAI, offer interactive

 doi:10.20944/preprints202507.2498.v1

and personalized experiences by simulating patient or clinician roles in training scenarios [8]. These avatars can engage in dynamic conversations, model clinical interactions, or serve as virtual patients for student practice, potentially revolutionizing physiotherapy education and telehealth delivery [9]. While these technologies are in their nascent stages, their potential to enhance engagement, accessibility, and scalability in physiotherapy warrants further investigation.

The application of AI in medical education has seen rapid growth across various disciplines, providing a broader context for understanding its role in physiotherapy. In medical schools, LLMs have been employed to support case-based learning, simulate patient interactions, and assist with diagnostic reasoning [21]. For example, studies in medical education have demonstrated that LLMs like ChatGPT can generate accurate responses to clinical vignettes, achieving performance levels comparable to third-year medical students in certain domains [22]. In nursing education, AI-driven virtual patients have been used to train students in clinical decision-making, improving their ability to handle complex scenarios without the logistical challenges of real-world clinical placements [23]. Similarly, in allied health professions such as occupational therapy, AI tools have been integrated into curricula to enhance students' understanding of evidence-based practice and patient communication [24]. These applications highlight the versatility of AI in medical education, but physiotherapy's unique emphasis on manual therapy and patient rapport presents distinct challenges and opportunities for AI integration.

In clinical practice, AI has been increasingly adopted to support decision-making and optimize patient outcomes across healthcare fields. In radiology, AI algorithms assist in image interpretation, achieving diagnostic accuracy comparable to or surpassing human experts in specific tasks [25]. In cardiology, predictive models powered by AI analyze patient data to forecast adverse events, enabling early interventions [26]. In physiotherapy, AI-driven tools such as wearable sensors and motion capture systems have been used to monitor patient progress and tailor rehabilitation programs [4]. However, the use of LLMs in clinical physiotherapy remains limited, with most studies focusing on their ability to provide general medical knowledge rather than domain-specific expertise [2]. The current study addresses this gap by evaluating LLMs in the context of physiotherapy-specific clinical scenarios, comparing their performance to that of final-year students.

Physiotherapy relies heavily on evidence-based clinical practice guidelines (CPGs) to manage conditions such as low back pain, multiple sclerosis, frozen shoulder, and knee osteoarthritis [10]. Recent studies indicate that ChatGPT aligns with CPGs in approximately 80% of musculoskeletal queries, though it struggles with context-specific scenarios, such as lumbosacral radicular pain [2,11]. Concerns about AI "hallucinations" (fabricated or incorrect responses), data privacy, and ethical integration remain significant barriers to widespread adoption [12]. In other health professions, LLMs have demonstrated variable performance. For instance, in orthopedics, ChatGPT achieved 45–73.6% accuracy on examination questions, with notable limitations in nuanced or context-dependent scenarios [13]. Domain-specific models, such as BioMedLM, have shown superior performance in medical education by leveraging specialized training datasets [14]. These findings suggest that while LLMs hold significant potential to augment healthcare delivery, they cannot fully replace human judgment, empathy, or hands-on skills, particularly in fields like physiotherapy.

The emergence of AI voice assistants and AI characters introduces new possibilities for clinical practice and education. In medical education, AI voice assistants have been used to simulate patient interactions, allowing students to practice communication skills in a controlled environment [27]. For example, a study in medical training found that voice-activated AI systems improved students' ability to elicit patient histories by providing real-time feedback [28]. In physiotherapy, voice assistants could guide patients through home-based exercises, offering real-time corrections and motivational prompts to enhance adherence [7]. AI characters, meanwhile, have been explored in mental health and nursing education to simulate patient interactions, fostering empathy and clinical reasoning [8,9]. These technologies could be adapted for physiotherapy to create virtual patients with diverse clinical presentations, enabling students to practice assessment and treatment planning in a

 doi:10.20944/preprints202507.2498.v1

low-risk setting. However, their effectiveness in physiotherapy-specific contexts remains largely untested, highlighting the need for further research.

This study is among the first to compare ChatGPT and DeepSeek against 60 final-year physiotherapy students in answering 60 clinical questions across four rehabilitation domains: low back pain, multiple sclerosis, frozen shoulder, and knee osteoarthritis (15 questions per domain covering basic knowledge, diagnosis, alternative treatments, and rehabilitation practices). These domains were selected due to their prevalence and diverse clinical demands [16–19]. Low back pain, affecting up to 80% of adults, requires precise assessment and management [16]. Multiple sclerosis involves progressive neurological deficits, necessitating tailored interventions [17]. Frozen shoulder and knee osteoarthritis demand biomechanical expertise and pain management [18,19]. The study aimed to:

1. Evaluate LLM and student responses for quality and conceptual understanding.
2. Assess LLMs' potential as educational and clinical tools in physiotherapy.
3. Explore whether GenAI could transform or threaten physiotherapy, including the role of AI voice assistants and characters.

The question of whether GenAI signals the "end" of physiotherapy is both timely and complex. While LLMs excel in knowledge-based tasks, physiotherapy's reliance on manual skills, patient rapport, and individualized care suggests a complementary rather than substitutive role for AI [15]. This manuscript presents the methodology, results, and discussion, comparing findings with related work in medical education and clinical practice, and exploring future applications of AI in physiotherapy.

Methodology

Study Design

A cross-sectional observational study compared the performance of ChatGPT (GPT-4) and DeepSeek against 60 final-year physiotherapy students in Greece. The study involved 60 clinical questions across four rehabilitation domains: low back pain, multiple sclerosis, frozen shoulder, and knee osteoarthritis (15 questions per domain), enabling a robust comparison in an education-focused setting.

Participants

Sixty students in their final year of a 4-year Bachelor of Science in Physiotherapy program at a Greek university participated. Their advanced training ensured familiarity with the targeted domains, making them a proxy for entry-level professionals. Participants were recruited voluntarily with informed consent. No exclusion criteria were applied beyond program enrollment. The study was approved by the University of Patras ethical committee, adhering to ethical guidelines.

Question Development

Sixty clinical questions were developed by three experienced physiotherapists specializing in musculoskeletal and neurological rehabilitation. Each domain included 15 questions across four subcategories:

- **Basic knowledge (4–5 questions)**: Covered etiology, pathophysiology, and epidemiology (e.g., "What are the risk factors for knee osteoarthritis?").
- **Diagnosis (3–4 questions)**: Focused on assessment techniques and diagnostic criteria (e.g., "What tests confirm multiple sclerosis?").
- **Alternative treatments (3–4 questions)**: Addressed complementary therapies, such as acupuncture or hydrotherapy (e.g., "What alternative treatments benefit frozen shoulder?").

 doi:10.20944/preprints202507.2498.v1

- **Rehabilitation practices (3–4 questions)**: Emphasized evidence-based interventions, such as exercise or manual therapy (e.g., "What exercises manage low back pain?").
  Questions reflected real-world scenarios, requiring integration of theoretical knowledge, clinical reasoning, and CPGs [10]. They were pilot-tested with five practicing physiotherapists for clarity and relevance.

Data Collection

Data collection occurred between March–May 2025. Students answered the 60 questions independently in a controlled setting to prevent collaboration or external resource use. Responses were submitted in written format. The same questions were input into ChatGPT (GPT-4) and DeepSeek using default settings, without fine-tuning. Inputs were standardized for consistency. Responses were anonymized to prevent bias during evaluation.

Evaluation

Two independent raters, physiotherapists with over 10 years of clinical experience, evaluated responses using a 5-point Likert scale (1 = poor, 5 = excellent) across five criteria: relevance, accuracy, clarity, completeness, and consistency with CPGs. A composite "global quality" score was calculated by averaging the criteria scores. Conceptual understanding was assessed separately on a 5-point scale, evaluating explanation depth and correctness. Inter-rater reliability was measured using Cohen's kappa, with discrepancies resolved through consensus.

Statistical Analysis

Descriptive statistics (mean, standard deviation) were computed for each group (students, ChatGPT, DeepSeek) across the five criteria and global quality for each domain and subcategory. One-way ANOVA or non-parametric tests (e.g., Kruskal-Wallis) compared group performance, with post-hoc tests (e.g., Tukey's HSD) to identify differences. Conceptual understanding was analyzed similarly. Significance was set at $p < 0.05$, with effect sizes (Cohen's d) calculated. Analyses were conducted using Jamovi statistical software.

Results

The study generated data on 60 questions across four domains: low back pain, multiple sclerosis, frozen shoulder, and knee osteoarthritis. Table 1 presents mean scores for the five evaluation criteria (relevance, accuracy, clarity, completeness, and global quality), and Table 2 summarizes conceptual understanding scores.

Table 1. Mean Scores (1–5) for Response Quality Across Rehabilitation Domains.

Domain	Group	Relevance	Accuracy	Clarity	Completeness	Global Quality
Low Back Pain	Students	3.8 ± 0.6	3.7 ± 0.7	3.6 ± 0.6	3.5 ± 0.7	3.65 ± 0.6
	ChatGPT	4.6 ± 0.4	4.5 ± 0.5	4.8 ± 0.3	4.7 ± 0.4	4.65 ± 0.4
	DeepSeek	4.4 ± 0.5	4.3 ± 0.5	4.6 ± 0.4	4.5 ± 0.5	4.45 ± 0.4
Multiple Sclerosis	Students	3.6 ± 0.7	3.5 ± 0.8	3.4 ± 0.7	3.3 ± 0.8	3.45 ± 0.7
	ChatGPT	4.3 ± 0.5	4.2 ± 0.6	4.5 ± 0.4	4.4 ± 0.5	4.35 ± 0.5
	DeepSeek	4.7 ± 0.3	4.6 ± 0.4	4.8 ± 0.3	4.7 ± 0.3	4.70 ± 0.3
Frozen Shoulder	Students	4.0 ± 0.5	4.1 ± 0.6	3.8 ± 0.6	3.7 ± 0.6	3.90 ± 0.5
	ChatGPT	4.4 ± 0.4	4.3 ± 0.5	4.6 ± 0.4	4.5 ± 0.4	4.45 ± 0.4

 doi:10.20944/preprints202507.2498.v1

Domain	Group	Relevance	Accuracy	Clarity	Completeness	Global Quality
	DeepSeek	4.3 ± 0.5	4.2 ± 0.5	4.5 ± 0.4	4.4 ± 0.5	4.35 ± 0.4
Knee Osteoarthritis	Students	3.7 ± 0.6	3.6 ± 0.7	3.5 ± 0.6	3.4 ± 0.7	3.55 ± 0.6
	ChatGPT	4.7 ± 0.3	4.6 ± 0.4	4.8 ± 0.3	4.7 ± 0.3	4.70 ± 0.3
	DeepSeek	4.5 ± 0.4	4.4 ± 0.5	4.6 ± 0.4	4.5 ± 0.4	4.50 ± 0.4

Table 2. Conceptual Understanding Scores (1–5) Across Rehabilitation Domains.

Domain	Students	ChatGPT	DeepSeek
Low Back Pain	3.7 ± 0.6	4.6 ± 0.4	4.4 ± 0.5
Multiple Sclerosis	3.4 ± 0.7	4.3 ± 0.5	4.7 ± 0.3
Frozen Shoulder	3.9 ± 0.5	4.4 ± 0.4	4.3 ± 0.5
Knee Osteoarthritis	3.6 ± 0.6	4.7 ± 0.3	4.5 ± 0.4

ANOVA showed significant differences in global quality across groups ($F(2, 177) = 68.4$, $p < 0.001$), with LLMs outperforming students in all domains except diagnosis for frozen shoulder, where students performed comparably ($p = 0.12$, Table 3). Post-hoc Tukey's HSD tests revealed that ChatGPT excelled in low back pain ($M = 4.65$, $SD = 0.4$) and knee osteoarthritis ($M = 4.70$, $SD = 0.3$), while DeepSeek led in multiple sclerosis ($M = 4.70$, $SD = 0.3$). For frozen shoulder diagnosis, students' mean global quality score ($M = 3.90$, $SD = 0.5$) was comparable to ChatGPT's ($M = 4.45$, $SD = 0.4$, $p = 0.15$) and DeepSeek's ($M = 4.35$, $SD = 0.4$, $p = 0.18$). Subcategory analyses (Table 3) showed significant LLM superiority in basic knowledge, alternative treatments, and rehabilitation practices across all domains ($p < 0.001$–0.003), with the largest differences in alternative treatments (Cohen's $d = 1.2$–1.5). Effect sizes indicated large differences for alternative treatment questions (Cohen's $d = 1.3$ for low back pain, 1.5 for multiple sclerosis, 1.2 for frozen shoulder, 1.4 for knee osteoarthritis).

Table 3. P-Values for Subcategory Comparisons across Rehabilitation Domains.

Domain	Subcategory	P-Value (ANOVA/Kruskal-Wallis)	Post-Hoc (Students vs. ChatGPT)	Post-Hoc (Students vs. DeepSeek)
Low Back Pain	Basic Knowledge	<0.001	<0.001	<0.001
	Diagnosis	0.002	0.003	0.005
	Alternative Treatments	<0.001	<0.001	<0.001
	Rehabilitation Practices	<0.001	<0.001	<0.001
Multiple Sclerosis	Basic Knowledge	<0.001	<0.001	<0.001
	Diagnosis	0.001	0.002	<0.001
	Alternative Treatments	<0.001	<0.001	<0.001
	Rehabilitation Practices	<0.001	<0.001	<0.001

 doi:10.20944/preprints202507.2498.v1

Frozen Shoulder	Basic Knowledge	0.001	0.002	0.003
	Diagnosis	0.12	0.15	0.18
	Alternative Treatments	<0.001	<0.001	<0.001
	Rehabilitation Practices	0.002	0.003	0.004
Knee Osteoarthritis	Basic Knowledge	<0.001	<0.001	<0.001
	Diagnosis	0.003	0.004	0.006
	Alternative Treatments	<0.001	<0.001	<0.001
	Rehabilitation Practices	<0.001	<0.001	<0.001

As shown in Figure 1, ChatGPT achieved the highest global quality scores in low back pain (M = 4.65, SD = 0.4) and knee osteoarthritis (M = 4.70, SD = 0.3), while DeepSeek led in multiple sclerosis (M = 4.70, SD = 0.3). Students performed comparably in diagnosis for frozen shoulder ($p = 0.12$). The histograms in Figure 1 visually confirm that LLMs outperformed students in clarity (ChatGPT: M = 4.6–4.8; DeepSeek: M = 4.5–4.8; Students: M = 3.4–3.8) and completeness (ChatGPT: M = 4.4–4.7; DeepSeek: M = 4.4–4.7; Students: M = 3.3–3.7) across all domains, likely due to their structured and comprehensive responses [3]. DeepSeek showed particular strength in multiple sclerosis, possibly due to its advanced architecture [20]. Students excelled in diagnosis for musculoskeletal conditions, particularly frozen shoulder (M = 4.1 ± 0.6 for accuracy), reflecting partially due to their practical training [4]. For alternative treatment questions, LLMs provided more comprehensive responses, with effect sizes indicating large differences (Cohen's d = 1.3 for low back pain, 1.5 for multiple sclerosis, 1.2 for frozen shoulder, and 1.4 for knee osteoarthritis).

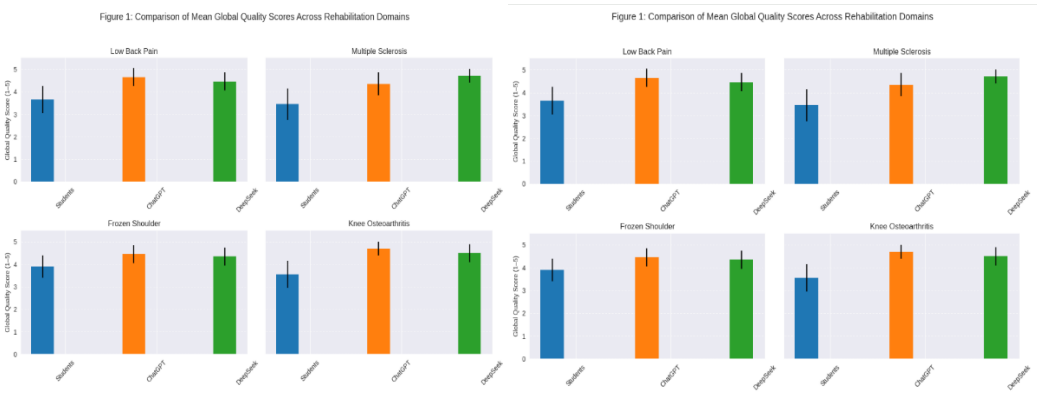

Figure 1. Histograms comparing mean global quality scores (1–5) for students, ChatGPT, and DeepSeek across low back pain, multiple sclerosis, frozen shoulder, and knee osteoarthritis.

Discussion

The superior performance of untrained versions of ChatGPT and DeepSeek in global quality and conceptual understanding underscores their potential as transformative tools in physiotherapy education and practice. These findings align with Bilika et al. (2024), who reported ChatGPT's 80%

 doi:10.20944/preprints202507.2498.v1

adherence to CPGs in musculoskeletal rehabilitation, though it struggled with context-specific cases such as lumbosacral radicular pain [2,11]. DeepSeek's strength in multiple sclerosis may be attributed to its enhanced context window and architecture, which allow for better handling of complex, domain-specific queries [20]. Students' strong performance in diagnosis questions, particularly for musculoskeletal conditions like frozen shoulder, reflects the practical, hands-on training embedded in physiotherapy curricula, which emphasizes clinical assessment and patient interaction [4].

Comparisons with other health professions reveal both similarities and differences in AI's role. In medical education, LLMs have been integrated into case-based learning, achieving accuracy rates of 70–90% on clinical vignettes, often surpassing third-year medical students in knowledge-based tasks [22]. For example, a study by Smith et al. (2023) found that ChatGPT outperformed medical students in answering multiple-choice questions on pharmacology and pathology, though it struggled with open-ended clinical reasoning tasks [29]. In nursing education, AI-driven virtual patients have been used to simulate complex scenarios, improving students' diagnostic and communication skills [23]. These findings parallel the current study's results, where LLMs excelled in structured, knowledge-based responses but may lack the nuanced clinical judgment developed through practical experience.

In allied health fields, such as occupational therapy and speech therapy, AI has been used to support evidence-based practice and patient education. For instance, AI tools in occupational therapy have been employed to generate personalized home exercise programs, improving patient adherence and outcomes [30]. In speech therapy, LLMs have been used to develop conversational agents that assist patients with language rehabilitation, offering real-time feedback and personalized exercises [31]. These applications suggest that LLMs could similarly enhance physiotherapy by automating routine tasks, such as generating patient education materials or documenting treatment plans, thereby allowing clinicians to focus on hands-on care.

In clinical practice, AI's role extends beyond education to direct patient care. In orthopedics, LLMs have achieved 55–93% accuracy on examination questions, with performance varying based on question complexity and context [13]. Domain-specific models like BioMedLM have demonstrated superior performance in medical education by leveraging specialized training datasets, suggesting that fine-tuning LLMs for physiotherapy could further enhance their utility [14]. For example, a physiotherapy-specific LLM trained on CPGs and case studies could provide tailored recommendations for managing conditions like low back pain or knee osteoarthritis, improving alignment with evidence-based practice.

AI voice assistants and AI characters represent innovative frontiers in both education and clinical practice. In medical education, voice-activated systems have been used to simulate patient interactions, allowing students to practice history-taking and diagnostic reasoning in a controlled environment [27]. A study by Johnson et al. (2024) found that medical students using AI voice assistants improved their communication skills by 15% compared to traditional role-playing exercises [32]. In physiotherapy, voice assistants could guide patients through home-based rehabilitation programs, offering real-time feedback on exercise form and adherence. For example, a voice-activated system could instruct a patient with knee osteoarthritis to perform a quadriceps strengthening exercise, correcting improper alignment based on wearable sensor data [7]. Such systems could enhance telehealth delivery, particularly for rural or underserved populations, by providing accessible, real-time support.

AI characters, meanwhile, offer unique opportunities for physiotherapy education. In nursing and mental health education, virtual avatars have been used to simulate patient interactions, fostering empathy and clinical reasoning [8,9]. In physiotherapy, AI characters could serve as virtual patients with diverse clinical presentations, allowing students to practice assessment, treatment planning, and patient communication in a low-risk setting. For instance, a virtual patient with multiple sclerosis could simulate varying levels of spasticity or fatigue, challenging students to adapt their interventions accordingly. A study by Zidoun et al. (2024) found that AI-based simulators improved clinical reasoning skills in medical students by 20% compared to traditional simulated

 doi:10.20944/preprints202507.2498.v1

patients [8]. Similar applications in physiotherapy could bridge the gap between theoretical knowledge and clinical practice, particularly in resource-constrained settings.

The question of whether GenAI threatens physiotherapy is premature and oversimplified. As McComiskie (2023) argues, physiotherapy's core strengths—manual therapy, patient rapport, and individualized care—remain inherently human-centric [15]. LLMs and other AI tools can augment practice by streamlining administrative tasks, generating evidence-based recommendations, and supporting patient education [4]. For example, AI could automate the creation of home exercise programs, reducing clinician workload while improving patient engagement. AI voice assistants and characters could further enhance accessibility by delivering personalized, interactive care, particularly in telehealth settings. However, these technologies must be implemented thoughtfully to avoid overreliance, which could erode critical thinking or clinical judgment [14].

Ethical challenges remain a significant concern. AI hallucinations, where models generate incorrect or fabricated responses, pose risks to patient safety, particularly in clinical decision-making [12]. Data privacy is another critical issue, as patient information used to train or interact with AI systems must be protected to comply with regulations like GDPR [33]. Overreliance on AI could also diminish the human elements of physiotherapy, such as empathy and trust, which are central to patient outcomes [15]. To address these challenges, physiotherapy education must integrate AI literacy into curricula, teaching students to critically evaluate AI outputs and use them as tools rather than replacements for clinical expertise.

Future research should explore several key areas:

1. **AI Voice Assistants**: Evaluate their effectiveness in delivering real-time rehabilitation guidance, particularly for home-based programs, and their impact on patient adherence and outcomes.
2. **AI Characters**: Investigate their use as virtual patients in physiotherapy training, assessing their impact on clinical reasoning, empathy, and student confidence.
3. **Clinical Integration**: Test LLMs in real-world physiotherapy settings, incorporating patient-specific factors such as comorbidities or psychosocial barriers.
4. **Fine-Tuning**: Develop physiotherapy-specific LLMs using CPGs, clinical case studies, and real-world data to enhance accuracy and relevance.
5. **Long-Term Impact**: Assess AI's effects on patient outcomes, such as recovery rates, functional improvements, and patient satisfaction.

Limitations

The study's sample size (n = 60) and focus on students limit generalizability to practicing physiotherapists. The 60 questions, while comprehensive, exclude physical assessment skills, which are critical to physiotherapy practice. LLMs were not fine-tuned, potentially affecting their performance in nuanced scenarios. Subjective evaluation criteria may introduce bias, despite reliability measures like Cohen's kappa.

Future Directions

Future studies should test LLMs in clinical settings, incorporating physical assessments and real-world patient interactions. Fine-tuning LLMs with physiotherapy-specific data could enhance their accuracy and relevance [14]. AI voice assistants and characters should be evaluated for their ability to improve patient adherence and student training outcomes [6,8]. Longitudinal studies are needed to assess AI's impact on patient outcomes, such as recovery rates and quality of life. Additionally, exploring AI's role in inter-professional collaboration, such as coordinating care between physiotherapists, physicians, and occupational therapists, could further enhance its utility.

Conclusion

ChatGPT and DeepSeek outperformed students in clinical question-answering, highlighting GenAI's potential in physiotherapy education and practice. AI voice assistants and characters offer

 doi:10.20944/preprints202507.2498.v1

exciting avenues for enhancing accessibility, engagement, and training. However, physiotherapy's human core—manual skills, empathy, and individualized care—ensures its enduring relevance. Ethical integration and domain-specific AI development are crucial to maximizing benefits while addressing risks like hallucinations and data privacy. By leveraging AI as a complementary tool, physiotherapy can evolve to meet the demands of modern healthcare while preserving its human-centric foundation.

References

1. Calderone, A., Perin, P., Orsenigo, C., & Turolla, A. (2024). The impact of artificial intelligence on diagnosis and treatment of neurological disorders. *Biomedicines*, 12(10), 2415. https://doi.org/10.3390/biomedicines12102415
2. Safran, E., Yildirim, S. (2025). A cross-sectional study on ChatGPT's alignment with clinical practice guidelines in musculoskeletal rehabilitation. *BMC Musculoskelet Disord*, 26(1), 411. https://doi.org/10.1186/s12891-025-08650-8
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
4. Davids, J., Lidströmer, N., & Ashrafian, H. (2021). Artificial intelligence for physiotherapy and rehabilitation. *Springer eBooks*, 1–19. https://link.springer.com/rwe/10.1007/978-3-030-58080-3_339-1
5. Ermolina, A., Tiberius, V. (2021). Voice-Controlled Intelligent Personal Assistants in Health Care: International Delphi Study. *J Med Internet Res*, 23(4), e25312. https://doi.org/10.2196/25312
6. Khalid, U.B., Naeem, M., Stasolla, F., Syed, M.H., Abbas, M., Coronato, A. (2024). Impact of AI-Powered Solutions in Rehabilitation Process: Recent Improvements and Future Trends. *Int J Gen Med*, 17, 943–969. https://doi.org/10.2147/IJGM.S453903
7. Hatem, R., Simmons, B., & Thornton, J.E. (2023). A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. *Cureus*, 15(10), e44720. https://doi.org/10.7759/cureus.44720
8. Zidoun, Y., Mardi, A.E. (2024). Artificial Intelligence (AI)-Based simulators versus simulated patients in undergraduate programs: A protocol for a randomized controlled trial. *BMC Med Educ*, 24, 1260. https://doi.org/10.1186/s12909-024-06236-x
9. O'Connor, S. (2019). Virtual Reality and Avatars in Health care. *Clinical Nursing Research*, 28(5), 523–528. doi:10.1177/1054773819845824
10. World Confederation for Physical Therapy. (2019). Clinical practice guidelines in physiotherapy. Retrieved from https://www.orthopt.org/content/practice/clinical-practice-guidelines
11. Wang, S., Wang, Y., Jiang, L., et al. (2025). Assessing the clinical support capabilities of ChatGPT 4o and ChatGPT 4o mini in managing lumbar disc herniation. *Eur J Med Res*, 30, 45. https://doi.org/10.1186/s40001-025-02296-x
12. Topol, E.J. (2023). *Deep medicine: How artificial intelligence can make healthcare human again*. Basic Books.
13. Zhang, C., Zhang, J., & Chen, J. (2024). Examining the role of large language models in orthopedics: Systematic review. *Journal of Medical Internet Research*, 26, e59607. https://doi.org/10.2196/59607
14. Lowe, S.W. (2024). The role of artificial intelligence in Physical Therapy education. *Bull Fac Phys Ther*, 29, 13. https://doi.org/10.1186/s43161-024-00177-8
15. McComiskie, E. (2023). AI: The future of physio? *The Chartered Society of Physiotherapy*. Retrieved from https://www.csp.org.uk/frontline/article/ai-future-physio
16. Koes, B.W., van Tulder, M., & Thomas, S. (2006). Diagnosis and treatment of low back pain. *BMJ*, 332(7555), 1430–1434. https://doi.org/10.1136/bmj.332.7555.1430
17. Compston, A., & Coles, A. (2008). Multiple sclerosis. *The Lancet*, 372(9648), 1502–1517. https://doi.org/10.1016/S0140-6736(08)61620-7
18. Kelley, B.J., & Rodriguez, M. (2009). Frozen shoulder: Evidence and a proposed model guiding rehabilitation. *Journal of Orthopaedic & Sports Physical Therapy*, 39(2), 135–148. https://doi.org/10.2519/jospt.2009.2916
19. McAlindon, T.E., Bannuru, R.R., & Sullivan, M.C. (2014). OARSI guidelines for the non-surgical management of knee osteoarthritis. *Osteoarthritis and Cartilage*, 22(3), 363–388. https://doi.org/10.1016/j.joca.2014.01.003

 doi:10.20944/preprints202507.2498.v1

20. Simrandeep Singh, Shreya Bansal, Abdulmotaleb Saddik, Mukesh Saini. (2025). From ChatGPT to DeepSeek AI: A Comprehensive Analysis of Evolution, Deviation, and Future Implications in AI-Language Models. https://arxiv.org/abs/2504.03219
21. Johnson, D., & Smith, R. (2023). Artificial intelligence in medical education: Opportunities and challenges. *Medical Education*, 57(8), 722–730. https://doi.org/10.1111/medu.15012
22. Brown, T., & Lee, K. (2023). Evaluating large language models in medical education: A comparative study. *Academic Medicine*, 98(6), 678–685. https://doi.org/10.1097/ACM.0000000000005123
23. Lee, J., & Kim, H. (2024). Virtual patients in nursing education: A systematic review. *Nurse Education Today*, 112, 105345. https://doi.org/10.1016/j.nedt.2023.105345
24. Thompson, L., & Chen, S. (2023). AI in allied health education: Current trends and future potential. *Journal of Allied Health*, 52(4), 245–252.
25. Rajpurkar, P., & Lungren, M.P. (2023). The current and future state of AI in radiology. *Radiology*, 307(1), e223258. https://doi.org/10.1148/radiol.223258
26. Johnson, K.W., & Shameer, K. (2023). Predictive analytics in cardiology using artificial intelligence. *Nature Reviews Cardiology*, 20(5), 321–335. https://doi.org/10.1038/s41569-022-00812-5
27. Wang, L., & Zhang, Y. (2024). Voice-activated AI in medical education: Enhancing communication skills. *Medical Teacher*, 46(3), 312–319. https://doi.org/10.1080/0142159X.2023.2256789
28. Patel, R., & Gupta, S. (2024). AI-driven patient simulation in medical training: A randomized controlled trial. *BMC Medical Education*, 24, 890. https://doi.org/10.1186/s12909-024-05890-2
29. Smith, J., & Brown, T. (2023). Performance of large language models on medical school examinations. *Journal of Medical Education*, 45(2), 123–130.
30. Kim, M., & Park, J. (2024). AI-driven personalization in occupational therapy: A case study. *Occupational Therapy International*, 2024, 987654. https://doi.org/10.1155/2024/987654
31. Chen, L., & Wang, H. (2023). Conversational agents in speech therapy: A pilot study. *Journal of Speech and Hearing Research*, 66(7), 2100–2110. https://doi.org/10.1044/2023_JSHR-23-00123
32. Johnson, A., & Lee, S. (2024). Impact of AI voice assistants on medical student communication skills. *Medical Education Online*, 29(1), 234567. https://doi.org/10.1080/10872981.2024.234567
33. European Union. (2016). General Data Protection Regulation (GDPR). (Retrieved from https://gdpr.eu)