Journal of Medical Systems (2025) 49:73
https://doi.org/10.1007/s10916-025-02207-x

RESEARCH

The Role of Artificial Intelligence Large Language Models in Personalized Rehabilitation Programs for Knee Osteoarthritis: An Observational Study

Ömer Alperen Gürses[1] · Anıl Özüdoğru[1] · Figen Tuncay[2] · Caner Kararti[1]

Received: 23 January 2025 / Accepted: 26 May 2025

Abstract

Background Large language models (LLMs) can contribute to treatment options and outcomes by assisting physiotherapists for conditions like osteoarthritis.

Aims The objective of this early-stage cross-sectional study is to assess the alignment of large language models with physiotherapists in designing physiotherapy and rehabilitation programs for knee osteoarthritis.

Methods Forty patients diagnosed with knee osteoarthritis were assessed using standardized clinical criteria. For each patient, individualized rehabilitation programs were created by three physiotherapists and by ChatGPT-4o and Gemini Advanced using structured prompts. The presence or absence of 50 clinically relevant rehabilitation parameters was recorded for each program. Chi-square tests were used to evaluate agreement rates between the LLMs and the physiotherapist-generated Consensus programs.

Results ChatGPT-4o achieved a 74% agreement rate with the physiotherapists' Consensus programs, while Gemini Advanced achieved 70%. Although both models showed high compatibility with general rehabilitation components, they demonstrated notable limitations in exercise specificity, including frequency, sets, and progression criteria. ChatGPT-4o performed as well as or better than Gemini in most phases, particularly in Phase 3, while Gemini showed lower consistency in balance and stabilization parameters.

Conclusions ChatGPT-4o and Gemini Advanced demonstrate promising potential in generating personalized rehabilitation programs for knee osteoarthritis. While their outputs generally align with expert recommendations, notable gaps remain in clinical reasoning and the provision of detailed exercise parameters. These findings underscore the importance of ongoing model refinement and the necessity of expert supervision for safe and effective clinical integration.

Keywords Artificial intelligence · Large language models · Physiotherapy · Rehabilitation program · Knee osteoarthritis

Introduction

Large language models (LLMs), a branch of artificial intelligence (AI), are advanced systems that leverage deep learning algorithms to process natural language and generate responses with human-like quality and consistency [1]. Notable LLMs such as ChatGPT and Gemini represent leading models in this field [2]. ChatGPT, developed by OpenAI, is based on the GPT architecture (versions 3.5 or 4.0) and functions both as a chatbot and a generative model, trained on multilingual datasets [3]. Gemini, in contrast, offers real-time data integration and is designed to interact with search engines, potentially reshaping information-seeking behavior [2].

✉ Ömer Alperen Gürses
omeralperengurses@gmail.com

Anıl Özüdoğru
aozudogru@hotmail.com

Figen Tuncay
figentuncay3206@hotmail.com

Caner Kararti
fzt.caner.92@gmail.com

[1] School of Physical Therapy and Rehabilitation, Department of Physiotherapy and Rehabilitation, Kırşehir Ahi Evran University, Merkez, Kırşehir 40100, Türkiye

[2] Faculty of Medicine, Department of Physical Medicine and Rehabilitation, Kırşehir Ahi Evran University, Merkez, Kırşehir 40100, Türkiye

Published online: 03 June 2025

The advent of LLMs has precipitated a surge in research endeavors exploring their potential applications in healthcare, clinical practice, and medical research [4, 5]. While conventional AI has exhibited limited involvement in clinical decision-making, LLMs, trained on extensive and diverse human-generated datasets, have catalyzed growing interest in their role in supporting clinical workflows, encompassing triage, diagnosis, and treatment planning [6]. Recent studies have explored the application of LLMs in osteoarthritis management, where models like ChatGPT have demonstrated moderate success in generating rehabilitation programs and aligning with clinical guidelines [7, 8]. Similar research in stroke rehabilitation has shown that LLMs are capable of mimicking clinical reasoning and creating structured treatment plans based on established principles [9]. Furthermore, investigations into conditions such as vestibular disorders, scoliosis, and musculoskeletal diagnostics suggest that these models may aid clinicians in patient communication, education, and treatment planning, particularly when supported by appropriate professional oversight [10–15]. By presenting medical information in a clear and patient-specific manner, LLMs can support clinicians in effectively communicating physiotherapy plans and explaining underlying conditions. This personalised approach may enhance patient understanding, improve adherence to treatment recommendations, and contribute to better health outcomes [16].

Osteoarthritis (OA) is one of the most common musculoskeletal diseases worldwide and has a significant effect on quality of life [17]. OA is a major focus of rehabilitation and multidisciplinary treatment approaches. By supporting both patients and physiotherapists, LLMs can contribute to better understanding, improved treatment strategies, and optimized outcomes for conditions like OA, addressing the needs of a broad patient population [18]. In particular, their potential to delineate fundamental rehabilitation strategies and lucidly expound treatment alternatives could assist physiotherapists in diminishing the time expended on repetitive documentation and preliminary programme design. Moreover, they could serve as a cautionary mechanism for interventions that are frequently overlooked during the planning process. Although studies on LLMs are present, research remains limited in two key areas: first, there are few studies of recent versions that outperform previous versions [7, 18, 19], and second, there are few studies that focus specifically on rehabilitation [9, 16, 20, 21]. This study represents an early-stage evaluation of large language model-based decision support tools by examining their alignment with physiotherapists in the development of personalized rehabilitation programs for knee osteoarthritis, thereby addressing a current gap in the literature.

Materials and Methods

Study Design

This cross-sectional study compared the physiotherapy programs developed by three experienced physiotherapists with at least five years of clinical experience in knee OA and those generated by ChatGPT-4o and Gemini Advanced, two large language models, for knee OA. The study was conducted between August and October 2024 at the physiotherapy outpatient clinic. The study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) and Reporting guideline for the early stage clinical evaluation of decision support systems driven by AI (DECIDE-AI) [22] to ensure high-quality reporting standards.

Participants

The study included 40 patients diagnosed with knee OA based on the American College of Rheumatology criteria by a physiatrist. The patients were aged between 40 and 65 years and diagnosed with grade 2 or 3 knee OA based on the Kellgren-Lawrence classification [23]. Exclusion criteria were previous knee surgery or joint injections within the last 6 months, history of any physiotherapy program, cognitive impairment, systemic diseases, neurological or orthopedic conditions affecting the lower extremities [24].

Measurements

Data on age, sex, body mass index (BMI), and educational status were recorded for all patients. Pain was evaluated using the numeric pain rating scale. The range of motion of the hip and knee joints in all directions was measured with a universal goniometer (Baseline-12-1000 Plastic 360 Degree ISOM), and the strength of the quadriceps and hamstring muscles was assessed using an isometric dynamometer (Lafayette Hand-Held Dynamometer). Functional status was assessed using the WOMAC and Lysholm scores. Physical performance was measured with the Timed Up and Go test, 40-meter fast walk test, 30-second sit-to-stand test, and stair climb test [25, 26]. Balance assessments included the single-leg stance test for static balance and the Four Square Step Test for dynamic balance. The selection of these performance tests aligns with recommendations from the Osteoarthritis Research Society International (OARSI), which endorses their use as standard reliable measures for evaluating functional outcomes in hip and knee OA [26].

Procedure

Two distinct approaches were used to create patient-based assessment and rehabilitation programs for all participants: one based on physiotherapist consensus and the other generated by LLMs.

Consensus physiotherapy programs: The data for each patient was independently evaluated by three physiotherapists, who subsequently developed preliminary programs. Through structured discussion, the physiotherapists were able to reach a Consensus on the most appropriate treatment program for each patient, resulting in a single standardized program for each patient.

AI-generated physiotherapy programs: The data obtained from patients after the evaluation were entered into ChatGPT-4o and Gemini Advanced using prompts written in Turkish. The prompts were originally in Turkish to simulate real-world usage scenarios in Turkey, where the primary audience includes Turkish-speaking physiotherapists and patients. The English translation of the prompts is provided in the main text, while the original Turkish prompts and the prompts related to patient data, in both Turkish and English versions, are included in Supplementary Material 1 to ensure transparency.

The following English translation of the prompt was used:

*"*Prepare a detailed three-phase physiotherapy program for a knee OA patient based on the provided evaluation parameters. The program should include the following components*:

Electrophysical agents: Specify appropriate modalities for each phase.

Thermal applications: Indicate whether hot or cold treatments are preferred based on the patient's needs.

Exercise applications: Provide a detailed exercise program for each phase, including repetitions, sets, and positions.

Phase transition criteria: Define specific criteria for progressing between phases in terms of pain, edema, balance, range of motion (ROM), muscle strength, and functionality.

Discharge criteria: Highlight the goals the patient should achieve by the end of the rehabilitation."*.

This structured prompt format was intentionally designed to guide the language models to generate outputs aligned with real-world physiotherapy program components. Therefore, the models were not producing responses entirely independently, but within a standardized and directive framework that ensured the inclusion of clinically relevant rehabilitation elements. ChatGPT-4o was prompted immediately after each patient's evaluation. A new conversation was initiated for each case to ensure independence of responses. According to OpenAI, this is sufficient to prevent prior context from influencing the model's output, even within the same session [1].

To ensure variety and avoid repetition, new conversations were initiated for each patient when querying ChatGPT-4 and Gemini Advanced. Each of the 40 patient profiles was individually presented to the language models, which were prompted to generate personalized rehabilitation programs tailored to the specific clinical characteristics of each case. For every rehabilitation parameter and phase, the presence or absence of a recommendation was recorded for each patient. These binary data were then aggregated to calculate the percentage of cases in which a given intervention was recommended, allowing for item-level comparison across models and with the Consensus group. Although reported as overall frequencies, the data structure was built on case-specific inputs and individualized AI responses. This approach ensured that each AI-generated rehabilitation plan was based on individualized clinical data, reflecting a tailored treatment structure for every patient.

Parameter selection: Initially, 58 parameters were identified across the rehabilitation programs developed by the physiotherapists and AI models. The final list of 50 parameters was determined based on their clinical relevance and frequency of application in knee OA rehabilitation, as supported by established guidelines and prior literature [27, 28]. Parameters that were deemed less relevant appeared only in one or two patient-specific outputs, or lacked support in evidence-based physiotherapy practices were excluded. This decision ensured that the analysis focused on the most meaningful and representative components of the rehabilitation programs, allowing for comparison grounded in widely accepted clinical standards. A detailed list of excluded parameters is provided in Supplementary Material 2.

Sample Size

The sample size was calculated using G*Power Software (version 3.1.9.7) to ensure sufficient statistical power for detecting differences in parameter-level agreement between AI-generated and Consensus programs. Based on a previous study in a similar field and using a chi-square test for goodness of fit, the required sample size was calculated as 40 patients to achieve 80% power with a 5% significance level [16].

Agreement Evaluation

For each rehabilitation parameter (e.g., type of exercise, modality, dosage), agreement between the AI-generated plans and the Consensus recommendations was assessed based on frequency of usage across 40 standardized patient profiles. Specifically, we calculated how often each intervention item was recommended by the Consensus group and by the AI model, expressed as a percentage of the total patient cases. Agreement was defined as a match in these usage frequencies, regardless of whether the recommendations were made for the same individual patients. Thus, the analysis reflects content-level agreement rather than case-specific alignment. This approach was selected to enable systematic comparison across a large dataset and to evaluate the general consistency of AI-generated outputs with expert-derived protocols.

Statistical Analysis

The rehabilitation programs for each patient were analyzed across the 50 selected parameters. Each parameter was recorded in SPSS Statistics 25 as either “present” or “absent” for the Consensus program, ChatGPT-4o, and Gemini Advanced. Chi-square (χ^2) tests were applied to evaluate the compatibility between AI-generated programs and the physiotherapists’ Consensus programs. Results were analyzed at a significance level of $p < 0.05$ and presented as absolute frequencies and percentages. This parameter-level analysis allowed for the identification of agreement and disagreement rates, providing a detailed evaluation of the alignment between the AI models and physiotherapist-developed programs.

Results

Of the 52 patients referred to the study, 9 did not meet the inclusion criteria and 3 declined to participate. Therefore, 40 patients who met the criteria and provided informed consent were enrolled. (age: 53.3 ± 7.17 years, height: 166 ± 9.05 cm, weight: 67.22 ± 11.7 kg, body mass index: 24.55 ± 4.99, Kellgren-Lawrence: 24 grade 2 and 16 grade 3)

Phase 1: Cold pack and TENS (Transcutaneous electrical nerve stimulation) were recommended across all groups with high levels of agreement (Consensus: 92.5%, ChatGPT4o: 97.5%, Gemini Advanced: 95%). No significant differences were identified for these modalities. Ultrasound and its duration/frequency demonstrated notable discrepancies, being recommended significantly less by Consensus compared to both ChatGPT4o and Gemini Advanced, with both differences being statistically significant ($p < 0.001$). Notable differences were observed in the recommendations for hip mobilization exercises and their set/frequency, with Consensus suggesting them far more often compared to both ChatGPT4o and Gemini Advanced, and both differences being statistically significant ($p < 0.001$). Hip abduction exercises and hamstring curls, along with their set/frequency, showed significant differences, with Consensus recommending them far more frequently than ChatGPT4o ($p < 0.001$) and Gemini Advanced ($p < 0.05$). Criteria for transitioning from Phase 1 to Phase 2 10% increase in quadriceps and hamstring muscle strength were recommended significantly less frequently by both ChatGPT4o and Gemini Advanced compared to Consensus, with both differences being statistically significant ($p < 0.001$). The findings for Phase 1 are summarized in Table 1.

Phase 2: TENS and interferential current, including their duration and frequency, were recommended significantly more frequently by Consensus compared to both ChatGPT4o and Gemini Advanced, with the differences being statistically significant ($p < 0.001$). Balance and proprioception exercises, along with their set and frequency, were similarly recommended more frequently by both ChatGPT4o and Consensus, while Gemini Advanced’s recommendations were notably lower, with statistically significant differences observed ($p < 0.001$). Hip stabilization exercises, including their set and frequency, were recommended far more frequently by Consensus compared to both ChatGPT4o and Gemini Advanced, with all differences reaching statistical significance ($p < 0.001$). Findings for these parameters are summarized in Table 2.

Phase 3: NMES (Neuromuscular electrical stimulation) and RUS current, including their duration and frequency, were recommended significantly more frequently by ChatGPT4o compared to both Consensus and Gemini Advanced, with all differences being statistically significant ($p < 0.001$). Dynamic balance exercises, along with their set and frequency, were recommended at similar levels by ChatGPT4o and Consensus, whereas Gemini Advanced’s recommendations were considerably lower, with statistically significant differences observed ($p < 0.001$). The findings for Phase 3 are summarized in Table 3.

When evaluating the overall performance across the 50 different parameters in the three phases, ChatGPT4o demonstrated discrepancies with Consensus in 13 out of 50 parameters, achieving an agreement rate of 74%. In comparison, Gemini Advanced exhibited discrepancies in 15 parameters, corresponding to an agreement rate of 70%. The average percentage of recommendations for each parameter has been calculated as follows: In Phase 1, 82.5% for Consensus, 75% for ChatGPT4o, and 76.11% for Gemini Advanced; in Phase 2, 90.66% for Consensus, 72.24% for ChatGPT4o, and 62.5% for Gemini Advanced; and in Phase 3, 82.15% for Consensus, 84.42% for ChatGPT4o, and 68.08% for Gemini Advanced. The findings are illustrated in Fig. 1.

Table 1 Phase 1 comparison of consensus and ChatGPT4o-Gemini advanced

Variable (available %)	Consensus	ChatGPT 4o	Gemini Advanced	p1	p2	p3
Coldpack	92.5	97.5	95	0.307	0.646	0.558
Coldpack duration/frequency	92.5	90	87.5	0.694	0.366	0.496
TENS/Interferans current	92.5	97.5	95	0.307	0.646	0.558
TENS/Interferans current duration/frequency	92.5	95	85	0.679	0.355	0.542
Ultrasound	5	52.5	82.5	**<0.001**	**<0.001**	**0.004**
Ultrasound duration/frequency	5	47.5	77.5	**<0.001**	**<0.001**	**0.005**
Knee ROM exercises	97.5	92.5	95	0.307	0.558	0.646
Knee ROM exercises set/frequency	97.5	92.5	90	0.307	0.598	0.694
Quadriceps isometric exercise	95	97.5	90	0.558	0.529	0.168
Quadriceps isometric exercise set/frequency	95	90	92.5	0.398	0.646	0.694
Hip mobilization exercises	85	27.5	20	**<0.001**	**<0.001**	0.288
Hip mobilization exercises set/frequency	85	25	17.5	**<0.001**	**<0.001**	0.471
Hip abduction exercises and hamstring curl	87.5	55	65	**<0.001**	**0.021**	0.364
Hip abduction exercises and hamstring curl set/frequency	87.5	50	62.5	**<0.001**	**0.010**	0.262
Criteria for transition from Phase 1 to Phase 2 (No significant knee swelling)	95	92.5	90	0.679	0.398	0.694
Criteria for transition from Phase 1 to Phase 2 (Pain NRS≤3)	95	97.5	92.5	0.558	0.646	0.307
Criteria for transition from Phase 1 to Phase 2 Knee ROM flexion≥90 extension ≤ -5 degree.	92.5	90	97.5	0.694	0.307	0.168
Criteria for transition from Phase 1 to Phase 2 Improvement in quadriceps and hamstring muscle strength.	92.5	60	35	**<0.001**	**<0.001**	**0.026**

Abbreviations: p1: Consensus-ChatGPT comparison; p2: Consensus-Gemini comparison; p3: ChatGPT-Gemini comparison; TENS: Transcutaneous electrical nerve stimulation; ROM: Range of motion; NRS: Numeric rating scale

Table 2 Phase 2 comparison of consensus and ChatGPT4o-Gemini advanced

Variable (available %)	Consensus	ChatGPT 4o	Gemini Advanced	p1	p2	p3
Hotpack	92.5	87.5	85	0.366	0.355	0.747
Hotpack duration/frequency	92.5	82.5	85	0.179	0.355	0.763
TENS/Interferans current	92.5	20	17.5	**<0.001**	**<0.001**	0.775
TENS/Interferans current duration/frequency	92.5	15	10	**<0.001**	**<0.001**	0.501
NMES/RUS current	87.5	77.5	90	0.242	0.725	0.132
NMES/RUS current duration/frequency	87.5	72.5	82.5	0.095	0.533	0.287
CKC exercises Quadriceps strengthening	97.5	95	92.5	0.558	0.307	0.646
CKC exercises Quadriceps strengthening set/frequency	97.5	92.5	87.5	0.646	0.366	0.458
Balance and proprioception exercises	90	92.5	27.5	0.694	**<0.001**	**<0.001**
Balance and proprioception exercises set/frequency	90	85	22.5	0.501	**<0.001**	**<0.001**
Lower extremity strengthening exercises	97.5	95	92.5	0.558	0.307	0.646
Lower extremity strengthening exercises set/frequency	97.5	87.5	92.5	0.091	0.773	0.458
Hip stabilization exercises	92.5	30	22.5	**<0.001**	**<0.001**	0.448
Hip stabilization exercises set/frequency	92.5	30	20	**<0.001**	**<0.001**	0.543
Hip abduction exercises and hamstring curl	70	67.5	47.5	0.844	0.092	0.137
Hip abduction exercises and hamstring curl set/frequency	70	62.5	42.5	0.562	0.041	0.144
Criteria for transition from Phase 2 to Phase 3 (No significant knee swelling)	95	92.5	80	0.646	0.071	0.158
Criteria for transition from Phase 2 to Phase 3 (Pain at a manageable level)	95	97.5	92.5	0.558	0.287	0.307
Criteria for transition from Phase 2 to Phase 3 Quadriceps and hamstring strength should support progression	92.5	90	97.5	0.694	0.773	0.168

Abbreviations: p1: Consensus-ChatGPT comparison; p2: Consensus-Gemini comparison; p3: ChatGPT-Gemini comparison; CKC: Closed Kinetic Chain; TENS: Transcutaneous electrical nerve stimulation; NMES: Neuromuscular electrical stimulation; ROM: Range of motion; NRS: Numeric rating scale

Table 3 Phase 3 comparison of consensus and ChatGPT4o-Gemini advanced

Variable (available %)	Consensus	Chat-GPT 4o	Gemini Advanced	p1	p2	p3
NMES/RUS current	17.5	65	15	**<0.001**	0.808	**<0.001**
NMES/RUS current duration/frequency	17.5	62.5	12.5	**<0.001**	0.619	**<0.001**
Quadriceps strengthening exercises	97.5	95	92.5	0.558	0.646	0.646
Quadriceps strengthening exercises set/frequency	97.5	92.5	87.5	0.646	0.122	0.366
Lower extremity strengthening exercises	90	80	85	0.286	0.572	0.639
Lower extremity strengthening set/frequency	90	77.5	80	0.132	0.286	0.823
Dynamic balance exercises	90	92.5	27.5	0.694	**<0.001**	**<0.001**
Dynamic balance exercises set/frequency	90	85	22.5	0.572	**<0.001**	**<0.001**
DLA exercises	90	87.5	97.5	0.725	0.168	0.091
DLA exercises set/frequency	90	87.5	85	0.725	0.572	0.747
Discharge criteria (pain-free functional activities)	95	92.5	97.5	0.739	0.739	0.739
Discharge criteria (muscle strength age and gender appropriate level)	95	97.5	90	0.558	0.398	0.168
Discharge criteria (significant improvement in balance and coordination ability)	95	82.5	92.5	0.078	0.646	0.179

Abbreviations: p1: Consensus-ChatGPT comparison; p2: Consensus-Gemini comparison; p3: ChatGPT-Gemini comparison; NMES: Neuromuscular electrical stimulation; DLA: Daily Life Activity

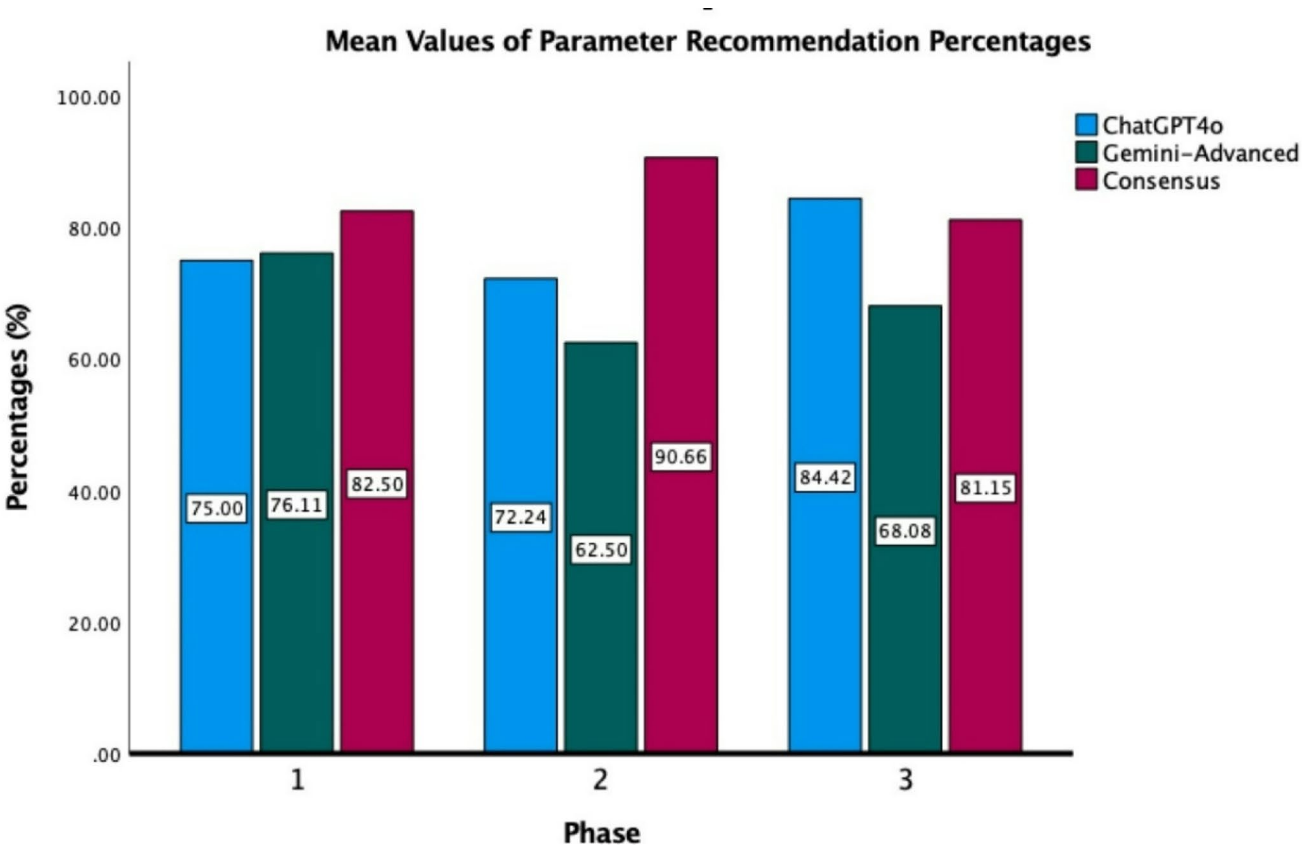

Fig. 1 Mean values of parameter recommendation percentages

Discussion

This study is the first to examine the potential of the large language models ChatGPT-4o and Gemini Advanced to create physiotherapy and rehabilitation programs for knee OA patients. Both models demonstrated strengths in generating general recommendations; however, they exhibited inconsistencies in terms of coverage and adherence to the Consensus. ChatGPT-4o showed greater consistency, while Gemini Advanced showed more variability in certain recommendations. These findings highlight the need for further refinement of such models to improve their reliability in clinical practice.

Treatment Recommendation

Studies on LLMs point out their potential in general knowledge distribution, while revealing limitations in certain clinical applications. In a study examining the potential of LLMs in knee OA treatment management, language models showed limited effectiveness in tasks requiring comprehensive medical knowledge. Their performance declined when moving from general to personalized tasks. In addition, it was determined that their performance can be notably improved through the inclusion of accurate information and explicit instructions [7].

One study showed that ChatGPT-4.0 demonstrates enhanced performance in responding to general OA-related queries, exhibiting higher accuracy and compatibility in providing rehabilitation information compared to other language models [18]. Similarly, in our study, although ChatGPT-4.0 and Gemini Advanced provided significantly fewer recommendations for hip mobilization and stabilization exercises compared to the Consensus, ChatGPT-4.0 demonstrated elevated accuracy and concordance rates in designing rehabilitation programs. This difference may be due to ChatGPT4o's and Gemini Advanced's tendency to focus primarily on the affected area when designing rehabilitation programs and thus their clinical reasoning may not have adequately recognized the importance of hip exercises in meeting broader rehabilitation needs.

For instance, ChatGPT has been shown to effectively design rehabilitation programs based on established frameworks such as FITT-VP (Frequency, Intensity, Time, Type, Volume, and Progression) [15]. In more complex scenarios such as scoliosis classification and treatment planning, ChatGPT-4 achieved high accuracy, whereas Gemini exhibited output inconsistencies that posed potential patient safety concerns. Other studies involving various musculoskeletal conditions—such as shoulder, spine, or vestibular disorders—have reported that while LLMs can provide general information, they often lack specificity and clinical reasoning in generating detailed exercise recommendations [11–14, 29].

In our study, the language models provided significantly fewer recommendations than the consensus, particularly for hip mobilisation, stabilisation, and abduction exercises and related parameters such as sets and frequency. These findings are consistent with the literature showing the limitations of LLMs in generating specific exercise recommendations, despite their optimistic level of accuracy in designing rehabilitation programmes.

A noteworthy finding in our study was the frequent recommendation of ultrasound in Phase 1 by the language models, although it was rarely included in the Consensus recommendations. This inconsistency may be attributed to limitations in the currentness of ChatGPT's and Gemini's training data, as well as the models' weaknesses in reviewing and integrating up-to-date literature [30]. Additionally, ChatGPT demonstrated strength in providing general recommendations, on the other hand it showed deficient consistency in certain exercise parameters (e.g., sets and repetitions). Similarly, while Gemini showed higher compatibility in certain clinical modalities, it struggled with defining adequate transition criteria and outcome measures. These findings underline the potential of language models as complementary tools in clinical practice and highlight the need for their optimization through the integration of more current and evidence-based datasets.

Compatibility with Clinical Consensus

Several studies have investigated the alignment of large language models (LLMs) with evidence-based clinical guidelines for osteoarthritis, revealing noteworthy findings regarding their level of concordance.

One study compared ChatGPT and Bard (Gemini) for concordance with the American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines for hip and knee OA in terms of alignment with clinical questions. It indicated that ChatGPT achieved an 80% concordance rate with AAOS guidelines. Moreover, ChatGPT outperformed Bard, which achieved a 60% concordance rate [8].

Beyond OA, LLMs have shown varying levels of guideline adherence across other clinical areas. In this context, one study evaluated the ability of LLMs to provide treatment recommendations for rotator cuff tears and anterior cruciate ligament injuries in accordance with AAOS clinical practice guidelines, finding that ChatGPT-4 achieved a concordance rate of 79.2%, outperforming other models [31]. Similarly, studies in other specialties—such as gastrointestinal oncology and pediatric orthopedics—have reported varying degrees of guideline alignment for both ChatGPT and Gemini, generally ranging between 67% and 77% [32, 33].

In this context, the results of our study demonstrated that ChatGPT-4.0 and Gemini Advanced exhibited both strengths and weaknesses in aligning with clinical practice consensus, consistent with findings in the current literature. ChatGPT-4.0 showed deficiencies in phase 3 NMES/RUS parameters, especially in terms of duration and frequency, whereas Gemini Advanced showed lower accuracy particularly in phase 2 balance and proprioception exercises, in phase 3 dynamic balance exercises, and in set and frequency parameters. Our findings revealed that ChatGPT-4.0 achieved a 74% concordance rate with Consensus guidelines, with 13 discrepancies out of 50 parameters, while Gemini Advanced achieved a 70% concordance rate with 15 discrepancies. These results confirm ChatGPT's relatively higher adherence to established Consensus. However, the close difference between the two models may be attributed to similar advances in the development of their premium versions. Taken together, both the literature and our findings indicate that ChatGPT-4o and Gemini Advanced demonstrate varying degrees of compatibility with clinical consensus, depending on the context. While both models perform adequately in general clinical applications, they exhibit notable limitations in specific parameters and individualized treatment planning. These findings highlight the need to enhance guideline adherence and improve the quality and currency of language model training datasets by integrating robust, evidence-based frameworks—particularly to strengthen their utility and reliability in physiotherapy and rehabilitation practice.

Clinician Support

Current literature suggests that LLMs like ChatGPT and Gemini hold strong potential in supporting clinical practice. A 2024 study showed that ChatGPT clearly presents treatment protocols and effectively explains surgical risks [34], while another found that both ChatGPT-4 and Gemini 1.5 Pro successfully simplified ultrasound findings with high accuracy and readability [35]. Other condition-specific studies reported varying effectiveness, with ChatGPT-4 performing well for low back pain and scoliosis, whereas Gemini, though faster, showed lower accuracy in some cases [36, 37]. Our findings support existing evidence that ChatGPT-4o and Gemini Advanced, each showing over 70% accuracy, can assist clinicians in developing patient-centered rehabilitation programs for knee osteoarthritis. These models may help with assessment, treatment planning, and patient communication; however, their variable precision underscores the need for clinician oversight to ensure safe and effective use in practice.

The literature demonstrates that the most up-to-date versions of LLMs tend to exhibit superior performance as compared with lower versions [19, 38–40]. In alignment with this, we hypothesize that the high concordance observed in our study with established Consensus may be attributed to the utilization of the most up-to-date and advanced version of the model. However, discrepancies observed in certain cases may be due to limitations of the models in accessing and integrating up-to-date literature [30]. Similar to the literature, the most important limitation identified in our study may be due to the decline in performance of language models when used in languages other than English [41–43].

Limitations and Future Work

The present study has several limitations. The best performance of language models is observed in English; however, in our study, the models were utilized in Turkish. While the treatment programs generated by AI language models raise ethical concerns, their application in clinical practice could provide clearer insights. Furthermore, our study focused solely on patients with knee osteoarthritis, limiting the generalizability of the findings. Additionally, the cross-sectional design of this study does not account for the rapid and continuous evolution of LLMs, particularly in premium versions that are regularly updated. Another limitation is that the stability of LLM-generated responses was not systematically assessed; the same input could potentially produce varying outputs at different times. Future research on the performance of language models in designing rehabilitation programs for other diseases and larger sample groups, in collaboration with physiotherapists, will allow for more comprehensive evaluations in this area. At the same time, the development of a language model specific to physiotherapy and rehabilitation may also be an extraordinary plan for the future.

Conclusion

In conclusion, the present study contributes to the growing evidence for the potential use of ChatGPT-4o and Gemini Advanced in clinical settings, with a specific focus on the design of individualized rehabilitation programs for knee osteoarthritis. These models may support physiotherapists, physicians, and clinicians in developing personalized treatment plans. However, it is evident that there are still limitations in terms of guideline adherence and data accuracy. Further refinements are required, including improvements in language-specific performance and integration of up-to-date evidence, in order to strengthen the role of these models in the development of rehabilitation programs and broader clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-025-02207-x.

Acknowledgments We would like to thank all participants for their volunteered participation.

Author Contributions Omer Alperen Gurses: Writing– review & editing, Writing– original draft, Methodology, Data curation, Conceptualization. Anıl Ozudogru: Writing– review & editing, Methodology, Supervision, Conceptualization. Figen Tuncay: Data curation. Caner Kararti: Methodology, Supervision, Formal analysis

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Ethical Approval The study was conducted in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of Kirsehir Ahi Evran University Faculty of Medicine Health Sciences Scientific Research(2024-13/110). Informed consent was obtained from all participants involved in the study.

Competing Interests The authors declare no competing interests.

Clinical Trial Number Not applicable.

References

1. OpenAI: Introducing ChatGPT. https://openai.com/blog/chatgpt/ (2023). Accessed.
2. Nazir T, Ahmad U, Mal M, Rehman MU, Saeed R, Kalia JS. Microsoft Bing vs. Google Bard in Neurology: A Comparative Study of AI-Generated Patient Education Material. medRxiv. 2023:2023.08. 25.23294641.
3. Ismail AMA. Chat GPT in tailoring individualized lifestyle-modification programs in metabolic syndrome: potentials and difficulties? Ann Biomed Eng. 2023;51(12):2634–5.
4. Goodman RS, Patrinely JR, Osterman T, Wheless L, Johnson DB. On the cusp: Considering the impact of artificial intelligence language models in healthcare. Med. 2023;4(3):139–40.
5. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. JMIR medical informatics. 2022;10(2):e32875.
6. Gaber F, Shaik M, Franke V, Akalin A. Evaluating large language model workflows in clinical decision support: referral, triage, and diagnosis. medRxiv. 2024:2024.09. 27.24314505.
7. Chen X, You M, Wang L, Liu W, Fu Y, Xu J, et al. Evaluating and Enhancing Large Language Models Performance in Domain-specific Medicine: Osteoarthritis Management with DocOA. arXiv preprint arXiv:240112998. 2024.
8. Yang J, Ardavanis KS, Slack KE, Fernando ND, Della Valle CJ, Hernandez NM. Chat Generative Pretrained Transformer (ChatGPT) and Bard: artificial intelligence does not yet provide clinically supported answers for hip and knee osteoarthritis. The Journal of Arthroplasty. 2024;39(5):1184–90.
9. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. Journal of Rehabilitation Medicine. 2023;55.
10. Alhur A. Redefining healthcare with artificial intelligence (AI): the contributions of ChatGPT, Gemini, and Co-pilot. Cureus. 2024;16(4). doi: https://doi.org/10.7759/cureus.57795.
11. 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. medRxiv. 2024:2024.01. 24.24301737. doi: https://doi.org/10.1101/2024.01.24.24301737.
12. Daher M, Koa J, Boufadel P, Singh J, Fares MY, Abboud JA. Breaking barriers: can ChatGPT compete with a shoulder and elbow specialist in diagnosis and management? JSES international. 2023;7(6):2534–41.
13. Fabijan A, Zawadzka-Fabijan A, Fabijan R, Zakrzewski K, Nowosławska E, Polis B. Assessing the Accuracy of Artificial Intelligence Models in Scoliosis Classification and Suggested Therapeutic Approaches. Journal of Clinical Medicine. 2024;13(14):4013.
14. Kim J-h. Search for medical information and treatment options for musculoskeletal disorders through an artificial intelligence chatbot: focusing on shoulder impingement syndrome. MedRxiv. 2022:2022.12. 16.22283512.
15. You M, Chen X, Liu D, Lin Y, Chen G, Li J. ChatGPT-4 and Wearable Device Assisted Intelligent Exercise Therapy for Co-existing Sarcopenia and Osteoarthritis (GAISO): A feasibility study and design for a randomized controlled PROBE non-inferiority trial. Journal of Orthopaedic Surgery and Research. 2024;19(1):635.
16. Bilika P, Stefanouli V, Strimpakos N, Kapreli EV. Clinical reasoning using ChatGPT: Is it beyond credibility for physiotherapists use? Physiotherapy Theory and Practice. 2023:1–20.
17. Vitaloni M, Botto-van Bemden A, Sciortino Contreras RM, Scotton D, Bibas M, Quintero M, et al. Global management of patients with knee osteoarthritis begins with quality of life assessment: a systematic review. BMC musculoskeletal disorders. 2019;20:1–12.
18. Cao M, Wang Q, Zhang X, Lang Z, Qiu J, Yung PS-H, et al. Large language models' performances regarding common patient questions about osteoarthritis: A comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and perplexity. Journal of Sport and Health Science. 2024:101016.
19. Uehara O, Morikawa T, Harada F, Sugiyama N, Matsuki Y, Hiraki D, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National Dental Examination. J Dent Educ. 2025;89(4):459–66.
20. Maggio MG, Tartarisco G, Cardile D, Bonanno M, Bruschetta R, Pignolo L, et al. Exploring ChatGPT's potential in the clinical stream of neurorehabilitation. Frontiers in Artificial Intelligence. 2024;7:1407905.
21. Rossettini G, Cook C, Palese A, Pillastrini P, Turolla A. Pros and cons of using artificial intelligence chatbots for musculoskeletal rehabilitation management. journal of orthopaedic & sports physical therapy. 2023;53(12):728–34.
22. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. bmj. 2022;377.

23. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clinical Orthopaedics and Related Research®. 2016;474:1886–93.
24. Mc H. Guidelines for the medical management of osteoarthritis. Part II. Osteoarthritis of the knee. Arthritis Rheum. 1995;38:1541–6.
25. Collins NJ, Misra D, Felson DT, Crossley KM, Roos EM. Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). Arthritis care & research. 2011;63(S11):S208-S28.
26. Dobson F, Hinman RS, Roos EM, Abbott JH, Stratford P, Davis AM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis and cartilage. 2013;21(8):1042–52.
27. Fransen M, McConnell S, Harmer AR, Van der Esch M, Simic M, Bennell KL. Exercise for osteoarthritis of the knee. Cochrane database of systematic reviews. 2015(1). doi: https://doi.org/10.1002/14651858.CD004376.pub3.
28. McAlindon TE, Bannuru RR, Sullivan M, Arden N, Berenbaum F, Bierma-Zeinstra S, et al. OARSI guidelines for the non-surgical management of knee osteoarthritis. Osteoarthritis and cartilage. 2014;22(3):363–88.
29. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Forte AJ. AI in Hand Surgery: Assessing Large Language Models in the Classification and Management of Hand Injuries. Journal of Clinical Medicine. 2024;13(10):2832.
30. Zimmermann R, Staab M, Nasseri M, Brandtner P. Leveraging large language models for literature review tasks-A case study using ChatGPT. International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability: Springer; 2024. p. 313–23.
31. Nwachukwu BU, Varady NH, Allen AA, Dines JS, Altchek DW, Williams III RJ, et al. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. Arthroscopy: The Journal of Arthroscopic & Related Surgery. 2024.
32. Pandya S, Palmer K, Meyer R, Htway ZM, Fujita M. AI at the Forefront: Navigating Oncologic Care for Six Gastrointestinal Cancers According to the NCCN Guidelines Utilizing Gemini-1.0 Ultra and ChatGPT-4.0 Tamir Bresler, MD, Tyler Wilson, MD, Tadevos Makaryan, MD.
33. Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and Gemini Provide Appropriate Recommendations for Pediatric Orthopaedic Conditions? Journal of Pediatric Orthopaedics. 2024:10.1097.
34. Duran A, Cortuk O, Ok B. Future Perspective of Risk Prediction in Aesthetic Surgery: Is Artificial Intelligence Reliable? Aesthetic Surgery Journal. 2024;44(11):NP839-NP49.
35. Güneş YC, Cesur T, Çamur E. Comparative Analysis of Large Language Models in Simplifying Turkish Ultrasound Reports to Enhance Patient Understanding. European Journal of Therapeutics. 2024;30(5):714–23.
36. Lang S, Vitale J, Galbusera F, Fekete T, Boissiere L, Charles YP, et al. Is the information provided by large language models valid in educating patients about adolescent idiopathic scoliosis? An evaluation of content, clarity, and empathy: the perspective of the European Spine Study Group. Spine Deform. 2025;13(2):361–72.
37. Scaff SP, Reis FJ, Ferreira GE, Jacob MF, Saragiotto BT. Assessing the performance of AI chatbots in answering patients' common questions about low back pain. Annals of the Rheumatic Diseases. 2024.
38. Hirosawa T, Harada Y, Tokumasu K, Ito T, Suzuki T, Shimizu T. Comparative study to evaluate the accuracy of differential diagnosis lists generated by gemini advanced, gemini, and bard for a case report series analysis: cross-sectional study. JMIR Medical Informatics. 2024;12:e63010.
39. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google Gemini and ChatGPT-4o. Clinical Rheumatology. 2024;43(11):3507–13.
40. O'Leary DE. Do ChatGPT 4o, 4, and 3.5 Generate "Similar" Ratings? Findings and Implications. IEEE Intelligent Systems. 2024;39(5):78–81.
41. Fang C, Wu Y, Fu W, Ling J, Wang Y, Liu X, et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. PLOS Digital Health. 2023;2(12):e0000397.
42. Howe NP. ChatGPT has a language problem-but science can fix it. Nature. 2024.
43. Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, Bui T, et al. Chatgpt beyond english: towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv: 2023;230405613.