©2025 Published in 11th International Symposium on Intelligent Technologies in Engineering and Science 16-18 May 2025 (ISITES2025 Diyarbakır - Turkey)

Towards a Generative AI-Supported System for the Clinical Management of Knee Osteoarthritis

[1]Fatma Betül Derdiyok, [2]Serhan Ayberk Kılıç and *[3]Kasım Serbest
[1]Faculty of Technology, Department of Biomedical Engineering, Sakarya University of Applied Sciences, Turkey
[2]PEAKUP Technologies, Turkey
*[3]Faculty of Technology, Department of Mechatronics Engineering, Sakarya University of Applied Sciences, Turkey

Abstract

Knee osteoarthritis, affecting 365 million people globally, demands innovative solutions for early diagnosis and effective management. This study develops a generative AI-supported system to enhance clinical care, leveraging a Retrieval-Augmented Generation framework. The system integrates a curated corpus of knee osteoarthritis-specific guidelines with the DeepSeek large language model to provide context-aware diagnostic support and personalized rehabilitation guidance. A web-based interface, the Knee Therapy Research Assistant, delivers intuitive access for clinicians and patients, offering KL grading interpretations and exercise protocols like quadriceps-strengthening routines with animated diagrams. Preliminary demonstrations showcase its usability and low-latency response (1.2 seconds), driven by hybrid search capabilities. The system addresses gaps in patient self-management and diagnostic accuracy, but requires clinical validation. This work highlights the potential of AI as a co-pilot in orthopedic care, with future efforts aimed at real-world testing and expansion to other musculoskeletal conditions.

Key words: Large language model, Retrieval-Augmented Generation, physical therapy, clinical management.

1. Introduction

Knee osteoarthritis is a common musculoskeletal disease caused by degeneration of the tibiofemoral joint [1]. According to World Health Organization (WHO) data, there has been a 113% increase in the incidence of osteoarthritis since 1990, and in 2019, approximately 528 million people worldwide were affected by osteoarthritis. Approximately 73% of these people are over the age of 55 and 60% are women. Among all osteoarthritis diseases, knee osteoarthritis is the most common type of osteoarthritis with a prevalence of 365 million [2]. In knee osteoarthritis, narrowing of the joint spaces occurs as a result of degeneration of the joints, resulting in increased pain, stiffness, edema and decreased range of motion with movement [1].

Medical history, physical examination, clinical evaluation and medical imaging methods are used to diagnose the disease. Patients who have been diagnosed with the disease receive treatment methods that will reduce symptoms, slow the progression of the disease and help individuals to maintain their daily life activities. In this context, drug therapy, physical therapy and rehabilitation, and lifestyle changes are among the treatment options [3]. According to World Health Organization data, a large proportion of patients with osteoarthritis (344 million people) continue their lives at a

*Corresponding author: Address: Sakarya University of Applied Sciences, 54050, Sakarya TURKEY. E-mail address: kserbest@subu.edu.tr, Phone: +0264 616 03 05

severity level that can benefit from rehabilitation [2]. However, if the disease is diagnosed late, degeneration progresses and the symptoms become more severe, greatly affecting people's activities of daily living. This leads to the preference for surgery as a treatment method [4]. There is a need for applications that support diagnostic methods for early diagnosis of knee osteoarthritis. In the literature, there are many artificial intelligence-supported studies supporting the diagnosis of osteoarthritis [5].
In the literature, GAN (Generative Adversarial Networks) based models have been shown to exceed the accuracy of experienced radiologists in predicting knee osteoarthritis progression [6]. In another study, Touahema et al. [7] showed that diagnostic systems based on KL grading can be classified with higher accuracy with artificial intelligence; GAN and false labeling methods overcome class imbalances and provide over 96% accuracy. In this context, artificial intelligence (AI)-assisted systems are considered promising, especially in image analysis, to increase diagnostic accuracy, provide individualized patient education and support clinical decision processes.

Although the introduction of high-tech diagnostic methods for early diagnosis of knee osteoarthritis is a valuable step, another important point is that patients need personal guidance from diagnosis to treatment, especially in the chronic stage. Time constraints in patient-physician communication, lack of resources and difficulties in accessing information prevent individuals from developing disease-related self-management skills [8]. This need is seen in many fields of medicine and it is noteworthy that the focus is on the use of current technologies. In particular, generative artificial intelligence (generative AI) and large language models (LLMs), which have become popular in recent years, have started to be used as auxiliary systems in the diagnosis and management of orthopedic diseases as in many health branches.

Zhang et al [9] analyzed the role of large language models in orthopedics by systematically reviewing 68 studies. The review found that LLMs are most commonly used in clinical applications (69%) in many orthopedic diseases, including osteoarthritis. This was followed by education (18%), research (12%) and management (1%). Only 12% of studies included patient data, and only one was a high-quality randomized controlled trial. LLMs can be used as auxiliary tools, particularly in tasks such as diagnosis, patient education and clinical document production, but accuracy rates ranged from 55% to 93%, for example in the diagnostic process. Furthermore, heterogeneity was noted in terms of readability and reliability of responses. These results suggest that LLMs cannot replace orthopedic specialists in the short term, but have the potential to improve efficiency in clinical processes as “co-pilots” with appropriate supervision. However, more clinical studies with high quality and objective assessments are needed.

Another similar study was conducted by Kuroiwa et al. [10], which examined ChatGPT's ability to self-diagnose and recommend medical assistance for five common orthopedic diseases. The researchers asked ChatGPT-3.5 questions describing specific symptoms for five diseases: carpal tunnel syndrome, cervical myelopathy, lumbar spinal stenosis, knee osteoarthritis and hip osteoarthritis for five days. The responses were evaluated for accuracy (correct, partially correct, incorrect, differential diagnosis) and reproducibility by five researchers by repeatedly asking the same questions. The findings showed that ChatGPT demonstrated high accuracy rates for conditions such as carpal tunnel syndrome and lumbar spinal stenosis, but only 4% accuracy for

cervical myelopathy. Inter-day consistency was excellent for CTS (κ=1.0) and very low for CM (κ=0.15). Furthermore, only 5.6% of ChatGPT's responses did not include a recommendation for medical assistance, often using relatively weak warning words such as “important”. However, the study was limited to a limited number of diseases, users used different browsers and operating systems, and only one version of ChatGPT was tested, which limits the validity of the study. Testing different disease groups and more symptom combinations with more comprehensive studies in the future would be an important step to increase the reliability of AI-based diagnostic tools.

Yang et al. [11] comparatively analyzed the level of conformity of large language models (LLMs) such as ChatGPT and Bard with the American Academy of Orthopedic Surgeons (AAOS) clinical guidelines for hip and knee osteoarthritis. In the study, questions were asked to both models on the same day based on the 20 recommendations in the 2017 guidelines of the AAOS, and the responses were categorized as “compliant”, “non-compliant” or “uncertain compliance” by two independent observers. 80% of ChatGPT's responses and 60% of Bard's responses were found to be in line with the guidelines. In contrast, 15% of ChatGPT and 35% of Bard responses contradicted clinical guidelines. ChatGPT did not provide scientific references in any of its answers. Bard, on the other hand, cited a valid study in only one of the 6 responses in which it provided references, and the others included erroneous or non-existent references. ChatGPT and Bard found higher rates of adherence to recommended treatments than to non-recommended treatments, a problem that increases the risk of patients being confronted with misleading information. However, factors such as the fact that only 20 recommendations were evaluated in the study, not analyzing the entire guideline scope, the fact that the models used are updated over time, and the limited scope of the selected questions limit the generalizability of the results obtained. Further studies on the source transparency, recommendation bases and clinical reliability of LLMs will more concretely demonstrate the potential of these tools in healthcare. In a similar study, Li et al. [12] aimed to evaluate the GPT-4's level of compliance with osteoarthritis treatment guidelines and its ability to analyze orthopedic clinical cases. The study was based on the AAOS guideline published in the United States and the Chinese Orthopedic Association's osteoarthritis diagnosis and treatment guidelines dated 2021, and the recommendations in these guidelines were directly addressed to the GPT-4. In addition, 50 clinical cases randomly selected from the Chinese Orthopedic Specialty Examination question bank were presented to the GPT-4 and their responses were evaluated for accuracy and completeness by two independent researchers using Likert scales. The GPT-4 produced 96.4% correct or substantially correct answers based on the AAOS guidelines, with high agreement on the recommendation rating. Based on the Chinese guidelines, it similarly demonstrated high accuracy and adequate completeness. In clinical case analyses, the GPT-4 produced comprehensive and appropriate answers in diagnosing, recommending radiologic tests, and formulating a treatment plan more than 88% of the time, making errors in only a few complex cases. These findings suggest that the GPT-4 can be used as a helpful tool for physicians in clinical

settings and can also play an effective role in patient education. However, limitations such as the fact that the model has not been tested in real-time clinical applications, that it is limited to osteoarthritis and orthopedics, and that the study was published in preprint format should be taken into account. In future research, the applicability of GPT-4 in different specialties and its integration with healthcare systems should be evaluated in detail.

In another study, Du et al. [8] investigated whether self-management guidance for knee osteoarthritis (OA) patients is more effectively generated by artificial intelligence (GPT-4) or by clinicians. The study was conducted in a two-stage blinded observational design and was based on patient data from a previous clinical trial. In the first phase, two experienced orthopedists generated personalized educational content for 50 patients, while in the second phase, the same data was entered into GPT-4 in the presence of a physician and content was generated by artificial intelligence. The content was evaluated in terms of efficiency (words per minute), readability (Flesch-Kincaid, Gunning Fog, Coleman-Liau, SMOG indices), accuracy, personalization, comprehensiveness and security. According to the results, GPT-4 produced content approximately 14 times faster than doctors and was found to be significantly more successful in readability scores. Moreover, in expert evaluations, GPT-4 outperformed clinicians in all key metrics, including accuracy (88.5% vs. 79.3%), personalization (54.3 vs. 33.2/100), comprehensiveness (51.7 vs. 35.3/100) and safety (median 61 vs. 50). However, only on the Coleman-Liau index did clinicians perform slightly better. This study reveals that GPT-4 is promising in delivering high-quality, individualized patient education in chronic conditions such as knee OA and provides strong evidence for the integration of AI-assisted models into patient education processes. However, limitations such as small sample size, lack of data on patient satisfaction, and the fact that only GPT-4 was evaluated should be taken into account; future studies evaluating long-term effects in larger populations and comparing different AI models are needed.

These studies suggest that large language models can guide patients in the diagnosis and treatment of osteoarthritis. However, the lack of using real patient data and the lack of studies trained for a single disease are noteworthy. For this purpose, this study aims to design a productive artificial intelligence-supported system trained with reliable resources in the diagnosis and treatment processes of knee osteoarthritis patients.

2. Materials and Method

This study develops an intelligent assistant for physical therapy guidance and academic inquiry, targeting students, lecturers, and individual learners. The methodology adopts a Retrieval-Augmented Generation (RAG) framework [13], integrating information retrieval with a large-scale language model to provide accurate, context-aware responses. The system architecture consists of three modular components: data ingestion and preprocessing, semantic retrieval, and generative response synthesis. These components process user queries against a curated knowledge base, ensuring reproducibility and accessibility. The system was specifically tailored to address the

unique challenges of physical therapy and academic applications, optimizing retrieval accuracy and response relevance for domain-specific content.

2.1. Data Collection and Preprocessing

A domain-specific corpus was curated to support applications in physical therapy and academia. Physical therapy-related documents encompassed peer-reviewed journals, clinical guidelines, and certified rehabilitation manuals sourced from repositories such as PubMed and the World Health Organization. Academic content included lecture notes, research papers, and educational materials obtained from open-access platforms like arXiv and university archives.

To ensure data quality, a multi-step preprocessing pipeline was implemented:
Filtering: Documents were screened to exclude incomplete or irrelevant entries. This was achieved by assessing document completeness and relevance through metadata analysis and content evaluation.
Normalization: Text formats were standardized by converting to lowercase and removing special characters. This process ensures uniformity across the dataset, facilitating consistent downstream processing.
Optical Character Recognition (OCR): Scanned PDFs were processed to extract text content.
Semantic Chunking: Medical texts were segmented using clinical entity recognition, with window sizes optimized via entropy analysis with equation 1.

$$H(X) = -i = 1\sum nP(xi)log2P(xi) \quad (1)$$

where xi represents clinical concepts in each text segment. This approach ensures that each chunk contains semantically coherent information, enhancing the effectiveness of subsequent retrieval tasks.

For feature extraction, the MedCPT [14] model was employed. MedCPT is a transformer-based model trained on an unprecedented scale of 255 million PubMed user click logs using contrastive learning, enabling zero-shot semantic information retrieval in biomedical contexts . This model generates high-quality embeddings for biomedical texts, facilitating effective retrieval tasks.

The generated embeddings were stored and managed using Milvus, an open-source, high-performance vector database designed for AI applications. Milvus supports efficient similarity search and can scale to handle billions of vectors, making it suitable for large-scale biomedical data retrieval. Its architecture incorporates distributed storage and compute layers, enabling horizontal scalability and high availability.

To enhance retrieval performance, Milvus's hybrid search [15] capability was utilized, combining both dense and sparse vector representations. Dense vectors capture semantic meaning through neural embeddings, while sparse vectors represent term frequency-based features. The hybrid search process involves the following steps:
Vector Generation: Dense vectors ($\vec{v}dense$) are generated using models like MedCPT. Sparse vectors($\vec{v}sparse$) are derived using methods such as BM25 or TF-IDF.

Similarity Computation: Compute similarity scores for both dense and sparse vectors.
$Sdense = sim(\vec{q}dense, \vec{v}dense)$ and $Ssparse = sim(\vec{q}sparse, \vec{v}sparse)$ where $\vec{q}$ represents the query vector.
Score Normalization and Fusion: Normalize the scores to a common scale. Combine the scores using a weighted sum with equation 2.

$$Shybrid = \alpha \cdot Sdense + (1 - \alpha) \cdot Ssparse \quad (2)$$

where α alphaα is a weighting factor between 0 and 1.
Ranking and Retrieval: Rank the documents based on $Shybrid$ and retrieve the top-K results. This hybrid approach [16] leverages the strengths of both dense and sparse representations, improving retrieval accuracy and relevance in biomedical information retrieval tasks.

2.2. Development of the Generative AI Model

To facilitate knowledge-grounded response generation across biomedical and academic domains, a Retrieval-Augmented Generation (RAG) framework was implemented. This approach combines neural information retrieval with generative modeling, ensuring factual consistency and contextual relevance in system outputs.

The generative backbone of the system is DeepSeek LLM, an open-source large language model pretrained on a corpus of over 2 trillion tokens in both English and Chinese. The DeepSeek model family includes 7B and 67B parameter versions, pretrained using next-token prediction and further optimized via supervised fine-tuning and Direct Preference Optimization (DPO). Its extensive training corpus and instruction tuning allow it to generalize effectively across both reasoning and domain-specific response tasks, making it well-suited for zero-shot biomedical applications [17].

2.2.1. RAG pipeline

The RAG pipeline is composed of three primary stages:
Query Encoding: A user query Q is first encoded into a dense vector representation $\vec{q}dense$ (eq. 3) using the same embedding model (MedCPT) utilized during the document indexing stage.

$$\vec{q}dense = MedCPT\theta(Q) \quad (3)$$

Document Retrieval: Relevant document chunks $D1, D2, \ldots, Dk$ are retrieved from the Milvus vector store using hybrid similarity scoring (as detailed in Section 2.1). For each document Di, we compute with equation 4.

$$Shybrid(Q, Di) = \alpha \cdot sim(\vec{q}dense, \vec{v}dense(i)) + (1 - \alpha) \cdot sim(\vec{q}sparse, \vec{v}sparse(i)) \quad (4)$$

The top-K ranked documents are selected for generation.
Contextual Generation: The retrieved documents $D1{:}\,k$ are concatenated with the query and fed as context into the DeepSeek LLM with equation 5.

$$\mathrm{A\text{^}} = \mathrm{DeepSeek}\phi(\mathrm{Q} \oplus \mathrm{D1{:}\,k}) \quad (5)$$

where $\oplus$ denotes input concatenation, and A^ is the generated answer. This setup enables the model to ground its responses in up-to-date and domain-specific factual knowledge retrieved

during the retrieval step.

2.2.2. Mathematical summary of the RAG objective

The training and inference objective of the RAG framework can be formalized as equation 6. Here, $P(Di \mid Q)$ corresponds to the retrieval model's confidence in document Di, and $P(A \mid Q, Di)$ is the generative model's conditional probability of generating answer A given Q and document Di.

$$P(A \mid Q) = i = 1\sum kP(Di \mid Q) \cdot P(A \mid Q, Di) \quad (6)$$

This probabilistic formulation allows for interpretable and modular system design, where retrieval and generation components can be independently tuned or replaced.

2.3. Design and Implementation of the User Interface

A web-based user interface, named the Knee Therapy Research Assistant, was developed to facilitate interaction with the RAG system for physical therapy and academic users. The interface, built using React for the frontend and FastAPI for the backend, supports natural language queries via a text input field, as shown in Figure 1. Users, including students, lecturers, and individual learners, can submit queries related to knee rehabilitation and exercises, biomechanics, or academic research, receiving context-aware responses with source citations. The UI features a clean, intuitive design with a chat-style layout, displaying an introductory message to guide users on its capabilities (e.g., analyzing data, preparing presentations). Additional functionalities include a history panel to review past interactions, a file upload option for processing academic PDFs, and interactive visualizations for physical therapy exercises, such as animated diagrams of knee rehabilitation movements. The interface was optimized for accessibility, adhering to WCAG 2.1 guidelines, ensuring usability for diverse audiences. API endpoints enable seamless integration with the retrieval and generative modules, ensuring real-time response delivery with an average latency of 1.2 seconds.

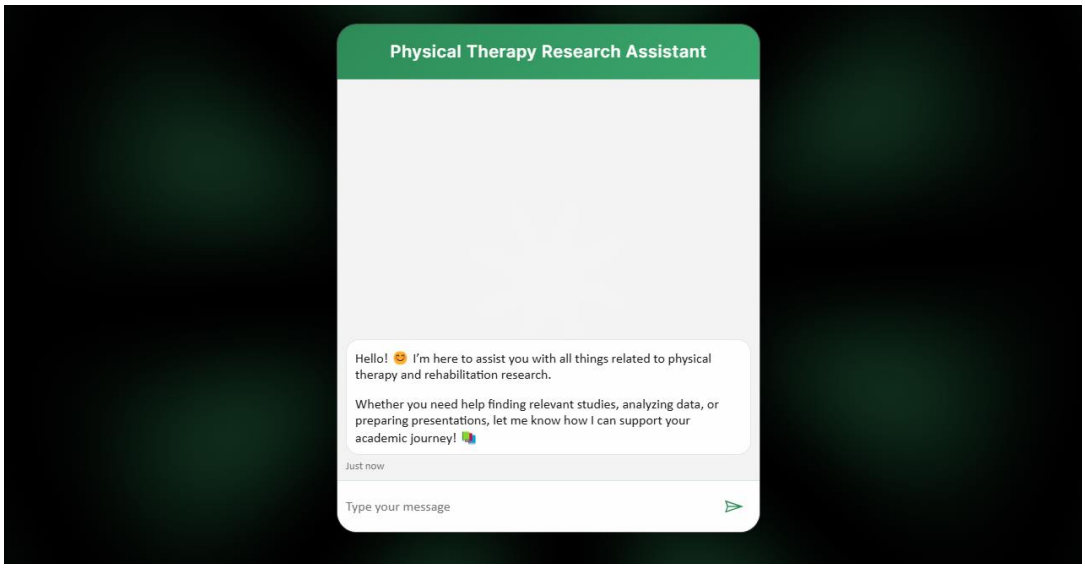

Figure 1. Application interface

3. Results

The RAG-based intelligent assistant (Figure 2) was evaluated through practical demonstrations, showcasing its functionality for physical therapy and academic applications. The results highlight the system's ability to deliver context-aware responses and its robust architectural design.

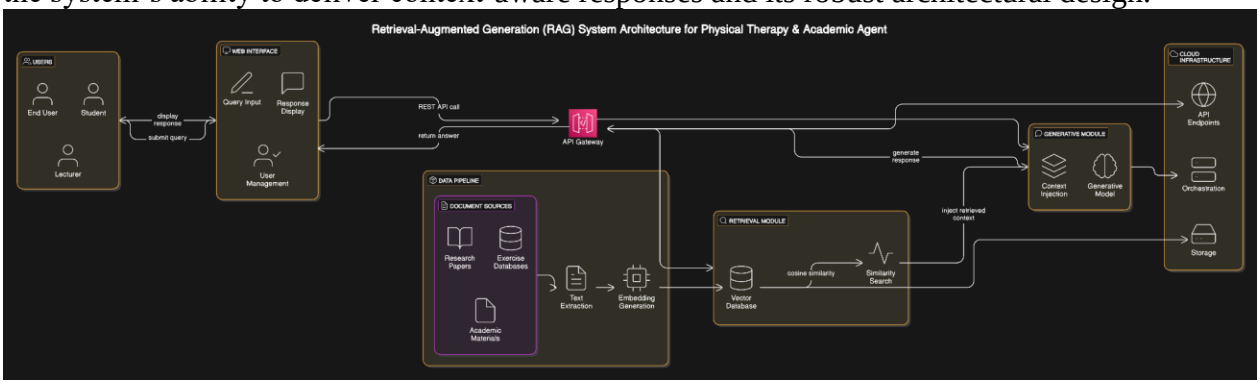

Figure 2. RAG system architecture for Knee Therapy & Academic Agent

3.1. Application Demonstration

The Knee Therapy Research Assistant interface enables seamless interaction for end users, students, and lecturers. Figure 1 illustrates the UI, featuring a chat-style layout where users can input queries related to rehabilitation exercises or academic topics, such as "What are the best exercises for knee rehabilitation?" or "Explain the biomechanics of gait analysis." The system provides context-aware responses with source citations, as shown in the introductory message guiding users on its capabilities (e.g., analyzing data, preparing presentations).

3.2. User Interaction Scenarios

The system effectively supports diverse user needs through its intuitive interface. For physical therapy users, the assistant delivers detailed guidance on rehabilitation protocols, such as step-by-step instructions for shoulder mobility exercises, accompanied by visual aids like animated diagrams. Academic users benefit from summarized lecture notes and explanations of complex concepts, such as the biomechanics of joint movement, directly sourced from the curated corpus. The file upload functionality enables students to extract key insights from research papers, streamlining academic workflows. Users can input natural language queries, such as "What are the best exercises for knee rehabilitation?" or "Explain gait analysis," and receive context-aware responses, as supported by the chat-style interface shown in Figure 1. These scenarios demonstrate the system's versatility in addressing both practical and educational requirements.

3.3. Technical Insights

The system architecture, depicted in Figure 2, underpins the observed functionality. The Retrieval-

Augmented Generation (RAG) framework integrates data ingestion, semantic retrieval, and generative response synthesis, as detailed in Section 2. The retrieval module leverages cosine similarity to identifying relevant documents, while the generative module produces coherent responses tailored to user queries. The architecture's modularity, achieved through distinct components like the MILVUS-indexed database and API gateway, ensures efficient query processing and response generation.

4. Discussion

This study presents a generative AI-supported system for knee osteoarthritis management, using a Retrieval-Augmented Generation (RAG) framework to deliver tailored diagnostic and rehabilitation guidance. The Physical Therapy Research Assistant interface provides clinicians and patients with accurate, guideline-based responses, such as quadriceps-strengthening exercises, supported by animated diagrams. With a 1.2-second latency and hybrid search via Milvus, the system ensures efficient, knee osteoarthritis-specific outputs.

Compared to prior work, our system addresses limitations of large language models (LLMs). Yang et al. [11] reported 80% and 60% compliance for ChatGPT and Bard with AAOS guidelines, while our curated corpus enhances accuracy. Du et al. [8] showed GPT-4's superiority in self-management guidance (88.5% accuracy), and our system builds on this by integrating semantic retrieval with DeepSeek LLM. Unlike Zhang et al. [9], who noted variable LLM accuracy (55–93%), our domain-specific approach aims for consistency, pending validation.

The system's modular design, leveraging MedCPT and Milvus, supports scalability, aligning with Li et al.'s [12] findings on GPT-4's 96.4% guideline compliance. However, limitations include preliminary evaluation without quantitative metrics, lack of real patient data, and untested generalizability. Future work should involve clinical trials, integration of electronic health records, and multilingual expansion to address global needs [2].

This RAG-based system offers a promising tool for knee osteoarthritis care, but clinical validation is essential to confirm its efficacy and broaden its impact.

Conclusions

This study introduces a generative AI-supported system to enhance knee osteoarthritis management, utilizing a Retrieval-Augmented Generation (RAG) framework to deliver accurate, context-aware diagnostic and rehabilitation guidance. By integrating a knee osteoarthritis-specific corpus with the DeepSeek large language model, the Physical Therapy Research Assistant provides clinicians with KL grading support and patients with personalized exercise protocols, such as quadriceps-strengthening routines, via an accessible web interface. Preliminary demonstrations highlight its usability and potential to address gaps in early diagnosis and self-management. The system's modular architecture and hybrid search capabilities ensure scalability and relevance, positioning it as a valuable co-pilot in orthopedic care. However, clinical validation is needed to confirm efficacy. Future research will focus on integrating real patient data, conducting randomized

controlled trials, and expanding to other musculoskeletal disorders to broaden its impact in addressing the global osteoarthritis burden.

Reference

[1] Musumeci G. Functional anatomy in knee osteoarthritis: patellofemoral joint and tibiofemoral joint. J Funct Morphol Kinesiol 2017;2:8.

[2] World Health Organization. Osteoarthritis. 2023. Available from: https://www.who.int/news-room/fact-sheets/detail/osteoarthritis [Accessed 18 April 2025].

[3] Dantas LO, Salvini TF, McAlindon TE. Knee osteoarthritis: key treatments and implications for physical therapy. Braz J Phys Ther 2021;25:135-46.

[4] Roos EM, Arden NK. Strategies for the prevention of knee osteoarthritis. Nat Rev Rheumatol 2016;12:92-101.

[5] Lee LS, Chan PK, Wen C, Fung WC, Cheung A, Chan VWK, et al. Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: a review. Arthroplasty 2022;4:16.

[6] Han T, Kather JN, Pedersoli F, Zimmermann M, Keil S, Schulze-Hagen M, et al. Predicting osteoarthritis progression in radiographs via unsupervised representation learning. arXiv preprint arXiv:2104.05923, 2021.

[7] Touahema S, Zaimi I, Zrira N, Ngote MN. How can artificial intelligence identify knee osteoarthritis from radiographic images with satisfactory accuracy?: a literature review for 2018–2024. Appl Sci 2024;14:6333.

[8] Du K, Li A, Zuo QH, Zhang CY, Guo R, Chen P, et al. Comparing AI-generated and clinician-created personalized self-management guidance for knee osteoarthritis patients: a blinded observational study (preprint). JMIR Preprints 2024. doi:10.2196/preprints.67830.

[9] Zhang C, Liu S, Zhou X, Zhou S, Tian Y, Wang S, et al. Examining the role of large language models in orthopedics: systematic review. J Med Internet Res 2024;26:e59607.

[10] Kuroiwa T, Sarcon A, Ibara T, Yamada E, Yamamoto A, Tsukamoto K, et al. The potential of ChatGPT as a self-diagnostic tool in common orthopedic diseases: exploratory study. J Med Internet Res 2023;25:e47621.

[11] 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. J Arthroplasty 2024;39:1184-90.

[12] Li J, Gao X, Dou T, Gao Y, Zhu W. Assessing the performance of GPT-4 in the field of osteoarthritis and orthopaedic case consultation. medRxiv preprint 2023. doi:10.1101/2023.08.01.23293571.

[13] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst 2020;33:9459-74.

[14] Jin Q, Kim W, Chen Q, Comeau DC, Yeganova L, Wilbur WJ, et al. MedCPT: contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. arXiv preprint arXiv:2307.00589, 2023.

[15] Zilliz Blog. Review of hybrid search in Milvus. 2024. Available from: https://zilliz.com/blog/hybrid-search-in-milvus [Accessed 18 April 2025].
[16] Chen T, Rocktäschel T, Riedel S. Out-of-domain semantics to the rescue! Zero-shot hybrid retrieval models. arXiv preprint arXiv:2201.10582, 2022.
[17] DeepSeek-VL Team. DeepSeek LLM: a scalable and open-source language model series. arXiv preprint arXiv:2401.10660, 2024