

Evaluating and Enhancing Large Language Models' Performance in Domain-specific Medicine: Explainable LLM with DocOA

Xi Chen, Li Wang, MingKe You, WeiZhi Liu, Yu Fu, Jie Xu, Shaiting Zhang, Gang Chen, Kang Li, Jian Li

Submitted to: Journal of Medical Internet Research
on: March 07, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	4
Supplementary Files.....	17
0.....	17
0.....	17
0.....	17
0.....	17
0.....	17
Figures	18
Figure 1.....	19
Figure 2.....	20
Figure 3.....	21
Figure 4.....	22
Figure 5.....	23
Figure 6.....	24
Related publication(s) - for reviewers eyes onlies	25
Related publication(s) - for reviewers eyes only 0.....	26
Related publication(s) - for reviewers eyes only 0.....	26
Related publication(s) - for reviewers eyes only 0.....	26

Evaluating and Enhancing Large Language Models' Performance in Domain-specific Medicine: Explainable LLM with DocOA

Xi Chen^{1*}; Li Wang^{1*}; MingKe You¹; WeiZhi Liu²; Yu Fu³; Jie Xu⁴; Shaiting Zhang⁴; Gang Chen¹; Kang Li⁵; Jian Li¹

¹Sports Medicine Center, West China Hospital, West Chian School of Medicine, Sichuan University, Chengdu, Sichuan, China Chengdu CN

²Sports Medicine Center, West China Hospital, West Chian School of Medicine, Sichuan University, Chengdu, Sichuan, China chengdu CN

³West China Hospital, West Chian School of Medicine, Sichuan University, Chengdu, Sichuan, China Chengdu CN

⁴Shanghai Artificial Intelligence Laboratory, OpenMedLab, Shanghai, 200030, China. Chengdu CN

⁵Shanghai Artificial Intelligence Laboratory, OpenMedLab, Shanghai, 200030, China Chengdu CN

*these authors contributed equally

Corresponding Author:

Jian Li

Sports Medicine Center, West China Hospital, West Chian School of Medicine, Sichuan University, Chengdu, Sichuan, China

Sports Medicine Center, West China Hospital, West Chian School of Medicine, Sichuan University, Chengdu, Sichuan, China

Chengdu

CN

Abstract

Background: The efficacy of large language models (LLMs) in domain-specific medicine, particularly for managing complex diseases such as osteoarthritis (OA), remains largely unexplored.

Objective: This study focused on evaluating and enhancing the clinical capabilities and explainability of LLMs in specific domains, using osteoarthritis (OA) management as a case study.

Methods: A domain specific benchmark framework was developed, which evaluate LLMs across a spectrum from domain-specific knowledge to clinical applications in real-world clinical scenarios. DocOA, a specialized LLM designed for OA management integrating retrieval-augmented generation (RAG) and instructional prompts, was developed. It can identify the clinical evidence upon which its answers are based through RAG, thereby demonstrating the explainability of those answers. The study compared the performance of GPT-3.5, GPT-4, and a specialized assistant, DocOA, using objective and human evaluations.

Results: Results showed that general LLMs like GPT-3.5 and GPT-4 were less effective in the specialized domain of OA management, particularly in providing personalized treatment recommendations. However, DocOA showed significant improvements.

Conclusions: This study introduces a novel benchmark framework which assesses the domain-specific abilities of LLMs in multiple aspects, highlights the limitations of generalized LLMs in clinical contexts, and demonstrates the potential of tailored approaches for developing domain-specific medical LLMs.

(JMIR Preprints 07/03/2024:58158)

DOI: <https://doi.org/10.2196/preprints.58158>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [a JMIR journal](#), my title and abstract will remain visible to all users.

Original Manuscript

Evaluating and Enhancing Large Language Models' Performance in Domain-specific Medicine: Explainable LLM with DocOA

Preprint
JMIR Publications

Abstract**Background**

The efficacy of large language models (LLMs) in domain-specific medicine, particularly for managing complex diseases such as osteoarthritis (OA), remains largely unexplored.

Objective

This study focused on evaluating and enhancing the clinical capabilities and explainability of LLMs in specific domains, using osteoarthritis (OA) management as a case study.

Methods

A domain specific benchmark framework was developed, which evaluate LLMs across a spectrum from domain-specific knowledge to clinical applications in real-world clinical scenarios. DocOA, a specialized LLM designed for OA management integrating retrieval-augmented generation (RAG) and instructional prompts, was developed. It can identify the clinical evidence upon which its answers are based through RAG, thereby demonstrating the explainability of those answers. The study compared the performance of GPT-3.5, GPT-4, and a specialized assistant, DocOA, using objective and human evaluations.

Results

Results showed that general LLMs like GPT-3.5 and GPT-4 were less effective in the specialized domain of OA management, particularly in providing personalized treatment recommendations. However, DocOA showed significant improvements.

Conclusion

This study introduces a novel benchmark framework which assesses the domain-specific abilities of LLMs in multiple aspects, highlights the limitations of generalized LLMs in clinical contexts, and demonstrates the potential of tailored approaches for developing domain-specific medical LLMs.

Key words: Large language model, Retrieval augmented generation, Domain-specific benchmark framework, Osteoarthritis management

Introduction

The rapid development of large language models (LLMs) has shown promising potential in the medical field, as demonstrated by their ability to pass the United States Medical Licensing Examination (USMLE) and diagnose clinical conditions¹⁻³. The promising performance of LLMs in the general medical field warrants further research and exploration of their clinical performance in domain-specific medical scenarios^{4,5}.

Osteoarthritis (OA) is one of the most prevalent and debilitating diseases that causes pain, disability, and loss of function⁶. The global prevalence of OA is approximately 7.6% (595 million people), as of 2020⁷. The management of OA requires complex strategies that encompass a variety of pharmacological treatments, lifestyle alterations, rehabilitation, and surgical interventions across multiple disciplines. Effective management of this condition necessitates the integration of extensive evidence-based medical data and the consideration of individual circumstances⁶.

Although some LLMs have achieved commendable results in general medical question-answering (QA) tasks, substantial limitations persist in their clinical capability, particularly in complex and multifaceted diseases such as OA⁸. However, the datasets used to train LLMs are predominantly composed of general medical knowledge and lack in-depth, domain-specific content. Existing research indicates that current training data and benchmarking methodologies may be inadequate for LLMs to acquire the necessary domain-specific knowledge and clinical capabilities⁴.

Additionally, LLMs may lack the ability to translate their knowledge into clinical proficiency. Despite possessing sound knowledge about certain diseases, effectively applying this knowledge to disease diagnosis remain challenging for LLMs⁹. This observation highlights the need to train and evaluate LLMs using datasets that are more closely aligned with clinical applications, thereby bridging the gap between theoretical knowledge and practical clinical usage.

To address these challenges, we proposed to build a dataset that focuses on specific medical diseases, which should encompass updated evidence-based medical knowledge capable of providing both physicians and patients with expert disease-related information. In addition, real-world cases featuring patient information and treatment decisions encountered in clinical practice should be included. This repository can serve as a benchmark for testing the performance of LLMs in specific medical domains, such as OA management.

Moreover, despite the fact that LLMs have demonstrated impressive capabilities, their internal mechanisms remain unclear. This lack of transparency poses unnecessary risks to downstream applications¹⁰ which is particularly crucial in the medical field and constitutes a significant ethical consideration.¹¹ Retrieval-Augmented Generation (RAG) offers a solution for explainability, as the RAG technique enables large models to identify the source of their answers when responding to questions. RAG is an AI framework that improves large language models by integrating relevant information from external knowledge bases, thus enhancing the accuracy and reliability of the model's responses while also providing efficient and cost-effective access to updated external data¹². Therefore, the integration of retrieval-augmented generation (RAG) and prompt engineering could enable the model to assimilate external knowledge bases and adhere to instructions to respond in a predetermined manner.

In general, we propose a dataset framework that encompasses updated evidence-based medical

knowledge, and real-world cases may effectively examine the capabilities of LLMs in clinical practice. The integration of RAG and prompt engineering may allow trained LLMs such as GPT-4 to acquire domain-specific abilities. Moreover, the management of OA serves as an ideal example in terms of its clinical significance and data volume on this research topic. Therefore, this study aimed to curate a dataset for OA management, evaluate knowledge of updated evidence-based medicine for LLMs and their capabilities in clinical scenarios, and adopt RAG and instruction prompts to enhance these capabilities.

Methods

This study curated an OA management dataset based on clinical guidelines and real-world cases. A benchmark was developed to evaluate the clinical knowledge and capabilities of LLMs for OA management. DocOA was built with instruction prompts and RAG and was tested along with other LLMs. **Figure 1** illustrates the flow diagram of the study.

Dataset

This dataset was developed based on key clinical guidelines and real-world patients. After the panel discussion, six well-acknowledged guidelines and data from 80 real-world patients were selected that included various aspects of OA management. The following guidelines were included: American Academy of Orthopedic Surgeons (AAOS) management of osteoarthritis of the knee (Non-arthroplasty)¹³; National Institute for Health and Care Excellence (NICE) guideline for osteoarthritis in over 16s¹⁴; Osteoarthritis Research Society International (OARSI) guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis¹⁵; Royal Australian College of General Practitioners Guideline for the management of knee and hip osteoarthritis¹⁶; American College of Rheumatology/Arthritis Foundation (ACR) Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee¹⁷; European League Against Rheumatism (EULAR) recommendations for the non-pharmacological core management of hip and knee osteoarthritis¹⁸. Between 1 April 2023 to 1 October 2023 80 patients diagnosed with osteoarthritis and who had received OA management at our hospital were randomly selected. The patient information, including age, sex, height, weight, body mass index (BMI), laterality of knee involvement, medical history, level of pain, mechanical symptoms, physical examination results, and radiographic findings, were retrieved. All identifiable information was concealed to maintain confidentiality.

The OA benchmark aims to test the clinical capabilities of LLMs at four levels within the context of evidence-based medicine, ranging from domain-specific knowledge to clinical capabilities. The benchmark assesses the performance of LLMs pertaining to OA knowledge, summarising the knowledge to formulate recommendations for specific management options, providing tailored management options for different patient populations, and formulating personalised management plans for real-world cases.

Assistant with RAG and Instruction Prompting

DocOA, a specialised assistant, was developed based on the GPT-4-1106-preview model, which integrates instruction prompts and RAG to enhance performance. The instruction prompt emphasised its role in providing evidence-based medical insights and personalised management programmes guided by evidence-based medicine. The DocOA strictly adheres to facts, avoids speculation, and clearly states its limitations. Moreover, it maintains a professional and informative tone suitable for medical discussions.

RAG has been used to respond to various OA-related queries. The RAG integrates a model's language generation capabilities with a retrieval system, enabling access to specific information from external sources¹⁹. Of the several RAG techniques and data structures tested, the retrieval function from OpenAI was adopted, and the most optimal data structure was selected and converted into the JavaScript Object Notation (JSON) format for optimal retrieval accuracy. In response to OA-related queries, the RAG enables the assistant to dynamically pull relevant data from the external dataset as it generates responses. The workflow of the assistant is illustrated in **Figure 2**.

The assistant block details the core of the DocOA system, utilising a base model of GPT-4-1106-preview. The description within the block serves as an instruction prompt that outlines the system's role. The Assistant's functionality includes a Retrieval tool that accesses external dataset.

The system receives input from thread blocks in the form of questions about osteoarthritis management and requests evaluation. This includes user messages with specific queries regarding osteoarthritis treatment and detailed instructions for the system to follow.

The execution command then triggers the processing of input data through the DocOA system. Through retrieval-augmented generation (RAG), DocOA incorporates external information from the knowledge database and follows the instructions to generate the final output. Simultaneously, based on instructional prompts, DocOA is capable of providing final outputs tailored to the user, whether they are patients or professional doctors, in the corresponding style.

Models Testing

DocOA and the two base models, GPT-3.5 and GPT-4, were tested against the OA benchmark. Each question was presented five times to each model to assess the robustness of its performance. Additionally, the zero-shot chain of thoughts (COT) prompt technique was tested for GPT-3.5 and GPT-4 to determine whether it outperformed the input-output (IO) technique.

Evaluation of LLMs' Performance

Objective Evaluation

The model-generated responses were compared with predefined correct answers for each subset of the benchmark. An answer was considered accurate if LLM provided correct knowledge (recommendation status and recommendation strength) about the treatment option and predicted the correct treatment recommendation (treatment appropriateness) for a specific patient profile or individual patient.

The human evaluation framework is an effective approach for identifying the gap between LLMs and clinical experts³. In this study, human evaluation was performed by both physicians and patients. Eighty items from the OA benchmark were randomly selected for a detailed human evaluation framework.

A panel of five physicians, each with a minimum of 10 years of experience in OA management, conducted the physician evaluation. The sequence of answers was randomized and the generating models were anonymized to ensure that the evaluation was conducted without any knowledge of the model that generated them. Specifically, each evaluator independently assessed the sampled responses. All sampled responses were compiled into an Excel sheet. The model labels corresponding to the responses were concealed. The evaluation metrics were established based on a previous study with modifications³. The physician assessed the quality of the responses in the

following domains: inaccurate content, relevance, hallucinations, missing content, likelihood of possible harm, extent of possible harm, and possibility of bias. The ability of LLM to achieve correct comprehension, retrieval, and reasoning was assessed using the method described in a previous study²⁰. Patient evaluation was conducted by assessing the user intent fulfilment, and helpfulness of the content. The detailed descriptions of each human evaluation metric are provided in Supplementary File 1.

Statistical Analysis

All statistical analyses were performed using the SPSS 25.0 software (IBM, Armonk, NY, USA) and GraphPad Prism version 8 (GraphPad Software, San Diego, CA, USA). Discontinuous data are expressed as incidence and rate and analysed using the chi-square test for differences. A P value less than 0.05 indicated statistical significance.

Results

OA Benchmark

The benchmark comprised four subsets of question-answer (QA) evaluations designed to test the performance of LLMs across a spectrum ranging from domain-specific knowledge to practical capability. Guideline-item QA (GIQA), which was developed based on specific items extracted from the clinical guidelines, evaluates the LLMs' knowledge of these well-established standards. The GIQA comprised 337 items. Management options QA (MOQA) included summarised recommendations for specific treatments from the included clinical guidelines. The MOQA, which comprised 145 items, evaluated LLMs' knowledge of specific treatment options, as well as their ability to summarise medical evidence. Treatment strategy QA (TSQA), which included treatment recommendations for different patient populations, provided treatment recommendations based on the patient's age, clinical presentation, and other factors. The TSQA, which comprised 216 items, evaluated the capability of LLMs to derive treatment recommendations for specific patient types. Real-case QA (RCQA) included treatment recommendations for 80 real-world patients. The RCQA, which comprised 80 items, evaluated LLMs' capability in formulating treatment recommendations in a more complicated scenario in which individual information is provided, mirroring real-world clinical decision-making. This dataset is available on GitHub <https://github.com/geteff1/Osteoarthritis-Benchmark>. Examples of each QA type are shown in Supplementary Figure 1.

Objective Evaluation

The accuracy of GPT-3.5 in GIQA, MOQA, TSQA and RCQA was 0.26, 0.22, 0.01 and 0.03, respectively. The accuracy of GPT-4 in GIQA, MOQA, TSQA, and RCQA was 0.38, 0.30, 0.07, and 0.01, respectively. The accuracy of DocOA in GIQA, MOQA, TSQA, and RCQA was 0.92, 0.87, 0.88, and 0.72, respectively. The accuracy of each model against the benchmark is presented in Table 1 and Figure 3a. As shown in Figure 3a, the degree of accuracy significantly decreased: GIQA > MOQA > TSQA > RCQA. As shown in Figure 3b, DocOA reported 111 failures in accessing the external dataset, which accounted for 12.4% of the inaccurate answers generated.

Table 1. Accuracy of each model against osteoarthritis benchmark

	GPT-3.5	GPT-4	DocOA	P Value
Osteoarthritis Benchmark	0.16	0.24	0.88	<0.001*

Guideline item QA (GIQA)	0.26	0.38	0.92	<0.001*
Management option QA (MOQA)	0.22	0.30	0.87	<0.001*
Treatment strategy QA (TSQA)	0.01	0.07	0.88	<0.001*
Real-case QA (RCQA)	0.03	0.01	0.72	<0.001#

*Further analysis showed all pairwise comparison had P value less than 0.05; #Further analysis showed P value is 0.056 for GPT-3.5 vs GPT-4, <0.001 for GPT-3.5 vs DocOA, <0.001 for GPT-4 vs DocOA.

Zero-shot chain of thoughts (COT) prompt techniques were adopted for GPT-3.5 and GPT-4. Compared with the input-output (I/O) prompt technique, no significant improvements in model performance were observed. The results are summarised in Table 2.

Table 2. Accuracy of different prompt techniques against osteoarthritis benchmark

	I/O*	GPT-3.5		I/O	GPT-4	
		COT#	P value		COT	P value
Osteoarthritis Benchmark	0.16	0.17	0.41	0.24	0.23	0.52
Guideline item QA (GIQA)	0.26	0.28	0.03	0.38	0.38	0.80
Management option QA (MOQA)	0.22	0.20	0.004	0.30	0.27	0.002
Treatment strategy QA (TSQA)	0.02	0.03	<0.001	0.07	0.07	0.79
Real-case QA (RCQA)	0.03	0.01	<0.001	0.01	0.01	0.20

* Input Output prompt technique; # zero-shot chain of thought prompt technique.

Human Evaluation Results

From each of the GIQA, MOQA, TSQA, and RCQA, 20 items were randomly selected along with the corresponding responses generated by each model. A total of 1200 outputs were evaluated by physicians and patients. The results of the human evaluations of GPT-3.5, GPT-4, and DocOA revealed distinct outcomes across several aspects. A few examples have been selected and summarized in **Supplementary File 5**, demonstrating how to apply the human evaluation framework to assess sampled responses.

Supplementary Figure 2 shows the human evaluation results for the models' output. The rate of inaccuracy was the highest for GPT-3.5 (57%), followed by GPT-4 and DocOA at 50%, and 19.3%, respectively. All the models achieved high relevance and infrequently produced hallucinatory content in their responses. GPT-3.5 had a higher proportion of responses with missing content (22%) than GPT-4 (16.4%) or DocOA (16.5%). GPT-3.5 presented a higher likelihood of generating harmful content (20%) than GPT-4 (11.3%) and DocOA (8.3%). Moreover, GPT-3.5 was associated with a higher risk of causing severe harm (10.5%) than GPT-4 (5.5%) and DocOA (3.5%). The likelihoods of potentially biased content were 13.3%, 9.5%, and 2.8% for GPT-3.5, GPT-4, and DocOA, respectively. The results of the human evaluation for each subset benchmark are listed in Supplementary Files 2, 3, and 4. The results showed a substantial decrease in performance in terms of inaccurate content and missing content (GIQA > MOQA > TSQA > RCQA).

Figure 4 shows the results of the LLMs' ability to assess correct comprehension, correct retrieval, and correct reasoning. Regarding the correct comprehension of the question, the response rate of DocOA was 91%, followed by GPT-4 (86%), and GPT-3.5 (82.5%). DocOA was able to correctly recall and present complete, relevant information in 65.8% of the responses, followed by GPT-4 (14.3%) and GPT-3.5 (12.0%). In terms of subset evaluation, the results showed comparable performance in

comprehension and reasoning among the different models, whereas a substantial performance decrease was found in correct retrieval across GIQA, MOQA, TSQA, and RCQA.

The results of patient evaluations are shown in Figure 5. DocOA achieved a success rate of 71.3% in fulfilling patient intention, with GPT-4 at 39.8% and GPT-3.5 at 36.5%. Of the responses generated by DocOA, 75.8% were considered to be at least somewhat helpful, compared to 47% for GPT-3.5 and 47.75% for GPT-4. For GPT-3.5 and GPT-4, the subset evaluation showed a substantial decrease in intent fulfilment and helpfulness as the tasks shifted from domain-specific knowledge to personalised treatment recommendations.

Discussion

This study introduced a benchmark framework to assess the performance of LLMs in specific medical domains. Using OA as a case study, this framework is the first to evaluate LLMs across a spectrum, from domain-specific knowledge to clinical applications in specific disease management. The incorporation of human evaluation provides multiple dimensions of assessment that are of considerable interest to clinical practitioners and patients, making it an essential tool for evaluating the clinical capabilities of LLMs.

The study found that the integration of RAG and instructional prompts substantially improved the domain-specific capabilities and explainability of general-purpose LLMs without additional training. By contrast, general-purpose models such as GPT-3.5 and GPT-4 exhibit unsatisfactory performance when benchmarked against OA management. Additionally, both models demonstrated a marked decline in performance as tasks shifted from domain-specific knowledge to personalized treatment recommendations. Overall, the findings of this study demonstrate a cost-effective method for evaluating and enhancing the capabilities of LLMs in specialized medical fields.

Domain-specific Medical Benchmark

Although benchmarks targeting general medical knowledge have been previously developed²¹⁻²³, recent research has suggested that these benchmarks are only preliminary indicators of medical knowledge. The absence of tailored benchmarks in specific domains remains a potential challenge for evaluating the clinical effectiveness of LLMs^{4,24}. Therefore, we developed a domain-specific benchmark focused on disease management for OA, which was selected for its prevalence, substantial disease burden, and complexity of its management strategies^{6,7}. This benchmark was designed to test the domain-specific knowledge and clinical capabilities of LLMs. The benchmark comprised four parts, each testing the ability of LLMs at different levels, including the ability to provide evidence-based knowledge, summarising knowledge to formulate recommendations, providing management recommendations for different patient populations, and formulating personalised management plans for real-world patients. The benchmark was constructed based on established clinical guidelines and real-world patient information. Clinical guidelines offer comprehensive reviews of updated evidence and expert opinions, making them reliable sources of domain-specific medical knowledge. Through panel discussions involving physicians and data scientists, the questions were designed in a hybrid format, integrating both definitive and interpretative elements. Using this benchmark, we confirmed that general-purpose LLMs exhibit suboptimal performance in specialised domains. A significant performance gap was observed between domain-specific knowledge and clinical proficiency. This highlights the challenges faced by general-purpose LLMs in effectively applying specialised knowledge to clinical scenarios.

Human Evaluation Framework

Human evaluation is a crucial component in assessing the medical capabilities of large language models (LLMs) and offers a multidimensional assessment of their clinical capabilities. In this study, the human evaluation framework was modified based on a previous study, and hallucinations and relevance were added as additional criteria^{3,4}. The evaluation criteria included accuracy, relevance, hallucinations, omissions, potential harm, and biased content. Moreover, the performance of LLMs in question comprehension, information retrieval, and medical reasoning was evaluated, as these are crucial abilities in tailoring patient-specific treatment. Patient evaluations primarily determine how responses address the user's intent and helpfulness. Although previous studies indicate a notable gap between objective benchmarking and human evaluation, our findings reveal a smaller discrepancy³. This could be attributed to the different knowledge domains and designs of the QA structure in this benchmark. The results of our study suggest that the GPT-3.5, GPT-4, and DocOA performed well in terms of hallucinations, comprehension, reasoning, and relevance. DocOA outperformed the other models in terms of accurate information, correct retrieval, and helpfulness as perceived by patients. This indicates that although generalised models are proficient in some areas, they remain inadequate in delivering the qualified responses required in a clinical context.

Augmenting LLM with Domain-specific Ability

Several techniques are available for developing medical LLMs, which primarily include integrating domain-specific knowledge during the training phase through techniques such as reinforcement learning with human feedback (RLHF)²⁵⁻²⁷. However, in this study, we focused on enhancing already-trained LLMs, such as GPT-4, by employing a suite of techniques, including RAG and instruction-based prompts. Similar methodologies have been applied to the development of specialised LLMs for chemical domain²⁸.

This approach was adopted for the following reasons: first, augmenting an existing model such as GPT-4 is more cost-effective than training a new model from scratch; second, advanced general-purpose models have been trained on diverse datasets, providing a broad base of general knowledge that can be beneficial for understanding and contextualising domain-specific information; third, techniques such as RAG and specialised prompting offer the convenience of being adjustable and refined over time, enabling easy adaptability to new evidence in the fast-evolving field of medicine.

Among these reasons, the role of RAG in this study needs to be emphasized. Explainability has always been a problem to be addressed in the application of LLMs, and it is also one of the ethical considerations in their application in the medical field. RAG enables the model to identify the source on which the generated answer was based on, which significantly improved the explainability of the model. In this study, the knowledge base was structured in a way that allows RAG to identify the clinical evidence that the answer was based on. When DocOA generates an OA relevant answer, it identifies the evidence on which the answer it is based on as well, makes it clear for professionals to evaluate the rationale and accuracy for its response. In general, provided with reliable and professional sources, RAG significantly improved the explainability and accuracy of LLMs in medical care.

Our results demonstrated that GPT-4 can effectively acquire domain-specific knowledge and clinical capabilities in the management of OA through a combination of approaches, including RAG

and instruction prompts. This strategy can also be applied cost-effectively to other medical domains.

Nonetheless, the efficacy of the RAG is contingent upon factors such as the size and quality of the data, retrieval techniques employed, and the underlying architecture of the LLM in use²⁹. The evaluation results showed that DocOA was able to correctly understand the clinical question and avoid hallucinations. The fundamental reason for the errors stems from the inherent limitations of the RAG technology, where the retrieved information may sometimes be incomplete or may not adequately address the question.

In RAG, the query process is the key determinant of the information retrieved from the external knowledge base. In the query process, text is converted into an array of floating points. The entire array corresponds to a point in an n-dimensional space, which is known as the text vector, also referred to as an embedding^{30,31}. The distance between vectors, which corresponds to semantic similarity, can be calculated. Therefore, by calculating semantic similarity, the query process retrieves the information from the external knowledge base that is considered most relevant to the query posted by the LLM. However, this process is sometimes uncertain because the most appropriate answer does not necessarily yield the highest similarity score. For example, in GIQA, the question is "Is Walking aids, assistive technology and adaptations at home and/or at work recommended for Physical Treatment of Knee osteoarthritis according to 2013 European League Against Rheumatism (EULAR) recommendations for the non-pharmacological core management of hip and knee osteoarthritis?" For this question, out of the five responses generated by LLM, four incorrectly state "recommend" when the correct answer should be "may be used." The error occurred because only a portion of the information was retrieved. However, neither the complete information nor the key information, which contains the answer, was retrieved. More specifically, if the retrieved information was "The frequent use of assistive technology and the high satisfaction rates with its use indicate that walking aids, assistive technology, and adaptations are important and useful for people with hip or knee OA." The model might be misled to generate the incorrect answer.

Generally, enhancing the effectiveness of RAG is an ongoing effort in the field of computer science, and the effectiveness of RAG is likely to keep increasing with the rapid development of large language models and RAG technologies.

The Potential Value of DocOA in Future Clinical Practice

DocOA has shown promising potential in clinical practice due to its proficient OA management knowledge and the ability to explain its answer. Previous studies have shown general-purpose LLMs to be limited in medical specialties such as nephrology, ophthalmology and osteoarthritis. The versatile and unverified training data could contain misleading medical information, which led to inaccurate answers. On the other hand, DocOA has shown good results in providing evidence-based OA management knowledge and formulating personalized treatment plans. To adopt DocOA in clinical practice, it is designed to answer patient queries and doctor queries differently, as shown in **Figure 6**. When generating answers for medical professionals, it is able to generate answer based on selected high quality clinical evidence and accurately identified the corresponding references, thus providing explainable and evidenced based answers.

Limitations

This study had several potential limitations that need to be addressed. First, OA management is

highly complex, and our current dataset remains limited and requires continuous supplementation and updating. Therefore, establishing specialised groups dedicated to building and updating these LLM databases is imperative for diverse medical applications. Second, our reliance on English sources could restrict the applicability and inclusivity of our findings across different linguistic and cultural contexts. Third, although there have been human evaluations of clinical case data, the augmented model has not yet been tested in a real-world clinical setting. These limitations highlight the necessity for ongoing development and a comprehensive, multidimensional approach for evaluating LLMs in the medical field.

Conclusion

In this study, we introduce a novel benchmark framework designed to evaluate the capabilities of LLMs in specific medical domains, with OA serving as a case study. This framework assesses LLMs in terms of medical knowledge, evidence summarisation, and clinical capabilities. Through a combination of objective measures and human evaluations, we identified the limitations of generalised LLMs in clinical contexts. Furthermore, our study demonstrated that DocOA, which integrates Retrieval-Augmented Generation (RAG) and instructional prompts, significantly improves both the domain-specific performance and the explainability of LLMs. This approach is a potentially cost-effective strategy for developing domain-specific medical LLMs.

References

- 1 Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. J. a. p. a. Capabilities of gpt-4 on medical challenge problems. (2023).
- 2 Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* **330**, 78-80, doi:10.1001/jama.2023.8288 (2023).
- 3 Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172-180, doi:10.1038/s41586-023-06291-2 (2023).
- 4 Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nature medicine* **29**, 1930-1940, doi:10.1038/s41591-023-02448-8 (2023).
- 5 Khera, R. *et al.* AI in Medicine-JAMA's Focus on Clinical Outcomes, Patient-Centered Care, Quality, and Equity. *Jama* **330**, 818-820, doi:10.1001/jama.2023.15481 (2023).
- 6 Hunter, D. J. & Bierma-Zeinstra, S. Osteoarthritis. *Lancet (London, England)* **393**, 1745-1759, doi:10.1016/s0140-6736(19)30417-9 (2019).
- 7 Steinmetz, J. D. *et al.* Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the Global Burden of Disease Study 2021. **5**, e508-e522 (2023).
- 8 Wang, L. *et al.* Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *NPJ digital medicine* **7**, 1-9, doi:10.1038/s41746-024-01029-4 (2024).
- 9 Li, J. *et al.* One is Not Enough: Multi-Agent Conversation Framework Enhances Rare Disease Diagnostic Capabilities of Large Language Models. (2023).
- 10 Zhao, H. *et al.* (arXiv, 2023).
- 11 Wang, C. *et al.* Ethical Considerations of Using ChatGPT in Health Care. *Journal of medical Internet research* **25**, e48009, doi:10.2196/48009 (2023).
- 12 Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. **33**, 9459-9474 (2020).
- 13 Brophy, R. H. & Fillingham, Y. A. J. J.-J. o. t. A. A. o. O. S. AAOS clinical practice guideline summary: management of osteoarthritis of the knee (nonarthroplasty). **30**, e721-e729 (2022).
- 14 Health, N. I. f. & Excellence, C. Osteoarthritis in over 16s: Diagnosis and management. (2022).
- 15 Bannuru, R. R. *et al.* OARSI guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis. *Osteoarthritis and cartilage* **27**, 1578-1589, doi:10.1016/j.joca.2019.06.011 (2019).
- 16 Practitioners, T. R. A. C. o. G. Guideline for the management of knee and hip

- osteoarthritis. *RACGP* (2018).
- 17 Kolasinski, S. L. *et al.* 2019 American College of Rheumatology/Arthritis Foundation Guideline for the Management of Osteoarthritis of the Hand, Hip, and Knee. *Arthritis & rheumatology (Hoboken, N.J.)* **72**, 220-233, doi:10.1002/art.41142 (2020).
- 18 Fernandes, L. *et al.* EULAR recommendations for the non-pharmacological core management of hip and knee osteoarthritis. *Annals of the rheumatic diseases* **72**, 1125-1135, doi:10.1136/annrheumdis-2012-202745 (2013).
- 19 Chen, J., Lin, H., Han, X. & Sun, L. J. a. p. a. Benchmarking large language models in retrieval-augmented generation. (2023).
- 20 Feng, S. Y., Khetan, V., Sacaleanu, B., Gershman, A. & Hovy, E. J. a. p. a. CHARD: Clinical Health-Aware Reasoning Across Dimensions for Text Generation Models. (2022).
- 21 Zhang, X., Wu, J., He, Z., Liu, X. & Su, Y. in *Proceedings of the AAAI conference on artificial intelligence*.
- 22 Peng, C. *et al.* A Study of Generative Large Language Model for Medical Research and Healthcare. (2023).
- 23 Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. **2**, e0000198 (2023).
- 24 Ayers, J. W. *et al.* Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA internal medicine* **183**, 589-596, doi:10.1001/jamainternmed.2023.1838 (2023).
- 25 Kraljevic, Z. *et al.* MedGPT: Medical concept prediction from clinical narratives. (2021).
- 26 Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. J. N. d. m. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. **4**, 86 (2021).
- 27 Tian, Y., Gan, R., Song, Y., Zhang, J. & Zhang, Y. J. a. p. a. ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences. (2023).
- 28 Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570-578, doi:10.1038/s41586-023-06792-0 (2023).
- 29 Zhang, P., Xiao, S., Liu, Z., Dou, Z. & Nie, J.-Y. J. a. p. a. Retrieve anything to augment large language models. (2023).
- 30 Lee, J. *et al.* The AAOS 2019 Clinical Practice Guidelines for the Management of Rotator Cuff Injuries Are Unbiased and Incorporate a Diverse Body of Literature. *Arthroscopy, sports medicine, and rehabilitation* **4**, e559-e565, doi:10.1016/j.asmr.2021.11.017 (2022).
- 31 Jang, B., Kim, I. & Kim, J. W. Effective Training Data Extraction Method to Improve Influenza Outbreak Prediction from Online News Articles: Deep Learning Model Study. *JMIR Medical Informatics* **9**, e23305, doi:10.2196/23305 (2021).

Supplementary Files

Supplementary file 2.

URL: <http://asset.jmir.pub/assets/59f62d24c8d425170e34abd767587e8f.pdf>

Supplementary file 1.

URL: <http://asset.jmir.pub/assets/b277fc3973f3c1c24ffe5f22ebddff56.pdf>

Supplementary file 3.

URL: <http://asset.jmir.pub/assets/d09ffaa04681e9935c517270424aaed8.pdf>

Supplementary file 4.

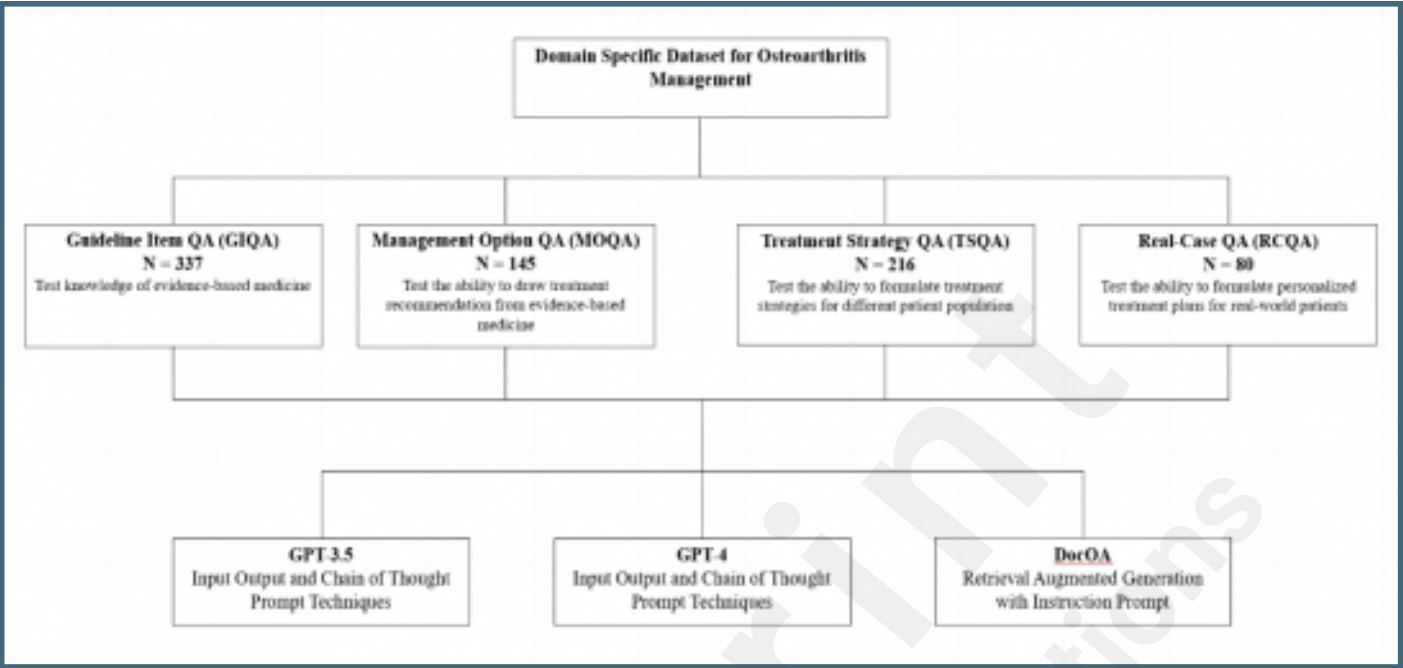
URL: <http://asset.jmir.pub/assets/4ea57b53cd4f6ecd08c9412e0ab5d341.pdf>

Supplementary File 5.

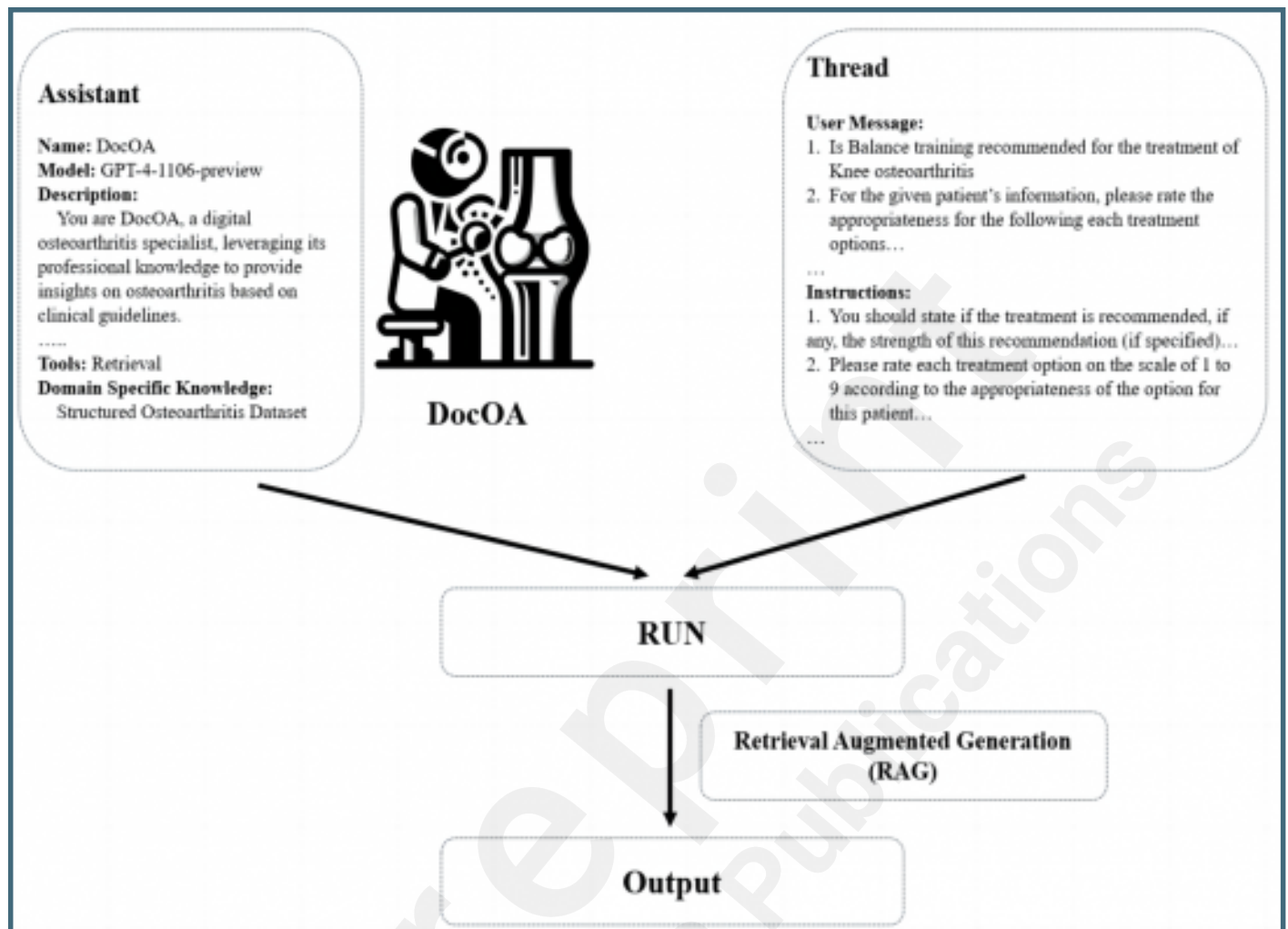
URL: <http://asset.jmir.pub/assets/6adb0742646903f1f1f5fc8afcc5c65f.pdf>

Figures

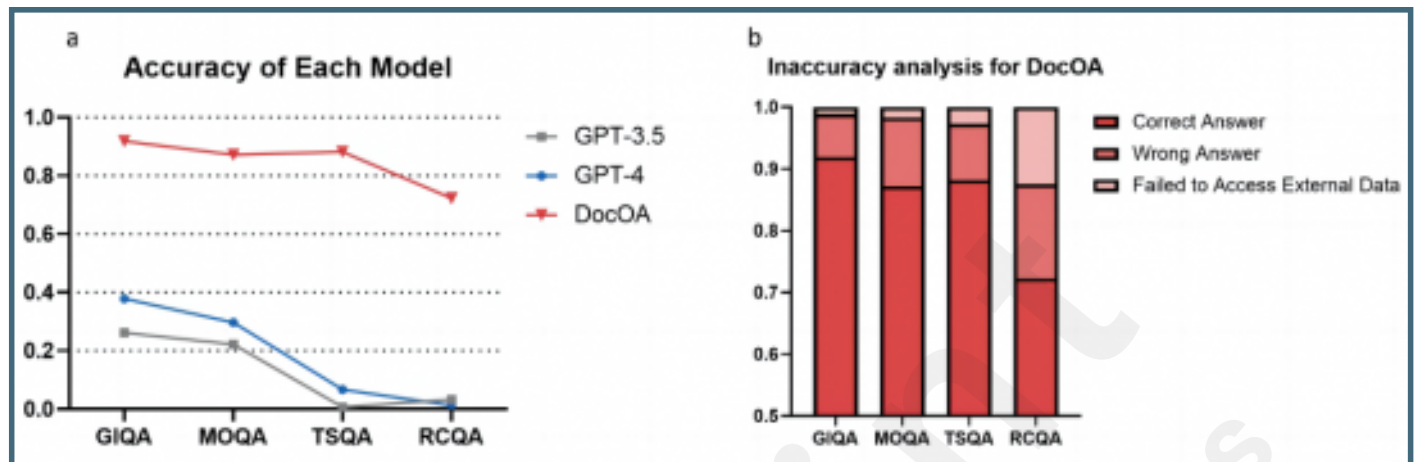
Study Flow diagram.



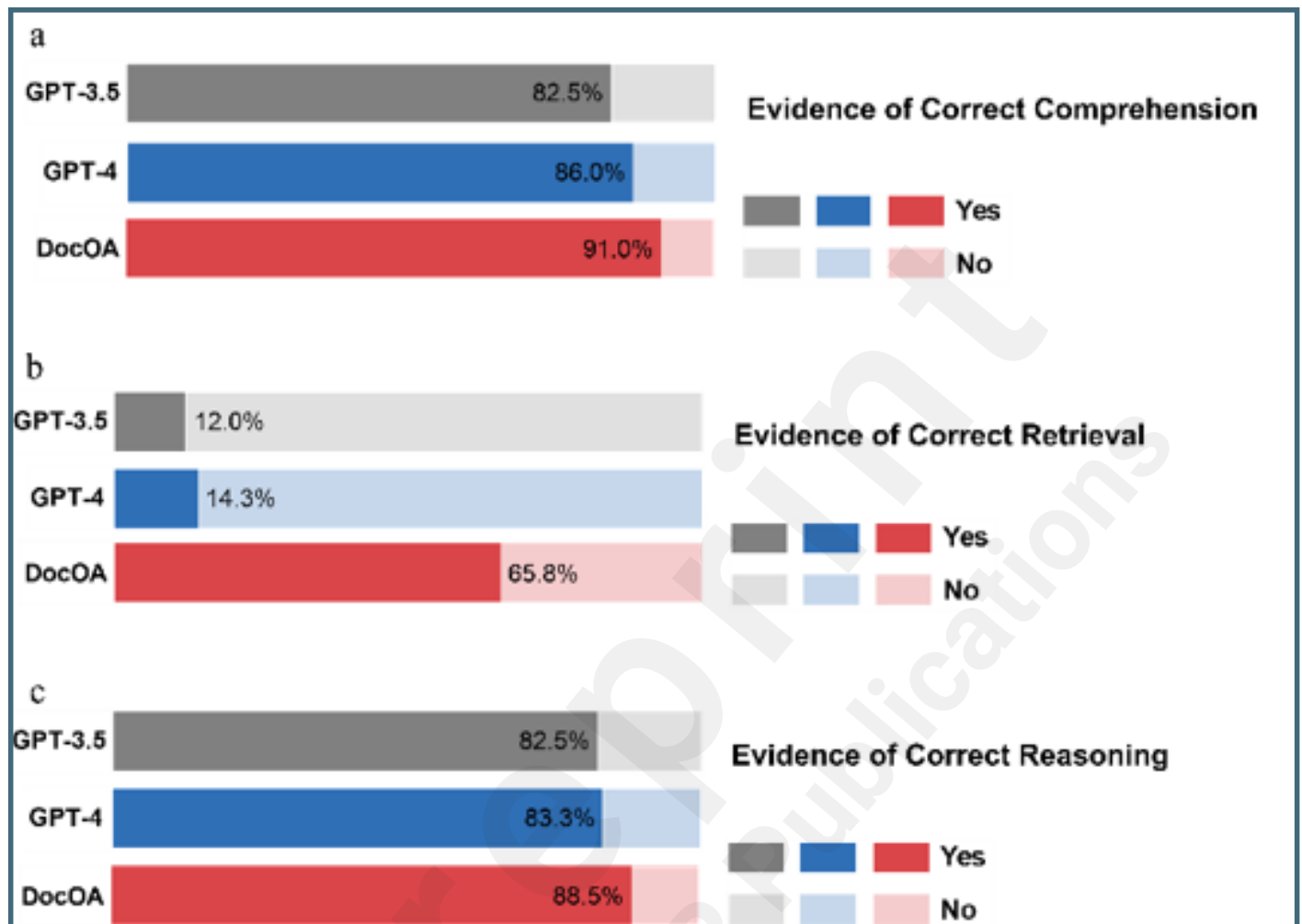
Workflow.



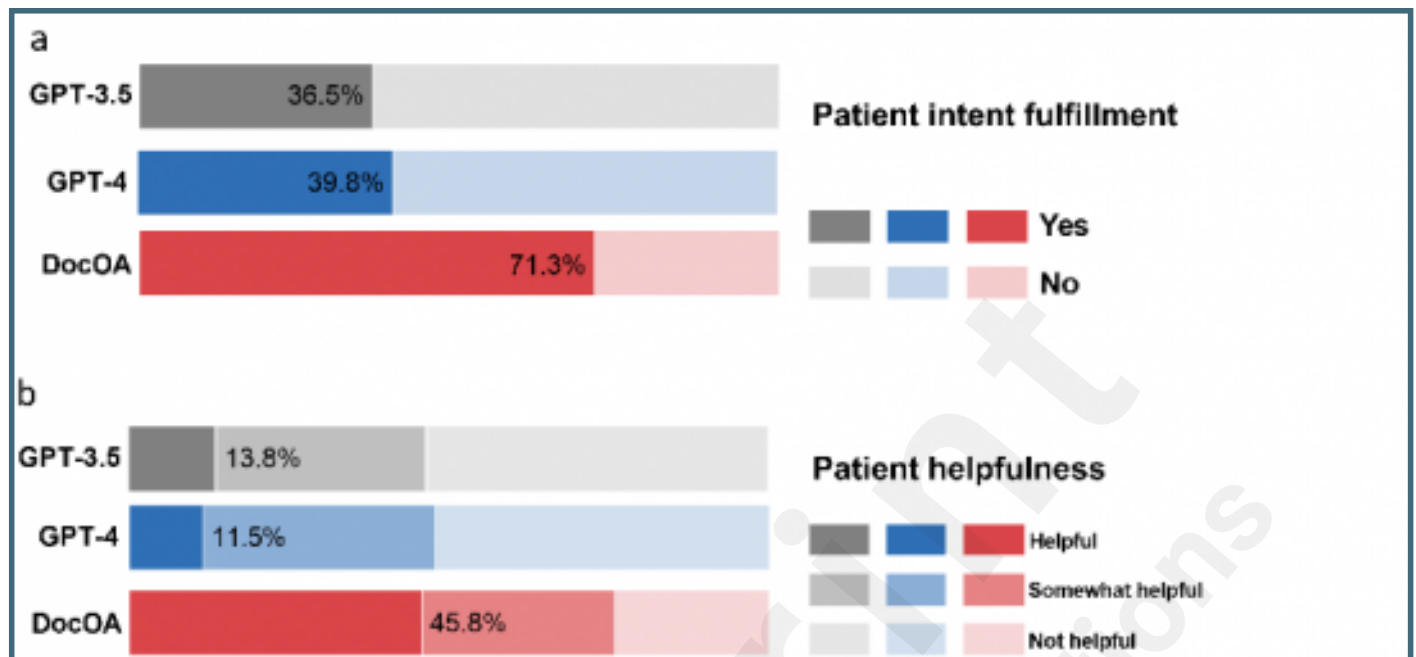
Results of objective evaluation a. accuracy of each model against each subset of benchmark; b. Inaccuracy analysis for DocOA due to wrong answer and failure in access external data.



Results of human evaluation for LLMs' comprehension, retrieval and reasoning ability a. evidence of correct comprehension; b. evidence of correct retrieval; c. evidence of correct reasoning.





Results of human evaluation by patients a. patient intent fulfillment; b. patient helpfulness.



Examples of clinical practice. a How DocOA provides exercise guidance for patients with OA. b How DocOA provides doctors with clinical evidence for exercise guidance in OA, as well as the clinical evidence the answer is based on.


User: I am a patient with OA and should I exercise even when my knee hurts?


 Searching my knowledge

 As a patient with osteoarthritis (OA), it is generally recommended to engage in land-based exercise to improve pain and function, regardless of your age, the severity of the disease, your functional status, or pain levels. This recommendation is based on the 2018 Royal Australian College of General Practitioners Guideline for the management of knee and hip osteoarthritis. It's important to note that while the strength of this recommendation is strong, the certainty of the evidence supporting it is considered low.
Source: 2018 Royal Australian College of General Practitioners Guideline

However, it's crucial to approach exercise with caution and consider personal factors such as your current pain levels and physical capabilities. It might be beneficial to consult with a healthcare professional to tailor an exercise program that suits your individual needs and limitations.

User: What exercise should I do?

 Searching my knowledge

 For managing knee osteoarthritis, it is recommended to engage in structured land-based exercise programs. These programs can include:


1. Strengthening exercises
2. Cardiovascular exercises
3. Balance training or neuromuscular exercises
4. Mind-body exercises, such as Tai Chi or Yoga


These exercises can be combined with dietary weight management for an integrated approach to treatment. The strength of this recommendation is strong, although the certainty of the evidence is not specified. This recommendation is based on the 2019 Osteoarthritis Research Society International (OARSI) guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis.
Source: 2019 OARSI Guideline

It's important to tailor these exercises to your individual needs and capabilities. If you are unsure about how to start or if you have specific health concerns, it would be beneficial to consult with a healthcare professional or a physiotherapist. They can help design a program that's right for you, considering your current level of fitness and any other health issues you might have.

a

User: I am an orthopedic surgeon, I want to know what is the latest evidence regarding the optimal exercise for osteoarthritis.

 Searching my knowledge

 The latest evidence on optimal exercises for osteoarthritis management, based on various guidelines, recommends the following approaches:

Structured Land-Based Exercise Programs: This includes strengthening, cardio, balance training, and neuromuscular exercises. These are strongly recommended for hip osteoarthritis, although the certainty of the evidence is not specified. This is in line with the 2019 Osteoarthritis Research Society International (OARSI) guidelines for the non-surgical management of knee, hip, and polyarticular osteoarthritis.
Source: 2019 OARSI Guideline

Walking, Muscle-Strengthening Exercises, and Tai Chi: These exercises are strongly recommended for knee osteoarthritis, based on the 2018 Royal Australian College of General Practitioners Guideline. The strength of the recommendation is strong, but the certainty of evidence is low.
Source: 2018 Royal Australian College of General Practitioners Guideline

General Exercise Recommendation: According to the 2021 American Academy of Orthopedic Surgeons (AAOS) Management of Osteoarthritis of the Knee (Non-Arthroplasty), exercise, in general, is recommended for knee osteoarthritis with a strong recommendation strength. The certainty of evidence for this is not specified.
Source: 2021 AAOS Guideline

Combination of Physical and Mind-Body Exercise: This approach includes strengthening, cardio, balance training, neuromuscular exercise, and mind-body exercise (like Tai Chi or Yoga), with or without dietary weight management. This is strongly recommended for knee osteoarthritis, as per the 2019 OARSI guidelines, although the certainty of the evidence is not specified.
Source: 2019 OARSI Guideline

Tailored Exercise: The 2022 National Institute for Health and Care Excellence (NICE) guideline emphasizes exercise tailored to the patient's needs for osteoarthritis treatment. This recommendation highlights the importance of therapeutic exercise in managing and reducing symptoms, improving, or maintaining physical functioning over the long term. However, the strength of this recommendation and the certainty of evidence are not specified.
Source: 2021 NICE Guideline

These recommendations highlight the importance of individualized exercise programs, considering the patient's specific needs, preferences, and physical capabilities. Exercise has been shown to have clinically important benefits for people with osteoarthritis, offering general health benefits and a safety profile superior to other common treatments like analgesia.

b

Related publication(s) - for reviewers eyes onlies

Supplementary Figure 2. Results of Human Evaluation for the Assessment of Responses a. inaccurate content; b. relevance; c. hallucination; d. missing content; e. likelihood of possible harm; f. extent of possible harm; g. possibility of bias.

URL: <http://asset.jmir.pub/assets/996e3dea44fc35f5c946c2aef12f0d96.pdf>

Supplementary Figure 1. Benchmark framework for osteoarthritis management.

URL: <http://asset.jmir.pub/assets/279c4d62a6136dacce9ea036cd8f74b8.pdf>

Supplementary File 6.

URL: <http://asset.jmir.pub/assets/b608b60575b07508b4336b3474ea3f9b.pdf>

