



OPEN

A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports

Daniel Truhn¹, Christian D. Weber², Benedikt J. Braun³, Keno Bressemer⁴, Jakob N. Kather^{5,6,7,8}, Christiane Kuhl¹ & Sven Nebelung¹✉

Large language models (LLMs) have shown potential in various applications, including clinical practice. However, their accuracy and utility in providing treatment recommendations for orthopedic conditions remain to be investigated. Thus, this pilot study aims to evaluate the validity of treatment recommendations generated by GPT-4 for common knee and shoulder orthopedic conditions using anonymized clinical MRI reports. A retrospective analysis was conducted using 20 anonymized clinical MRI reports, with varying severity and complexity. Treatment recommendations were elicited from GPT-4 and evaluated by two board-certified specialty-trained senior orthopedic surgeons. Their evaluation focused on semiquantitative gradings of accuracy and clinical utility and potential limitations of the LLM-generated recommendations. GPT-4 provided treatment recommendations for 20 patients (mean age, 50 years \pm 19 [standard deviation]; 12 men) with acute and chronic knee and shoulder conditions. The LLM produced largely accurate and clinically useful recommendations. However, limited awareness of a patient's overall situation, a tendency to incorrectly appreciate treatment urgency, and largely schematic and unspecific treatment recommendations were observed and may reduce its clinical usefulness. In conclusion, LLM-based treatment recommendations are largely adequate and not prone to 'hallucinations', yet inadequate in particular situations. Critical guidance by healthcare professionals is obligatory, and independent use by patients is discouraged, given the dependency on precise data input.

Abbreviations

AI Artificial intelligence
LLM Large language model
MRI Magnetic resonance imaging

Large language models (LLMs) have recently spread into virtually all aspects of life, including medicine. Within the first two months of its launch, chatGPT, the most popular LLM and the archetype of dialogue-based artificial intelligence, attracted more than 100 million users and, in 2023, averaged more than 13 million daily visitors^{1,2}.

While chatGPT (based on the GPT-3.5-model) performed at or near the passing threshold of 60% accuracy when undergoing the three standardized examinations of the United States Medical Licensing Exam (USMLE)³, its successor GPT-4 as the latest state-of-the-art LLM performed considerably better⁴. In addition to exceeding the USMLE passing threshold by over 20 percentage points, GPT-4 performs medical reasoning similarly to well-studied experts⁵. Beyond taking tests, chatGPT has diagnostic and triage abilities close to practicing

¹Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Pauwels Street 30, 52074 Aachen, Germany. ²Department of Orthopaedics and Trauma Surgery, University Hospital RWTH Aachen, Aachen, Germany. ³University Hospital Tuebingen on Behalf of the Eberhard-Karls-University Tuebingen, BG Hospital, Schnarrenbergstr. 95, Tuebingen, Germany. ⁴Department of Radiology, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Hindenburgdamm 30, 12203 Berlin, Germany. ⁵Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany. ⁶Department of Medicine I, University Hospital Dresden, Dresden, Germany. ⁷Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany. ⁸Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany. ✉email: snebelung@ukaachen.de

physicians when dealing with case vignettes of conditions with variable severity^{6,7}. GPT-4 has a clinical value in diagnosing challenging geriatric patients⁸. It also generates broadly appropriate recommendations for common questions about cardiovascular disease prevention⁹, breast cancer prevention, and screening¹⁰. It can also generate structured radiologic reports from written (prosaic) text¹¹, and draft Impressions sections of radiologic reports, even though the GPT-4-generated Impressions are not (yet) as good as the radiologist-generated ones regarding coherence, comprehensiveness, and factual consistency¹². For a comprehensive overview of LLMs and their utilization in radiology, the reader is referred to recent review articles^{13–15}.

Despite the growing popularity of LLMs, concerns have arisen regarding the validity and reliability of their recommendations, particularly in medicine. LLMs tend to produce convincing but factual incorrect text (commonly referred to as “hallucinations”), which raises the question if LLMs are sufficiently sophisticated to be used as resources for health advice or may pose a potential danger¹⁶. In particular, patients may rely on information provided by artificial intelligence without consulting healthcare professionals^{17,18}.

Likely, LLMs will also be extensively used by radiologists in the future. In this era of complex interdisciplinary patient management, a radiologist’s work often does not end with submitting a report of an imaging study. Addressing patients’ concerns and questions and communicating appropriately with non-radiologist colleagues requires solid knowledge of treatment options, prioritization, and limitations.

Consequently, the present study aims to investigate the validity of treatment recommendations provided by GPT-4, explicitly focusing on common orthopedic conditions, where accurate diagnosis and appropriate treatment are crucial for patients’ recovery and long-term well-being^{19–21}. By analyzing the treatment recommendations derived from clinical MRI reports, we evaluate whether the advice given by GPT-4 is scientifically sound and clinically safe. We hypothesize that GPT-4 produces largely accurate treatment recommendations yet is at substantial risk of hallucinations and may thus pose a potential risk for patients seeking health advice.

Materials and methods

Study design and dataset characteristics

The local ethical committee (Medical Faculty, RWTH Aachen University, Aachen, Germany, reference number 23/111) approved this retrospective study on anonymized data and waived the requirement to obtain individual informed consent. All methods were carried out in accordance with relevant guidelines and regulations. Following local data protection regulations, the board-certified senior musculoskeletal radiologist with ten years of experience (SN) screened all knee and shoulder MRI studies and associated clinical reports produced during the clinical routine at our tertiary academic medical center (University Hospital Aachen, Aachen, Germany) during February and March of 2023. Ninety-four knee MRI studies and 38 shoulder MRI studies were available for selection. We selected ten studies per joint, ensuring various conditions with variable severity and complexity. Table 1 provides a synopsis of the selected imaging studies with patient demographics, referring disciplines, reasons for the exam, principal diagnoses and treatment recommendations, and a statement on whether the treatment recommendations were considered problematic. Supplementary Table 1 provides more details on the reported diagnoses. Intentionally, we included MRI reports from patients with different demographic characteristics (i.e., age and sex) and referrals from various clinical disciplines. The diagnosis was checked for coherence and consistency using the associated clinical documentation (e.g., history and physical findings) and other non-imaging findings (e.g., laboratory values, intra-operative findings, functional tests, and others). Consequently, MRI studies were disregarded if additional findings were incoherent, inconsistent, or contradictory with the reference diagnosis.

The selected MRI reports were extracted from the local Picture Archiving and Communication System (iSite, Philips Healthcare, Best, Netherlands) as intended for clinical communication, i.e., in German. The MRI reports were anonymized by removing the patient’s name, age, sex, and reference to earlier imaging studies. In the history and reason-for-exam section, any reference that may influence treatment recommendations, e.g., “preoperative evaluation [requested]”, was removed, too.

GPT-4 Encoding and Prompting

GPT-4 was accessed online (<https://chat.openai.com/>) on April 11th and 12th, 2023, and operated as the chatGPT March 23 version. Prompts were provided in a standardized format and the following sequence:

- Prompt #1: Please translate the following MRI report into English.
- Prompt #2: This is the MRI report of a [numerical age]-year-old [sex, woman/man]. Do the conditions need to be treated? And if so, how? Please be as specific as possible.
- Prompt #3: This is too unspecific. Please advise the patient on what to do. Imagine you are the treating physician. Prioritize your treatment recommendations—begin with what is most sensible and relevant.

Consequently, GPT-4 was provided with the patient’s age and sex only. The translated (English) version of the clinical MRI report was checked for overall quality and in terms of accuracy, consistency, fluency, and context by the senior musculoskeletal radiologist (SN), who holds the certificate of the Educational Commission for Foreign Medical Graduates (ECFMG). A new chat session was started for each patient to avoid memory retention bias.

Alongside the MRI reports, the treatment recommendations made by GPT-4 following the initial (prompt #2) and the follow-up request (prompt #3) were saved.

Figure 1 provides an overview of the workflow.

MRI Study	Age	Sex	Referral	Reason for exam	Principal reported diagnoses	Principal treatment recommendation	Treatment recommendation(s) problematic?	
Knee	1	75	Male	INT	Undulating fever and swollen and painful knee	Joint infection with concomitant osteomyelitis and advanced degeneration	Address infection and inflammation (antibiotics) after blood tests	Yes
	2	68	Male	OTS	Worsening symptoms. Medial osteoarthritis	Medial compartmental osteoarthritis with bone changes and meniscus degeneration and chondropathy elsewhere in the joint	Conservative measures. If failing, surgical measures	No
	3	62	Male	OTS	Medial pain since fall several weeks ago	Insufficiency fracture of the medial femoral condyle and cartilage intact	Rest and avoid weight-bearing. Use crutches or cane	Yes
	4	62	Male	GP	Clinical suspicion of medial meniscus lesion	Horizontal posterior root tear of the medial meniscus	Conservative measures. If failing, surgical measures	Yes
	5	52	Female	ED	Joint pain since fall	Medial meniscus tear of posterior horn	Conservative measures. If failing, surgical measures	No
	6	48	Male	OTS	S/p knee dislocation. After reduction	Rupture of the ACL, medial collateral ligament, bucket handle tear of the medial meniscus (O'Donoghue triad), and rupture of the PCL, medial retinaculum, and popliteus muscle	Surgery. Conservative measures while waiting for surgery	Yes
	7	31	Female	OTS	Clinical suspicion of ACL or meniscus damage after skiing accident	Partial rupture/overstretching of the proximal ACL and the MCL. Tibial bone bruise	Conservative measures	No
	8	25	Female	OTS	Radiographic suspicion of bony avulsion of ACL	Bony avulsion of the ACL, impression fracture of the tibial plateau, and radial tear of the lateral meniscus posterior horn	Surgery	Yes
	9	14	Male	PED	S/p ACL reconstruction. Pain after knee distortion during fall	Partial ACL graft rupture and s/p lateral patellar dislocation	Conservative measures	Yes
	10	14	Female	PED	Regular follow-up	Constant osteochondroma of the distal femur (cartilage cap: 0.5 cm)	No treatment if asymptomatic. Surgical removal if symptomatic	Yes
Shoulder	1	68	Female	OTS	Palpable mass over shoulder. No pain	Subcutaneous lipoma and advanced attritive changes of the glenohumeral joint (rheumatic disease)	Conservative measures	Yes
	2	68	Female	GP	Impingement syndrome	Degenerative changes of the GH and AC joints. Partial ruptures of the SSP and SSC tendons	Conservative measures. If failing, surgical measures	Yes
	3	63	Male	OTS	Clinical suspicion of calcific tendinitis. Biceps pathology?	Calcific tendinitis and ISP tendinopathy	Conservative measures	No
	4	61	Male	OTS	S/p second dislocation	Hill-Sachs lesion, glenoid bone loss of 8%, HAGL and GLAD lesions	Conservative measures. If failing, surgical measures	No
	5	60	Male	OTS	Traumatic injury three months ago with supraspinatus tendon tear on ultrasound	Massive rotator cuff tear involving SSP, SSC, and ISP tendons with volume atrophy and fatty infiltration, as well as activated AC joint arthritis	Conservative measures. If failing, surgical measures	No
	6	60	Male	OTS	Acute-on-chronic shoulder pain. Post-instability osteoarthritis	Advanced osteoarthritis of GH and AC joints with SSP and ISP tendinopathy	Conservative measures. If failing, surgical measures	No
	7	34	Male	GP	S/p second dislocation	Hill-Sachs lesion, glenoid bone loss of 15%, ALPSA and GLAD lesions	Conservative measures. If failing, surgical measures	No
	8	47	Female	GP	Clinical suspicion of adhesive capsulitis	Adhesive capsulitis	Conservative measures. If failing, surgical measures	No
	9	32	Male	PS	Palpable swelling over the shoulder. No pain	No pathology, just a strong muscular build	Conservative measures	No
	10	53	Female	INT	Impingement syndrome for six months. Loss of abduction strength	Intratendinous lesion of SSP tendon with bursitis	Conservative measures. If failing, surgical measures	No

Table 1. Synopsis of demographics, referring departments, patient history, and principal diagnoses and treatment recommendations as well as a brief statement on whether the treatment recommendations were considered problematic. Abbreviations: AC—acromioclavicular, ACL—anterior cruciate ligament, ALPSA— anterior labroligamentous periosteal sleeve avulsion of the labrum, ER—emergency department, GH—glenohumeral, GLAD—glenolabral articular disruption, GP—general practitioner, HAGL—humeral avulsion of the glenohumeral ligament, INT—Internal Medicine, ISP—infraspinatus, OTS—orthopedic and trauma surgery, PCL—posterior cruciate ligament, PED – Pediatrics, PS—Plastic Surgery, RD—Referring department, SSC—subscapularis, SSP – supraspinatus. Supplementary Tables 1 and 2 provide further details on the reported diagnoses and treatment recommendations, respectively.

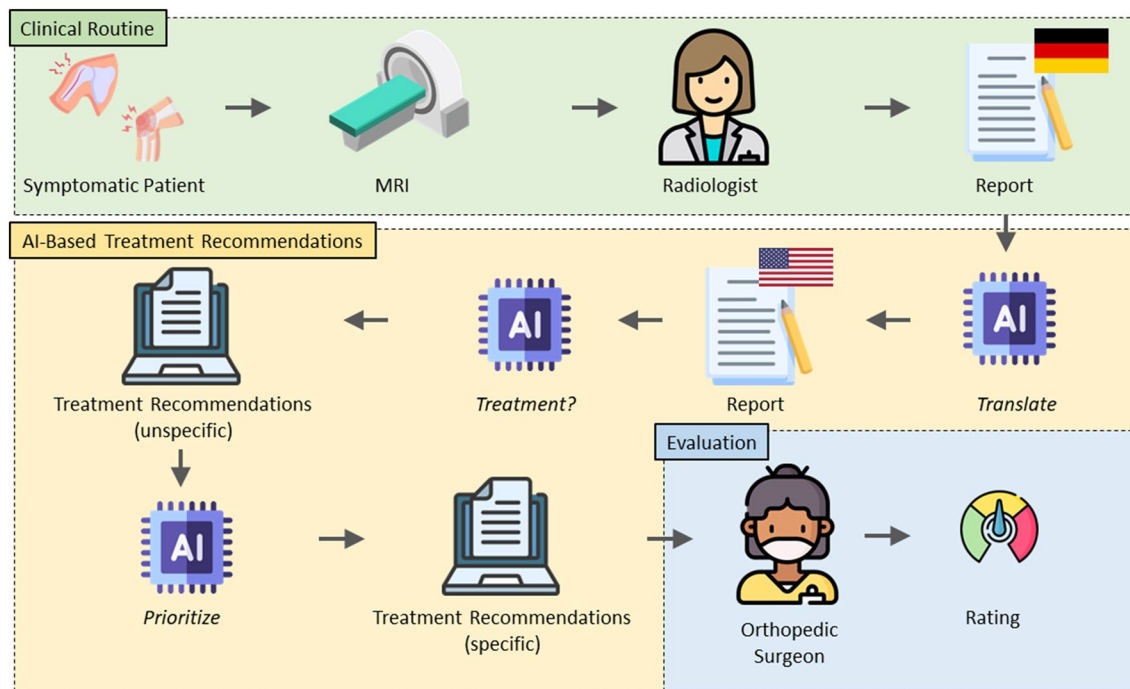


Figure 1. Workflow of the Artificial Intelligence-powered MRI-to-treatment recommendation pipeline. GPT-4, denoted as the AI icon, was prompted three times to translate the MRI report, provide general treatment recommendations, and prioritize its recommendations. Two experienced orthopedic surgeons rated the patient-specific treatment recommendations.

Evaluation of treatment recommendations

Two board-certified and specialty-trained senior orthopedic surgeons with ten (BB) and 12 (CW) years of clinical experience in orthopedic and trauma surgery, evaluated the treatment recommendations made by GPT-4.

Both raters evaluated the treatment recommendations separately by answering the itemized questions in Table 2. Treatment recommendations were rated on Likert scales extending from 1 (*poor or strongly disagree*) to 5 (*excellent or strongly agree*) regarding overall quality, scientific and clinical basis, and clinical usefulness and relevance. Whether the treatment recommendations are up-to-date and consistent was rated on a binary basis, i.e., yes or no.

Afterward, both raters held a consensus meeting where discrepant ratings were discussed until a consensus was reached. Only the consented scores were registered and subsequently analyzed.

Results

In all responses, GPT-4 consistently disclaimed it was not a doctor. GPT-4 offered some general information on the conditions and potential treatment options, yet would not be willing to provide specific medical advice. It repetitiously stressed the importance of consulting with healthcare professionals for personalized treatment recommendations.

GPT-4 explained the MRI findings separately using layman’s language. It continuously worked down the list of findings when formulating its treatment recommendations, following the radiologist’s prioritization.

Question to evaluate	Possible answers
The overall quality of the treatment recommendations is	Poor [1]—Fair [2]—Good [3]—Very good [4]—Excellent [5]
Treatment recommendations are based on scientific and clinical evidence	Strongly disagree [1]—Disagree [2]—Neutral [3]—Agree [4]—Strongly Agree [5]
Treatment recommendations are clinically useful and relevant	Strongly disagree [1]—Disagree [2]—Neutral [3]—Agree [4]—Strongly Agree [5]
Treatment recommendations are up to date	Yes–no
Treatment recommendations are consistent	Yes–no

Table 2. Itemized questions used to rate the treatment recommendations for each MRI report. Two experienced orthopedic surgeons used Likert scales (1 to 5) or binary schemes (yes or no). Additionally, raters were asked to provide (free-text) comments for each patient.

The overall quality of the treatment recommendations was rated as good or better for the knee and shoulder. Similarly, the recommendations were mainly up-to-date and consistent, adhering to clinical and scientific evidence and clinically useful/relevant (Fig. 2). Notably, the treatment recommendations provided for the shoulder were rated more favorably. We did not find signs of hallucinations, i.e., seemingly correct responses that (i) were non-sensical when considered against common knowledge in radiology or orthopedic surgery/traumatology or (ii) inconsistent with framework information or conditions stated in the radiologist's request. Moreover, we did not find signs of speculations or oversimplifications.

GPT-4's treatment recommendations generally followed a schematic approach. In most cases, conservative treatment was recommended initially, regularly accompanied by physical therapy. Surgical treatment was considered a potential option for those patients where conservative treatment, including physical therapy, would not yield satisfactory results. Representative MR images, MRI report findings, and GPT-4-based treatment recommendations are provided for the knee (Fig. 3) and shoulder (Fig. 4).

The two orthopedic surgeons agreed that some recommendations could have been more specific. In numerous patients, GPT-4 provided general advice instead of tailoring the treatment recommendations to the particular condition or patient, thereby limiting their clinical usefulness. Furthermore, GPT-4 tended to err on the side of caution, recommending more conservative treatment options and leaving the decision for surgery to the specialists to be consulted. Supplementary Table 2 provides further details on the treatment recommendations for each patient/MRI report and associated comments by the two orthopedic surgeons.

Discussion

Our study suggests that GPT-4 can produce valuable treatment recommendations for common knee and shoulder conditions. The recommendations were largely up-to-date, consistent, clinically useful/relevant, and aligned with the most recent clinical and scientific evidence.

We observed signs of reasoning and inference across multiple key findings. For example, GPT-4 correctly deduced that meniscus tears may be associated with bone marrow edema (as a sign of excessive load transmission). Hence, its recommendation to "address focal bone marrow edema: As this issue could be related to the medial meniscus tear [...]" was entirely plausible.

Similarly, GPT-4 demonstrated considerable foresight as it recommended organizing post-surgical care and rehabilitation for the patient with multi-ligament knee injuries and imminent surgery. Whether this recommendation can be regarded as "planning" is questionable, though, as true planning abilities in the non-medical domain are still limited^{4,16}. Instead, these recommendations are likely based on the schematic treatment regime that GPT-4 encountered in its training data.

Interestingly, GPT-4 recommended lifestyle modifications, i.e., weight loss and low-impact exercise, and assistive devices (such as braces, canes, or walkers) for shoulder degeneration. While these are sensible and appropriate recommendations for knee osteoarthritis, such recommendations are of doubtful value in shoulder osteoarthritis. In patients with shoulder osteoarthritis or degeneration, exercises to improve the range of motion were not recommended, even though they are indicated²². Again, this observation is likely attributable to the statistical modeling behavior of GPT-4, given the epidemiologic dominance of knee OA over shoulder OA.

Additional limitations of GPT-4 became apparent when the model was tasked to make treatment recommendations for patients with complex conditions or multiple relevant findings.

Critically, the patient with septic arthritis of the knee was not recommended to seek immediate treatment. This particular treatment recommendation, or rather the failure to stress its urgency, is negligent and dangerous. Septic arthritis constitutes a medical emergency, which may lead to irreversible joint destruction, morbidity, and

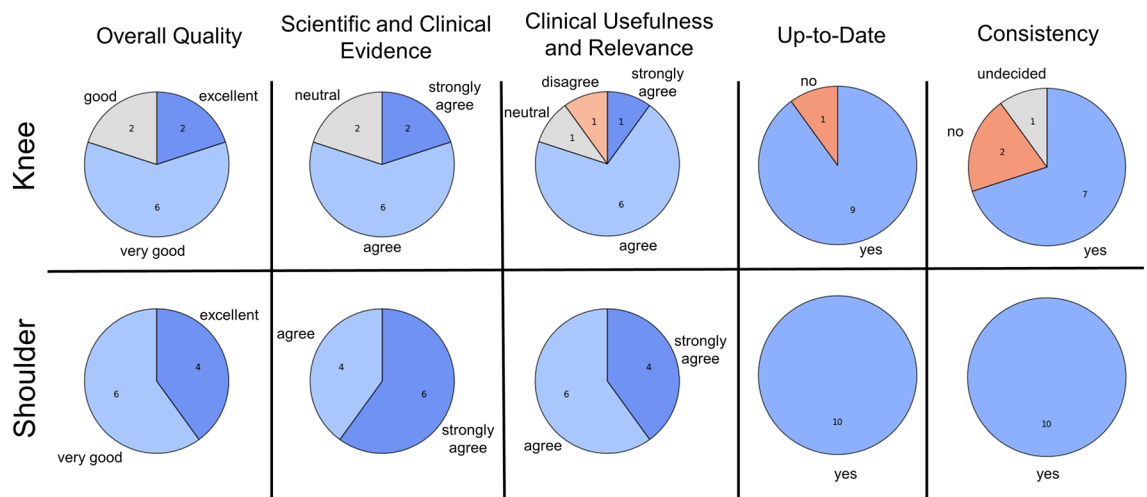


Figure 2. Multidimensional ratings of the treatment recommendations provided by GPT-4. In a consensus meeting, two experienced orthopedic surgeons evaluated the treatment recommendations for various knee and shoulder conditions derived from clinical MRI reports. Ratings were based on five-item Likert scales, and counts were provided only for selected answers.

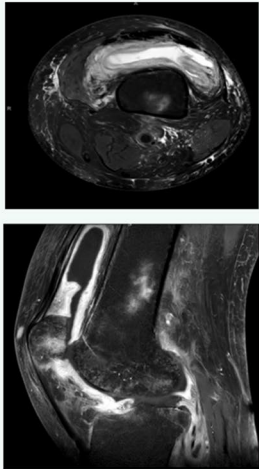
MRI Study	Treatment Recommendations (by GPT-4)
 <p data-bbox="453 667 560 688">Knee MRI #1</p> <ul data-bbox="355 695 635 762" style="list-style-type: none"> • Joint infection and osteomyelitis • Chronic synovitis • Advanced degenerative changes 	<p data-bbox="683 180 1474 233">As a treating physician, I would first perform a thorough physical examination and review the patient's medical history to get a complete understanding of their condition. Based on the MRI findings, my treatment recommendations, prioritized from most sensible and relevant, would include:</p> <p data-bbox="683 247 986 268">Address the suspected infection and inflammation:</p> <ul data-bbox="722 268 1474 338" style="list-style-type: none"> ○ Prescribe appropriate antibiotics for osteomyelitis after obtaining blood tests and potentially a bone biopsy to identify the causative organism and tailor the antibiotic treatment. ○ Consider prescribing nonsteroidal anti-inflammatory drugs (NSAIDs) or corticosteroids to manage pain and inflammation related to chronic synovitis, provided there are no contraindications. <p data-bbox="683 352 1018 373">Manage degenerative joint changes and meniscal tears:</p> <ul data-bbox="722 373 1474 478" style="list-style-type: none"> ○ Recommend a trial of conservative management, including rest, ice, compression, and elevation (RICE), as well as pain-relief medications such as NSAIDs, if not contraindicated. ○ Refer the patient to a physical therapist for a personalized program to improve joint mobility, reduce pain, and strengthen surrounding muscles. ○ If conservative treatments are not effective or the patient's condition worsens, discuss the option of arthroscopic surgery to address the meniscal tears and degenerative changes. <p data-bbox="683 493 842 514">Address cartilage damage:</p> <ul data-bbox="722 514 1474 562" style="list-style-type: none"> ○ Based on the severity of the retropatellar cartilage denudation and focal cartilage defect, discuss potential treatments, such as hyaluronic acid injections, platelet-rich plasma (PRP) therapy, or surgical procedures like cartilage grafting or joint replacement, if conservative treatments fail to provide adequate relief. <p data-bbox="683 577 986 598">Monitor the reactive popliteal lymphadenopathy:</p> <ul data-bbox="722 598 1474 632" style="list-style-type: none"> ○ As the lymphadenopathy may be related to the infection or inflammation in the knee joint, continue to monitor it and reassess after addressing the underlying causes. <p data-bbox="683 646 1002 667">Follow-up and adjust the treatment plan as needed:</p> <ul data-bbox="722 667 1474 701" style="list-style-type: none"> ○ Schedule regular follow-up appointments to monitor the patient's progress, adjust the treatment plan, and address any new concerns or complications that may arise. <p data-bbox="683 716 1474 749">Please note that I am an AI language model and not a licensed physician. The patient should consult with their healthcare provider for personalized advice and treatment recommendations.</p>

Figure 3. Representative knee joint MR images of a patient with a joint infection, key MRI report findings, and specific treatment recommendations by GPT-4. Axial proton density-weighted fat-saturated image above the patella (upper image) and sagittal post-contrast T1-weighted fat-saturated image through the central femur diaphysis (lower image). Of all 20 MRI studies/reports and associated treatment recommendations, these treatment recommendations were rated lowest.

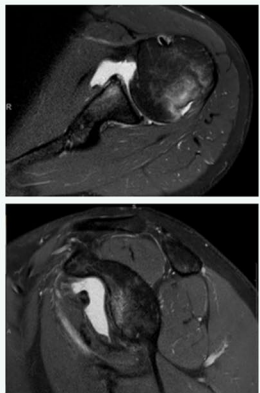
MRI Study	Treatment Recommendations (by GPT-4)
 <p data-bbox="435 1465 568 1486">Shoulder MRI #7</p> <ul data-bbox="355 1493 647 1631" style="list-style-type: none"> • Hill-Sachs lesion (post dislocation) • Anterior labroligamentous periosteal sleeve avulsion (ALPSA) lesion • Glenolabral articular disruption (GLAD) lesion 	<p data-bbox="683 1054 1474 1140">As an AI language model, I am not a physician and cannot provide specific medical advice. However, I can outline a potential approach that a treating physician might consider, given the conditions mentioned in the MRI report. It is essential to consult with an orthopedic surgeon or specialist for personalized advice and treatment options based on the patient's specific case, medical history, and activity level.</p> <p data-bbox="683 1155 1474 1220">Initial conservative treatment: As a first step, the physician might recommend conservative measures, including rest, pain management with anti-inflammatory medications, and activity modification to prevent further injury. This approach allows time for healing and assessing the patient's response to these measures.</p> <p data-bbox="683 1234 1474 1320">Physical therapy and rehabilitation: If the patient shows improvement with conservative treatment, the physician may recommend a targeted physical therapy program. The program would aim to strengthen the shoulder muscles, improve range of motion, and enhance shoulder stability. This phase may also involve proprioception and neuromuscular training to minimize the risk of future dislocations.</p> <p data-bbox="683 1335 1474 1400">Re-assessment and further imaging: After a period of conservative treatment and physical therapy, the physician may re-assess the patient's symptoms and stability. Further imaging (e.g., repeat MRI or CT scan) might be necessary to evaluate the healing progress and determine the next steps in the treatment plan.</p> <p data-bbox="683 1415 1474 1501">Surgical intervention: If the conservative treatment and physical therapy do not provide sufficient improvement in the patient's symptoms or stability, the physician may consider surgical intervention. The choice of surgery would depend on the specific lesions and their severity, as well as the patient's activity level and expectations. Potential surgical options may include:</p> <ol data-bbox="722 1501 1474 1587" style="list-style-type: none"> Arthroscopic Bankart repair for the ALPSA lesion Bone grafting, Latarjet procedure, or remplissage for addressing the Hill-Sachs lesion and glenoidal bone loss Arthroscopic repair for the GLAD lesion <p data-bbox="683 1602 1474 1642">Remember, it is essential to consult with an orthopedic surgeon or specialist for personalized advice and treatment options based on the patient's specific case, medical history, and activity level.</p>

Figure 4. Representative shoulder joint MR images of a patient after re-dislocation, key MRI report findings, and specific treatment recommendations by GPT-4. Axial and parasagittal proton density-weighted fat-saturated images through the humeral head and glenoid, respectively.

mortality. Literature studies report mortality rates of 4% to 42%^{23–25}. Furthermore, because of the stated cartilage damage in this patient, GPT-4 also recommended cartilage resurfacing treatment. However, doing so in a septic joint is contraindicated and medical malpractice²⁶.

GPT-4 was similarly unaware of the patient's overall situation after knee dislocation. Even though the surgical treatment recommendations for multi-ligament knee injuries were plausible, a potential concomitant popliteal

artery injury was not mentioned. It occurs in around 10% of knee dislocations and may dramatically alter treatment².

Remarkably, we did not find signs of so-called "hallucinations", i.e., GPT-4 "inventing" facts and confidently stating them. Even though speculative at this stage, the absence of such hallucinations may be due to the substantial and highly specific information provided in the prompt (i.e., the entire MRI report per patient) and our straightforward prompting strategy compared to more suggestive promptings of other studies¹⁶.

No patient is treated on the basis of the MR images or the MRI report. Nonetheless, using real-patient (anonymized) MRI reports rather than artificial data, increases our study's applicability and impact.

However, while GPT-4 offered treatment recommendations, it is crucial to understand that it is not a replacement for professional medical evaluation and management. The accuracy of its recommendations is largely contingent upon the input's specificity, correctness, and reasoning, which is typically not how a patient would phrase the input and prompt the tool. Therefore, LLMs, including GPT-4, should be used as supplementary resources by healthcare professionals only, as they provide critical oversight and contextual judgment. Optimally, healthcare professionals know a patient's constitution and circumstances to provide effective, safe, and nuanced diagnostic and treatment decisions. Consequently, we caution against the use of GPT-4 by laypersons for specific treatment suggestions.

Along similar lines, integrating LLMs into clinical practice warrants ethical considerations, particularly regarding medical errors. First and foremost, their use does not obviate the need for professional judgment from healthcare professionals who are ultimately responsible for interpreting the LLM's output. As with any tool applied in the clinic, LLMs should only assist (rather than replace) healthcare professionals. However, the safe and efficient application of LLMs requires a thorough understanding of their capabilities and limitations. Second, developers must ensure that their LLMs are rigorously tested and validated for clinical use and that potential limitations and errors are communicated, necessitating ongoing performance monitoring. Third, healthcare institutions integrating LLMs into their clinical workflows should establish governance structures and procedures to monitor performance and manage errors. Fourth, the patient (as a potential end-user) must be made aware of the potential for hallucinations and erroneous and potentially harmful advice. Our study highlights the not-so-theoretical occurrence of harmful advice—in that case, we advocate a framework of shared responsibility. The healthcare professional is immediately responsible for patient care if involved in alleged malpractice. Simultaneously, LLM developers and healthcare institutions share an ethical obligation to maximize the benefits of LLMs in medicine while minimizing the potential for harm. While there is no absolute safeguard against medical errors, informed patients make informed decisions—this applies to LLMs as to any other health resource utilized by patients seeking medical advice.

Importantly, LLMs, including GPT-4, are currently not approved as medical devices by regulatory bodies. Therefore, LLMs cannot and should not be used in the clinical routine. However, our study indicates that the capability of LLMs to make complex treatment recommendations should be considered in their regulation.

Moreover, the recent advent of multimodal LLMs such as GPT-4Vision (GPT-4V) has highlighted the (potentially) vast capacities of multimodal LLMs in medicine. In practice, the text prompt (e.g., original MRI report) could be supplemented by select MR images or additional clinical parameters such as laboratory values. Recent literature evidence studying patients in intensive care confirmed that models trained on imaging and non-imaging data outperformed their counterparts trained on only one data type.²⁷ Consequently, future studies are needed to elucidate the potentially enhanced diagnostic performance as well as the concomitant therapeutic implications.

When evaluating the original MRI report (in German) and its translated version (in English), we observed them to be excellently aligned regarding accuracy, consistency, fluency, and context. This finding is confirmed by earlier literature, indicating an excellent quality of GPT-4-based translations, at least for high-resource European languages such as English and German²⁸. Inconsistent taxonomies in MRI reports may be problematic for various natural language processing tasks but did not affect the quality of report translations in this study.

Our study has limitations. First, we studied only a few patients, i.e., ten patients each for the shoulder and knee. Consequently, our investigation is a pilot study with preliminary results and lacks a solid quantitative basis for statistical analyses. Consequently, no statistical analysis was attempted based on our dataset. Second, to enhance its depth and relevance to clinical scenarios, GPT-4's predictions need to be more specific. Additional 'fine-tuning' and domain-specific training using medical datasets, clinical examples, and multimodal data may enhance its robustness and specificity as well as its overall value as a supplementary resource in healthcare. Third, the patient spectrum was broad. A more thorough performance assessment would require substantially more patients with rare conditions and subtle findings to be included. Fourth, treatment recommendations were qualitatively judged by two experienced orthopedic surgeons. Given the excellent level of inter-surgeon agreement, we consider the involvement of two surgeons sufficient, yet involving three or more surgeons could have strengthened the outcome basis even further. Fifth, the tendency of GPT-4 to give generic and unspecific answers and to err on the side of caution rendered it challenging to assess its adherence to guidelines or best practices exactly. Sixth, we used a standardized and straightforward way of prompting GPT-4. After more extensive modifications of these prompts, outcomes may be different.

In summary, common conditions and associated treatment recommendations were well handled by GPT-4, whereas the quality of the treatment recommendations for rare and more complex conditions remains to be studied. Most treatment recommendations provided by GPT-4 were largely consistent with the expectations of the evaluating orthopedic surgeons. The schematic approach used by GPT-4 often aligns well with the typical treatment progression in orthopedic surgery and sports medicine, where conservative treatments are usually attempted first, and surgical intervention is considered only after the failure of conservative treatments.

Conclusion

In conclusion, GPT-4 demonstrates the potential to provide largely accurate and clinically useful treatment recommendations for common orthopedic knee and shoulder conditions. Expert surgeons rated the recommendations at least as "good", but the patient's situation and treatment urgency were not fully considered. Therefore, patients need to consult healthcare professionals for personalized treatment recommendations, while GPT-4 may be a supplementary resource rather than a replacement for professional medical advice after regulatory approval.

Data availability

Data generated or analyzed during the study (i.e., the original MRI reports) are available from the corresponding author upon reasonable request.

Received: 11 August 2023; Accepted: 14 November 2023

Published online: 17 November 2023

References

- Ruby, D. *ChatGPT Statistics for 2023 (New Data + GPT-4 Facts)*, <<https://www.demandsage.com/chatgpt-statistics/>> (2023).
- Naziri, Q. *et al.* Knee dislocation with popliteal artery disruption: A nationwide analysis from 2005 to 2013. *J. Orthop.* **15**, 837–841 (2018).
- Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2**, e0000198 (2023).
- Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4. <http://arxiv.org/abs/2303.12712> (2023).
- Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. <http://arxiv.org/abs/2303.13375> (2023).
- Rao, A. S. *et al.* Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv* <https://doi.org/10.1101/2023.02.21.23285886> (2023).
- Levine, D. M. *et al.* The diagnostic and triage accuracy of the GPT-3 artificial intelligence model. *medRxiv* <https://doi.org/10.1101/2023.01.30.23285067> (2023).
- Shea, Y.-F., Lee, C. M. Y., Ip, W. C. T., Luk, D. W. A. & Wong, S. S. W. Use of GPT-4 to analyze medical records of patients with extensive investigations and delayed diagnosis. *JAMA Netw. Open* **6**, e2325000–e2325000 (2023).
- Sarraju, A. *et al.* Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* **329**, 842–844 (2023).
- Haver, H. L. *et al.* Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* **307**(4), e230424 (2023).
- Adams, L. C. *et al.* Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: A multilingual feasibility study. *Radiology* **307**, e230725. <https://doi.org/10.1148/radiol.230725> (2023).
- Sun, Z. *et al.* Evaluating GPT-4 on impressions generation in radiology reports. *Radiology* **307**, e231259 (2023).
- Stanzione, A. *et al.* Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. In *Diagnostic and Interventional Radiology (Ankara, Turkey)* (2023).
- Tippareddy, C., Jiang, S., Bera, K. & Ramaiya, N. Radiology reading room for the future: Harnessing the power of large language models like ChatGPT. *Current Probl. Diagn. Radiol.* <https://doi.org/10.1067/j.cpradiol.2023.08.018> (2023).
- Bera, K., O'Connor, G., Jiang, S., Tirumani, S. H. & Ramaiya, N. Analysis of ChatGPT publications in radiology: Literature so far. *Current Probl. Diagn. Radiol.* <https://doi.org/10.1067/j.cpradiol.2023.10.013> (2023).
- Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
- Richardson, J. P. *et al.* Patient apprehensions about the use of artificial intelligence in healthcare. *NPJ. Digit. Med.* **4**, 140 (2021).
- Ayers, J. W. *et al.* Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Int. Med.* <https://doi.org/10.1001/jamainternmed.2023.1838> (2023).
- Vallier, H. A., Wang, X., Moore, T. A., Wilber, J. H. & Como, J. J. Timing of orthopaedic surgery in multiple trauma patients: Development of a protocol for early appropriate care. *J. Orthop. Trauma* **27**, 543–551 (2013).
- Shan, L., Shan, B., Suzuki, A., Nouh, F. & Saxena, A. Intermediate and long-term quality of life after total knee replacement: A systematic review and meta-analysis. *JBJS* **97**, 156–168 (2015).
- Vuurberg, G. *et al.* Diagnosis, treatment and prevention of ankle sprains: Update of an evidence-based clinical guideline. *Br. J. Sports Med.* **52**, 956–956 (2018).
- Chillemi, C. & Franceschini, V. Shoulder osteoarthritis. *Arthritis* **2013** (2013).
- Coakley, G. *et al.* BSR & BHP, BOA, RCGP and BSAC guidelines for management of the hot swollen joint in adults. *Rheumatology* **45**, 1039–1041 (2006).
- Kaandorp, C. J., Krijnen, P., Moens, H. J., Habbema, J. D. & van Schaardenburg, D. The outcome of bacterial arthritis: A prospective community-based study. *Arthritis Rheum* **40**, 884–892. <https://doi.org/10.1002/art.1780400516> (1997).
- Fangtham, M. & Baer, A. N. Methicillin-resistant Staphylococcus aureus arthritis in adults: Case report and review of the literature. *Semin Arthritis Rheum* **41**, 604–610. <https://doi.org/10.1016/j.semarthrit.2011.06.018> (2012).
- Olsen, A. S. & Shah, V. M. Surgical Approaches to Advanced Knee OA (TKA, UKA, Osteotomy). In *Principles of Orthopedic Practice for Primary Care Providers*, 425–435 (2021).
- Khader, F. *et al.* Multimodal deep learning for integrating chest radiographs and clinical parameters: A case for transformers. *Radiology* **309**, e230806 (2023).
- Jiao, W., Wang, W., Huang, J., Wang, X. & Tu, Z. Is ChatGPT a good translator? Yes with GPT-4 as the engine. <http://arxiv.org/abs/2301.08745> (2023)

Acknowledgements

In accordance with the COPE (Committee on Publication Ethics) position statement of 13 February 2023 (<https://publicationethics.org/cope-position-statements/ai-author>), the authors hereby disclose the use of the following artificial intelligence models during the writing of this article. GPT-4 (OpenAI) for checking spelling and grammar.

Author contributions

Guarantors of integrity of entire study, all authors; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting, D.T.; manuscript revision for important intellectual content,

all authors; approval of final version of submitted manuscript, all authors; literature research, D.T., S.N.; experimental studies, D.T., C.K., C.D.W., B.J.B., S.N., D.T.; data analysis, D.T., S.N.; manuscript editing: all authors.

Funding

Open Access funding enabled and organized by Projekt DEAL. JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMV11-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERANET Transcan), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (Transplant.KI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312) and the National Institute for Health and Care Research (NIHR, NIHR213331) Leeds Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. DT is supported by the European Union's Horizon Europe programme (ODELIA, 101057091), by grants from the Deutsche Forschungsgemeinschaft (DFG) (TR 1700/7-1), and the German Federal Ministry of Education and Research (SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A). SN is funded by grants from the Deutsche Forschungsgemeinschaft (DFG) (NE 2136/3-1). KB is supported by the European Union's Horizon Europe programme (COMFORT, 101079894).

Competing interests

JNK declares consulting services for Owkin, France; DoMore Diagnostics, Norway and Panakeia, UK and has received honoraria for lectures by Eisai, Roche, MSD and Fresenius. The other authors report no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-47500-2>.

Correspondence and requests for materials should be addressed to S.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023, corrected publication 2024