BMJ Open

Quantitative evaluation of GPT-4's performance on US and Chinese osteoarthritis treatment guideline interpretation and orthopaedic case consultation

Juntan Li ,[1,2] Xiang Gao,[3] Tianxu Dou,[4] Yuyang Gao,[4] Xu Li,[3] Wannan Zhu[1]

To cite: Li J, Gao X, Dou T, *et al*. Quantitative evaluation of GPT-4's performance on US and Chinese osteoarthritis treatment guideline interpretation and orthopaedic case consultation. *BMJ Open* 2024;**14**:e082344. doi:10.1136/bmjopen-2023-082344

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-082344).

Received 21 November 2023
Accepted 28 November 2024

[1]Jinzhou Medical University, Jinzhou, Liaoning, China
[2]The First Affiliated Hospital of China Medical University, Shenyang, Liaoning, China
[3]Department of Orthopedics, Fourth Affiliated Hospital of China Medical University, Shenyang, Liaoning, China
[4]Department of Orthopedics, The First Hospital of China Medical University, Shenyang, China

Correspondence to
Dr Wannan Zhu;
17173854@qq.com

ABSTRACT

Objectives To evaluate GPT-4's performance in interpreting osteoarthritis (OA) treatment guidelines from the USA and China, and to assess its ability to diagnose and manage orthopaedic cases.

Setting The study was conducted using publicly available OA treatment guidelines and simulated orthopaedic case scenarios.

Participants No human participants were involved. The evaluation focused on GPT-4's responses to clinical guidelines and case questions, assessed by two orthopaedic specialists.

Outcomes Primary outcomes included the accuracy and completeness of GPT-4's responses to guideline-based queries and case scenarios. Metrics included the correct match rate, completeness score and stratification of case responses into predefined tiers of correctness.

Results In interpreting the American Academy of Orthopaedic Surgeons and Chinese OA guidelines, GPT-4 achieved a correct match rate of 46.4% and complete agreement with all score-2 recommendations. The accuracy score for guideline interpretation was 4.3±1.6 (95% CI 3.9 to 4.7), and the completeness score was 2.8±0.6 (95% CI 2.5 to 3.1). For case-based questions, GPT-4 demonstrated high performance, with over 88% of responses rated as comprehensive.

Conclusions GPT-4 demonstrates promising capabilities as an auxiliary tool in orthopaedic clinical practice and patient education, with high levels of accuracy and completeness in guideline interpretation and clinical case analysis. However, further validation is necessary to establish its utility in real-world clinical settings.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ The study uses a systematic, quantitative approach to evaluate GPT-4's performance in interpreting treatment guidelines from two different healthcare systems.
⇒ The methodology includes a rigorous assessment of accuracy and completeness, based on predefined scoring systems for guideline interpretation and case response analysis.
⇒ Comprehensive assessment was performed, considering not only GPT-4's accuracy but also the completeness of its responses, providing a holistic evaluation of its capabilities.
⇒ Evaluations were conducted in a simulated environment using text prompts, which may not fully reflect GPT-4's performance in real-world clinical settings.
⇒ The study's focus was limited to a narrow subset of orthopaedic knowledge, potentially limiting the generalisability of findings across the entire orthopaedic field or other medical disciplines.

INTRODUCTION

Large language models (LLMs) refer to a type of machine learning algorithm designed to generate text that mimics human-like semantic and syntactic structures. These models, trained on large data sets of internet text, use transformer-based algorithms, such as the Generative Pretrained Transformer (GPT) series pioneered by OpenAI.[1] Using patterns learnt during training, these models interpret contextual input and predict the next word in a sentence.[2 3] LLMs have shown great potential in various applications. A notable example is ChatGPT, which demonstrates impressive human-like expression and reasoning. Its use cases span tasks such as drafting emails, writing code, creative writing and even translating complex medical terminology into simple language understandable by laypeople.[4 5] Furthermore, it has been used as a tool to prepare for medical board exams, showing its great potential in education.[6 7]

GPT-4, as the most recent version in the GPT series initiated by OpenAI, constitutes a notable progress in the sphere of LLMs.[8 9] Compared with its predecessor, GPT-4 has shown improved performance in numerous tasks.[10 11] Research has shown that GPT-4 surpasses ChatGPT in medical board exam simulations, demonstrating higher

precision and better comprehension of complex, high-level questions. This infers enhanced abilities on the part of GPT-4 in context comprehension and problem resolution.[12 13] Despite these remarkable abilities, it is important to acknowledge that LLMs, including GPT-4, do not understand text in the same way humans do. They lack consciousness, and any statements they generate about the world require fact-checking for accuracy. As a result, the model may produce incorrect information because of its inherent 'illusions'.

Osteoarthritis (OA) is a chronic degenerative joint disease that poses a significant public health challenge due to its high prevalence and disability rate.[14 15] There are multiple treatment options for OA, including non-pharmacological approaches like physical therapy and lifestyle changes. In addition, pharmacological treatments and surgical procedures are also available.[16] Self-education plays a crucial role in managing OA as well-informed patients are more likely to actively participate in their care, follow treatments and achieve better health outcomes.[17]

This study investigated the potential of GPT-4 in the field of OA. We assessed the accuracy and completeness of GPT-4's responses by comparing them with established treatment guidelines from both China and the USA. A primary objective was to evaluate the feasibility of using GPT-4 as a tool to support patient education and assist clinicians. Additionally, we examined GPT-4's performance in diagnosing and recommending treatment for orthopaedic conditions.

METHODS

Data source

The present study used multiple data sources to evaluate GPT-4's performance. These sources include the *Evidence-Based Clinical Practice Guideline for the Management of Osteoarthritis of the Knee (Non-Arthroplasty)*, issued by the American Academy of Orthopaedic Surgeons (AAOS) in 2021, which provides 28 recommendations for OA management, organised into four-star categories for clarity and visualisation.[18] Additionally, we used the *2019 Chinese Guidelines for Osteoarthritis Diagnosis and Treatment*, developed by the Chinese Orthopaedic Association (COA), which includes 30 recommendations addressing key clinical concerns and categorises them into A, B and C levels based on recommendation strength.[19] Finally, 50 case analysis questions were selected from the Chinese Orthopaedic Specialist Examination question repository through random sampling, using a computer-generated random number. Figure 1 illustrates the study flow chart.

Prompt and response generation

A prompt acts as the steering wheel in language models, guiding the direction of the generated response and affecting the quality, relevance and safety of the output. In this study, GPT-4 was not explicitly instructed to refer to specific guidelines within the prompt. The prompts for the AAOS guideline and COA are provided in English.

Within the context of the AAOS guidelines, GPT-4 is directly interrogated based on the specifics of these recommendations. An exemplar query could be, 'Are canes recommended for improving function and quality of life for osteoarthritis patients?' Considering the formidable reasoning and logical capabilities of GPT-4, we further probe, 'Given that a 4-star rating represents the pinnacle of recommendation, how many stars would you accord this particular recommendation?' Subsequently, the responses generated by GPT-4 are compared with the established guidelines for comparison. In relation to the Chinese OA guidelines, GPT-4 was directly queried using the 15 clinically pertinent questions outlined within these

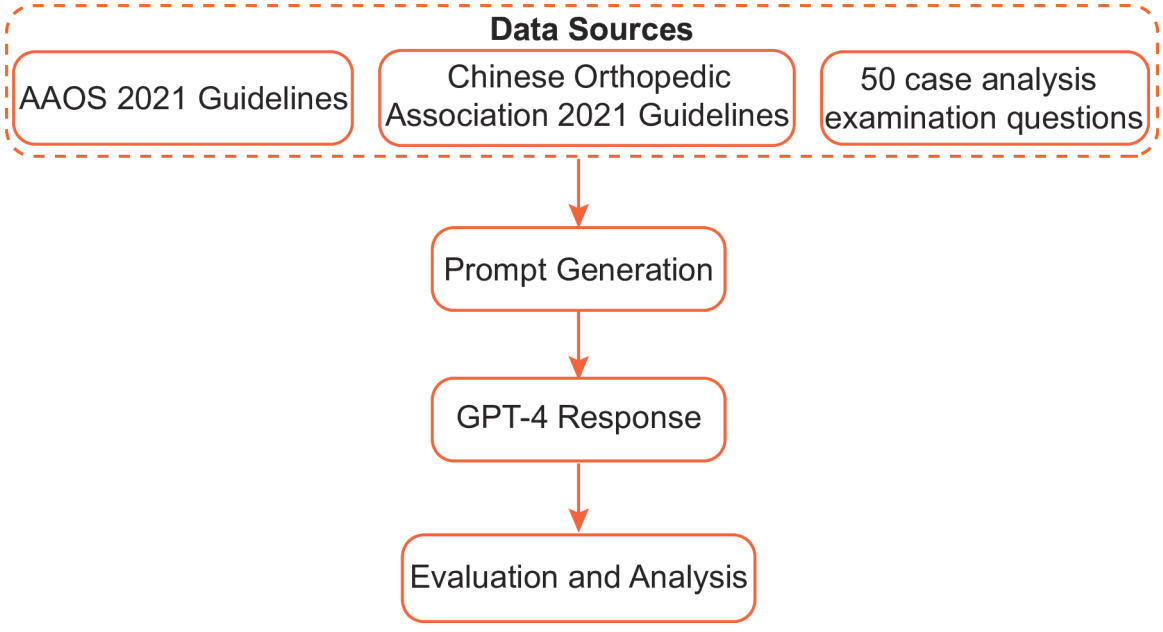

Figure 1 Flow chart of the study. AAOS, American Academy of Orthopaedic Surgeons.

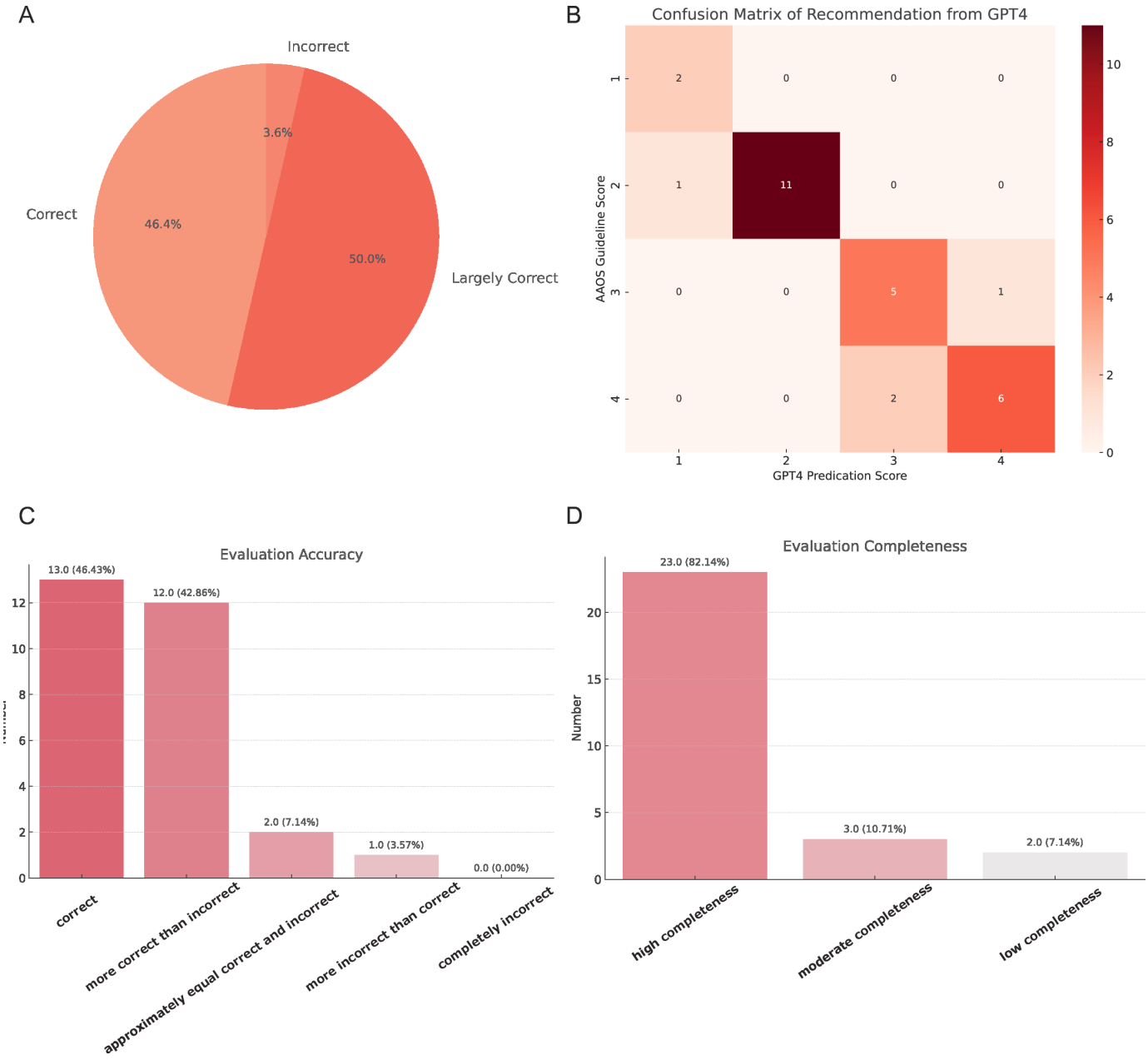

Figure 2 (A) The pie chart shows the accuracy rate of recommended level predicted by GPT-4. (B) The confusion matrix compares guideline-recommended level with those predicted by GPT-4. (C) The bar chart shows Likert scale score distribution of accuracy. (D) The bar chart shows Likert scale score distribution of completeness.

guidelines. The subsequent analysis focused on delineating the differences between GPT-4's responses and the recommendations explicitly enumerated in the guidelines. With respect to testing the case inquiry abilities, we initially provide case information, after which GPT-4 is assigned to respond to these cases concerning further radiological examinations, primary diagnoses and therapeutic strategies. This procedure is intended to assess its potential effectiveness as an adept assistant in the field of orthopaedics.

Patient and Public Involvement

No patient was involved in the study. Two independent evaluators, Senior Physical Therapist Wannan Zhu and Associate Professor Xiang Gao, each with over 10 years of clinical experience in OA, assessed the accuracy and completeness of GPT-4's responses. In cases of discrepancies, Professor Xu Li, with over 20 years of clinical experience, was consulted to determine the final ranking.

For accuracy, responses were deemed 'accurate' if they aligned with the GCP guidelines and 'inaccurate' if there were any deviations. A 5-point Likert scale was used to evaluate accuracy (5=correct, 4=more correct than incorrect, 3=equal parts correct and incorrect, 2=more incorrect than correct, 1=completely incorrect). For completeness, a set of key points was defined. Responses were marked 'complete' if they included all necessary

elements and 'incomplete' if any were missing. A 3-point Likert scale was employed to measure completeness (3=high completeness, 2=moderate completeness, 1=low completeness). The standard of evaluation is shown in online supplemental table 1.

In case inquiries, GPT-4's responses were classified into four tiers: 4=comprehensive, 3=correct but inadequate, 2=mixed with correct and incorrect or outdated data and 1=completely incorrect. This classification helped evaluate GPT-4's ability to identify orthopaedic pathologies.

Statistical analysis

In our statistical analysis, the comparative data were systematically organised using Excel, facilitating a clear delineation of GPT-4's responses across specific categories. With the aid of GPT-4's advanced data analysis module (ChatGPT August 3 version), we were able to compute essential descriptive statistics such as means, SD, frequencies and percentages. For a more in-depth understanding, we employed the same module to generate comprehensive visualisations, prominently featuring pie charts, confusion matrices, and bar graphs.

RESULTS

AAOS guideline

In the AAOS guidelines, recommendations related to OA are ranked from 1 to 4, and GPT-4 also assigns ratings to recommendations on a similar scale of 1–4 (online supplemental table 2). Occasionally, GPT-4 may provide a neutral rating, such as 2 or 3. In such instances, we categorise it as 'largely correct'. If the ratings completely match, they are deemed 'correct', while completely different ratings are labelled 'incorrect'. As depicted in figure 2A, the correct match is at 46.4%, while largely correct ratings account for 50%. Figure 2B presents a confusion matrix comparing guideline rankings with those predicted by GPT-4. Specifically, when the AAOS guideline suggested a score of 1, GPT concurred in 66.7% of the cases. Impressively, for an AAOS recommendation of score 2, GPT-4 showed complete agreement, matching the score in 100% of instances. Similarly, with an AAOS recommendation of score 3, GPT-4 aligned in 83.3% of the cases. When the AAOS guideline indicated a score of 4, GPT-4 mirrored this recommendation in 75% of the instances.

Figure 2C, D delineates the distribution of Likert scores for both accuracy and completeness. Out of the 28 responses generated by GPT-4, the average score for precision was 4.3±1.6, and the average score for completeness stood at 2.8±0.6 (table 1). The scores pertaining to accuracy did not exhibit significant variances across different levels of evidence or recommendation gradings.

Chinese guideline for diagnosis and treatment of osteoarthritis

In the COA, 15 key questions were proposed with respect to which experts succinctly formulated 30 recommendations. In this study, these 15 questions were directly input into GPT-4 to explore the accuracy and completeness of its answers in relation to the 30 recommendations (online supplemental table 3). Among the 30 recommendations, 11 were rated as A-level, 11 as B-level and 8 as C-level. In terms of accuracy, the average scores of the three levels in GPT-4's responses were 4.0±0.6, 4.5±0.6 and 4.5±0.7, respectively. In terms of completeness, the average scores of the three levels were 2.9±0.3, 2.3±0.9 and 2.1±1.0, respectively (table 1). As shown in figure 3A, B, most of the responses possess high accuracy, suggesting that GPT-4 provides comprehensive and precise answers to questions related to OA, reflecting a thorough understanding of OA. A confusion matrix visually presenting the results of evaluation by the two assessors is provided in online supplemental figure 1.

Table 1 GPT-4 accuracy and completeness against osteoarthritis guideline from the USA and China

Osteoarthritis guideline	Accuracy (5 points)	Completeness (3 points)
American Academy of Orthopaedic Surgeons	4.3±1.6	2.8±0.6
Chinese guideline		
Grade A	4.0±0.6	2.9±0.3
Grade B	4.5±0.6	2.3±0.9
Grade C	4.5±0.7	2.1±1.0

Case inquiry ability evaluation

In this study, we randomly selected 50 common orthopaedic clinical cases, using GPT-4 for case analysis to generate responses regarding further radiological examinations, diagnosis and treatment (online supplemental table 4). Across the three categories, GPT-4's average scores were 3.78±0.52, 3.82±0.48 and 3.8±0.6, respectively. Figure 4A displays the quality of GPT-4's responses, revealing a high level of performance across all categories, with over 88% of responses being comprehensive. GPT-4 only committed an error in the 'Further radiological examinations' category in the case of peroneal nerve paralysis post-knee arthroplasty. Additionally, a diagnostic error occurred in the case of lumbar tuberculosis, which subsequently led to an incorrect treatment suggestion.

DISCUSSION

The advent of artificial intelligence, specifically GPT-4, offers transformative potential across various fields, including medicine.[20–22] As an emerging innovation, GPT-4 requires thorough exploration and validation before being integrated into patient healthcare services. In this study, we aimed to evaluate GPT-4's efficacy in accordance with OA treatment guidelines from the USA and China, as well as its ability to address orthopaedic case inquiries.

The results demonstrate the potential utility and effectiveness of GPT-4 in orthopaedics, particularly in

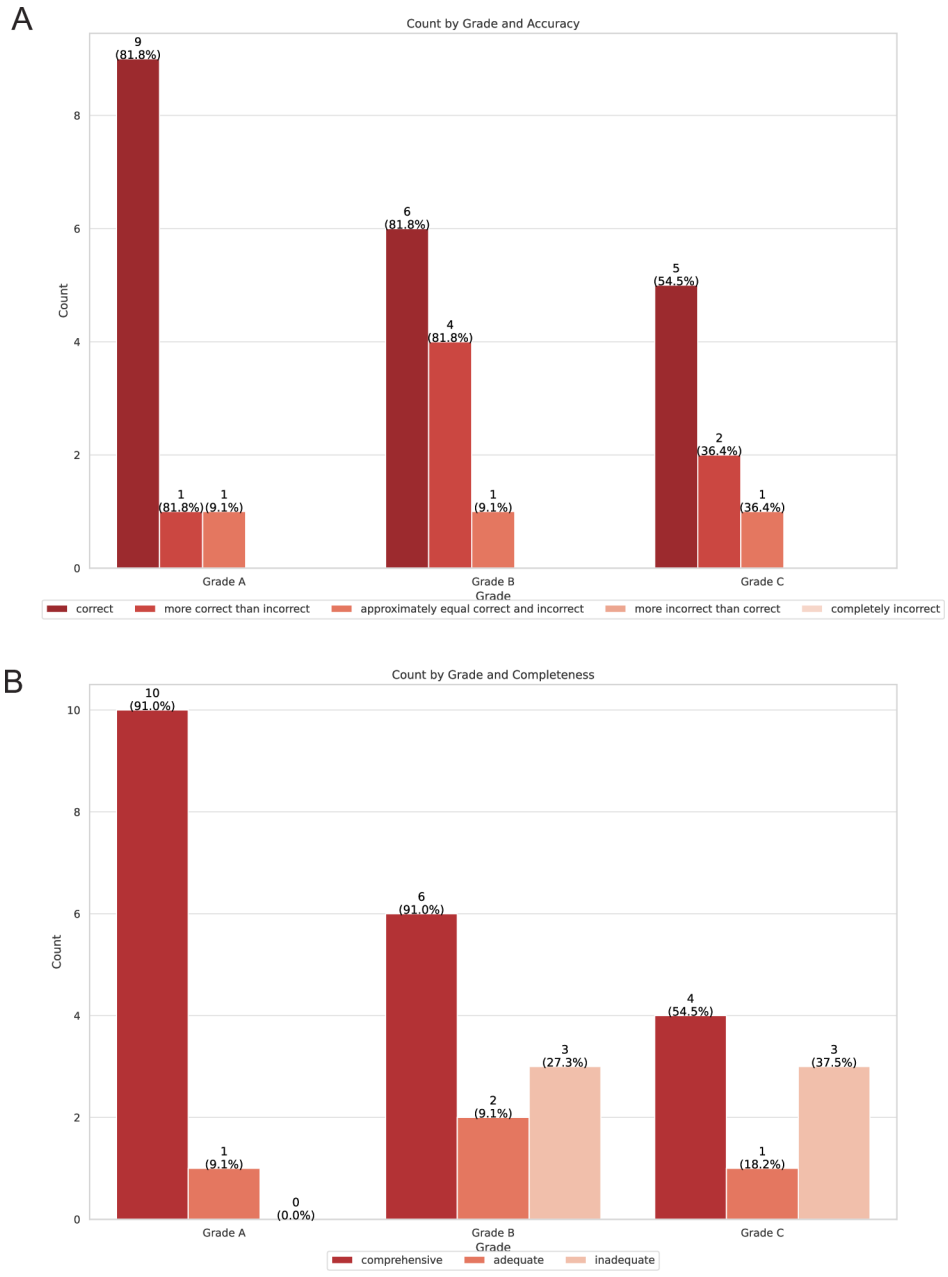

Figure 3 (A) The clustered bar graph illustrates the accuracy of GPT-4's responses in proposing suggestions for Grade A, B and C levels.(B) The clustered bar graph illustrates the completeness of GPT-4's responses in proposing suggestions for Grade A, B and C levels.

managing OA. GPT-4's impressive performance in interpreting OA guidelines, answering questions and handling clinical cases highlights its potential as a valuable tool in orthopaedic practice. The evaluation of its case inquiry ability further underscores its potential for clinical case analysis. Although there were some errors, GPT-4's overall performance in recommending radiological examinations, providing diagnoses and suggesting treatment plans was highly commendable. It is important to note that GPT-4 sometimes provides citations for its viewpoints, such as referencing American College of Rheumatology (ACR) guidelines. However, its viewpoints sometimes do not align with the cited sources and may even include incorrect information. Therefore, we cannot fully rely on its responses yet.

GPT-4 indeed exhibits remarkable outcomes. For instance, it evinces a profound comprehension of the utility of traditional Chinese medicine and herbal therapies in the investigation and management of OA. In undertaking additional assessments for instances of post-joint replacement infection, GPT-4 explicitly articulates that C reactive protein and erythrocyte sedimentation rate tests are required in conjunction with radiological examinations. Notably, through text-based case analysis alone, it possesses the capability to diagnose Felty's syndrome accurately, a rare autoimmune disorder typically prevalent

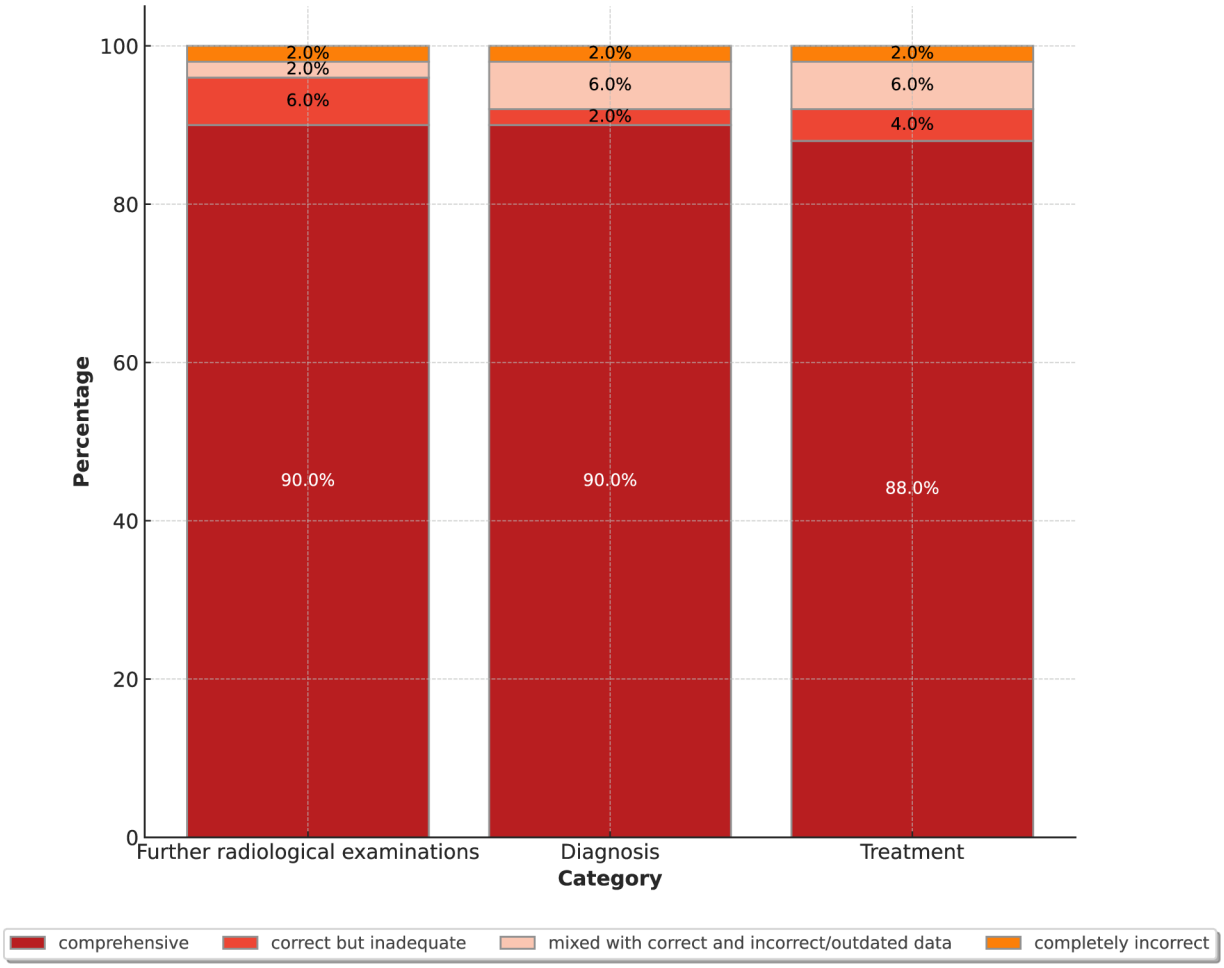

Figure 4 The stacked bar chart shows the comprehensive level of GPT-4's answers in the areas of further radiological examinations, primary diagnoses and treatment.

among individuals suffering from severe rheumatoid arthritis. However, GPT-4's response on this topic presents some discrepancies. While the model correctly identifies acetaminophen as a commonly used over-the-counter medication for pain relief with a favourable safety profile, it inaccurately references the 2019 guidelines from the ACR and the Arthritis Foundation, suggesting that acetaminophen is conditionally recommended against for managing OA of the hand, hip and knee. After a thorough review, we found that acetaminophen remains recommended in the 2019 ACR guidelines, which highlights a gap between the AI-generated response and the actual evidence-based recommendations. This inconsistency underscores the importance of verifying AI-generated medical information, particularly when it seems well-founded but diverges from established guidelines. In addition, GPT-4 demonstrated a consistent pattern of inaccurate diagnoses when applied to bone tumour cases, including conditions like osteochondroma and osteosarcoma. While the model appropriately recognised the need for additional diagnostic tests, its ultimate diagnostic recommendations were frequently incorrect. This discrepancy may be due to the relatively low incidence of bone tumours, leading to limited exposure in the model's training data. As a result, GPT-4's diagnostic reliability in this area appears compromised, indicating that its performance may be more robust in more common orthopaedic conditions and weaker in rarer, less represented cases. This highlights the importance of further refinement and data set enhancement to improve GPT-4's diagnostic capabilities in rare orthopaedic diseases.

GPT-4's capabilities can offer valuable support in various clinical scenarios, as demonstrated in the following applications: (1) acting as a helpful tool for orthopaedic surgeons to quickly understand and apply treatment guidelines, aiding in evidence-based clinical decision-making; (2) enhancing the clinical knowledge and case analysis skills of orthopaedic physicians through case-based training and (3) using GPT-4 to improve patient education by providing clear explanations of medical conditions and treatment plans. In the distinctive healthcare landscape of China, characterised by healthcare disparities and resource constraints, GPT-4 could play a pivotal role in bolstering healthcare delivery, particularly in primary care settings and rural clinics. In China's unique healthcare landscape, marked by disparities and resource limitations, GPT-4 could play a crucial role in improving healthcare delivery, particularly in primary

care and rural clinics. It can assist rural physicians and grassroots hospitals in the initial assessment and diagnosis of OA. Furthermore, the uneven levels of medical education across different universities in China highlight GPT-4's potential in narrowing the educational gap. GPT-4 could provide medical students and clinicians with valuable resources for understanding clinical guidelines and analysing cases, thereby raising the overall standard of medical education and practice. However, in real-world applications, careful supervision of GPT-4's recommendations by physicians is essential to avoid over-reliance on its automated outputs, ensuring accurate and personalised healthcare services.

Other researchers from various medical fields have also explored the response capabilities of GPT-4, resulting in a myriad of perspectives. For example, Yoshiyasu *et al* evaluated GPT-4's accuracy and completeness against the International Consensus Statement on Allergy and Rhinology: Rhinosinusitis.[23] However, only 54% of GPT-4's responses achieved full marks in accuracy, and 71% received full marks in completeness. Yeo *et al* used ChatGPT (GPT-3.5) to inquire about two diseases, cirrhosis and hepatocellular carcinoma.[5] Both diseases achieved over 70% accuracy full marks and more than 40% completeness full marks. The authors believe that ChatGPT may serve as an adjunct informational tool for patients and physicians to improve outcomes. An innovative study demonstrates that the integration of ChatGPT enables surgeons to confidently and calmly manage mpox (monkeypox) patients and future epidemics, thereby enhancing clinical decision-making and improving patient outcomes.[24] Another study highlights that the integration of ChatGPT/GPT-4 in spinal surgery practice enhances perioperative management, improves communication, supports real-time decision-making and assists in postoperative rehabilitation, leading to improved patient outcomes and more efficient clinical workflows.[25]

In the field of orthopaedics, although specific data are not yet available, there are already scholars who have made a certain degree of forecasts. For instance, GPT-4 can assist doctors in five areas of joint replacement: scientific research, disease diagnosis, treatment options, preoperative planning, intraoperative support and postoperative rehabilitation.[26] In sports medicine, GPT-4 can contribute to diagnostic imaging, exercise prescription, medical supervision, surgical treatment, sports nutrition and scientific research.[27] The author believes that while GPT-4 will not replace doctors, it could become an indispensable scientific assistant for sports doctors in the future.

However, while these findings are promising, it is important to approach the integration of AI tools like GPT-4 in healthcare with caution. A few errors identified in the case analysis suggest that the tool is not infallible and should not be relied on blindly. Human oversight and supervision remain essential, particularly in complex and nuanced clinical scenarios. It is also important to consider that the tool's performance could be influenced by the quality and specificity of the input data provided. Therefore, continued research and monitoring of GPT-4's performance in different clinical situations and contexts is necessary. In future research, we plan to evaluate GPT-4 as a patient education tool by comparing it with traditional verbal education methods provided by doctors and nurses. This comparison aims to provide a more comprehensive assessment of GPT-4's potential impact in real-world medical settings, particularly in improving patient understanding and engagement.

LIMITATIONS

Despite the promising results, this study has certain limitations that should be acknowledged. Since evaluations were conducted in a simulated environment with textual prompts, the real-world clinical performance of GPT-4 remains unclear. Its use in more complex patient cases could reveal limitations that were not evident in this initial analysis. Furthermore, the study focused on a narrow subset of orthopaedic knowledge, and its capabilities across the full field have yet to be fully explored. GPT-4's performance also depends heavily on the quality of training data, and biases in the data may affect its effectiveness, requiring continuous updates. Additionally, the subjectivity of Likert scale assessments and the small number of evaluators may affect the reliability of the results. Future research should include broader clinical scenarios, larger reviewer samples and objective measures to enhance validity. The real-world integration of GPT-4 into orthopaedic care must be approached cautiously, with expert supervision essential to mitigate potential errors.

CONCLUSION

In conclusion, this study offers initial evidence of GPT-4's potential as an orthopaedic assistant, showing strong performance in interpreting OA guidelines and analysing clinical cases. The results suggest that GPT-4 could be useful for patient education, training junior physicians and supporting clinical decision-making. However, errors in complex cases underscore the importance of caution and expert oversight before real-world implementation. While promising, further technical refinement and thorough validation across various clinical settings are crucial to understanding the full capabilities and limitations of LLMs like GPT-4 in healthcare. Expert supervision remains essential due to the risk of inaccuracies.

Acknowledgements We extend our sincere appreciation to Mr. Bingqiang Zhan from Aigclink for generously providing us with access to a GPT-4 account.

Contributors JL is responsible for the overall content as guarantor and was responsible for the experimental design. TD and YG took charge of data collection. XG, WZ, XL evaluated the results generated by GPT-4. All authors participated in drafting and revising the manuscript and have read and approved the final version of the manuscript. GPT-4 was employed in two capacities in this study: first, as a research tool to evaluate its performance in interpreting osteoarthritis treatment guidelines from the USA and China and in orthopaedic case consultation; second, to enhance the language and clarity of this manuscript.

Funding This work was supported by the Basic Scientific Research Project of the Liaoning Provincial Department of Education, Project Number LJ212410159035.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. Not applicable. All data are included within the article or uploaded as supplementary information.

ORCID iD

Juntan Li http://orcid.org/0000-0003-1487-4297

REFERENCES

1 Wang SH. OpenAI - explain why some countries are excluded from ChatGPT. *Nature New Biol* 2023;615:34.
2 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. advances in neural information processing systems. 2017.30.
3 Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care* 2023;27:75.
4 Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proc Natl Acad Sci U S A* 2023;120:e2305016120.
5 Yeo YH, Samaan JS, Ng WH, *et al*. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023;29:721–32.
6 Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel)* 2023;11:887.
7 Ali R, Tang OY, Connolly ID, *et al*. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* 2023;93:1090–8.
8 Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N Engl J Med* 2023;388:1233–9.
9 Graham F. Daily briefing: What scientists think of GPT-4, the new AI chatbot. *Nature New Biol* 2023.
10 Sun Z, Ong H, Kennedy P, *et al*. Evaluating GPT4 on Impressions Generation in Radiology Reports. *Radiology* 2023;307:e231259.
11 Bhayana R, Bleakney RR, Krishna S. GPT-4 in Radiology: Improvements in Advanced Reasoning. *Radiology* 2023;307:e230987.
12 Kumah-Crystal Y, Mankowitz S, Embi P, *et al*. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? *J Am Med Inform Assoc* 2023;30:1558–60.
13 Lyu Q, Tan J, Zapadka ME, *et al*. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art* 2023;6:9.
14 Glyn-Jones S, Palmer AJR, Agricola R, *et al*. Osteoarthritis. *Lancet* 2015;386:376–87.
15 Wood G, Neilson J, Cottrell E, *et al*. Osteoarthritis in people over 16: diagnosis and management-updated summary of NICE guidance. *BMJ* 2023;380:24.
16 Arden NK, Perry TA, Bannuru RR, *et al*. Non-surgical management of knee osteoarthritis: comparison of ESCEO and OARSI 2019 guidelines. *Nat Rev Rheumatol* 2021;17:59–66.
17 Bennell KL, Lawford BJ, Keating C, *et al*. Comparing Video-Based, Telehealth-Delivered Exercise and Weight Loss Programs With Online Education on Outcomes of Knee Osteoarthritis : A Randomized Trial. *Ann Intern Med* 2022;175:198–209.
18 Brophy RH, Fillingham YA. AAOS Clinical Practice Guideline Summary: Management of Osteoarthritis of the Knee (Nonarthroplasty), Third Edition. *J Am Acad Orthop Surg* 2022;30:e721–9.
19 Zhang Z, Huang C, Jiang Q, *et al*. Guidelines for the diagnosis and treatment of osteoarthritis in China (2019 edition). *Ann Transl Med* 2020;8:1213.
20 The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Dig Health* 2023;5:e102.
21 Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6:1169595.
22 Temsah O, Khan SA, Chaiah Y, *et al*. Overview of Early ChatGPT's Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts. *Cureus* 2023;15:e37281.
23 Yoshiyasu Y, Wu F, Dhanda AK, *et al*. GPT-4 accuracy and completeness against International Consensus Statement on Allergy and Rhinology: Rhinosinusitis. *Int Forum Allergy Rhinol* 2023;13:2231–4.
24 He Y, Wu H, Chen Y, *et al*. Can ChatGPT/GPT-4 assist surgeons in confronting patients with Mpox and handling future epidemics? *Int J Surg* 2023;109:2544–8.
25 He Y, Tang H, Wang D, *et al*. Will ChatGPT/GPT-4 be a Lighthouse to Guide Spinal Surgeons? *Ann Biomed Eng* 2023;51:1362–5.
26 Cheng K, Li Z, Li C, *et al*. The Potential of GPT-4 as an AI-Powered Virtual Assistant for Surgeons Specialized in Joint Arthroplasty. *Ann Biomed Eng* 2023;51:1366–70.
27 Cheng K, Guo Q, He Y, *et al*. Artificial Intelligence in Sports Medicine: Could GPT-4 Make Human Doctors Obsolete? *Ann Biomed Eng* 2023;51:1658–62.