Clinical Interventions in Aging

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Exploring and Comparing the Use of Large Language Models in Supporting Osteoporosis Health Consultations

Xin Li[1,*], Gen Li[2,*], Yue Zhao[3], Yixin Liang[4], Yuefu Dong[1], Jian Zhang[1]

[1]Department of Orthopedics, The First People's Hospital of Lianyungang, Lianyungang, Jiangsu, People's Republic of China; [2]Department of Orthopedics, The Second Affiliated Hospital of Xuzhou Medical University, Xuzhou, Jiangsu, People's Republic of China; [3]Department of Nursing, Lianyungang Maternity and Child Health Hospital, Lianyungang, Jiangsu, People's Republic of China; [4]Department of Osteoporosis, The First People's Hospital of Lianyungang, Lianyungang, Jiangsu, People's Republic of China

*These authors contributed equally to this work

Correspondence: Yuefu Dong; Jian Zhang, Department of Orthopedics, The First People's Hospital of Lianyungang, Lianyungang, Jiangsu, People's Republic of China, Email dongyuefu@163.com; lygyyzj@163.com

Purpose: To compare the medical accuracy and content comprehensiveness of three large language models (LLMs) in generating responses to frequently asked osteoporosis-related questions and to determine their potential role in clinical support.

Methods: Twenty-five questions covering six clinical domains were submitted to each model in isolated sessions. Five senior orthopedic physicians, each with over 25 years of clinical experience, independently rated the medical accuracy of each response using a 5-point Likert scale. Responses rated as "acceptable" or above were further evaluated for content comprehensiveness. Statistical analysis included the Kruskal–Wallis test and Dunn's post hoc test with Bonferroni correction.

Results: A total of 75 unique responses (25 questions × 3 models) were evaluated by five orthopedic experts, yielding 375 ratings. ChatGPT-4o achieved the highest accuracy score (median: 4.6; IQR: 4.4–4.8), significantly outperforming Gemini-2.5 Pro ($p=0.039$) and DeepSeek-R1 ($p<0.001$). For content comprehensiveness, both ChatGPT-4o and Gemini-2.5 Pro had a median score of 4.4, higher than DeepSeek-R1 (median: 4.2), though differences did not reach statistical significance ($p=0.0536$). Gemini-2.5 Pro was noted for its fluent and user-friendly language but lacked clinical depth in some responses. DeepSeek-R1, despite offering source citations, demonstrated greater inconsistency.

Conclusion: LLMs have clear potential as tools for patient education in osteoporosis. ChatGPT-4o demonstrated the most balanced and clinically reliable performance. Nonetheless, expert medical oversight remains essential to ensure safe and context-appropriate use in healthcare settings.

Keywords: large language models, osteoporosis, patient education, AI in healthcare, clinical consultation support

Introduction

Osteoporosis is a systemic metabolic bone disease characterized by reduced bone mass and deterioration of bone microarchitecture.[1] The condition significantly increases the risk of fractures, impairs quality of life, and is associated with higher rates of disability and mortality.[2–4] With the global trend of population aging, the prevalence of osteoporosis continues to rise, presenting substantial challenges for its prevention and management.[5,6] The global prevalence among adults aged ≥50 years is approximately 19.7%, with higher rates in women (24.3%) than in men (11.9%).[7]

In clinical practice, patients with osteoporosis exhibit broad and sustained informational needs concerning the disease's pathophysiology, risk factors, diagnostic procedures, treatment options, medication side effects, and prognosis management.[8] However, due to constraints such as limited consultation time, heavy clinician workload, and patients' varying levels of health literacy, traditional face-to-face communication often falls short in meeting expectations for

Received: 30 July 2025
Accepted: 13 November 2025
Published: 21 November 2025

continuous and personalized health education.[9] Therefore, exploring efficient and reliable supplementary tools to enhance the quality and accessibility of patient education has become increasingly important.

In recent years, the rapid advancement of artificial intelligence (AI)—particularly the emergence of large language models (LLMs) such as OpenAI's ChatGPT series, Google's Gemini, and domestic models like DeepSeek-R1—has sparked growing interest in their potential applications in medical health education.[10–13] These models generate semantically coherent answers to natural language inputs, enabling applications in patient Q&A and health education. However, the accuracy, completeness, and clinical practicality of LLM-generated content remain largely unverified.[14]

To date, there has been a lack of direct comparative studies evaluating the performance of mainstream LLMs in addressing common questions from osteoporosis patients, particularly in terms of clinical applicability. Given the significant differences in their data sources, training sets, and technical parameters, these models may perform variably in real-world clinical contexts.[15,16] Therefore, rigorous clinical simulation studies are necessary to quantitatively assess and compare the medical accuracy and content completeness of their responses, in order determine their potential as valuable tools in patient education.

This study aims to compare the performance of three mainstream LLMs—ChatGPT-4o, Gemini-2.5 Pro, and DeepSeek-R1—in responding to common questions posed by osteoporosis patients. The primary objective is to evaluate their strengths and limitations in terms of medical correctness and content comprehensiveness, and to explore their feasibility and applicability in clinical consultation support and health education.

Methods

Study Design

The overall workflow of this study is illustrated in Figure 1. The research was conducted in the Department of Orthopedics at Lianyungang First People's Hospital, China, from May 3 to May 25, 2025. A group of orthopedic experts and clinical researchers collaboratively designed a set of 25 common questions related to osteoporosis. These questions were initially sourced from multiple authoritative online health information platforms and were subsequently screened and revised by the expert panel to reflect the most frequently asked questions by patients and their primary care providers in clinical practice (see Supplementary Figure 1). Although designed to ensure clinical relevance and content validity, the set of 25 questions was not derived from a previously validated or standardized questionnaire.

To further analyze the performance of each LLM across different topical domains, the questions were categorized into six thematic areas based on prior literature: pathogenesis, risk factors, clinical manifestations, diagnosis, treatment and prevention, and prognosis. A designated researcher (XL) generated all model responses between May 20 and May 24, 2025, using three large language models—ChatGPT-4o (OpenAI, https://chat.openai.com), Gemini-2.5 Pro (Google DeepMind, https://gemini.google.com), and DeepSeek-R1 (DeepSeek AI, https://chat.deepseek.com). ChatGPT-4o and Gemini-2.5 Pro were accessed through paid professional subscriptions, whereas DeepSeek-R1 was accessed via its free public interface with real-time web retrieval enabled. Each question was submitted in an independent chat window without prior context, and no post-processing or manual editing of model outputs was performed.

DeepSeek-R1, a free model with real-time web access and source citation capabilities, is characterized by its timeliness and contextual relevance. In contrast, GPT-4o and Gemini-2.5 Pro are subscription-based models with higher parameter counts and computational power, theoretically enabling more complex and in-depth responses. To eliminate contextual interference, each question was submitted in a new chat window, ensuring that the models answered independently without prior conversational context.

This report follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines, which provide a standardized framework for reporting key elements of observational studies, such as objectives, methods, results, and limitations.

Accuracy Assessment

The scoring panel consisted of five senior orthopedic physicians, each with over 25 years of clinical experience in diagnosing and treating musculoskeletal disorders, including osteoporosis. To ensure objectivity, model identities were

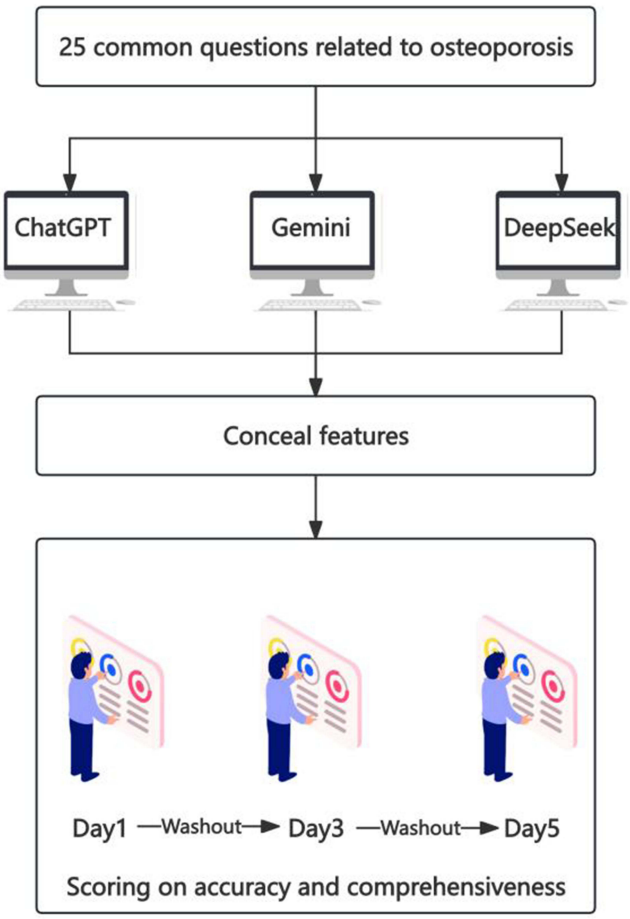

Figure 1 Study design and evaluation workflow. Overview of the study procedures, including question selection, model response generation, expert scoring, and statistical analysis steps.

fully concealed during the evaluation, and all metadata or clues that might indicate the source model were removed. Each rater independently assessed the medical accuracy of the LLM-generated answers using a 5-point scale. All model responses were anonymized and randomized prior to expert evaluation to prevent order or source-related bias (Table 1). Each question–answer pair was independently scored by five raters. In cases of disagreement, scores were reviewed collectively, and the final rating was determined by consensus among the experts. The 5-point Likert scale and the cutoff

Table 1 Rating Criteria for Medical Response Evaluation

Rating	Score	Definition
Poor	1	The response contains serious medical errors and inaccurate content, likely to mislead patients and cause harm.
Fair	2	The response has moderate factual errors that may cause misunderstandings but are unlikely to result in direct harm; clarification is needed.
Acceptable	3	The response is mostly accurate with minor factual inaccuracies; unlikely to mislead but clarification is advisable.
Good	4	The response is medically accurate, comprehensive, and clearly presented; suitable for direct patient education with minimal need for clarification.
Excellent	5	The response is entirely accurate, scientifically rigorous, well-structured, and informative, meeting the standards of ideal clinical communication without requiring any further clarification.

https://doi.org/10.2147/CIA.S551572

for "acceptable" performance (≥3) were adapted from previous studies evaluating the medical accuracy of large language model outputs in healthcare, where a score of 3 ("acceptable") or higher indicated clinically adequate content quality.[17,18]

Comprehensiveness Evaluation

For responses rated "acceptable" (Likert score≥3), raters also evaluated content comprehensiveness using the following 5-point scale (Table 2). The final comprehensiveness score for each answer was calculated as the mean of the five raters' individual scores.

Statistical Analysis

All statistical analyses were performed using R software. For each question–model pair, the mean of the five ratings was calculated, and the aggregated accuracy and comprehensiveness metrics (median and interquartile range) were computed across the 75 averaged responses. Shapiro–Wilk tests were applied to assess the normality of continuous variables. As word count data were normally distributed, differences among models were analyzed using one-way ANOVA, followed by Tukey's post hoc test. Accuracy and comprehensiveness scores did not meet normality assumptions; therefore, the Kruskal–Wallis test was used for overall comparisons, with Dunn's test and Bonferroni correction applied for post hoc pairwise analyses. To assess inter-rater reliability, intraclass correlation coefficients (ICCs) were computed using a two-way random-effects model. We reported both single-rater and average-rater ICCs, with particular focus on the ICC3k, which showed the degree of consistency across expert ratings. To compare the distribution of categorical ratings across models, Pearson's chi-square was used as appropriate. A two-tailed p-value of < 0.05 was considered statistically significant. All figures were generated using the ggplot2 and pheatmap packages.

Results

Response Length

The average response lengths varied substantially across the LLMs (see Supplementary Tables 1–3). Gemini-2.5 Pro generated the longest responses, with an average word count of 568.9, followed by DeepSeek-R1 (275.5) and GPT-4o (218.7) (Table 3). While longer responses might suggest greater elaboration, length did not consistently correlate with higher scores in either accuracy or comprehensiveness, indicating that verbosity alone is not a reliable proxy for quality.

Table 2 Scoring Criteria for Comprehensiveness Evaluation

Rating	Score	Definition
Not comprehensive	1	Severely lacking essential information.
Slightly comprehensive	2	Covers only basic or minimal content.
Moderately comprehensive	3	Includes substantial detail; relatively complete.
Comprehensive	4	Covers all major aspects; well-rounded content.
Very comprehensive	5	Thorough and detailed response with excellent coverage.

Table 3 Summary of Performance Metrics Across Three Language Models

Model	Average Word Count	Accuracy (Median [IQR])	Comprehensiveness (Median [IQR])
ChatGPT-4o	218.7	4.6 [4.4–4.8]	4.4 [4.2–4.4]
Gemini-2.5 Pro	568.9	4.4 [4.2–4.6]	4.4 [4.2–4.6]
DeepSeek-R1	275.5	4.2 [4.0–4.2]	4.2 [4.0–4.4]

Inter-Rater Reliability

To assess the consistency of expert ratings across models, we computed intraclass correlation coefficients (ICCs) based on a two-way random-effects model. The ICC for single raters (ICC3) was 0.25 (95% CI: 0.156–0.37), indicating fair agreement. In contrast, the ICC for average ratings across five raters (ICC3k) was 0.63 (95% CI: 0.480–0.75), reflecting good inter-rater reliability. The relatively low single-rater ICC reflects expected variability in subjective expert judgment, whereas the average-rater ICC indicates acceptable overall consistency when aggregated across multiple raters.

Accuracy of Medical Information

Statistically significant differences in medical accuracy scores were observed among the three language models, as determined by the Kruskal–Wallis test (χ^2 = 34.90, df = 2, p < 0.001). GPT-4o achieved the highest accuracy, with a median score of 4.6 and an interquartile range (IQR) of 4.4–4.8, followed by Gemini-2.5 Pro [median: 4.4; IQR: 4.2–4.6] and DeepSeek-R1 [median: 4.2; IQR: 4.0–4.2] (Table 3).

Post hoc pairwise comparisons using Dunn's test with Bonferroni correction revealed that GPT-4o performed significantly better than both Gemini-2.5 Pro (p = 0.039) and DeepSeek-R1 (p < 0.001). Additionally, Gemini-2.5 Pro showed significantly higher accuracy scores than DeepSeek-R1 (p = 0.002).

These statistical results were corroborated by visual assessment of the boxplot distributions (Figure 2). GPT-4o not only exhibited the highest median and upper quartile values but also demonstrated the narrowest IQR, indicating greater consistency and reduced variability across rated responses. In contrast, DeepSeek-R1 displayed a lower median score with a broader distribution, reflecting greater performance inconsistency and the presence of lower-rated outliers.

Comprehensiveness of Content

Differences in content comprehensiveness among the models were modest, with numerical trends favoring GPT-4o but without reaching statistical significance. Both GPT-4o and Gemini-2.5 Pro achieved identical median scores of 4.4, with interquartile ranges (IQR) of 4.2–4.4 and 4.2–4.6, respectively. DeepSeek-R1 had a slightly lower median score of 4.2,

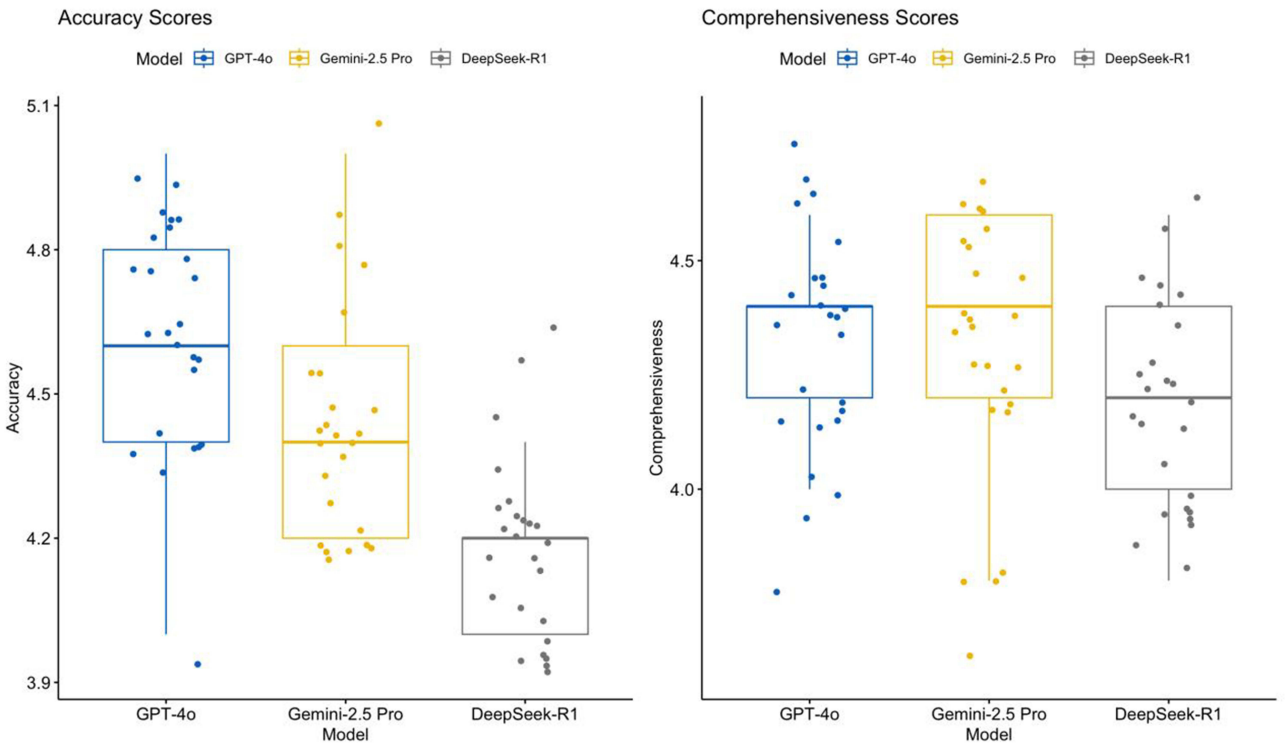

Figure 2 Boxplots of accuracy and comprehensiveness scores across large language models. Distribution of expert ratings (median ± IQR) for ChatGPT-4o, Gemini-2.5 Pro, and DeepSeek-R1; overall group differences tested with the Kruskal–Wallis test followed by Dunn's post hoc comparison.

https://doi.org/10.2147/CIA.S551572

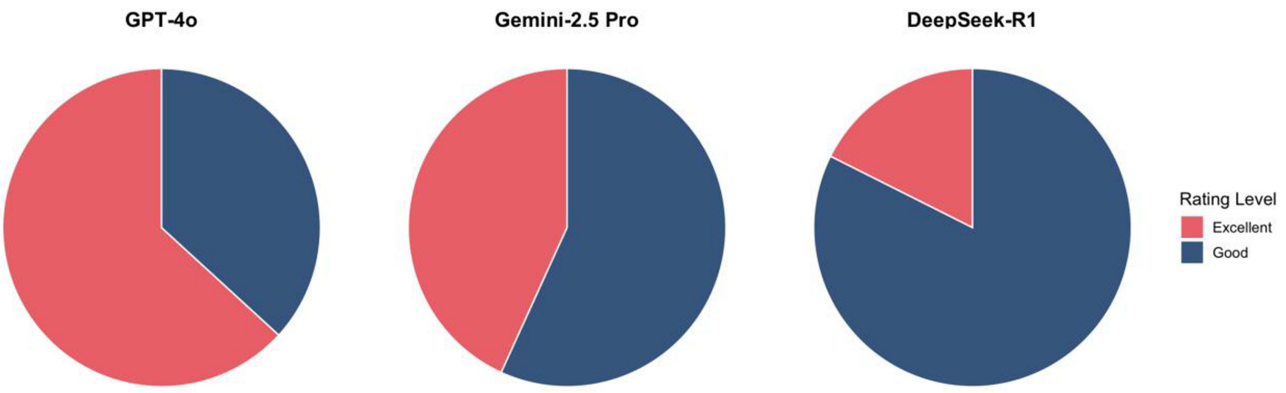

Figure 3 Proportion of "Excellent" and "Good" accuracy ratings by model. Stacked bar chart showing the percentage of high-quality responses (Likert ≥ 4) for each model.

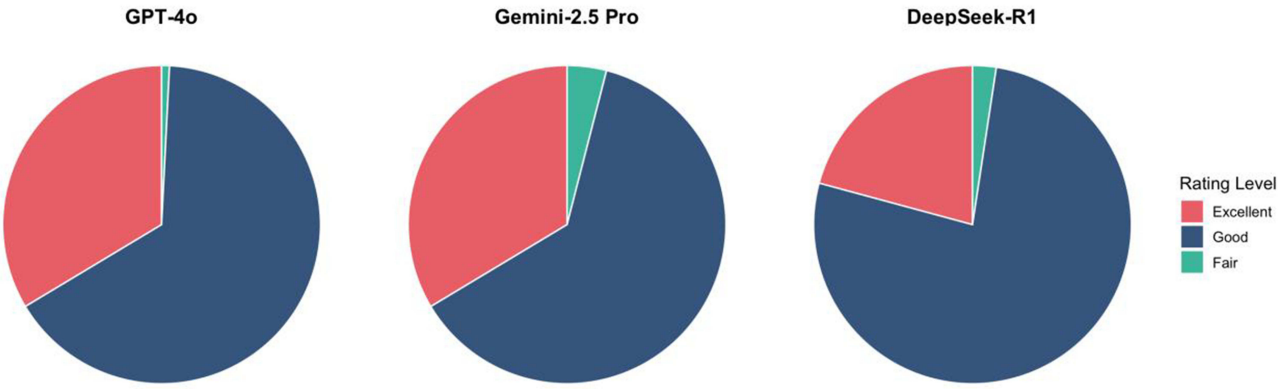

Figure 4 Proportion of accuracy and comprehensiveness ratings by model. Comparison of categorical distributions across models; percentages calculated based on five-rater consensus ratings.

with an IQR of 4.0–4.4 (Table 3). Although the Kruskal–Wallis test did not reach conventional significance ($\chi^2 = 5.85$, $p = 0.0536$), the observed numerical trends and visual distributions in Figures 3 and 4 suggest a relative advantage for GPT-4o in comprehensiveness.

Post hoc pairwise comparisons using Dunn's test with Bonferroni adjustment revealed no statistically significant differences between any model pairs ($p > 0.09$ for all), indicating that despite small observed differences, overall content comprehensiveness was comparable across models. This may reflect a potential ceiling effect or generally high baseline performance among current-generation LLMs in generating medically complete responses for osteoporosis-related questions.

Response Consistency Across Questions

Heatmap visualization of Accuracy and Comprehensiveness scores across 25 unique osteoporosis-related questions (Figure 5) showed that GPT-4o maintained consistently high performance, particularly in topics involving clinical diagnosis, fracture risk assessment, and treatment guidelines. Gemini-2.5 Pro demonstrated moderate variability, with lower performance on pharmacologic management questions. DeepSeek-R1 exhibited the greatest inconsistency across questions, with marked fluctuations in both accuracy and completeness, suggesting limited generalizability of its outputs across the medical domain.

Discussion

This study provides a comprehensive comparative analysis of three prominent LLMs—ChatGPT-4o, Gemini-2.5 Pro, and DeepSeek-R1—in addressing common questions frequently asked by patients with osteoporosis. Our findings revealed

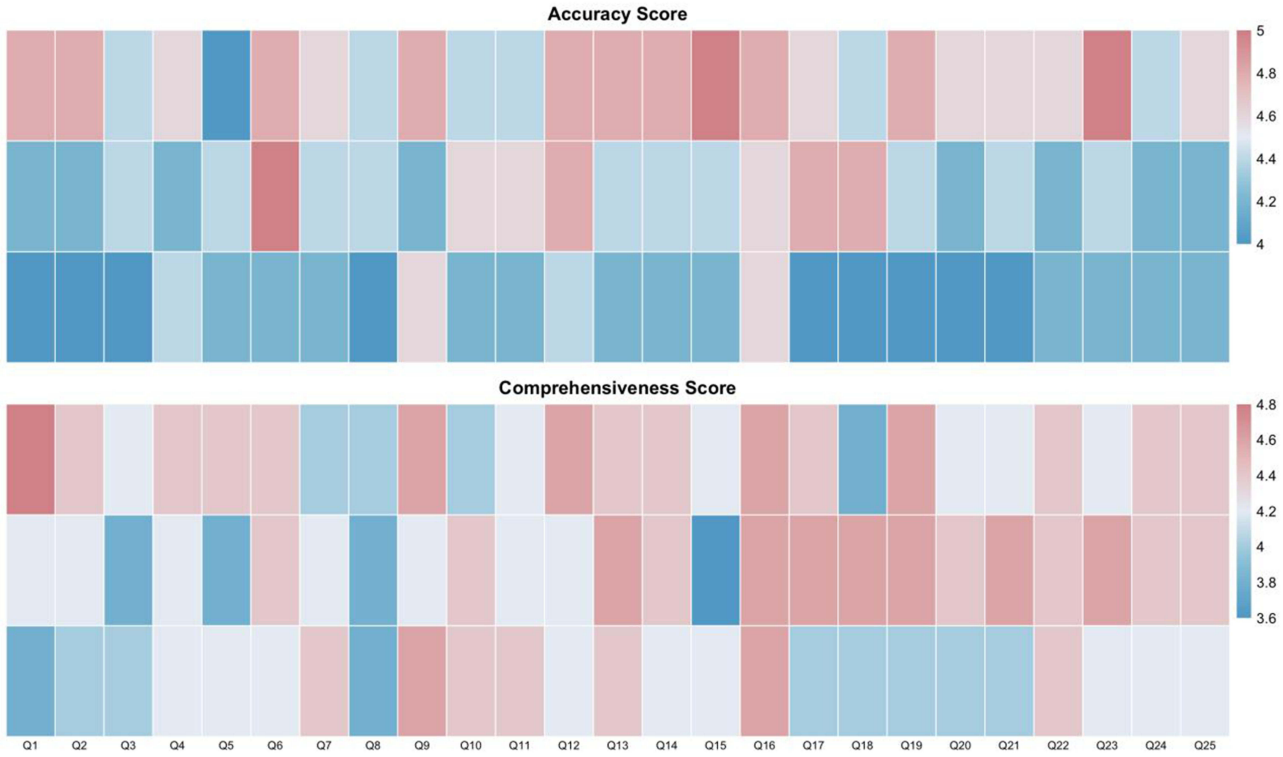

Figure 5 Heatmaps of accuracy and comprehensiveness scores across 25 questions. Color intensity represents mean scores for each question–model pair, with darker shades indicating higher accuracy and content depth.

that all three LLMs exhibited high fluency and coherence in natural language generation. Among them, ChatGPT-4o demonstrated significantly superior performance in both medical accuracy and content comprehensiveness, suggesting it is currently the most promising model for real-world application in patient health education within the context of osteoporosis.

ChatGPT-4o's answers were consistently rated as "Good" or "Excellent", with high lexical precision and adequate clinical depth. It covered key domains such as pathophysiology, diagnostic pathways, treatment modalities, and lifestyle interventions. Gemini-2.5 Pro was strong in linguistic fluency and user-oriented expression but occasionally lacked specificity in technical domains. DeepSeek-R1, while offering real-time web-based citations, was prone to factual inconsistencies and terminological inaccuracies, which compromised its reliability in medical communication. Because DeepSeek-R1 retrieves information directly from the internet, its performance may also be affected by real-time data variability, representing a potential confounder when comparing it with models operating on static knowledge bases.

Interpretation in Light of Other Evidence

The findings of this study align with and expand upon an emerging body of research evaluating the role of LLMs in healthcare communication.[19] Previous investigations have consistently shown that advanced models like GPT-4 perform well in general medical question answering and medical licensing exams.[20,21] The results of our study reinforce these observations by demonstrating that ChatGPT-4o, an advanced successor to GPT-4, exhibits superior performance in a real-world, domain-specific scenario. Unlike test-based studies, our design simulates patient inquiries that often require both factual precision and an empathetic communication style. The fact that ChatGPT-4o maintained high performance across both clinical depth and patient-oriented readability illustrates its adaptability beyond exam-style content.

Our evaluation of Gemini-2.5 Pro aligns with previous observations regarding Google's large language models, particularly their emphasis on producing fluent and natural-sounding text.[17,22] In our study, Gemini-2.5 Pro consistently generated responses with high linguistic quality, making it well-suited for general health education and user-facing applications. However, we also observed that its performance in tasks requiring precise clinical reasoning and factual

consistency was somewhat less robust than that of ChatGPT-4o. This suggests that although Gemini-2.5 Pro delivers user-friendly responses, it may require further optimization to match the clinical precision achieved by GPT-4o.

Interestingly, DeepSeek-R1's performance diverges from the theoretical advantage expected of models with real-time internet access. Prior research has hypothesized that access to up-to-date information may improve relevance, especially in rapidly evolving fields.[23,24] However, our findings suggest that information retrieval does not inherently equate to medical accuracy, particularly when models lack robust content validation pipelines. This aligns with concerns raised in AI safety literature regarding the "hallucination" phenomenon in LLMs, where confidently stated but incorrect information may be generated.[18,25,26] In medical contexts, such hallucinations can be especially hazardous, reinforcing the need for controlled content sources and domain-specific fine-tuning.[27–29]

Recent investigations in musculoskeletal disorders have highlighted the potential of large language models to enhance orthopedic patient communication and clinical decision-making. Studies evaluating these systems in total hip and shoulder arthroplasty demonstrated that they could provide generally accurate and comprehensible responses to common patient inquiries, though challenges remain regarding readability, source verification, and contextual precision.[30,31] Similar evaluations in fracture management further indicated that advanced language models can achieve high diagnostic accuracy and recall when analyzing structured clinical information yet still require professional oversight to ensure safety and reliability.[32,33] Collectively, these findings illustrate the growing integration of generative AI tools into musculoskeletal health care while underscoring the continued necessity of human supervision and rigorous validation before clinical adoption.[34,35]

Building upon these prior applications in musculoskeletal contexts, our study addresses a relative gap in the literature: the evaluation of LLMs in chronic disease-specific communication. While prior work has often focused on general medicine or acute care scenarios, few studies have examined performance in conditions like osteoporosis, which require long-term education, risk mitigation, and patient adherence support. The ability of ChatGPT-4o to maintain high ratings across multifaceted domains—pathogenesis, prevention, diagnosis, and prognosis—underscores its potential to serve not only as a question-answering system, but also as a sustainable companion for long-term health management.

Clinical Implications

The results of this study underscore the considerable promise of LLMs in augmenting both clinical communication and patient education, particularly in the context of chronic, high-prevalence conditions such as osteoporosis. The superior performance of ChatGPT-4o in delivering medically accurate, well-structured, and patient-friendly responses suggests that LLMs can function as effective supplementary tools in routine clinical workflows. In outpatient settings, where consultation time is often limited and patients may struggle to retain verbal explanations, AI-generated summaries and clarifications can serve as valuable reinforcements of key health messages.

A key strength of LLMs is their ability to provide access to easy-to-understand health information, which can help bridge the gap between communication-impaired healthcare providers and patients with varying levels of health literacy.[36,37] For osteoporosis—a condition that requires long-term self-management, medication adherence, and lifestyle modification—continuous access to reliable educational support is particularly important.[38] LLMs such as ChatGPT-4o may help reinforce physician guidance, address follow-up questions outside of clinical hours, and assist in patient decision-making regarding pharmacologic and non-pharmacologic interventions.

From an educational standpoint, the models may also be deployed in community health programs or telemedicine services, where access to specialists is limited. Their ability to deliver language-appropriate and contextually tailored responses makes them especially useful in rural or underserved populations.[39] For example, patients with early-stage osteoporosis or those identified via population screening could use AI-powered tools to learn about risk factors, preventive behaviors, and when to seek medical evaluation—potentially reducing delays in diagnosis and improving treatment initiation rates.

Furthermore, LLMs can serve as educational aids for non-physician healthcare workers, including nurses, pharmacists, and caregivers who frequently interact with patients but may not have in-depth orthopedic training. By providing accurate and rapid access to medical explanations, these models can enhance the quality of indirect patient education and improve interdisciplinary care coordination.

Despite these advantages, it is important to emphasize that LLMs cannot replace clinical judgment.[40] They are unable to take into account patient-specific variables such as renal function, polypharmacy, fracture history, or psychosocial status, all of which can significantly influence their role in personalized counseling or diagnostic reasoning. Additionally, their responses, while generally well-organized, may lack the nuanced emotional support and motivational interviewing techniques often needed in managing chronic illnesses such as osteoporosis. Clinician-patient dialogue remains irreplaceable for ensuring that medical recommendations are not only delivered but also contextualized, accepted, and acted upon. Failure to do so could erode patient trust, delay appropriate care, or inadvertently promote harmful behaviors.

In the long term, LLMs may play a pivotal role in patient engagement and behavior change, provided they are aligned with evidence-based guidelines and human oversight. For osteoporosis care in particular, the incorporation of LLMs into fracture liaison services, bone health clinics, and post-fracture rehabilitation programs may enhance patient understanding, reduce information asymmetry, and foster adherence to long-term treatment plans.

Strengths and Limitations

This study offers a structured and clinically grounded evaluation of three prominent large language models in addressing patient-centered questions in the context of osteoporosis, a domain where accurate, comprehensible, and sustained health education is vital for long-term disease management. Several methodological strengths enhance the validity and relevance of our findings.

A key strength lies in the real-world applicability of the question set. Rather than relying on academic or synthetic test items, the 25 questions were derived from actual patient queries as encountered on widely used health information platforms and in outpatient clinical practice. This ensures that the evaluation reflects the types of information patients genuinely seek and the communication challenges that clinicians commonly face.

Another major strength is the use of a rigorous and blinded expert rating system, involving five orthopedic specialists with over 25 years of clinical experience. All answers were anonymized, and scorers were unaware of which model generated each response. The dual assessment of both medical accuracy and content comprehensiveness, using a well-defined 5-point Likert scale, enabled us to capture not only factual correctness but also the depth and structure of the information provided. Moreover, the inclusion of multiple thematic domains—pathogenesis, diagnosis, treatment, prognosis, etc.—allowed us to evaluate performance across a spectrum of clinical subtopics.

The study also contributes to the literature by examining model-specific variability, which is often overlooked in broader AI performance benchmarks. Our findings highlight not just aggregate performance metrics but also inconsistencies and domain-specific weaknesses that may have important implications for clinical deployment.

However, the study also has several limitations that should be considered when interpreting the results.

First, the evaluation was conducted using a static, single-turn prompt-response format, which does not fully capture the dynamic, iterative nature of real-world patient–AI interactions. In clinical settings, patients often ask follow-up questions, express concerns, or require clarification—tasks that demand contextual memory and conversational continuity, which were not assessed in this study.

Second, while the accuracy ratings were conducted by clinical experts, the study did not incorporate feedback from actual patients, which limits the understanding of how end-users perceive and interpret the AI-generated responses. Patient-centered outcomes such as perceived trust, emotional tone, cultural appropriateness, and behavioral influence remain unexplored and warrant further investigation in future studies.

In addition, the threshold defining an "acceptable" response (Likert score ≥ 3) was not derived from a formally validated instrument but was adapted from prior LLM evaluation literature, which may limit comparability with other studies.

Lastly, although we evaluated three high-profile LLMs, the rapid evolution of AI models means that their capabilities may change significantly in short timeframes. This underscores the need for ongoing model surveillance and periodic re-evaluation.

Conclusion

In summary, this study demonstrates that large language models—particularly ChatGPT-4o—show strong potential in supporting osteoporosis-related patient education. Among the three models evaluated, ChatGPT-4o consistently provided the most accurate and comprehensive responses, suggesting it may be a valuable tool for enhancing health communication in clinical and educational settings. While Gemini-2.5 Pro offered accessible and fluent language suitable for general audiences, and DeepSeek-R1 showed strengths in citing external sources, both models had limitations in medical precision. Future work should focus on improving domain-specific accuracy, testing patient usability, and integrating these tools into supervised healthcare systems to maximize their benefit and minimize risks.

Ethics Approval and Informed Consent

This study was reviewed and approved by the Research Ethics Committee of the First People's Hospital of Lianyungang (No: LW-20251017001). All procedures performed in this study involving human participants were conducted in accordance with the ethical standards of the institutional research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

Acknowledgments

We would like to thank all of the doctors who participated in this study for their cooperation.

Funding

This work was supported by the Medical Technology Support Program of the First People's Hospital of Lianyungang (YK202306).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Liu Y, Yang Y, Li Y, et al. Association between nutritional and inflammatory status and mortality outcomes in patients with osteoporosis and osteopenia. *J Nutr Biochem*. 2025;143:109936. doi:10.1016/j.jnutbio.2025.109936
2. Huo R, Wei C, Huang X, et al. Mortality associated with osteoporosis and pathological fractures in the United States (1999-2020): a multiple-cause-of-death study. *J Orthop Surg Res*. 2024;19:568. doi:10.1186/s13018-024-05068-1
3. Dovjak P, Iglseder B, Rainer A, et al. Prediction of fragility fractures and mortality in a cohort of geriatric patients. *J Cachexia Sarcopenia Muscle*. 2024;15(6):2803–2814. doi:10.1002/jcsm.13631
4. Chen H, Lou Y, Fei S, et al. Association between physical activity and mortality in patients with osteoporosis: a cohort study of NHANES. *Osteoporos Int*. 2024;35:2195–2202. doi:10.1007/s00198-024-07280-5
5. US Preventive Services Task Force; Nicholson WK, Silverstein M, Wong JB, et al. Screening for osteoporosis to prevent fractures: us preventive services task force recommendation statement. *JAMA*. 2025;333:498–508. doi:10.1001/jama.2024.27154
6. Nguyen A, Lee P, Rodriguez EK, et al. Addressing the growing burden of musculoskeletal diseases in the ageing US population: challenges and innovations. *Lancet Healthy Longev*. 2025;6:100707. doi:10.1016/j.lanhl.2025.100707
7. Burnett-Bowie S-AM, Wright NC, Yu EW, et al. The American society for bone and mineral research task force on clinical algorithms for fracture risk report. *J Bone Miner Res*. 2024;39:517–530. doi:10.1093/jbmr/zjae048
8. Lems WF, Raterman HG. Critical issues and current challenges in osteoporosis and fracture prevention. An overview of unmet needs. *Ther Adv Musculoskelet Dis*. 2017;9:299–316. doi:10.1177/1759720X17732562
9. Oster A, Wiking E, Nilsson GH, Olsson CB. Patients' expectations of primary health care from both patients' and physicians' perspectives: a questionnaire study with a qualitative approach. *BMC Prim Care*. 2024;25:128. doi:10.1186/s12875-024-02389-2
10. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. *Nat Med*. 2025;31:1233–1238. doi:10.1038/s41591-024-03456-y
11. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31:943–950. doi:10.1038/s41591-024-03423-7
12. Doo FX, Savani D, Kanhere A, et al. Optimal large language model characteristics to balance accuracy and energy use for sustainable medical applications. *Radiology*. 2024;312:e240320. doi:10.1148/radiol.240320
13. Iqbal U, Tanweer A, Rahmanti AR, et al. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J Biomed Sci*. 2025;32:45. doi:10.1186/s12929-025-01131-z
14. Menz BD, Kuderer NM, Bacchi S, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ*. 2024;384:e078538. doi:10.1136/bmj-2023-078538

15. Telenti A, Auli M, Hie BL, et al. Large language models for science and medicine. *Eur J Clin Invest*. 2024;54(6):e14183. doi:10.1111/eci.14183
16. Javed H, El-Sappagh S, Abuhmed T. Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. *Artif Intell Rev*. 2024;58:12. doi:10.1007/s10462-024-11005-9
17. Scaff SPS, Reis FJJ, Ferreira GE, et al. Assessing the performance of AI chatbots in answering patients' common questions about low back pain. *Ann Rheum Dis*. 2025;84:143–149. doi:10.1136/ard-2024-226202
18. Hao G, Wu J, Pan Q, Morello R. Quantifying the uncertainty of LLM hallucination spreading in complex adaptive social networks. *Sci Rep*. 2024;14:16375. doi:10.1038/s41598-024-66708-4
19. Abdelhamid AA, El-Kenawy E-SM, Ibrahim A, et al. Innovative feature selection method based on hybrid sine cosine and dipper throated optimization algorithms. *IEEE Access*. 2023;11:79750–79776. doi:10.1109/ACCESS.2023.3298955
20. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written german medical licensing examination: observational study. *JMIR Med Educ*. 2024;10:e50965. doi:10.2196/50965
21. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023. doi:10.1227/neu.0000000000002632
22. Cao C, Sang J, Arora R, et al. Development of prompt templates for large language model-driven screening in systematic reviews. *Ann Intern Med*. 2025;178:389–401. doi:10.7326/ANNALS-24-02189
23. Aftab W, Apostolou Z, Bouazoune K, Straub T. Optimizing biomedical information retrieval with a keyword frequency-driven prompt enhancement strategy. *BMC Bioinf*. 2024;25:281. doi:10.1186/s12859-024-05902-7
24. Yu H, Fan L, Li L, et al. Large language models in biomedical and health informatics: a review with bibliometric analysis. *J Healthc Inform Res*. 2024;8:658–711. doi:10.1007/s41666-024-00171-8
25. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*. 2025;43:1–55. doi:10.1145/3703155
26. Ahmadi A. Unravelling the mysteries of hallucination in large language models: strategies for precision in artificial intelligence language generation. *AJCST*. 2024;13:1–10. doi:10.70112/ajcst-2024.13.1.4144
27. Liu F, Zhou H, Gu B, et al. Application of large language models in medicine. *Nat Rev Bioeng*. 2025;3:445–464. doi:10.1038/s44222-025-00279-5
28. Yang Y, Jin Q, Zhu Q, et al. Beyond multiple-choice accuracy: real-world challenges of implementing large language models in healthcare. *Annu Rev Biomed Data Sci*. 2025;8(1):305–316. doi:10.1146/annurev-biodatasci-103123-094851
29. Yu P, Xu H, Hu X, Deng C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. *Healthcare*. 2023;11:2776. doi:10.3390/healthcare11202776
30. Mika AP, Martin JR, Engstrom SM, et al. Assessing ChatGPT responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. 2023;105(19):1519–1526. doi:10.2106/JBJS.23.00209
31. Lack BT, Mouhawasse E, Childers JT, et al. Can ChatGPT answer patient questions regarding reverse shoulder arthroplasty? *J ISAKOS*. 2024;9:100323. doi:10.1016/j.jisako.2024.100323
32. Marcaccini G, Seth I, Xie Y, et al. Breaking bones, breaking barriers: chatGPT, deepseek, and gemini in hand fracture management. *J Clin Med*. 2025;14:1983. doi:10.3390/jcm14061983
33. El-Kenawy E-SM, Khodadadi N, Mirjalili S, et al. Metaheuristic optimization for improving weed detection in wheat images captured by drones. *Mathematics*. 2022;10:4421. doi:10.3390/math10234421
34. Alkanhel R, M. El-kenawy E-S, A. Abdelhamid A. Network intrusion detection based on feature selection and hybrid metaheuristic optimization. *Comput Mater Continua*. 2023;74(2):2677–2693. doi:10.32604/cmc.2023.033273
35. Atteia G, M. El-kenawy E-S, Abdel Samee N. Adaptive dynamic dipper throated optimization for feature selection in medical data. *Comput Mater Continua*. 2023;75(1):1883–1900. doi:10.32604/cmc.2023.031723
36. Tripathi S, Alkhulaifat D, Muppuri M, et al. Large language models for global health clinics: opportunities and challenges. *J Am College Radiol*. 2025:S1546144025002054. doi:10.1016/j.jacr.2025.04.007
37. Raghu Subramanian C, Yang DA, Khanna R. Enhancing health care communication with large language models—the role, challenges, and future directions. *JAMA Network Open*. 2024;7:e240347. doi:10.1001/jamanetworkopen.2024.0347
38. Rubæk M, Hitz MF, Holmberg T, et al. Effectiveness of patient education for patients with osteoporosis: a systematic review. *Osteoporos Int*. 2022;33(5):959–977. doi:10.1007/s00198-021-06226-5
39. Farías H, González Aroca J, Ortiz D. Chatbot based on large language model to improve adherence to exercise-based treatment in people with knee osteoarthritis: system development. *Technologies*. 2025;13:140. doi:10.3390/technologies13040140
40. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med*. 2024;30:2613–2622. doi:10.1038/s41591-024-03097-1