Journal of Orthopaedic Surgery and Research

RESEARCH ARTICLE Open Access

# Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis

Linjian Tong[1†], Chaoyang Zhang[2†], Rui Liu[1], Jia Yang[1] and Zhiming Sun[1*]

## Abstract

**Backgrounds** The use of large language models (LLMs) in medicine can help physicians improve the quality and effectiveness of health care by increasing the efficiency of medical information management, patient care, medical research, and clinical decision-making.

**Methods** We collected 34 frequently asked questions about glucocorticoid-induced osteoporosis (GIOP), covering topics related to the disease's clinical manifestations, pathogenesis, diagnosis, treatment, prevention, and risk factors. We also generated 25 questions based on the 2022 American College of Rheumatology Guideline for the Prevention and Treatment of Glucocorticoid-Induced Osteoporosis (2022 ACR-GIOP Guideline). Each question was posed to the LLM (ChatGPT-3.5, ChatGPT-4, and Google Gemini), and three senior orthopedic surgeons independently rated the responses generated by the LLMs. Three senior orthopedic surgeons independently rated the answers based on responses ranging between 1 and 4 points. A total score (TS) > 9 indicated 'good' responses, 6 ≤ TS ≤ 9 indicated 'moderate' responses, and TS < 6 indicated 'poor' responses.

**Results** In response to the general questions related to GIOP and the 2022 ACR-GIOP Guidelines, Google Gemini provided more concise answers than the other LLMs. In terms of pathogenesis, ChatGPT-4 had significantly higher total scores (TSs) than ChatGPT-3.5. The TSs for answering questions related to the 2022 ACR-GIOP Guideline by ChatGPT-4 were significantly higher than those for Google Gemini. ChatGPT-3.5 and ChatGPT-4 had significantly higher self-corrected TSs than pre-corrected TSs, while Google Gemini self-corrected for responses that were not significantly different than before.

**Conclusions** Our study showed that Google Gemini provides more concise and intuitive responses than ChatGPT-3.5 and ChatGPT-4. ChatGPT-4 performed significantly better than ChatGPT3.5 and Google Gemini in terms of answering general questions about GIOP and the 2022 ACR-GIOP Guidelines. ChatGPT3.5 and ChatGPT-4 self-corrected better than Google Gemini.

[†]Linjian Tong and Chaoyang Zhang contributed equally to this work.

*Correspondence:
Zhiming Sun
szhm0618@163.com

Full list of author information is available at the end of the article

**Keywords** Large language models, AI, ChatGPT, Google Gemini, Glucocorticoid-Induced osteoporosis

## Introduction

Glucocorticoid-induced osteoporosis (GIOP) is a form of osteoporosis caused by long-term or high-dose use of glucocorticoid medications in patients with a variety of inflammatory and autoimmune diseases [1–3]. Glucocorticoids decrease bone formation, increase bone resorption, and interfere with calcium absorption and excretion, thereby leading to bone loss [4, 5]. GIOP evolves more rapidly than other types of osteoporosis, with severe bone density loss sometimes observed within a few months of glucocorticoid administration; furthermore, while symptoms might be absent in the early stages of the disease, bone pain, loss of height, or fracture might occur as the disease progresses [6, 7].

The rate of new fractures after one year of glucocorticoid therapy can reach 17%, and GIOP is more likely to affect the spine, especially the vertebrae [1, 8]. Therefore, GIOP can lead to compression fractures. Fractures occur in 30–50% of patients receiving long-term glucocorticoid therapy, and fractures are usually asymptomatic [1, 8, 9]. These fractures may occur as little as three months after starting steroid therapy at doses as low as 2.5 mg of prednisone per day [1]. In addition, people of any age and sex can develop osteoporosis with glucocorticoid use, and the risk is higher among older adults, postmenopausal women, and those with other risk factors for o steoporosis [10–12]. Consequently, understanding the characteristics of GIOP is essential for the prevention and management of this condition [13, 14].

Natural language processing (NLP) is a form of artificial intelligence (AI) that is dedicated to enabling computers to understand, interpret, and respond to human language. NLP combines methods from computer science, AI, and linguistics to analyze, understand, and generate natural language [15, 16]. Large language models (LLMs) are a subfield of NLP that focuses on developing large-scale machine learning models to process, understand, and generate natural language [17]. LLMs are typically built by training a model on large amounts of textual data, and they are able to capture the complexity and nuances of language [18]. Currently, the most advanced LLM chatbots are ChatGPT-3.5 and ChatGPT-4, which were developed by the OpenAI Foundation, and Google Gemini [19–21].

The use of LLMs in the medical field can help physicians improve the quality and effectiveness of health care by increasing the efficiency of medical information management, patient care, medical research, and clinical diagnosis [22–26]. However, in real applications, different versions and implementations of LLM chatbots may have different levels of performance, so it is also essential to choose the right model for a particular task [27–30]. In the study by Zhi Wei Lim et al., ChatGPT-4 showed excellent accuracy in answering questions about myopia care, with 80.6% of the responses rated as "good," compared to 61.3% for ChatGPT-3.5 and 54.8% for Google Gemini [31]. ChatGPT has also been found to be reasonably accurate in answering general questions about osteoporosis, but the responses to questions based on the National Osteoporosis Guidelines Group guidelines were only 61.3% accurate [32]. The purpose of this study was to evaluate and compare the performance of three publicly available LLMs, namely, OpenAI's ChatGPT-3.5 and ChatGPT-4, as well as Google Gemini, in answering questions related to GIOP and the 2022 ACR-GIOP Guidelines. These findings will help to determine which model performs better in a particular task or application scenario, thus enabling users to make more informed choices.

## Methods

### Study design

A set of 34 general questions related to GIOP (Supplementary Table 1a) were curated from reputable online health information sources, including UpToDate, the American College of Rheumatology (ACR), the National Center for Biotechnology Information (NCBI), and Endocrine News. Subsequently, for further optimization, questions were selected based on their applicability to common clinical settings. To deepen the understanding of the strengths and weaknesses of different LLM chatbots in addressing various topics, we categorized these questions into 6 critical fields, namely, clinical manifestation, pathogenesis, diagnosis, treatment, prevention and risk factors. We also prepared 25 questions based on the 2022 ACR-GIOP Guidelines (Supplementary Table 1b). Answers to these question queries were generated from March 13 to March 25, 2024, by using two versions of ChatGPT (versions ChatGPT-3.5 and ChatGPT-4, OpenAI, California) and Google Gemini (Google LLC, Alphabet Inc., California). Each question was entered as a separate conversation, and the conversations were reset after each query to collect the content of the replies. The content of the LLM chatbot replies was converted to plain text format, any information in the text that identified the LLM chatbot was removed, and the responses were rated by three orthopedic surgeons experienced in treating osteoporosis. Figure 1 shows the overall design of this study.

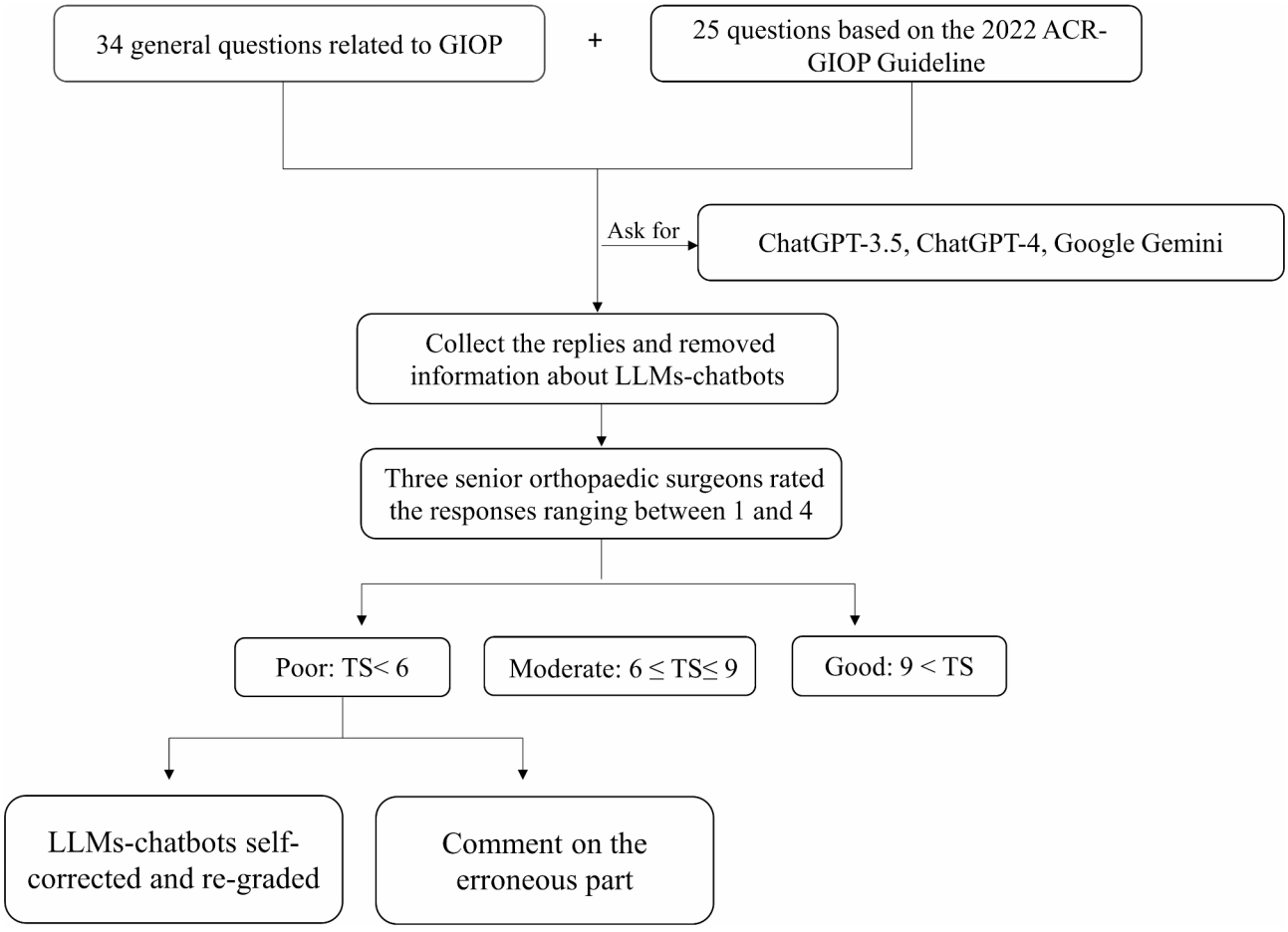


**Fig. 1** Flowchart of the overall study design

#### Accuracy assessment

Three senior orthopedic surgeons independently rated the answers based on responses ranging between 1 and 4 points (1 indicates that the answer is completely incorrect, 2 indicates that part of the answer is correct but contains incorrect information, 3 indicates a correct but inadequate answer, and 4 indicates a correct and adequate answer). The consistency of the three senior orthopedic surgeons' ratings of the ChatGPT-3.5, ChatGPT-4, and Google Gemini responses to the questions was assessed using the Fleiss' Kappa coefficient. A total score (TS) > 9 indicated 'good' responses, 6 ≤ TS ≤ 9 indicated 'moderate' responses, and TS < 6 indicated 'poor' responses.

#### Re-evaluating the accuracy of LLM chatbot self-correcting

The questions that were recognized as 'poor' were subjected to further questioning, where the incorrect parts were explained through an orthopedic specialist pointing out incorrect or inaccurate sentences within the content of the responses. The answers to these questions were also self-corrected in the LLM chatbot chat program: "This doesn't seem quite right. Can you answer it again?". Subsequently, the responses were collected and converted to plain text format, information identifying that LLM chatbot was removed from the text, the order was disrupted, and the corrected content was reevaluated by the three raters. This round of reassessment was conducted one week after the previous round of scoring. During this round of reassessment, the scorers were not informed that the responses were self-correcting versions.

#### Statistical analysis

SPSS 26 software (IBM Corp. Released 2021) was used for the data analysis. Normally distributed data are expressed as the mean ± standard deviation, nonnormally distributed data are presented as the median (percentile$_{25}$-percentile$_{75}$) (M($P_{25}$-$P_{75}$)), and the Kruskal–Wallis *H* test was used for multiple comparisons to determine the significance of the differences between ChatGPT-3.5, ChatGPT-4 and Google Gemini. Paired t tests were used to compare the initial TS and self-corrected TS, and Pearson's chi-squared test was used to compare initial accuracy ratings and self-corrected accuracy ratings.

$P<0.05$ was considered to indicate a statistically significant difference. Fleiss' Kappa was using for assessing the consistency of the responses to the questions ratings of ChatGPT-3.5, ChatGPT-4, and Google Gemini scores by the three senior orthopedic surgeons. Fleiss' Kappa values between 0 and 1. The degree of consistency is poor from 0 to 0.2; moderate from 0.2 to 0.4; medium from 0.4 to 0.6; strong from 0.6 to 0.8; and very strong from 0.8 to 1.0.

## Results

### Length of responses from LLM chatbot

Table 1 shows the length of words and characters generated by the LLM chatbots to the GIOP-related general questions. The mean±standard deviation of the word count was 346.50±68.19 for ChatGPT-3.5, 303.91±43.56 for ChatGPT-4, and the M($P_{25}$-$P_{75}$) of the word count was 308.50 (266.75-350.25) for Google Gemini (Fig. 2a). The number of words generated by ChatGPT-4 and Google Gemini was significantly higher than that generated by ChatGPT-3.5 ($P<0.05$). The number of characters generated by ChatGPT-3.5 was 2445.65±467.72, the number of characters generated by ChatGPT-4 was 2119.29±300.83, and the number of words generated by Google Gemini was 2206.00 (1888.50-2546.25) (Fig. 2b). The number of characters generated by ChatGPT-4 was significantly lower than that generated by ChatGPT-3.5 ($P<0.05$).

Table 2 shows the length of words and characters generated by the LLM chatbots in response to the questions related to the 2022 ACR-GIOP Guideline. The M($P_{25}$-$P_{75}$) word counts were 378.00 (303.00-407.00), 328.00 (257.00-361.00) and 317.00 (269.50–350.00) for ChatGPT-3.5, ChatGPT-4 and Google Gemini (Fig. 2c), respectively. The M($P_{25}$-$P_{75}$) word counts were 2,783.00 (2,173.00–2,967.50), 2407.00 (1875.00-2564.00) and 2273.00 (2016.50-2451.50) for ChatGPT-3.5, ChatGPT-4 and Google Gemini (Fig. 2d), respectively. The number of words and characters generated by Google Gemini was significantly higher than that of ChatGPT-3.5 ($P<0.05$). The number of words and characters per question generated by the LLM chatbots is shown in Supplementary Table 3,4a-c.

### Accuracy and grading of LLMs chatbot responses

Table 3 shows the TSs of the LLM chatbot responses to the different topics within the GIOP-related general questions. Regarding pathological mechanisms, the TS of ChatGPT-4 [10.00 (9.00-10.50)] was significantly higher than that of ChatGPT-3.5 [8.00 (6.50-8.00)] ($P<0.05$). Table 4 shows the TSs of the LLM chatbot in terms of the 2022 ACR-GIOP Guideline-related questions. Google Gemini [8.00 (8.00–11.00)] had a lower TS than ChatGPT-4 [10.00 (9.00–11.00)] ($P<0.05$).

Table 5 shows the accuracy ratings of the LLM chatbot responses to the different topics within the GIOP-related general questions. Regarding pathological mechanisms, ChatGPT-3.5 was significantly worse ($P<0.05$) than ChatGPT-4 and Google Gemini. Overall, the ChatGPT-4 performed excellently in answering GIOP-related general questions, with no 'poor' responses, and it was more effective in addressing the topic of clinical presentation, with a 100% probability of responding 'good'. Table 6 shows the accuracy ratings of the responses of the LLM chatbots to questions related to the 2022 ACR-GIOP Guidelines. ChatGPT-4 had the highest percentage of 'good' answers, accounting for 64%. Google Gemini had the lowest percentage of 'good' answers, accounting for 32%. Google Gemini and ChatGPT-3.5 had four poor answers, but overall, there was no significant difference among the three LLM chatbots ($P>0.05$). The raw responses to each question generated by the LLM chatbot are shown in Supplementary Tables 2a-c.

### Self-correcting capacity of LLM chatbots

Table 7 shows the changes in ChatGPT-3.5 after self-correcting for responses with a TS<6. The average TS for the initial responses was 4.00±0.89, and the average TS for the self-corrected responses was 6.67±2.16, which was significantly higher ($P<0.05$). Table 8 shows the changes in ChatGPT-4 after self-correcting for responses with a TS<6. The self-corrected TS was significantly higher than the initial TS (4.00±1.00 vs. 11.00±1.00, $P<0.05$). Table 9 shows the changes in Google Gemini after self-correcting for responses with a TS<6. However, there was no significant difference in the TS or ratings between the initial responses and the self-corrected responses; these findings suggest that Google Gemini's self-correction abilities are worse than those of ChatGPT-3.5 and

**Table 1** Length of LLMs-chatbots' responses to general questions about GIOP

| | Response length (words) | Response length (characters) |
|---|---|---|
| ChatGPT-3.5, $(\bar{x} \pm \mathrm{sd})$ | 346.50±68.19 | 2445.65±467.72 |
| ChatGPT-4, $(\bar{x} \pm \mathrm{sd})$ | 303.91±43.56* | 2119.29±300.83** |
| Google Gemini, M($P_{25}$-$P_{75}$) | 308.50(266.75-350.25) ** | 2206.00(1888.50-2546.25) |
| *P* value | 0.005 | 0.006 |

*$P<0.05$, **$P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; ^$P<0.05$, ^^$P<0.01$, ChatGPT-4 vs. Google Gemini

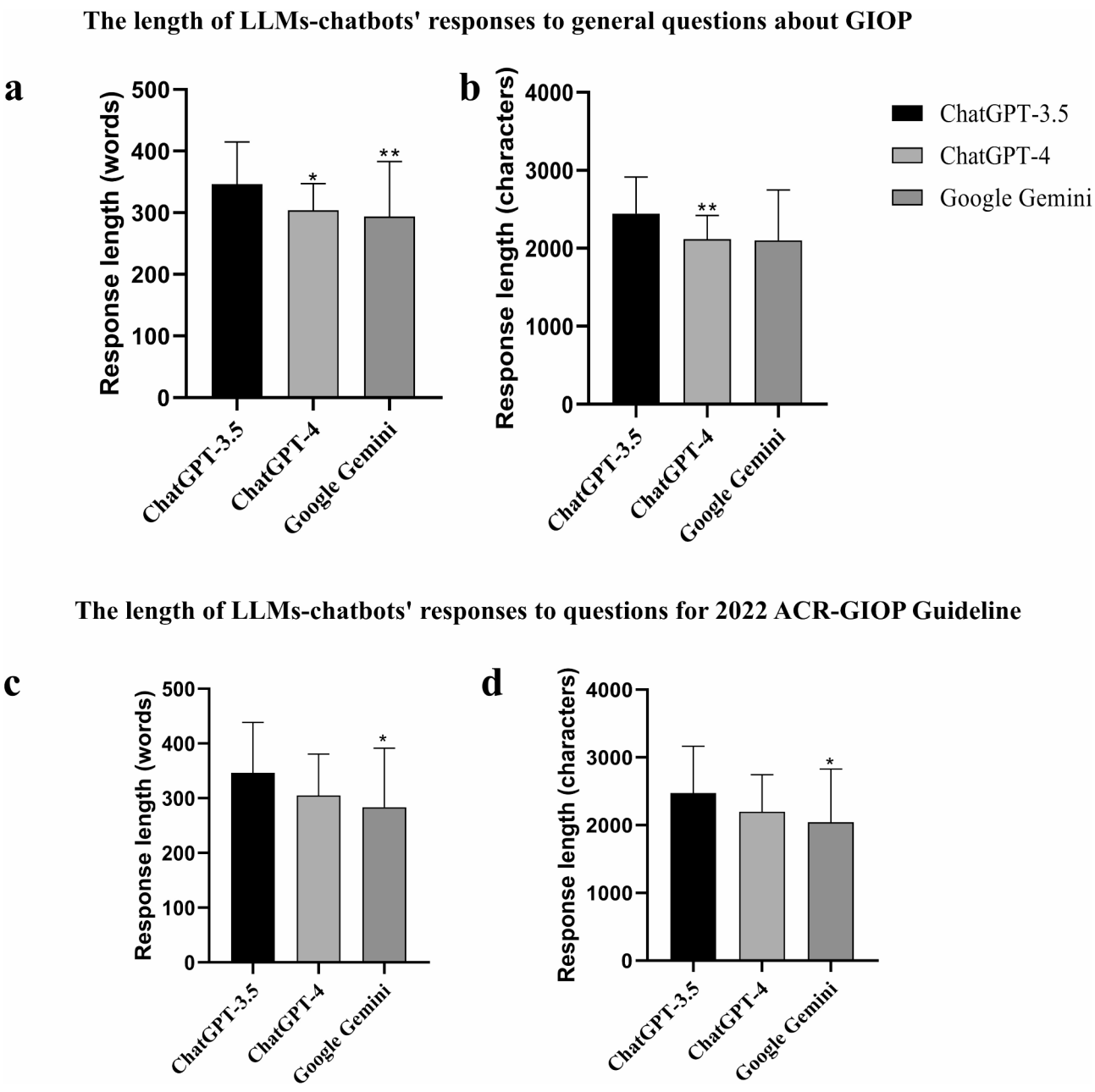


**Fig. 2** **a**, **b** The length of words and characters generated by LLM chatbots to GIOP-related general questions; **c**, **d** The length of words and characters generated by LLM chatbots to questions related to the 2022 ACR-GIOP Guidelines. $^{*}P<0.05$, $^{**}P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; $^{\wedge}P<0.05$, $^{\wedge\wedge}P<0.01$, ChatGPT-4 vs. Google Gemini

**Table 2** Length of LLMs-chatbots' responses to questions for 2022 ACR-GIOP Guideline

| | Response length (words) | Response length (characters) |
|---|---|---|
| ChatGPT-3.5, M($P_{25}$-$p_{75}$) | 378.00(303.00-407.00) | 2783.00(2173.00-2967.50) |
| ChatGPT-4, M($P_{25}$-$p_{75}$) | 328.00(257.00-361.00) | 2407.00(1875.00-2564.00) |
| Google Gemini, M($P_{25}$-$p_{75}$) | 317.00(269.50–350.00) * | 2273.00(2016.50-2451.50) * |
| *P* value | 0.017 | 0.031 |

$^{*}P<0.05$, $^{**}P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; $^{\wedge}P<0.05$, $^{\wedge\wedge}P<0.01$, ChatGPT-4 vs. Google Gemini

**Table 3** Differences in LLMs-chatbots' TS of response to general questions about GIOP

| Topic | ChatGPT-3.5, M($P_{25}$-$p_{75}$) | ChatGPT-4, M($P_{25}$-$p_{75}$) | Google Gemini, M($P_{25}$-$p_{75}$) | *P* value |
|---|---|---|---|---|
| Clinical Manifestation | 10.00(7.75–10.25) | 10.50(10.00-11.25) | 10.00(8.00-10.25) | 0.158 |
| Pathogenesis | 8.00(6.50-8.00) | 10.00(9.00-10.50) * | 11.00(7.50–11.50) | 0.017 |
| Diagnosis | 10.00(7.00–12.00) | 9.00(9.00–9.00) | 10.00(10.00–10.00) | 0.853 |
| Treatment | 10.50(8.25-12.00) | 9.00(9.00-9.25) | 8.00(5.25–10.50) | 0.422 |
| Prevention | 8.00(6.50–11.50) | 10.00(9.00–11.00) | 8.00 (3.50–11.00) | 0.492 |
| Risk Factor | 9.00(6.50–12.00) | 9.00(8.00–12.00) | 12.00(8.00–12.00) | 0.833 |
| All Questions | 9.00(7.00–11.00) | 10.00 (9.00–11.00) | 10.00 (7.75-11.00) | 0.236 |

*$P<0.05$, **$P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; ^$P<0.05$, ^^$P<0.01$, ChatGPT-4 vs. Google Gemini

**Table 4** Differences in LLMs-chatbots' TS of response to questions about 2022 ACR-GIOP Guideline

| Topic | ChatGPT-3.5 | ChatGPT-4 | Google Gemini | *P* value |
|---|---|---|---|---|
| 2022 ACR-GIOP Guideline | 9.00(8.00–11.00) | 10.00 (9.00–11.00) | 8.00 (8.00–11.00) ^ | 0.012 |

*$P<0.05$, **$P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; ^$P<0.05$, ^^$P<0.01$, ChatGPT-4 vs. Google Gemini

**Table 5** Accuracy ratings of LLMs-chatbots' responses to general questions related to GIOP

| Topic | Total, *n* | ChatGPT-3.5, *n*(%) | | | ChatGPT-4, *n*(%) | | | Google Gemini, *n*(%) | | | *P* value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Poor | Moderate | Good | Poor | Moderate | Good | Poor | Moderate | Good | |
| Clinical Manifestation | 6 | 0(0) | 2(33) | 4(67) | 0(0) | 0(0) | 6(100) | 0(0) | 2(33) | 4(66) | 0.471 |
| Pathogenesis | 9 | 1(11) | 7(78) | 1(11) | 0(0) | 4(44) | 5(56) | 0(0) | 4(44) | 5(56) | 0.023* |
| Diagnosis | 3 | 0(0) | 1(33) | 2(67) | 0(0) | 2(67) | 1(33) | 0(0) | 0(0) | 3(100) | 0.143 |
| Treatment | 6 | 0(0) | 3(50) | 3(50) | 0(0) | 5(83) | 1(17) | 1(17) | 3(50) | 2(33) | 0.250 |
| Prevention | 5 | 0(0) | 3(60) | 2(40) | 0(0) | 2(40) | 3(60) | 2(40) | 1(20) | 2(40) | 0.517 |
| Risk Factor | 5 | 1(20) | 2(40) | 2(40) | 0(0) | 3(60) | 2(40) | 0(0) | 2(40) | 3(60) | 1.000 |
| All Questions | 34 | 2(6) | 18(53) | 14(41) | 0(0) | 18(53) | 16(47) | 3(9) | 12(35) | 19(56) | 0.663 |

*$P<0.05$, **$P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; ^$P<0.05$, ^^$P<0.01$, ChatGPT-4 vs. Google Gemini

**Table 6** Accuracy ratings of LMs-chatbots' responses to questions related to 2022 ACR-GIOP Guideline

| Topic | Total, *n* | ChatGPT-3.5, *n*(%) | | | ChatGPT-4, *n*(%) | | | Google Gemini, *n*(%) | | | *P* value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Poor | Moderate | Good | Poor | Moderate | Good | Poor | Moderate | Good | |
| 2022 ACR-GIOP Guideline | 25 | 4(16) | 9(36) | 12(48) | 3(12) | 6(24) | 16(64) | 4(16) | 13(52) | 8(32) | 0.244 |

*$P<0.05$, **$P<0.01$, ChatGPT-3.5 vs. ChatGPT-4 and Google Gemini; ^$P<0.05$, ^^$P<0.01$, ChatGPT-4 vs. Google Gemini

**Table 7** Self-correcting capacity of ChatGPT-3.5

| Topic | Question No. | TS | | Accuracy ratings | |
|---|---|---|---|---|---|
| | | Initial | Self-correction | Initial | Self-correction |
| Pathogenesis | 8. Can glucocorticoids cause osteoporosis by affecting the gonads? | 4 | 8 | Poor | Moderate |
| risk factors | 5. How often does the risk assessment of fractures caused by glucocorticoid induced-osteoporosis need to be tested? | 4 | 6 | Poor | Moderate |
| 2022 ACR-GIOP Guideline | 4. In the way that fracture risk is assessed in patients treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months), is it possible that age leads to a different assessment? | 5 | 10 | Poor | Good |
| | 5. When should fracture risk be assessed in patients treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months)? How long do they need to be evaluated at intervals after that? | 3 | 7 | Poor | Moderate |
| | 16. What is the FRAX GC correction? | 3 | 4 | Poor | Poor |
| | 24.What has been updated in the 2022 GIOP Guidelines for the Prevention and Treatment of ACRs compared to the previous ones? | 5 | 5 | Poor | Poor |
| | | 4.00 ± 0.89 | 6.67 ± 2.16 | - | |
| ***P* value** | | 0.021* | | 0.061 | |

*$P<0.05$, **$P<0.01$, Initial TS vs. Self-corrected TS; ^$P<0.05$, ^^$P<0.01$, Initial accuracy ratings vs. Self-corrected accuracy ratings

**Table 8** Self-correcting capacity of ChatGPT-4

| Topic | Question No. | TS | | Accuracy ratings | |
|---|---|---|---|---|---|
| | | Initial | Self-correction | Initial | Self-correction |
| 2022 ACR-GIOP Guideline | 4. In the way that fracture risk is assessed in patients treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months), is it possible that age leads to a different assessment? | 4 | 11 | Poor | good |
| | 5. When should fracture risk be assessed in patients treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months)? How long do they need to be evaluated at intervals after that? | 5 | 10 | Poor | good |
| | 22.What are the indications for romocizumab in the treatment of glucocorticoid-induced osteoporosis? | 3 | 12 | Poor | good |
| | | 4.00 ± 1.00 | 11.00 ± 1.00 | - | |
| ***P* value** | | 0.026* | | 0.05 | |

*$P<0.05$, **$P<0.01$, Initial TS vs. Self-corrected TS; ^$P<0.05$, ^^$P<0.01$, Initial accuracy ratings vs. Self-corrected accuracy ratings

**Table 9** Self-correcting capacity of Google Gemini

| Topic | Question No. | TS | | Accuracy ratings | |
|---|---|---|---|---|---|
| | | Initial | Self-correction | Initial | Self-correction |
| treatments | 4. What are the dosing regimens for each of the drugs commonly used to treat glucocorticoid-induced osteoporosis? For which populations are they indicated? By which route of administration? | 3 | 3 | Poor | Poor |
| prevention | 2.Which medication can prevent glucocorticoid-induced osteoporosis? | 3 | 3 | Poor | Poor |
| | 5. How glucocorticoid-induced osteoporosis should be prevented in children? | 4 | 8 | Poor | Moderate |
| 2022 ACR-GIOP Guideline | 4. In the way that fracture risk is assessed in patients treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months), is it possible that age leads to a different assessment? | 4 | 9 | Poor | Moderate |
| | 5. When should fracture risk be assessed in patients treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months)? How long do they need to be evaluated at intervals after that? | 5 | 10 | Poor | Good |
| | 8.What are the therapeutic recommendations for patients at low fracture risk treated with glucocorticoids (≥ 2.5 mg/d for more than 3 months)? | 3 | 3 | Poor | Poor |
| | 16. What is the FRAX GC correction? | 3 | 3 | Poor | Poor |
| | | 3.00(3.00–8.00) | 3.00(3.00–9.00) | - | |
| ***P* value** | | 0.284 | | 0.192 | |

*$P<0.05$, **$P<0.01$, Initial TS vs. Self-corrected TS; ^$P<0.05$, ^^$P<0.01$, Initial accuracy ratings vs. Self-corrected accuracy ratings

ChatGPT-4. Supplementary Tables 5a-c show the LLM chatbot responses with TSs < 6. The specific parts of the responses that contain errors are highlighted in yellow. In addition, these tables provide further explanations of the errors identified by professional orthopedic physicians.

## Discussion

GIOP is caused by long-term use of glucocorticoid medications (usually defined as more than 3 months) in patients who suffer from a variety of inflammatory and autoimmune diseases, such as asthma, rheumatoid arthritis, and lupus erythematosus. It is characterized by a decrease in bone mineral density and susceptibility to fracture, a lack of obvious symptoms in the early stages, and a higher risk of osteoporosis in older patients, females, and patients who use higher doses of glucocorticosteroids [1, 5, 13, 33]. Long-term use of glucocorticoids may also cause or exacerbate other health problems, such as muscle loss, weight gain, high blood pressure, diabetes, and eye problems (e.g., cataracts) [7, 13, 34]. Based on these characteristics, the patient's quality of life may be affected, and the ability to perform daily activities may be reduced [35]. Regular monitoring of bone density and individualized risk assessments are crucial for patients who are using or need to use glucocorticosteroids for a long period of time [1, 8, 14, 36].

With the development of AI, LLM chatbots, such as ChatGPT-3.5, ChatGPT-4, and Google Gemini, have been widely applied in the medical field [19, 30, 37]. According to a study by Giovanni Maria Iannantuono et al., LLM chatbots can quickly provide cancer patients with medical knowledge, drug information, disease symptoms and treatments, and other relevant information [38]. According to a study by Giacomo Rossettini et

al., LLM chatbots (e.g., ChatGPT, Microsoft Bing, and Google Gemini) play a role in musculoskeletal rehabilitation by providing health counseling, medication management and reminders, and psychological support to patients [39]. In a study by Zhi Wei Lim et al., the ability of ChatGPT-3.5, ChatGPT-4, and Google Gemini to provide accurate responses to common myopia-related queries was evaluated, and the results showed that ChatGPT-4 is more able to provide accurate and comprehensive responses to myopia-related queries than the other LLMs [31]. According to Cigdem Cinar's study, ChatGPT had high accuracy in responses to general questions about osteoporosis and reduced accuracy in responses about osteoporosis guidelines [32]. There are no studies that have tested the performance of LLM chatbots in answering questions related to osteoporosis caused by glucocorticoids.

When answering the general GIOP-related questions ChatGPT-4, Google Gemini provided more concise answers than ChatGPT-3.5, and when answering the questions related to the 2022 ACR-GIOP Guidelines, number of Google Gemini generated shorter responses than ChatGPT-3.5 in terms of both words and characters, thus suggesting that Google Gemini may be more focused on providing concise and direct answers to improve the efficiency of information delivery (Tables 1 and 2). The above results suggest that due to the technical and algorithmic differences in LLM chatbots, they perform differently in information processing and question answering and that Google Gemini and ChatGPT-4 may focus more on providing concise and direct answers to improve the efficiency of information delivery. However, through the content of Google Gemini's specific answers, Google Gemini did not provide a clear answer for some questions, leading to a reduction in the length of the answer (Supplementary Table 2, 3a-c). This difference may be related to the different LLM chatbot algorithms used [40]. In contrast, ChatGPT-3.5 may provide more detailed information, including background information, multiple perspectives, or additional explanations, which increases the number of characters and words. ChatGPT-4, which is an iteration of ChatGPT-3.5 that takes into account user feedback and improvement, adopts a more advanced linguistic representation and uses a larger and more diverse dataset; thus, it better captures linguistic patterns and meets users' needs for high-quality answers, which is a sign of continuous progress in the field of natural language processing [38, 41, 42].

We also commissioned three professional orthopedic experts to rate the accuracy of responses generated by different LLM chatbots (Table 3; Supplementary Table 1, 2a-c). In terms of pathological mechanisms, the TS of ChatGPT-3.5 was significantly lower than that of ChatGPT-4 ($P<0.05$), and the accuracy of ChatGPT-3.5 was also significantly lower than that of ChatGPT-4 and Google Gemini. However, there was no significant difference in the ratings of three different LLM chatbots on the remaining topics. In response to questions related to the 2022 ACR-GIOP Guidelines, Google Gemini's TS was significantly lower than that of ChatGPT-4 ($P<0.05$), and there was no significant difference between the ratings of ChatGPT-4 and ChatGPT-3.5 (Table 4; Supplementary Table 1, 3a-c). The difference in scores between ChatGPT-3.5 and ChatGPT-4, Google Gemini could be due to a number of factors. ChatGPT-3.5 was trained on data available up to January 2022, and ChatGPT-4 was launched by OpenAI in March 2023. Building on the foundation of ChatGPT-3.5, ChatGPT-4 adopts a more advanced model architecture and richer training data in the medical domain. Furthermore, ChatGPT-4 is able to perform domain-specific fine-tuning and improve the contextual understanding of specialized medical terminology. More importantly, ChatGPT-4 improves the accuracy of answering questions in specific medical domains by iteratively improving the user feedback of ChatGPT-3.5 [43]. Therefore, in our study, the accuracy of ChatGPT-4 was significantly higher than that of ChatGPT-3.5 in terms of answering questions about the pathological mechanisms of GIOP. Google Gemini, which was developed by Google and leverages Google's long experience in search, natural language processing, and other AI areas, is fundamentally different from ChatGPT-4 (developed by OpenAI) in the way it processes information and answers questions [44–47]. In our study, we found that ChatGPT-4 performed better than Google Gemini in answering questions such as the 2022 ACR-GIOP Guideline. It is possible that ChatGPT-4's dataset contains more professional literature or guidelines related to newer glucocorticoids and osteoporosis, and thus, it will be more accurate in processing related questions.

In our study, we also compared the self-correcting updating ability of three different LLM chatbots by prompting questions with a "poor" answer rating and then compared the self-correcting ability with the ratings of a professional orthopedic surgeon. Our study showed that ChatGPT3.5 had a total of six responses rated as "poor" on all questions, two of which were related to general GIOP information and four of which were related to the 2022 ACR-GIOP Guidelines. The overall performance was "poor", with an average TS of $4.00\pm0.89$ and an average TS of $6.67\pm2.16$ after correction, which was significantly higher than pre-correction TS ($P<0.05$). However, there was no significant improvement in the accuracy after correction ($P>0.05$). There were three questions that yielded "poor" responses from ChatGPT4, all of which were questions about the 2022 ACR-GIOP Guidelines. The average pre-correction TS was $4.00\pm1.00$ before correction; after correction, the

responses were rated as "good", with an average TS of 11.00±1.00 ($P<0.05$). There were seven questions that received a "poor" response from Google Gemini, three of which were related to general information about GIOP and four of which were related to the 2022 ACR-GIOP Guidelines. The results showed that Google Gemini's scores and grades did not change significantly between pre-correction and post-correction ($P>0.05$). The three different LLM chatbots performed poorly in answering such questions about the 2022 ACR-GIOP Guidelines, with all of them generating 3–4 responses with poor ratings. According to professional orthopedic surgeons, the "poor" responses were mostly due to a lack of specificity in the details, a failure to answer according to the guidelines, and the inability to respond professionally to the questions asked. These findings indicate that these LLM chatbots have poor knowledge of the 2022 ACR-GIOP Guidelines, thus suggesting that the chatbots may have limited utility for patients with GIOP. Furthermore, these findings indicate that appropriate management and timely assessment of GIOP by a professional health care team are essential for different conditions.

### Strengths and limitations

The selected questions we chose might not have been comprehensive enough and might have biased the ultimate answers generated by the three LLMs. The scoring system used in this study was developed by ourselves. The Fleiss' Kappa values of three orthopaedic surgeons scoring the responses generated by ChatGPT-3.5, ChatGPT-4, and Google Gemini were 0.384, 0.350, and 0.340, respectively, suggesting that our scoring system is generally reliable across evaluators. This result may be related to the professional background and experience of the evaluator. Our study has some time limitations; new, relevant content about GIOP will be constantly uploaded to the internet in the future. The iterative updating of LLMs, such as ChatGPT-3.5 and ChatGPT-4.0, is quick and may soon outpace the models evaluated in this study. As a result, the benchmark comparisons made in this thesis may quickly become outdated. However, the insights we have gained in applying these models to specific diseases are still of great value and provide a foundation for future studies using more advanced models. In addition, our studies were conducted in English, and we did not study dialogs in other languages. Finally, we did not assess the comprehensibility of the responses to the three LLMs because the comprehensibility of LLMs varies across education levels. There are no relevant articles reporting the use of LLMs in daily care for GIOP patients or caregivers. However, the functions of LLMs, such as disease education, treatment recommendations, and patient counseling, can improve the quality and efficiency of care, and LLMs can be extended to other potential applications in GIOP care, such as the development of personalized treatment plans and the implementation of real-time monitoring and alerts. Technical improvements are needed due to the technical limitations of LLMs in health care applications, such as improving the accuracy of the model, increasing the training of medically specific datasets, ensuring the security and privacy of patient data to avoid data misuse, and ensuring that the application of the technology complies with the ethical and legal frameworks of health care. These improvements would enhance the applicability of LLMs in GIOP care.

## Conclusion

Our study showed that ChatGPT-4 and Google Gemini provided more concise and intuitive answers than CChatGPT3.5 and that ChatGPT-4 performed significantly better than did CChatGPT3.5 and Google Gemini in terms of answering general questions related to GIOP. Our findings also showed that ChatGPT3.5 and ChatGPT-4 self-corrected better than Google Gemini. This finding might be related to differences in design patterns, training database updates and application algorithms between the LLMs.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13018-024-04996-2.

Supplementary Material 1

#### Acknowledgements

Not applicable.

#### Author contributions

LT, CZ and ZS designed the study. ZS provided the funding. LT contributed to the data collection. LT and CZ wrote the manuscript. LT and CZ provided resources and participated in the data analysis. JY, RL and ZS independently rated the responses. LT and CZ performed the data validation and edited the manuscript. LT and ZS confirmed the authenticity of all the raw data. All authors have read and approved the final manuscript.

#### Funding

This work was supported by the Natural Science Foundation of Tianjin (grant number 23JCYBJC01740).

#### Data availability

No datasets were generated or analysed during the current study.

## Declarations

#### Ethics approval and consent to participate

This study did not involve the use of human participants, human data, or human tissue.

#### Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

#### Competing interests

The authors declare no competing interests.

**Author details**
[1]Clinical College of Neurology, Neurosurgery and Neurorehabilitation, Tianjin Medical University, Tianjin 300070, China
[2]Department of Orthopedics, Tianjin Medical University Baodi Hospital, Tianjin 301800, China

Received: 25 April 2024 / Accepted: 12 August 2024
Published online: 18 September 2024

**References**

1. Humphrey MB, Russell L, Danila MI, Fink HA, Guyatt G, Cannon M, Caplan L, Gore S, Grossman J, Hansen KE, et al. 2022 American College of Rheumatology Guideline for the Prevention and Treatment of Glucocorticoid-Induced osteoporosis. Arthritis Rheumatol. 2023;75(12):2088–102. https://doi.org/10.1002/art.42646.
2. Migliorini F, Colarossi G, Eschweiler J, Oliva F, Driessen A, Maffulli N. Antiresorptive treatments for corticosteroid-induced osteoporosis: a bayesian network meta-analysis. Br Med Bull. 2022;143(1):46–56. https://doi.org/10.1093/bmb/ldac017.
3. Cheng CH, Chen LR, Chen KH. Osteoporosis due to hormone imbalance: an overview of the effects of Estrogen Deficiency and glucocorticoid overuse on bone turnover. Int J Mol Sci. 2022;23(3). https://doi.org/10.3390/ijms23031376.
4. Wang T, Liu X, He C. Glucocorticoid-induced autophagy and apoptosis in bone. Apoptosis. 2020;25(3–4):157–68. https://doi.org/10.1007/s10495-020-01599-0.
5. den Uyl D, Bultink IE, Lems WF. Advances in glucocorticoid-induced osteoporosis. Curr Rheumatol Rep. 2011;13(3):233–40. https://doi.org/10.1007/s11926-011-0173-y.
6. Rizzoli R, Biver E. Glucocorticoid-induced osteoporosis: who to treat with what agent? Nat Rev Rheumatol. 2015;11(2):98–109. https://doi.org/10.1038/nrrheum.2014.188.
7. Silverman SL, Lane NE. Glucocorticoid-induced osteoporosis. Curr Osteoporos Rep. 2009;7(1):23–6. https://doi.org/10.1007/s11914-009-0005-4.
8. Buckley L, Guyatt G, Fink HA, Cannon M, Grossman J, Hansen KE, Humphrey MB, Lane NE, Magrey M, Miller M, et al. 2017 American College of Rheumatology Guideline for the Prevention and Treatment of Glucocorticoid-Induced osteoporosis. Arthritis Rheumatol. 2017;69(8):1521–37. https://doi.org/10.1002/art.40137.
9. Adami G, Saag KG. Glucocorticoid-induced osteoporosis: 2019 concise clinical review. Osteoporos Int. 2019;30(6):1145–56. https://doi.org/10.1007/s00198-019-04906-x.
10. Migliorini F, Colarossi G, Baroncini A, Eschweiler J, Tingart M, Maffulli N. Pharmacological management of postmenopausal osteoporosis: a Level I evidence based - Expert Opinion. Expert Rev Clin Pharmacol. 2021;14(1):105–19. https://doi.org/10.1080/17512433.2021.1851192.
11. Migliorini F, Maffulli N, Colarossi G, Eschweiler J, Tingart M, Betsch M. Effect of drugs on bone mineral density in postmenopausal osteoporosis: a bayesian network meta-analysis. J Orthop Surg Res. 2021;16(1):533. https://doi.org/10.1186/s13018-021-02678-x.
12. Migliorini F, Giorgino R, Hildebrand F, Spiezia F, Peretti GM, Alessandri-Bonetti M, Eschweiler J, Maffulli N. Fragility fractures: risk factors and management in the Elderly. Med (Kaunas). 2021;57(10). https://doi.org/10.3390/medicina57101119.
13. Anastasilaki E, Paccou J, Gkastaris K, Anastasilakis AD. Glucocorticoid-induced osteoporosis: an overview with focus on its prevention and management. Horm (Athens). 2023;22(4):611–22. https://doi.org/10.1007/s42000-023-00491-1.
14. Cho SK, Sung YK. Update on glucocorticoid Induced osteoporosis. Endocrinol Metab (Seoul). 2021;36(3):536–43. https://doi.org/10.3803/EnM.2021.1021.
15. Pruneski JA, Pareek A, Nwachukwu BU, Martin RK, Kelly BT, Karlsson J, Pearle AD, Kiapour AM, Williams RJ 3. Natural language processing: using artificial intelligence to understand human language in orthopedics. Knee Surg Sports Traumatol Arthrosc. 2023;31(4):1203–11. https://doi.org/10.1007/s00167-022-07272-0.
16. Arivazhagan N, Van Vleck TT. Natural Language Processing Basics. Clin J Am Soc Nephrol. 2023;18(3):400–1. https://doi.org/10.2215/cjn.0000000000000081.
17. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2.
18. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, Naugler C. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inf Decis Mak. 2024;24(1):72. https://doi.org/10.1186/s12911-024-02459-6.
19. Shieh A, Tran B, He G, Kumar M, Freed JA, Majety P. Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. Sci Rep. 2024;14(1):9330. https://doi.org/10.1038/s41598-024-58760-x.
20. Wang G, Gao K, Liu Q, Wu Y, Zhang K, Zhou W, Guo C. Potential and limitations of ChatGPT 3.5 and 4.0 as a source of COVID-19 information: Comprehensive Comparative Analysis of Generative and authoritative information. J Med Internet Res. 2023;25:e49771. https://doi.org/10.2196/49771.
21. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large Language models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study. JMIR Med Inf. 2024;12:e55318. https://doi.org/10.2196/55318.
22. Amin KS, Mayes LC, Khosla P, Doshi RH. Assessing the efficacy of large Language models in Health literacy: a comprehensive cross-sectional study. Yale J Biol Med. 2024;97(1):17–27. https://doi.org/10.59249/ztoz1966.
23. Carlà MM, Gambini G, Baldascino A, Boselli F, Giannuzzi F, Margollicci F, Rizzo S. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. Graefes Arch Clin Exp Ophthalmol. 2024. https://doi.org/10.1007/s00417-024-06470-5.
24. Bazzari FH, Bazzari AH. Utilizing ChatGPT in Telepharmacy. Cureus. 2024;16(1):e52365. https://doi.org/10.7759/cureus.52365.
25. Athavale A, Baier J, Ross E, Fukaya E. The potential of chatbots in chronic venous disease patient management. JVS Vasc Insights. 2023;1. https://doi.org/10.1016/j.jvsvi.2023.100019.
26. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, Wong WM, Wang X, Jin Tan MC, Chang Koh VT, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. 2023;26(11):108163. https://doi.org/10.1016/j.isci.2023.108163.
27. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of Diagnostic and Triage Accuracy of Ada Health and WebMD Symptom checkers, ChatGPT, and Physicians for patients in an Emergency Department: Clinical Data Analysis Study. JMIR Mhealth Uhealth. 2023;11:e49995. https://doi.org/10.2196/49995.
28. Posner KM, Bakus C, Basralian G, Chester G, Zeiman M, O'Malley GR, Klein GR. Evaluating ChatGPT's capabilities on Orthopedic Training examinations: an analysis of New Image Processing features. Cureus. 2024;16(3):e55945. https://doi.org/10.7759/cureus.55945.
29. Revilla-León M, Barmak BA, Sailer I, Kois JC, Att W. Performance of an Artificial Intelligence-based Chatbot (ChatGPT) answering the European certification in Implant Dentistry exam. Int J Prosthodont. 2024;37(2):221–4. https://doi.org/10.11607/ijp.8852.
30. Kim TW. Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: a narrative review. J Educ Eval Health Prof. 2023;20(38). https://doi.org/10.3352/jeehp.2023.20.38.
31. Lim ZW, Pushpanathan K, Yew SME, Lai Y, Sun CH, Lam JSH, Chen DZ, Goh JHL, Tan MCJ, Sheng B, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. 2023;95:104770. https://doi.org/10.1016/j.ebiom.2023.104770.
32. Cinar C. Analyzing the performance of ChatGPT about osteoporosis. Cureus. 2023;15(9):e45890. https://doi.org/10.7759/cureus.45890.
33. Conti V, Russomanno G, Corbi G, Toro G, Simeon V, Filippelli W, Ferrara N, Grimaldi M, D'Argenio V, Maffulli N, et al. A polymorphism at the translation start site of the vitamin D receptor gene is associated with the response to anti-osteoporotic therapy in postmenopausal women from southern Italy. Int J Mol Sci. 2015;16(3):5452–66. https://doi.org/10.3390/ijms16035452.
34. Migliorini F, Maffulli N, Spiezia F, Tingart M, Maria PG, Riccardo G. Biomarkers as therapy monitoring for postmenopausal osteoporosis: a systematic review. J Orthop Surg Res. 2021;16(1):318. https://doi.org/10.1186/s13018-021-02474-7.
35. Chiodini I, Merlotti D, Falchetti A, Gennari L. Treatment options for glucocorticoid-induced osteoporosis. Expert Opin Pharmacother. 2020;21(6):721–32. https://doi.org/10.1080/14656566.2020.1721467.

36. Migliorini F, Maffulli N, Spiezia F, Peretti GM, Tingart M, Giorgino R. Potential of biomarkers during pharmacological therapy setting for postmenopausal osteoporosis: a systematic review. J Orthop Surg Res. 2021;16(1):351. https://doi.org/10.1186/s13018-021-02497-0.
37. Shen OY, Pratap JS, Li X, Chen NC, Bhashyam AR. How does ChatGPT Use Source Information compared with Google? A Text Network Analysis of Online Health Information. Clin Orthop Relat Res. 2024;482(4):578–88. https://doi.org/10.1097/corr.0000000000002995.
38. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol. 2023;13:1268915. https://doi.org/10.3389/fonc.2023.1268915.
39. Rossettini G, Cook C, Palese A, Pillastrini P, Turolla A. Pros and cons of using Artificial Intelligence Chatbots for Musculoskeletal Rehabilitation Management. J Orthop Sports Phys Ther. 2023;53(12):1–7. https://doi.org/10.2519/jospt.2023.12000.
40. 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. https://doi.org/10.3389/frai.2023.1169595.
41. Xu D, Zhao J, Liu R, Dai Y, Sun K, Wong P, Ming SLS, Wearn KL, Wang J, Xie S, et al. ChatGPT4's proficiency in addressing patients' questions on systemic lupus erythematosus: a blinded comparative study with specialists. Rheumatology (Oxford). 2024. https://doi.org/10.1093/rheumatology/keae238.
42. Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, Staubli SM. Reliability of Medical Information provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res. 2023;25:e47479. https://doi.org/10.2196/47479.
43. Delsoz M, Madadi Y, Munir WM, Tamm B, Mehravaran S, Soleimani M, Djalilian A, Yousefi S. Performance of ChatGPT in Diagnosis of Corneal Eye Diseases. *medRxiv* 2023. https://doi.org/10.1101/2023.08.25.23294635
44. Mihalache A, Grad J, Patil NS, Huang RS, Popovic MM, Mallipatna A, Kertes PJ, Muni RH. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. Eye (Lond). 2024. https://doi.org/10.1038/s41433-024-03067-4.
45. Masalkhi M, Ong J, Waisberg E, Lee AG. Google DeepMind's gemini AI versus ChatGPT: a comparative analysis in ophthalmology. Eye (Lond). 2024. https://doi.org/10.1038/s41433-024-02958-w.
46. Carlà MM, Gambini G, Baldascino A, Giannuzzi F, Boselli F, Crincoli E, D'Onofrio NC, Rizzo S. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. Br J Ophthalmol. 2024. https://doi.org/10.1136/bjo-2023-325143.
47. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-Head comparison of ChatGPT Versus Google search for Medical Knowledge Acquisition. Otolaryngol Head Neck Surg. 2023. https://doi.org/10.1002/ohn.465.

## Publisher's Note