BMC Musculoskeletal Disorders

RESEARCH

Open Access

Comparative analysis of ChatGPT-4o mini, ChatGPT-4o and Gemini Advanced in the treatment of postmenopausal osteoporosis

Rui Liu[1†], Jian Liu[2†], Jia Yang[1], Zhiming Sun[1*] and Hua Yan[1*]

Abstract

Background Osteoporosis is a sex-specific disease. Postmenopausal osteoporosis (PMOP) has been the focus of public health research worldwide. The purpose of this study is to evaluate the quality and readability of artificial intelligence large-scale language models (AI-LLMs): ChatGPT-4o mini, ChatGPT-4o and Gemini Advanced for responses generated in response to questions related to PMOP.

Methods We collected 48 PMOP frequently asked questions (FAQs) through offline counseling and online medical community forums. We also prepared 24 specific questions about PMOP based on the Management of Postmenopausal Osteoporosis: 2022 ACOG Clinical Practice Guideline No. 2 (2022 ACOG-PMOP Guideline). In this project, the FAQs were imported into the AI-LLMs (ChatGPT-4o mini, ChatGPT-4o, Gemini Advanced) and randomly assigned to four professional orthopedic surgeons, who independently rated the satisfaction of each response via a 5-point Likert scale. Furthermore, a Flesch Reading Ease (FRE) score was calculated for each of the LLMs' responses to assess the readability of the text generated by each LLM.

Results When it comes to addressing questions related to PMOP and the 2022 ACOG-PMOP guidelines, ChatGPT-4o and Gemini Advanced provide more concise answers than ChatGPT-4o mini. In terms of the overall FAQs of PMOP, ChatGPT-4o has a significantly higher accuracy rate than ChatGPT-4o mini and Gemini Advanced. When answering questions related to the 2022 ACOG-PMOP guidelines, ChatGPT-4o mini vs. ChatGPT-4o have significantly higher response accuracy than Gemini Advanced. ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced all have good levels of self-correction.

Conclusions Our research shows that Gemini Advanced and ChatGPT-4o provide more concise and intuitive answers. ChatGPT-4o responds better in answering frequently asked questions related to PMOP. When answering questions related to the 2022 ACOG-PMOP guidelines, ChatGPT-4o mini and ChatGPT-4o responded significantly better than Gemini Advanced. ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced have demonstrated a strong ability to self-correct.

†Rui Liu and Jian Liu contributed equally to the study.

*Correspondence:
Zhiming Sun
renzushanxiaoguai@163.com
Hua Yan
yanhua20042007@sina.com

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

Clinical trial number Not applicable.

Keywords Postmenopausal osteoporosis, Artificial intelligence large-scale Language models, ChatGPT-4o mini, ChatGPT-4o, Gemini advanced

Background

PMOP represents a systemic bone disease in which women experience increased bone resorption and reduced bone formation [1]. This is a consequence of the decrease in estrogen levels that occurs after menopause, resulting in a reduction in bone mass, destruction of the bone microstructure, increased bone fragility and susceptibility to fracture [2, 3]. Approximately one in two women older than 50 years will experience an osteoporotic fracture [4]. A study revealed that the prevalence rate of PMOP was 11.4% for women aged < 65 years and 26.4% for those aged ≥ 60–69 years [5]. In recent years, there has been a notable increase in the attention and focus on the treatment and prevention of PMOP within the medical community. System-level and national health care programs have been implemented worldwide. However, studies have demonstrated that 80–90% of adults do not receive appropriate osteoporosis management, even in secondary prevention [6, 7]. Additionally, many postmenopausal women do not take timely preventive measures, such as calcium and vitamin D supplementation or regular bone density testing [8]. This suggests that there is a lack of awareness about the severity of the disease, its prevention methods, and the potential health risks associated with it. It can be reasonably deduced that increased public awareness of the disease will contribute to the implementation of early prevention and intervention strategies, which will in turn reduce the associated health risks.

AI-LLMs are sophisticated neural network computer programs based on deep learning techniques that are capable of reading and comprehending text, as well as learning through the analysis of vast quantities of textual data, thereby continuously increasing their capacity to understand and generate language [9–11] AI-LLMs have the potential to suggest personalized treatment plans and assist physicians in making more appropriate treatment decisions [12]. In a study conducted by Mohammad Delsoz, AI-LLMs demonstrated the potential to assist physicians in the primary triage of glaucoma patients and in the clinical practice of eye care by analyzing the history and symptoms of glaucoma patients [13]. In a study conducted by Potapenko et al., AI-LLMs were utilized to respond to queries pertaining to prevalent retinal conditions [14]. These findings indicated that ChatGPT furnished more precise responses. In a study conducted by Grünebaum et al., AI-LLMs were utilized to respond to queries pertaining to obstetrics and gynecology [15]. These applications markedly increase the efficacy and caliber of health care services, offering robust technical assistance for the advancement of personalized and precision medicine [16]. Nevertheless, it remains uncertain whether AI-LLMs are capable of providing up-to-date information and making clinical decisions in the context of postmenopausal osteoporosis. The objective of this study was to evaluate the performance of the latest AI-LLMs (ChatGPT-4o, ChatGPT-4o, Gemini Advanced) in generating professional and clinically accurate responses to common clinical questions about postmenopausal osteoporosis and in accordance with the 2022 ACOG Guidelines. This study was to also ascertain whether AI-LLMs can facilitate improvements in postmenopausal osteoporosis patients' comprehension of the disease, self-management abilities, the advancement of personalized medicine, and the alleviation of the scarcity of health care resources.

Methods

The FAQs were selected for the purpose of investigating the applicability of different AI-LLMs to common clinical settings. The 48 FAQs (Supplementary Table 1a) related to PMOP were divided into six domains: clinical manifestation, diagnosis, pathogenesis, treatment, prevention, and risk factors. This division was performed to explore the ability of different AI-LLMs to address different PMOP conditions. The sources of these questions include MedlinePlus, the Cochrane Library, the National Osteoporosis Foundation, the Mayo Clinic, UpToDate, WebMD, and offline patient counseling. To further examine the comprehension of different AI-LLMs regarding the Specialized PMOP's Guidelines, 24 additional questions were formulated on the basis of the 2022 ACOG-PMOP Guidelines (Supplementary Table 1b). The most current versions of the AI-LLMs were utilized in this study. The AI-LLMs utilized in this study were ChatGPT-4o mini (July 18, 2024), ChatGPT-4o (May 13, 2024), and Gemini Advanced (Gemini 1.5 Pro, June 24, 2024). When queries were posed to the AI-LLMs, a new dialog box was generated for each question, and the responses were collated at the conclusion of the interaction. Any references to the AI-LLMs were removed, the responses were aggregated, and they were then randomly assigned to four orthopedic specialists with expertise in the treatment of osteoporosis for separate dialog inputs for Likert scale scoring. Each dialog was subsequently reset after each query to collate the content of the replies. The content of the AI-LLM replies was converted to plain text format, and any information in the text identifying the AI-LLMs

was removed. In addition, we counted the characters, total words, total syllables, total sentences, and FRE score (206.835–1.015 (total words/total sentences)-84.6 (total syllables/total words)) for each response content(90–100: Very easy to read; 80–89: Easy to read; 70–79: Easier to read; 60–69: Standard reading difficulty; 50–59: Difficult to read, suitable for college students or professionals; 30–49: Difficult to read, suitable for experts or readers in a specific field; 0–29: Very difficult to read, usually an academic paper or legal document [17]). Figure 1 illustrates the design flow of the present study.

Four orthopedic surgeons with expertise in specialized fields evaluated the AI-LLM responses via a 5-point Likert scale [18, 19] (1 for fully disagree, 2 for partially disagree, 3 for neither agree nor disagree, 4 for partially agree, 5 for fully agree). AS ≤ 2 is indicative of poor performance, 2 < AS ≤ 3 is indicative of fair performance, 3 < AS ≤ 4 is indicative of good performance, and AS > 4 is indicative of excellent performance. The consistency of the responses of the four specialized orthopedic surgeons to ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced was evaluated via Fleiss's kappa coefficient.

We further explored the ability of AI-LLMs to self-correct. Questions with an AS ≤ 2 that were identified as 'poor', where the incorrect part was pointed out by an orthopedic specialist, were subject to further questioning along the lines of 'You do not seem to have answered that correctly, can you answer it again?' Replies were collected and converted to plain text format, and any information in the text identifying the LLM chatbot was removed and randomly assigned to four raters to reevaluate the corrected content. This round of evaluation was completed two weeks after the final round of scoring.

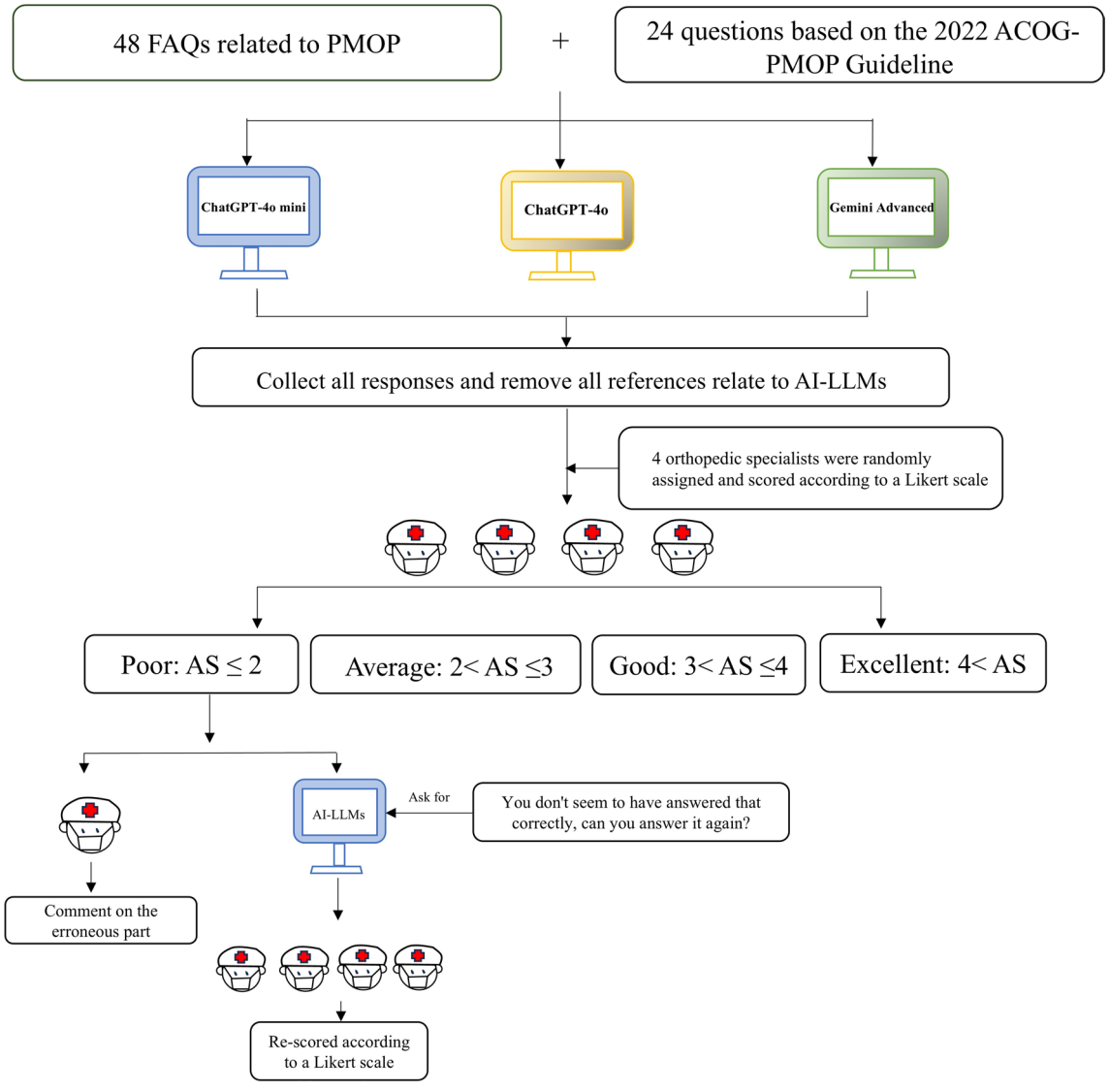

Fig. 1 Flowchart of the overall study design

During the initial round of re-evaluation, the raters were not informed that the responses were self-correcting versions.

Statistical analysis

The data analysis was conducted via SPSS 26 software (released by IBM Corp. in 2021). Normally distributed data are expressed as the mean ± standard deviation, whereas non-normally distributed data are expressed as the median (25th–75th percentile) ($M(P_{25}\text{-}P_{75})$). Statistical comparisons were performed via the Kruskal–Wallis H test to determine the significance of differences in the FRE score and AS between ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced. When significant differences among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced were detected, Dunn's test with Bonferroni correction was applied to identify specific pairwise differences. Paired t tests were employed to evaluate the initial AS and self-corrected AS. For categorical outcomes, ratings were dichotomized into 'excellent' vs. 'other', statistical comparisons were performed via the Pearson's chi-square tests to determine the significance of differences in ratings the between ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced. When significant differences among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced were detected, Bonferroni correction (adjusted $\alpha = \frac{0.05}{3}$) was applied to identify specific pairwise differences. The consistency of the responses from the four advanced orthopedic surgeons to the question ratings on the ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced scores was assessed via Fleiss's kappa. Fleiss's kappa coefficient ranged between 0 and 1. According to the established criteria, consistency is classified as poor when the coefficient is between 0 and 0.2, moderate when it is between 0.2 and 0.4, moderate when it is between 0.4 and 0.6, strong when it is between 0.6 and 0.8, and very strong when it is between 0.8 and 1.0.

Results

Length and FRE score of the responses from ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced

Table 1 shows the average total characters, total words, total syllables, total sentence lengths, and FRE scores generated by the AI-LLMs for the FAQs of the PMOPs in the different subject areas. The total characters, total words, total syllables, total sentences, and FRE score responses to the individual questions on the AI-LLMs are shown in Supplementary Tables 2a-c. The *P*-values corrected using Dunn's test and Bonferroni are shown in Supplementary Tables 2d-g. There was no significant difference in the total number of characters, total words, total syllables, or total sentence lengths or FRE scores generated by the AI-LLMs for the topics of "Clinical Manifestation", "Treatment", and "Prevention", and the FRE scores were not significantly different. In the FAQ responses related to the topic of "Diagnosis", Gemini Advance's total number of characters (1484.88 ± 377.67 vs. 2355.00 ± 796.14), total words (227.00 ± 55.68 vs. 338.13 ± 110.78), total syllables (425.00 ± 124.67 vs. 677.25 ± 225.17), and total sentences (13.63 ± 3.62 vs. 25.25 ± 10.18) were significantly less than that of ChatGPT-4o mini ($P < 0.05$), and the FRE score of ChatGPT-4o was significantly greater than that of ChatGPT-4o mini (47.55 ± 17.08 vs. 23.24 ± 5.10) ($P < 0.05$). In the FAQ responses with the topic "Pathogenesis", ChatGPT-4o with Gemini Advanced had significantly fewer total syllables than ChatGPT-4o mini (384.88 ± 158.79, 368.88 ± 133.67 vs. 623.00 ± 266.78) ($P < 0.05$), and the total number of sentences with ChatGPT-4o was significantly less than those of ChatGPT-4o mini (11.00 ± 3.82 vs. 23.13 ± 10.45) ($P < 0.05$). In the FAQ responses concerning the topic of "Risk Factor", Gemini Advanced had a significantly greater FRE score than ChatGPT-4o mini (55.11 ± 15.21 vs. 24.06 ± 12.63) ($P < 0.05$). Summarizing all the FAQ responses revealed that Gemini Advanced had significantly fewer total syllables and total sentences than ChatGPT-4o mini (409.90 ± 168.96 vs. 562.56 ± 230.88, 16.08 ± 8.27 vs. 22.31 ± 9.84) ($P < 0.05$). The FRE score of ChatGPT-4o was significantly greater than that of ChatGPT-4o mini (37.23 ± 13.18 vs. 28.88 ± 11.69) ($P < 0.05$). Table 2 shows the length and FRE score, total characters, total words, total syllables, and total sentences of the AI-LLMs' responses to the questions related to the 2022 ACOG-PMOP Guideline. The results revealed that the total words and total syllables of Gemini Advanced were significantly lower than those of ChatGPT-4o and ChatGPT-4o mini (221.67 ± 57.33 vs. 316.04 ± 120.65, 329.96 ± 128.78; 344.79 ± 91.34 vs. 618.13 ± 237.32, 642.46 ± 258.46) ($P < 0.05$), and the FRE score of Gemini Advanced was significantly greater than that of ChatGPT-4o and ChatGPT-4o mini (63.40 ± 2.58 vs. 27.00 ± 9.46, 25.77 ± 13.33) ($P < 0.05$).

AS and grading of the ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced responses

Table 3 shows the ASon the Likert scale of the ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced's responses to the FAQs about PMOP on different topics. The Likert scale scores for the four reviewers' responses to the individual questions on the AI-LLMs are shown in Supplementary Tables 2a-c. The *P*-values corrected using Dunn's test and Bonferroni are shown in Supplementary Table 2 h. In terms of "Diagnosis", Gemini Advanced's AS (2.53 ± 1.15) was significantly lower than that of ChatGPT-4o (4.19 ± 0.78) ($P < 0.05$). In terms of "Pathogenesis", ChatGPT-4o mini had a significantly lower AS (2.84 ± 0.86) than ChatGPT-4o (4.28 ± 0.47) ($P < 0.05$). In terms of "Risk Factor", Gemini Advanced had a

Table 1 Length and FRE of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced's responses to faqs about PMOP

	ChatGPT-4o mini, ($\bar{x} \pm sd$)	ChatGPT-4o, ($\bar{x} \pm sd$)	Gemini Advanced, ($\bar{x} \pm sd$)	*P* value
Clinical Manifestation				
Total characters	1633.88 ± 876.35	1528.75 ± 1114.82	1158.75 ± 263.91	0.512
Total words	242.25 ± 119.61	235.63 ± 172.25	150.25 ± 40.32	0.281
Total syllables	468.50 ± 248.18	442.50 ± 315.96	268.88 ± 80.44	0.179
Total sentences	15.88 ± 6.62	16.25 ± 10.73	8.38 ± 2.26	0.052
FRE	28.49 ± 15.46	30.53 ± 12.05	37.33 ± 18.01	0.608
Diagnosis				
Total characters	2355.00 ± 796.14	1971.75 ± 671.95	1484.88 ± 377.67*	0.019
Total words	338.13 ± 110.78	310.25 ± 100.11	227.00 ± 55.68*	0.022
Total syllables	677.25 ± 225.17	491.25 ± 139.00	425.00 ± 124.67*	0.019
Total sentences	25.25 ± 10.18	18.75 ± 8.24	13.63 ± 3.62**	0.012
FRE	23.24 ± 5.10	47.55 ± 17.08*	31.70 ± 18.02	0.027
Pathogenesis				
Total characters	2186.00 ± 911.15	1457.38 ± 505.70	1561.50 ± 671.83	0.087
Total words	317.88 ± 127.21	219.00 ± 77.64	212.00 ± 83.41	0.077
Total syllables	623.00 ± 266.78	368.88 ± 133.67*	384.88 ± 158.79*	0.043
Total sentences	23.13 ± 10.45	11.00 ± 3.82*	14.25 ± 7.46	0.024
FRE	28.00 ± 10.74	39.72 ± 11.57	31.79 ± 20.94	0.145
Treatment				
Total characters	2086.50 ± 629.51	2220.75 ± 632.69	2229.13 ± 771.09	0.867
Total words	300.50 ± 88.90	319.63 ± 92.79	279.13 ± 122.15	0.886
Total syllables	583.88 ± 172.19	606.00 ± 165.48	553.63 ± 234.78	0.998
Total sentences	27.00 ± 8.35	25.88 ± 8.22	22.63 ± 10.95	0.640
FRE	30.32 ± 10.41	32.50 ± 11.83	23.29 ± 11.55	0.242
Prevention				
Total characters	1978.50 ± 787.73	2343.63 ± 856.17	1795.13 ± 519.30	0.300
Total words	294.38 ± 113.46	349.38 ± 127.22	273.50 ± 67.37	0.391
Total syllables	541.88 ± 226.26	624.13 ± 229.81	470.25 ± 115.20	0.296
Total sentences	23.25 ± 8.58	25.75 ± 10.62	18.50 ± 3.55	0.195
FRE	39.19 ± 9.01	40.90 ± 7.75	41.73 ± 16.88	0.746
Risk Factor				
Total characters	1755.13 ± 956.98	1455.88 ± 564.12	1524.13 ± 714.82	0.811
Total words	246.88 ± 126.38	217.63 ± 79.87	218.13 ± 91.97	0.878
Total syllables	480.88 ± 235.89	405.75 ± 150.42	356.75 ± 146.10	0.665
Total sentences	19.38 ± 12.51	17.00 ± 12.35	19.13 ± 10.37	0.871
FRE	24.06 ± 12.63	32.20 ± 12.05	55.11 ± 15.21*	0.005
All Questions				
Total characters	1999.17 ± 825.82	1829.69 ± 802.40	1625.58 ± 643.02	0.060
Total words	275.25 ± 119.59	275.25 ± 119.59	226.67 ± 88.22	0.067
Total syllables	562.56 ± 230.88	489.75 ± 212.52	409.90 ± 168.96**	0.004
Total sentences	22.31 ± 9.84	19.10 ± 10.38	16.08 ± 8.27**	0.007
FRE	28.88 ± 11.69	37.23 ± 13.18*	36.82 ± 18.97	0.015

P < 0.05 indicates significance, *P* < 0.01 indicates high significance. The *P*-value is obtained using the Kruskal–Wallis H test for comparisons among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced. (*$P < 0.05$, **$P < 0.01$, ChatGPT-4o mini vs. ChatGPT-4o and Gemini Advanced; ^$P < 0.05$, ^^$P < 0.01$, ChatGPT-4o vs. Gemini Advanced. The *P*-value is obtained using Dunn's test with Bonferroni correction)

significantly greater AS (4.09 ± 0.79) than ChatGPT-4o mini (2.34 ± 0.72) ($P < 0.05$). Overall, the AS of ChatGPT-4o was significantly greater than that of ChatGPT-4o mini and Gemini Advanced according to the ASs of all the FAQs related to PMOP (4.01 ± 0.81 vs. 3.21 ± 1.02, 3.42 ± 1.16). Table 4 shows the mean scores on the Likert scale for the ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced for the 2022 ACOG-PMOP Guideline guideline-related questions. Google Gemini's AS was significantly lower than that of ChatGPT-4o mini, ChatGPT-4o (2.56 ± 0.90 vs. 3.49 ± 0.98, 3.90 ± 0.79) ($P < 0.05$).

Tables 5 and 6 shows the chi-square test for the overall comparisons of the ratings for ChatGPT-4o, ChatGPT-4o mini, and Gemini Advanced across different

Table 2 Length and FRE of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced's responses to questions for 2022 ACOG-PMOP guideline

2022 ACOG-PMOP Guideline				
	ChatGPT-4o mini, ($\bar{x} \pm sd$)	ChatGPT-4o, ($\bar{x} \pm sd$)	Gemini Advanced, ($\bar{x} \pm sd$)	*P* value
Total characters	2244.58 ± 868.36	2201.75 ± 857.17	1889.33 ± 620.33	0.129
Total words	329.96 ± 128.78	316.04 ± 120.65$^{\wedge\wedge}$	221.67 ± 57.33**	0.001
Total syllables	642.46 ± 258.46	618.13 ± 237.32$^{\wedge\wedge}$	344.79 ± 91.34**	8.000E-06
Total sentences	24.88 ± 13.98	24.38 ± 11.15	19.71 ± 6.78	0.289
FRE	25.77 ± 13.33	27.00 ± 9.46$^{\wedge\wedge}$	63.40 ± 2.58**	4.987E-11

P < 0.05 indicates significance, *P* < 0.01 indicates high significance. The *P*-value is obtained using the Kruskal–Wallis H test for comparisons among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced. ($^{*}P < 0.05$, $^{**}P < 0.01$, ChatGPT-4o mini vs. ChatGPT-4o and Gemini Advanced; $^{\wedge}P < 0.05$, $^{\wedge\wedge}P < 0.01$, ChatGPT-4 vs. Gemini Advanced. ChatGPT-4o vs. Gemini Advanced. The *P*-value is obtained using Dunn's test with Bonferroni correction)

Table 3 Likert scale' AS of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced's responses to faqs about PMOP

Topic	ChatGPT-4o mini, ($\bar{x} \pm sd$)	ChatGPT-4o, ($\bar{x} \pm sd$)	Gemini Advanced ($\bar{x} \pm sd$)	*P* value
Clinical Manifestation	4.03 ± 0.91	3.78 ± 0.54	3.25 ± 1.08	0.195
Diagnosis	3.13 ± 0.52	4.19 ± 0.78$^{\wedge\wedge}$	2.53 ± 1.15	0.006
Pathogenesis	2.84 ± 0.86	4.28 ± 0.47**	3.40 ± 1.16	0.011
Treatment	3.41 ± 1.19	4.13 ± 0.70	4.19 ± 0.89	0.165
Prevention	3.53 ± 1.12	4.09 ± 0.99	3.06 ± 1.24	0.188
Risk Factor	2.34 ± 0.72	3.56 ± 1.19	4.09 ± 0.79*	0.013
All Questions	3.21 ± 1.02	4.01 ± 0.81$^{**\wedge}$	3.42 ± 1.16	0.001

P < 0.05 indicates significance, *P* < 0.01 indicates high significance. The *P*-value is obtained using the Kruskal–Wallis H test for comparisons among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced. ($^{*}P < 0.05$, $^{**}P < 0.01$, ChatGPT-4o mini vs. ChatGPT-4o and Gemini Advanced; $^{\wedge}P < 0.05$, $^{\wedge\wedge}P < 0.01$, ChatGPT-4 vs. Gemini Advanced. ChatGPT-4o vs. Gemini Advanced. The *P*-value is obtained using Dunn's test with Bonferroni correction)

topics. The *P*-values corrected using Bonferroni correction (**adjusted α** = $\frac{0.05}{3}$) are shown in Table 7. In terms of "Pathogenesis", ChatGPT-4o was significantly better than ChatGPT-4o mini ($P < 0.0167$). In terms of "Risk Factor", Gemini Advanced performed significantly better than ChatGPT-4o mini ($P < 0.0167$). Overall, ChatGPT-4o performed well in answering the FAQs about PMOP, significantly better than ChatGPT-4o mini and Gemini Advanced ($P < 0.0167$), with only two "poor" responses and 28 "excellent" ratings (Fig. 2a). As for the responses to questions from the 2022 ACOG-PMOP Guideline. ChatGPT-4o similarly outperformed Gemini Advanced ($P < 0.0167$). ChatGPT-4o had the highest percentage of "excellent" answers, at 62.5%. Google Gemini had the lowest percentage of "excellent" answers. Google Gemini had the lowest percentage of "excellent" answers, at 12.5% (Fig. 2b). The specific content of the AI-LLMs' answers to all the questions is shown in Supplementary Tables 3a-b.

Self-correcting capacity of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced

Table 8 shows the changes after self-correction of the ChatGPT-4o mini scale for questions with an AS ≤ 2. ChatGPT-4o mini had a mean AS (1.78 ± 0.15) for initial responses and 4.22 ± 0.26 for self-corrected responses, which was significantly greater, and the ratings increased significantly ($P < 0.05$). Table 9 shows the changes after self-correction for questions ChatGPT-4o with an AS ≤ 2. The mean AS after self-correction was significantly greater than the initial mean AS (4.08 ± 0.14 vs. 1.75 ± 0.25, $P < 0.05$). Table 10 shows the changes after Gemini Advanced self-correction for questions with an AS ≤ 2. The mean AS after Gemini Advanced self-correction was significantly greater than the initial mean AS (4.03 ± 0.61 vs. 1.75 ± 0.24), and the ratings were also significantly improved ($P < 0.05$). These findings suggest that ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced all have strong self-correcting abilities. Supplementary Tables 4a-c show the post-self-correction content of ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced for questions with AS ≤ 2, respectively. Specific parts of the initial responses that contained errors are highlighted in yellow. In addition, a professional orthopedic surgeon evaluated and prompted the parts of the initial content that were incorrect.

Table 4 Likert scale' AS of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced's responses to questions for 2022 ACOG-PMOP guideline

Topic	ChatGPT-4o mini	ChatGPT-4o	Gemini Advanced	*P* value
2022 ACOG-PMOP Guideline	3.49 ± 0.98	3.90 ± 0.79 $^{\wedge\wedge}$	2.56 ± 0.90**	4.30E-05

P < 0.05 indicates significance, *P* < 0.01 indicates high significance. The *P*-value is obtained using the Kruskal–Wallis H test for comparisons among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced. ($^{*}P < 0.05$, $^{**}P < 0.01$, ChatGPT-4o mini vs. ChatGPT-4o and Gemini Advanced; $^{\wedge}P < 0.05$, $^{\wedge\wedge}P < 0.01$, ChatGPT-4 vs. Gemini Advanced. ChatGPT-4o vs. Gemini Advanced. The *P*-value is obtained using Dunn's test with Bonferroni correction)

Table 5 The gradings of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced' responses to faqs about PMOP

Topic	Total, *n*	ChatGPT-4o mini, *n*(%)				ChatGPT-4o, *n*(%)				Gemini Advanced *n*(%)				*P* value
		Poor	Average	Good	Excellent	Poor	Average	Good	Excellent	Poor	Average	Good	Excellent	
Clinical Manifestation	8	0(0)	1(12.5)	2(25)	5(62.5)	0(5)	1(12.5)	5(62.5)	2(25)	1(12.5)	4(50)	0(0)	3(37.5)	0.058
Diagnosis	8	0(0)	5(62.5)	3(37.5)	0(0)	0(0)	1(12.5)	2(25)	5(62.5)	3(37.5)	3(37.5)	0(0)	2(25)	0.007
Pathogenesis	8	1(12.5)	5(62.5)	1(12.5)	1(12.5)	0(0)	0(0)	2(25)	6(75)	2(25)	1(12.5)	1(12.5)	4(50)	0.028
Treatment	8	1(12.5)	3(37.5)	0(0)	4(50)	0(0)	1(12.5)	1(12.5)	6(75)	1(12.5)	0(0)	0(0)	7(87.5)	0.263
Prevention	8	1(12.5)	2(25)	1(12.5)	4(50)	1(12.5)	0(0)	2(25)	5(62.5)	3(37.5)	1(12.5)	3(37.5)	1(12.5)	0.354
Risk Factor	8	4(50)	2(25)	2(25)	0(0)	1(12.5)	2(25)	1(12.5)	4(50)	0(0)	1(12.5)	1(12.5)	6(75)	0.019
All Questions	48	7(14.5)	18(37.5)	9(18.8)	14(29.2)	2(4.2)	5(10.4)	13(27.1)	28(58.3)	10(20.8)	10(20.8)	5(10.4)	23(48)	0.022

P < 0.05 indicates significance, *P* < 0.01 indicates high significance. The *P*-value is obtained using the chi-square test for comparisons among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced

Table 6 The gradings of ChatGPT-4o Mini, ChatGPT-4o, and gemini advanced' responses to questions for 2022 ACOG-PMOP guideline

Topic	Total, *n*	ChatGPT-4o mini, *n*(%)				ChatGPT-4o, *n*(%)				Gemini Advanced *n*(%)				*P* value
		Poor	Average	Good	Excellent	Poor	Average	Good	Excellent	Poor	Average	Good	Excellent	
Guideline	24	2(8.3)	9(37.5)	4(16.7)	9(37.5)	1 (4.2)	3(12.5)	5(20.8)	15(62.5)	8(33.3)	11(45.9)	2(8.3)	3(12.5)	0.001

P < 0.05 indicates significance, *P* < 0.01 indicates high significance. The *P*-value is obtained using the chi-square test for comparisons among ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced

Table 7 The chi-square test for pairwise comparisons between ratings of ChatGPT-4o, ChatGPT-4omini, and gemini advanced on different topics, with *P*-values before and after bonferroni correction (adjusted $\alpha = \frac{0.05}{3}$)

Topic		Original *P* value	Bonferroni correction (adjusted $\alpha = \frac{0.05}{3}$).	Significant
Clinical Manifestation	ChatGPT-4o mini vs. ChatGPT-4o	0.429	0.0167	No
	ChatGPT-4o mini vs. Gemini Advanced	0.184	0.0167	No
	ChatGPT-4o vs. Gemini Advanced	0.068	0.0167	No
Diagnosis	ChatGPT-4o mini vs. ChatGPT-4o	0.029	0.0167	No
	ChatGPT-4o mini vs. Gemini Advanced	0.032	0.0167	No
	ChatGPT-4o vs. Gemini Advanced	0.096	0.0167	No
Pathogenesis	ChatGPT-4o mini vs. ChatGPT-4o*	0.011	0.0167	Yes
	ChatGPT-4o mini vs. Gemini Advanced	0.180	0.0167	No
	ChatGPT-4o vs. Gemini Advanced	0.452	0.0167	No
Treatment	ChatGPT-4o mini vs. ChatGPT-4o	0.413	0.0167	No
	ChatGPT-4o mini vs. Gemini Advanced	0.200	0.0167	No
	ChatGPT-4o vs. Gemini Advanced	1.000	0.0167	No
Prevention	ChatGPT-4o mini vs. ChatGPT-4o	0.765	0.0167	No
	ChatGPT-4o mini vs. Gemini Advanced	0.347	0.0167	No
	ChatGPT-4o vs. Gemini Advanced	0.184	0.0167	No
Risk Factor	ChatGPT-4o mini vs. ChatGPT-4o	0.108	0.0167	No
	ChatGPT-4o mini vs. Gemini Advanced*	0.005	0.0167	Yes
	ChatGPT-4o vs. Gemini Advanced	0.452	0.0167	No
All Questions	ChatGPT-4o mini vs. ChatGPT-4o*	0.001	0.0167	Yes
	ChatGPT-4o mini vs. Gemini Advanced	0.108	0.0167	No
	ChatGPT-4o vs. Gemini Advanced*	0.012	0.0167	Yes
Guideline	ChatGPT-4o mini vs. ChatGPT-4o	0.166	0.0167	No
	ChatGPT-4o mini vs. Gemini Advanced	0.066	0.0167	No
	ChatGPT-4o vs. Gemini Advanced*	1.13E-4	0.0167	Yes

*$P < 0.0167$ indicates significance. The *P*-value is obtained using chi-square test with Bonferroni correction

Discussion

Compared with premenopausal women, perimenopausal women are at an earlier risk of developing osteoporosis than men are due to the rapid decline in estrogen levels and significantly accelerated bone loss, leading to an increased risk of fracture [20]. According to the European Vertebral Osteoporosis Study (EVOS), the prevalence of vertebral fractures in women aged 50–79 years is approximately 12.0%, and after 50 years of age, the prevalence of vertebral fracture increases with age [21]. The National Health and Nutrition Examination Survey (NHANES) suggested that in the United States, 6.2% of adults aged 65 years and over had osteoporosis at the lumbar spine or femur neck. The prevalence of osteoporosis at either skeletal site was higher among women (24.8%) than among men (5.6%) [22]. Based on Cummings SR et al., an epidemiological survey of postmenopausal osteoporosis in white women estimated that a 50-year-old woman has a 15–20% lifetime risk of hip fracture and a 50% risk of any osteoporotic fracture [23]. Hip fractures can result in poor quality of life, a dependent living situation, and an increased risk of death [24]. Postmenopausal osteoporosis seriously affects women's work and quality of life, and it has become a public health problem that urgently needs to be solved [25]. In recent years, postmenopausal osteoporosis treatment and prevention have been increasingly considered and valued by the medical field [26].

With the development of artificial intelligence, AI-LLMs have become more widely used in medical fields, such as radiology, medical care, and medical education [27–29]. According to a study by Yunus Balel et al., ChatGPT-4o is a valuable tool for suggesting topics for the evaluation of oral and maxillofacial surgery systems [30]. The ability of AI-LLMs (ChatGPT-4o and Claude 3-Opus) to process images can also help medical researchers quickly diagnose the benign and malignant nature of tumors, showing promise for future applications in medical imaging [31]. In a study by Enes Efe Is et al., the performance of ChatGPT-4o and Google Gemini in answering questions at the rheumatology board level was evaluated, and the results revealed that ChatGPT-4o answered the questions significantly more accurately than Google Gemini did; however, Google Gemini was more self-correcting than ChatGPT-4o was. This result suggests that AI-LLMs perform differently when faced with different questions and prompts [32]. However, no study has tested the performance of AI-LLM chatbots in answering questions related to postmenopausal osteoporosis.

When general FAQs about PMOP were answered, in "Diagnosis", Gemini Advanced had more concise answers

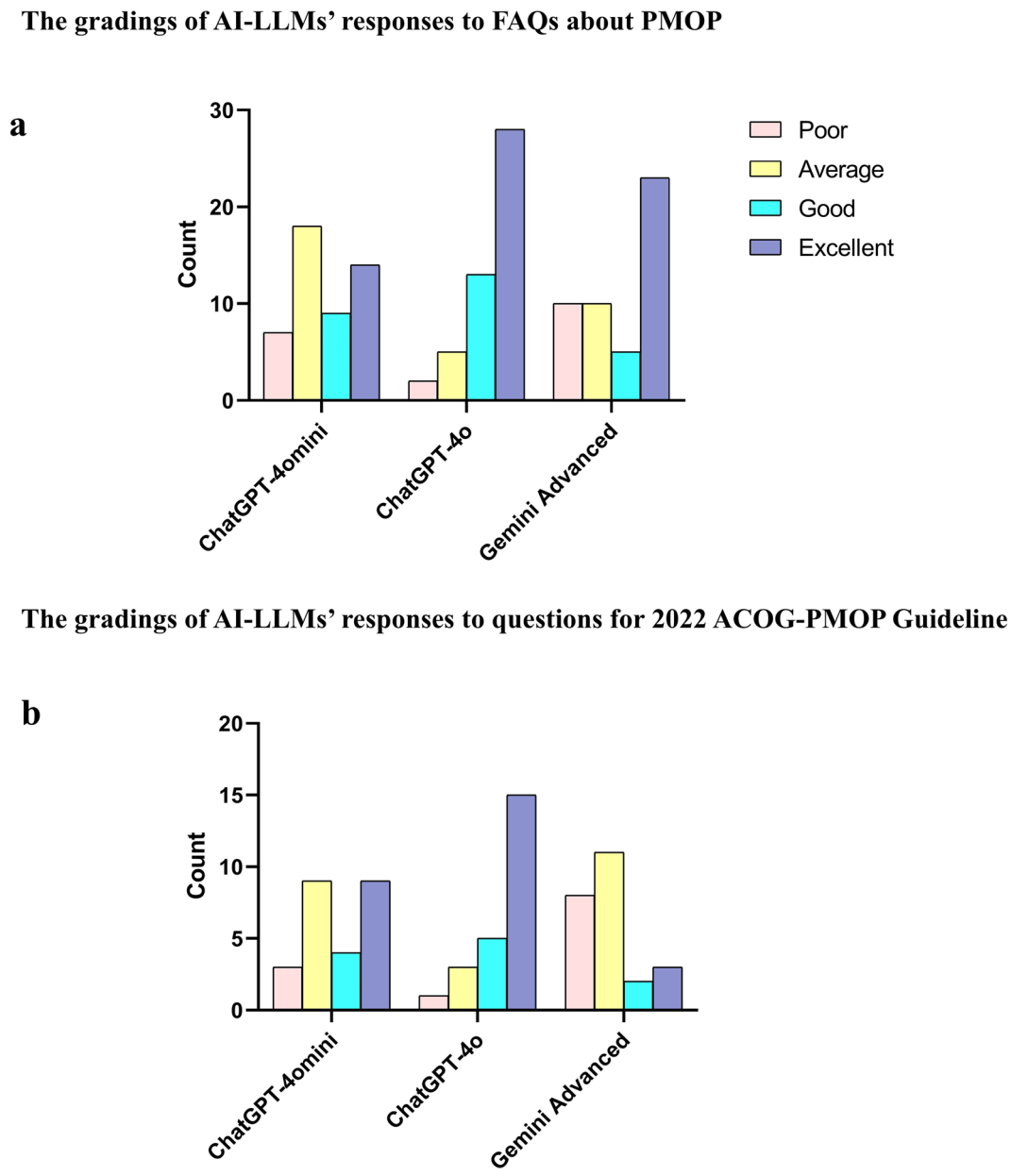

Fig. 2 **a** Gradings of ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced's responses to FAQs about PMOP; **b** gradings of ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced's responses to questions for 2022 ACOG-PMOP Guideline

than ChatGPT-4o mini, but ChatGPT-4o was significantly more readable than ChatGPT-4o mini. In terms of "Pathogenesis", ChatGPT-4o and Gemini Advanced had significantly fewer total syllables than ChatGPT-4o mini. In terms of "Risk Factor", Gemini Advanced has significantly better readability than ChatGPT-4o mini, and in answering general FAQs about PMOP, Gemini Advanced has significantly fewer total syllables and total sentences than ChatGPT-4o mini, and ChatGPT-4o's overall readability is better than that of ChatGPT-4o mini. In answering questions related to the 2022 ACOG-PMOP Guideline, Gemini Advanced is more concise and has better readability than ChatGPT-4o mini and ChatGPT-4o in terms of total words and total syllables. The above results suggest that, owing to the different algorithms and versions of AI-LLMs, they perform differently in information processing and Q&A, and Google Gemini and ChatGPT-4 may focus more on providing concise and direct answers to improve the efficiency of information delivery. In addition, ChatGPT-4o is more readable, which may be related to the fact that ChatGPT-4o is an optimized version of ChatGPT-4 that focuses on efficiency and performance, and may use a simpler sentence structure and common vocabulary, avoiding complex terminology and

Table 8 Self-correcting capacity of ChatGPT-4o mini

Topic	Question No.	TS		Accuracy ratings	
		Initial	Self-correction	Initial	Self-correction
Pathogenesis	6. Is the pathogenesis of postmenopausal osteoporosis reversible?	2	4.25	Poor	Excellent
Treatment	1. How is postmenopausal osteoporosis treated?	1.5	4.5	Poor	Excellent
Prevention	1.How often should perimenopausal women have a bone density test?	1.75	4.25	Poor	Excellent
Risk factors	1.What factors increase the risk of postmenopausal osteoporosis?	1.75	3.75	Poor	Good
	2.Whether a family history increases the risk of postmenopausal osteoporosis?	1.75	4.25	Poor	Excellent
	3. Does early menopause increase the risk of postmenopausal osteoporosis?	1.75	4.5	Poor	Excellent
	8. Can bone density tests predict the risk of fractures with postmenopausal osteoporosis?	1.75	4.5	Poor	Excellent
2022 ACOG-PMOP Guideline	4.What is the ACOG recommendation for the treatment of bisphosphonates in postmenopausal osteoporosis patients?	2	4	Poor	Good
	12.After assessing for remediable secondary causes, what criteria is recommended for postmenopausal osteoporosis in patients who meet the criteria?	1.75	4	Poor	Good
		1.78 ± 0.15	4.22 ± 0.26		
***P*value**		2.7384E-8		4.1E-5	

lengthy expressions to make the content more understandable. Notably, Gemini's Advanced responses are also very concise, with illustrations in some paragraphs to help readers understand the text more fully, but according to the reviewers, who rated Gemini Advanced's responses as "poor," Gemini Advanced's responses were too concise. Gemini Advanced's responses were too concise to provide clear answers to some questions, resulting in shorter answers (Supplementary Table 2, 3a-c). In contrast, ChatGPT-4o mini provided more comprehensive information, attempting to cover more context and detail to ensure that the user was fully understood, resulting in an increased number of characters and words, and conversely, a much reduced readability.

In terms of "Diagnosis", Gemini Advanced patients had significantly lower AS than the ChatGPT-4o ($P<0.05$). In terms of "Pathogenesis", ChatGPT-4o mini was also significantly less accurate than ChatGPT-4o. In terms of "Risk Factor", Gemini Advanced had a significantly greater AS than both ChatGPT-4o mini. For the FAQs concerning PMOP overall, ChatGPT-4o had a significantly higher AS than ChatGPT-4o mini and Gemini Advanced. In response to questions related to the 2022 ACOG-PMOP Guideline, Gemini Advanced had a significantly lower AS than ChatGPT-4o mini and ChatGPT-4o ($P<0$ 05) (Table 4). The difference in scores between ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced, may be due to a number of factors. ChatGPT-4o, introduced by OpenAI in May 2024, is an optimized version of ChatGPT-4 with more parameters and computational power, and is more focused on efficiency and performance. In addition, ChatGPT-4o mini is a smaller model of the most cost-effective ChatGPT-4o introduced by OpenAI in July 2024, supporting a wide range of tasks with its low cost and low latency, with limited exposure to a limited amount of data, especially in specialized domains, which tend to miss some of the most recent or more detailed medical information, however, it has academic benchmarks in textual intelligence and multimodal inference that both outperform GPT-3.5 Turbo and other smaller models. Gemini Advanced (Gemini 1.5 Pro) is a new generation of AI-LLMs released by Google in February 2024, with a context window of millions of tokens capable of comprehending long texts, audio, and videos; it specializes in logical reasoning and code generation, and accessing up-to-date web information. It is fundamentally different from ChatGPT-4omini and ChatGPT-4o (developed by OpenAI) in that it processes information and answers questions. However, in our study, we found that except for the "Risk Factor", where Gemini Advanced has a higher AS than ChatGPT-4o mini and ChatGPT-4o, the rest of the questions answered by Gemini Advanced are not very satisfactory, especially in regard to answering questions related to the 2022 ACOG-PMOP Guideline. For the guideline, it performed worse than ChatGPT-4o mini and ChatGPT-4o. This may be related to the fact that ChatGPT-4o may focus more on the knowledge understanding and reasoning ability of the medical domain in terms of model architecture and optimization strategy, for example, it optimizes specifically for medical terminology and logical relationships, and thus will be more accurate in dealing with related questions.

Table 9 Self-correcting capacity of ChatGPT-4o

Topic	Question No.	TS		Accuracy ratings	
		Initial	Self-correction	Initial	Self-correction
Prevention	1. How often should perimenopausal women have a bone density test?	1.75	4	Poor	Good
Risk factors	3. Does early menopause increase the risk of postmenopausal osteoporosis?	1.5	4	Poor	Good
2022 ACOG-PMOP Guideline	6.What are the treatment recommendations for postmenopausal osteoporosis patients with cardiovascular disease?	2	4.25	Poor	Excellent
		1.75 ± 0.25	4.08 ± 0.14		
***P*value**		0.001		0.100	

In our study, we compared the ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced self-correcting abilities for questions rated "poor". Our study revealed that the ChatGPT-4o mini had a total of nine responses rated "poor" across all the questions, with a mean AS of 1.78 ± 0.15 and a significantly greater mean AS of 4.22 ± 0.26 after correction ($P < 0.05$). The ChatGPT-4o had three questions with responses of "poor", with a mean AS of 1.75 ± 0.25 before correction and 4.08 ± 0.14 after correction, indicating significantly higher scores and levels ($P < 0.05$). Sixteen questions from Gemini Advanced received "poor" responses. The results revealed that Gemini Advanced scores and grades changed significantly before and after correction ($P < 0.05$). For responses rated as "poor," according to professional orthopedic surgeons, the "poor" responses were primarily due to a lack of specificity in detail, failure to follow guidelines, and an inability to professionally answer the questions posed. These findings suggest that these AI-LLMs failed to cover the latest medical advances and treatment options, dealt with somewhat complex and specialized medical issues, and had limited reasoning ability and insufficient depth. However, after correction, the scores and ratings of all the AI-LLMs increased significantly, thus suggesting that these AI-LLMs may try to make better use of their knowledge base to generate more accurate responses when asked a second time, by using more effective information retrieval strategies, and by "rethinking" their previous responses, ChatGPT ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced all possess strong self-correcting abilities.

Overall, this study evaluates the performance of three AI-LLMs (ChatGPT-4o mini, ChatGPT-4o, and Gemini Advanced) in helping to answer FAQs about PMOP and 2022 ACOG-PMOP Guideline. The results showed that the quality of ChatGPT-4o's responses was superior than that of other models overall, with a higher proportion of "excellent" ratings ($P < 0.05$). ChatGPT-4o's superior performance provides more solid evidence that more advanced models will provide more reliable and clinically relevant output. Our findings are similar to previous studies showing AI-LLMs' potentials in healthcare is consistent. For example, Wang et al. found that advanced AI-LLMs such as ChatGPT-4, showed greater consistency and accuracy in answering specialized medical questions about osteoarthritis [16]. Similarly, in the study of Zhi Wei Lim et al., ChatGPT-4o has a higher potential in providing accurate and comprehensive answers to myopia-related queries [33]. In our study, ChatGPT-4o demonstrated a greater ability to generate evidence-based treatment recommendations, which is critical to support clinical decision-making in PMOP's care, and these findings have important implications for AI-LLMs into clinical practice. For example, ChatGPT-4o can serve as a valuable tool for clinicians seeking quick access to evidence-based guidelines or patient education materials. However, caution is needed. Clinicians should eliminate uncertainties in diagnosis and treatment by repeating questions or cross-referencing to verify that the AI -LLMs' results are available with credible resources, especially in complex cases.

Strengths and limitations

Although we collected questions about PMOP from multiple sources and had a small number of questions, the coverage of the questions was perhaps incomplete and this study is based on evaluation of simulated Q&A

Table 10 Self-correcting capacity of gemini advanced

Topic	Question No.	TS		Accuracy ratings	
		Initial	Self-correction	Initial	Self-correction
Clinical Manifestation	8.What is the difference between postmenopausal osteoporosis and other types of osteoporosis	2	4.25	Poor	Excellent
Diagnosis	5.How to distinguish the symptoms of postmenopausal osteoporosis from arthritis?	1.5	4.5	Poor	Excellent
	6.What are the diagnostic criteria for fractures due to postmenopausal osteoporosis?	1.5	4.75	Poor	Excellent
	7.What is the difference between the diagnosis of postmenopausal osteoporosis and other types of osteoporosis?	1.5	4	Poor	Good
Pathogenesis	5.Why does bone loss intensify after menopause?	1.75	4.25	Poor	Excellent
Treatment	3.What is the preferred medication for treating postmenopausal osteoporosis?	2	2.75	Poor	Average
Prevention	6.How to prevent postmenopausal osteoporosis through diet?	2	4.25	Poor	Excellent
	7.How to prevent fractures caused by postmenopausal osteoporosis?	1.75	3.74	Poor	Good
2022 ACOG-PMOP Guideline	1.What assessments should be done before drug treatment for patients with postmenopausal osteoporosis?	1.75	4.25	Poor	Excellent
	4.What is the ACOG recommendation for the treatment of bisphosphonates in postmenopausal osteoporosis patients?	1.75	4.25	Poor	Excellent
	6.What are the treatment recommendations for postmenopausal osteoporosis patients with cardiovascular disease	1.5	4.5	Poor	Excellent
	7.How often does the ACOG recommend dual energy X-ray absorptiometry in postmenopausal osteoporosis patients during drug therapy?	2	4	Poor	Good
	11.What are the recommended osteoporosis risk assessment tools for breast cancer patients prior to starting aromatase inhibitors and chemotherapy?	2	4.5	Poor	Excellent
	12.After assessing for remediable secondary causes, what criteria is recommended for postmenopausal osteoporosis in patients who meet the criteria?	1.5	4	Poor	Good
	14.What are the bisphosphonates approved in the FDA for the treatment of postmenopausal osteoporosis and what are their indications?	1.75	4	Poor	Good
	15.What are the adverse effects of bisphosphate in the treatment of postmenopausal osteoporosis? What are the indications and methods of discontinuation?	1.75	2.5	Poor	Average
		1.75 ± 0.24	4.03 ± 0.61		
***P*value**		1.1085E-9		3.3273E-9	

scenarios, rather than real clinical data. In the future, comparisons should be made to distinguish them from actual patient questions and answers in order to enhance the generalizability of our results. The scoring system used in this study was a Likert scale. Even among experienced orthopedic specialists, there may be differences in their understanding of, and emphasis on, criteria such as accuracy, completeness, and clarity. For example, some experts may be more concerned with the accuracy of the answer, whereas others may be more concerned with whether the answer is easy for the patient to understand. In this regard, we conducted a Fleiss's kappa coefficient test on the ratings given by the four orthopedic specialists, and the Fleiss's kappa values were 0.238, 0.105, and 0.290, indicating that the consistency of the ratings given by the four orthopedic specialists was relatively low, which may be related to the individual specialists. This may be related to the different clinical experiences and research directions of individual experts. Moreover, owing to the time factor limitations of our study, the training data and algorithms of the AI model were constantly updated, so the test results only reflected the performance of the model at a specific point in time. The training data of each AI-LLM may introduce regional, temporal, or demographic biases, which may affect the performance and applicability of the model in different clinical environments. And our study focused on limited AI-LLMS, not including other large AI-LLMs (e.g., DEEPseek, etc.). The performance of the model may change over time. Additionally, due to model updates, different results may be obtained even when testing with the same problem, which can affect the reproducibility of the study. In conclusion, as technology continues to evolve, large-scale language models will play an increasingly

important role in the treatment of postmenopausal osteoporosis. They can assist physicians in diagnosis and treatment, provide patient education and support, facilitate medical research, and promote telemedicine. To better realize these perspectives, there is a need to further improve the accuracy, reliability, and safety of the models, as well as to enhance their integration with the health care system. Moreover, there is a need to focus on ethical and social implications to ensure that AI technologies are applied in accordance with human values and interests.

Conclusion

Our study revealed that ChatGPT-4o and Gemini Advanced's answers to PMOP-related questions were more concise, clearer, and understandable, and Gemini Advanced's answers featured illustrations that helped patients fully comprehend the text, however Gemini Advanced's shortcomings were that the answers were too concise and not tailored to the relevant questions, which needs to be improved. ChatGPT-4o significantly outperformed ChatGPT4o mini and Gemini Advanced in answering PMOP-related FAQs. ChatGPT4o mini and ChatGPT-4o were significantly better at answering 2022 ACOG-PMOP Guideline-related questions. ChatGPT4o mini and ChatGPT-4o significantly outperformed Gemini Advanced, and our results also suggest that ChatGPT4o mini, ChatGPT-4o, and Gemini Advanced have stronger self-corrective abilities, a finding that may be related to the fact that the current AI-LLMs have stronger feedback mechanisms and dynamic adaptabilities.

Abbreviations

PMOP	Postmenopausal osteoporosis
AI-LLMs	Artificial intelligence large-scale language models
FAQs	Frequently asked questions
2022 ACOG-PMOP Guideline	Management of postmenopausal osteoporosis:2022 ACOG clinical practice guideline No. 2
FRE	Flesch reading ease
NHANES	National health and nutrition examination survey

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12891-025-08601-3.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

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

Funding

This work was supported by the Key Program of the Natural Science Foundation of Tianjin (award number S24YBL069).

Data availability

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Declarations

Ethics approval

Not applicable.

Consent for publication

Not applicable.

Human ethics and consent to participate declarations

Not applicable.

The name of the approval committee or the internal review board (IRB)

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Clinical College of Neurology, Neurosurgery and Neurorehabilitation, Tianjin Medical University, Tianjin, China
[2]College of Computer Science, Nankai University, Tianjin 300350, China

Received: 23 January 2025 / Accepted: 31 March 2025
Published online: 16 April 2025

References

1. Walker MD, Shane E. Postmenopausal osteoporosis. N Engl J Med. 2023;389(21):1979–91. https://doi.org/10.1056/NEJMcp2307353.
2. Porter JL, Varacallo M. Osteoporosis. In: *StatPearls*. Treasure Island (FL) ineligible companies. Disclosure: Matthew Varacallo declares no relevant financial relationships with ineligible companies.: StatPearls Publishing. Copyright. © 2024, StatPearls Publishing LLC.; 2024.
3. Ramchand SK, Leder BZ. Sequential therapy for the Long-Term treatment of postmenopausal osteoporosis. J Clin Endocrinol Metab. 2024;109(2):303–11. https://doi.org/10.1210/clinem/dgad496.
4. Reid IR. A broader strategy for osteoporosis interventions. Nat Rev Endocrinol. 2020;16(6):333–9. https://doi.org/10.1038/s41574-020-0339-7.
5. Zhang X, Wang Z, Zhang D, Ye D, Zhou Y, Qin J, Zhang Y. The prevalence and treatment rate trends of osteoporosis in postmenopausal women. PLoS ONE. 2023;18(9):e0290289. https://doi.org/10.1371/journal.pone.0290289.
6. Management of Postmenopausal Osteoporosis. ACOG clinical practice guideline 2. Obstet Gynecol. 2022;139(4):698–717. https://doi.org/10.1097/aog.0000000000004730.
7. Eastell R, Rosen CJ, Black DM, Cheung AM, Murad MH, Shoback D. Pharmacological management of osteoporosis in postmenopausal women: an endocrine society** clinical practice guideline. J Clin Endocrinol Metab. 2019;104(5):1595–622. https://doi.org/10.1210/jc.2019-00221.
8. LeBoff MS, Greenspan SL, Insogna KL, Lewiecki EM, Saag KG, Singer AJ, Siris ES. The clinician's guide to prevention and treatment of osteoporosis. Osteoporos Int. 2022;33(10):2049–102. https://doi.org/10.1007/s00198-021-05900-y.
9. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language models in medicine: the potentials and pitfalls: A narrative review. Ann Intern Med. 2024;177(2):210–20. https://doi.org/10.7326/m23-2772.
10. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, Löffler CML, Schwarzkopf SC, Unger M, Veldhuizen GP, et al. The future landscape of large Language models in medicine. Commun Med (Lond). 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1.
11. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large Language models in medicine. JAMA. 2023;330(9):866–9. https://doi.org/10.1001/jama.2023.14217.

12. Ellaway RH, Tolsgaard M. Artificial scholarship: LLMs in health professions education research. Adv Health Sci Educ Theory Pract. 2023;28(3):659–64. https://doi.org/10.1007/s10459-023-10257-4.
13. 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
14. Potapenko I, Malmqvist L, Subhi Y, Hamann S. Artificial Intelligence-Based ChatGPT responses for patient questions on optic disc Drusen. Ophthalmol Ther. 2023;12(6):3109–19. https://doi.org/10.1007/s40123-023-00800-2.
15. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. 2023;228(6):696–705. https://doi.org/10.1016/j.ajog.2023.03.009.
16. Wang L, Chen X, Deng X, Wen H, You M, Liu W, Li Q, Li J. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digit Med. 2024;7(1):41. https://doi.org/10.1038/s41746-024-01029-4.
17. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV information on Google versus AI (ChatGPT). Otolaryngol Head Neck Surg. 2024;170(6):1504–11. https://doi.org/10.1002/ohn.506.
18. 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;262(9):2945–59. https://doi.org/10.1007/s00417-024-06470-5.
19. Lee Y, Shin T, Tessier L, Javidan A, Jung J, Hong D, Strong AT, McKechnie T, Malone S, Jin D, et al. Harnessing artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and bard in generating clinician-level bariatric surgery recommendations. Surg Obes Relat Dis. 2024;20(7):603–8. https://doi.org/10.1016/j.soard.2024.03.011.
20. Clynes MA, Harvey NC, Curtis EM, Fuggle NR, Dennison EM, Cooper C. The epidemiology of osteoporosis. Br Med Bull. 2020;133(1):105–17. https://doi.org/10.1093/bmb/ldaa005.
21. O'Neill TW, Felsenberg D, Varlow J, Cooper C, Kanis JA, Silman AJ. The prevalence of vertebral deformity in European men and women: the European vertebral osteoporosis study. J Bone Min Res. 1996;11(7):1010–8. https://doi.org/10.1002/jbmr.5650110719.
22. Johnston CB, Dagar M. Osteoporosis in older adults. Med Clin North Am. 2020;104(5):873–84. https://doi.org/10.1016/j.mcna.2020.06.004.
23. Cummings SR, Black DM, Rubin SM. Lifetime risks of hip, Colles', or vertebral fracture and coronary heart disease among white postmenopausal women. Arch Intern Med. 1989;149(11):2445–8.
24. Cummings SR, Melton LJ. Epidemiology and outcomes of osteoporotic fractures. Lancet. 2002;359(9319):1761–7. https://doi.org/10.1016/s0140-6736(02)08657-9.
25. Bhatnagar A, Kekatpure AL. Postmenopausal osteoporosis: A literature review. Cureus. 2022;14(9):e29367. https://doi.org/10.7759/cureus.29367.
26. Jeong HG, Kim MK, Lim HJ, Kim SK. Up-to-Date knowledge on osteoporosis treatment selection in postmenopausal women. J Menopausal Med. 2022;28(3):85–91. https://doi.org/10.6118/jmm.22007.
27. Daungsupawong H, Wiwanitkit V. LLMs in radiology through prompt engineering: Comment. *Rofo* 2024. https://doi.org/10.1055/a-2295-3839
28. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on large Language models (LLMs). NPJ Digit Med. 2024;7(1):183. https://doi.org/10.1038/s41746-024-01157-x.
29. Benítez TM, Xu Y, Boudreau JD, Kow AWC, Bello F, Van Phuoc L, Wang X, Sun X, Leung GK, Lan Y, et al. Harnessing the potential of large Language models in medical education: promise and pitfalls. J Am Med Inf Assoc. 2024;31(3):776–83. https://doi.org/10.1093/jamia/ocad252.
30. Balel Y, Zogo A, Yıldız S, Tanyeri H. Can ChatGPT-4o provide new systematic review ideas to oral and maxillofacial surgeons? J Stomatol Oral Maxillofac Surg. 2024;125(5s2):101979. https://doi.org/10.1016/j.jormas.2024.101979.
31. Chen Z, Chambara N, Wu C, Lo X, Liu SYW, Gunda ST, Han X, Qu J, Chen F, Ying MTC. Assessing the feasibility of ChatGPT-4o and Claude 3-Opus in thyroid nodule classification based on ultrasound images. Endocrine. 2024. https://doi.org/10.1007/s12020-024-04066-x.
32. Is EE, Menekseoglu AK. Comparative performance of artificial intelligence models in rheumatology board-level questions: evaluating Google gemini and ChatGPT-4o. Clin Rheumatol. 2024;43(11):3507–13. https://doi.org/10.1007/s10067-024-07154-5.
33. 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.

Publisher's note