Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Reliability and Readability of AI Chatbot Responses on Venous Thromboembolism After Fracture Surgery: A Comparative Study

Mengyue Xie
Peking University Shenzhen Hospital
Li Yi
Peking University Shenzhen Hospital
Chuanliang Zhang
Peking University Shenzhen Hospital
Ming Xiao
Peking University Shenzhen Hospital
Xiaohua Wu
Peking University Shenzhen Hospital
Yingsi Zhang
852068629@qq.com

Peking University Shenzhen Hospital

Research Article

Keywords: Venous thromboembolism, Fracture surgery, Artificial intelligence, Chatbots, Patient education

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11424655/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Venous thromboembolism (VTE) is an important complication after fracture surgery, and patients increasingly seek health information through internet based resources and artificial intelligence (AI) chatbots. However, the reliability and readability of AI generated information about postoperative VTE remain unclear. This study evaluated public information needs regarding VTE and assessed the reliability and readability of information provided by four widely used AI chatbots.

Methods

Google Trends was used to characterize worldwide public search interest in VTE from August 2021 to August 2026. Twenty patient oriented questions were developed based on popular searches and clinical experience and submitted verbatim to ChatGPT-5.6 Sol, Claude Sonnet 5, DeepSeek-V4-Pro, and Gemini 3.6 Flash. Reliability was assessed using DISCERN, EQIP, GQS, and JAMA criteria, while readability was evaluated using six established indices.

Results

Global search interest in VTE increased markedly from mid 2025 and peaked in July 2026. Reliability differed significantly among the four chatbots across all four assessment tools, with ChatGPT showing the highest median DISCERN, EQIP, and JAMA scores; all omnibus comparisons remained significant after Holm–Šídák correction (adjusted $P < 0.0001$). Readability remained poor across all models, with grade level indices well above the recommended sixth grade benchmark and FRES values ranging from 28.00 to 36.35, far below the recommended threshold of 80.

Conclusion

Public interest in VTE has increased in recent years, and contemporary AI chatbots differ in the reliability of information provided about VTE after fracture surgery. ChatGPT showed higher reliability scores across several domains, whereas source transparency remained limited across models. Responses from all four chatbots were substantially more complex than the recommended reading level for patient education. Improvements in reliability, source transparency, and readability are needed to enhance the usefulness of AI generated information for patient education.

1 Introduction

Venous thromboembolism (VTE), which includes deep vein thrombosis (DVT) and pulmonary embolism (PE), is an important complication after fracture surgery and can contribute substantially to postoperative morbidity and mortality. Patients undergoing surgical treatment for fractures may be particularly vulnerable because trauma, tissue injury, reduced mobility, and patient related factors can act together to increase thrombotic risk. A large systematic review and meta analysis of more than three million patients undergoing fracture surgery identified

several factors associated with postoperative VTE, including advanced age, immobility, obesity, previous thromboembolic events, cardiovascular disease, and other comorbidities[1]. Clinical guidelines therefore emphasize appropriate assessment and prevention of perioperative VTE in surgical and orthopedic populations[2]. Importantly, thromboembolic risk does not necessarily end at hospital discharge. Studies in trauma and operatively treated fracture populations have shown that clinically relevant VTE events may continue to occur during the post discharge period[3, 4]. These features make clear patient education regarding VTE prevention, symptom recognition, medication use, and timely medical evaluation an important component of postoperative care.

Patients increasingly obtain health information outside traditional clinical encounters, particularly through internet based resources. Search engine data can provide insight into topics that attract public attention and the types of health information that people actively seek. Google Trends has consequently been widely used in health research to characterize temporal and geographic patterns of online information seeking[5]. However, access to information alone does not ensure that it is useful to patients. Health information must also be understandable and presented at an appropriate reading level. This is particularly relevant in orthopedics, where previous studies have consistently found that many online patient education materials exceed recommended readability levels. For example, most patient education materials provided by the American Academy of Orthopaedic Surgeons were written above the sixth grade reading level commonly recommended for patient materials[6]. Similar readability concerns have been reported for trauma related orthopedic information[7].

The rapid development of large language models has introduced AI chatbots as another readily accessible source of medical and health information. Unlike conventional web searches, these systems can generate direct, conversational responses to individual patient questions. Previous studies have demonstrated the potential of AI chatbots to provide useful responses to public health and medical questions[8]. Nevertheless, the clinical quality of generated information remains variable. Chatbot responses may omit important details, provide insufficient source attribution, or present information with a level of complexity that is difficult for the general public to understand[9, 10]. A recent systematic review and meta analysis found that AI chatbot responses used for patient education were generally appropriate and accurate, but their average readability remained near the college level, limiting their accessibility to many patients[11]. Thus, reliability and readability represent separate but complementary considerations when evaluating AI generated patient information.

Interest in the use of AI for orthopedic patient education has also increased. Recent studies have evaluated chatbot responses to questions about individual fracture types and postoperative management. For example, AI chatbots have been assessed for patient questions following distal radius fracture surgery, with generally acceptable information quality but reading levels substantially above recommended patient education standards[12]. Similar findings have been reported for fracture related questions in other orthopedic settings, suggesting that clinically useful content does not necessarily translate into easily understandable information[13]. However, evidence regarding AI generated information specifically addressing VTE after fracture surgery remains limited, and comparative evaluations across several contemporary AI chatbots are lacking. Therefore, this study aimed to evaluate the reliability and readability of information about VTE after fracture surgery generated by four widely used AI chatbots. Public information needs were characterized using Google Trends and clinical expertise to develop 20 patient oriented core questions. The responses were subsequently assessed using multiple established measures of information quality and readability, with the aim of determining their suitability as sources of patient health information.

2 Methods

Study design

This study evaluated the reliability and readability of health information about venous thromboembolism (VTE) after fracture surgery provided by four artificial intelligence chatbots. Twenty core questions were developed based on Google Trends data and physicians' clinical experience, and each question was submitted to the four chatbots for response generation. The generated responses were subsequently evaluated using established reliability and readability assessment tools. The overall study workflow is shown in Fig. 1.

Development of Core Public Education Questions and Response Collection

Google Trends was used to identify common public information needs related to venous thrombosis. We searched for the topic "venous thrombosis", for which Google Trends incorporates conceptually equivalent search terms across languages to reflect global search interest. We extracted the 10 most popular related queries worldwide during the period from August 2021 to August 2026. The Google Trends data were accessed in August 2026, and the exported list of related queries is provided in Supplementary Table S1.

The identified queries were reviewed by a multidisciplinary clinical panel consisting of senior orthopedic surgeons and nurses. Drawing on their clinical experience, the panel refined and expanded these commonly searched topics into 20 patient oriented core questions. The questions were organized into four domains: Basic Understanding and Risk, Symptoms and Seeking Medical Care, Prevention, and Medication Safety and Daily Life Management. Together, these domains were intended to reflect both public information needs and clinically important safety concerns related to venous thromboembolism after fracture surgery. The final 20 questions are presented in Table 1.

Table 1
Twenty Core Public Education Questions.

Number	Question
1	What is venous thromboembolism (VTE), and how is it related to deep vein thrombosis (DVT) and pulmonary embolism (PE)?
2	Why are patients at increased risk of developing VTE after fracture surgery?
3	Which types of fracture surgery are associated with a higher risk of VTE?
4	Which patients are more likely to develop VTE after fracture surgery?
5	When is the risk of VTE highest after fracture surgery, and does the risk continue after hospital discharge?
6	What symptoms usually occur when a patient develops DVT after fracture surgery?
7	Leg pain and swelling are common after fracture surgery. How can a patient distinguish normal postoperative symptoms from possible DVT?
8	What symptoms can occur when a patient develops PE after fracture surgery?
9	Which symptoms after fracture surgery require an immediate visit to the emergency department or a call to emergency medical services?
10	If DVT or PE is suspected after fracture surgery, what tests are usually performed at the hospital?
11	What measures can patients take to prevent VTE after fracture surgery?
12	When should patients begin walking after fracture surgery, and how much activity is helpful for preventing VTE?
13	Can ankle pump exercises and other leg exercises help prevent VTE after fracture surgery, and how should patients perform them?
14	Can compression stockings and intermittent pneumatic compression devices prevent VTE after fracture surgery, and how long should patients use them?
15	Does every patient need aspirin or an anticoagulant to prevent VTE after fracture surgery?
16	What are the differences among aspirin, low molecular weight heparin, apixaban, and rivaroxaban for preventing VTE after fracture surgery?
17	How long should patients usually take medication to prevent VTE after fracture surgery?
18	Which signs of bleeding require patients taking anticoagulants after fracture surgery to contact a doctor or seek immediate medical attention?
19	What should patients do if they miss a dose of an anticoagulant after fracture surgery? Should they double the next dose?
20	If a patient develops DVT or PE after fracture surgery, how is it usually treated, and how might it affect rehabilitation and future physical activity?

Each question was then submitted verbatim to four AI chatbots: ChatGPT-5.6 Sol, Claude Sonnet 5, DeepSeek-V4-Pro, and Gemini 3.6 Flash. No additional instructions, role prompts, or contextual information were provided. The full response generated by each chatbot was collected for subsequent reliability and readability analyses. This minimal prompting approach was used to approximate how patients may typically seek health information

from AI chatbots in real world settings. For the public education analysis, each question was submitted once to each model to represent a typical single user interaction. Therefore, the consistency of responses across repeated trials was not assessed in this study.

Before each query session, browser data, including cookies, cache, and browsing history, were cleared to reduce potential effects from personalization or previous interactions. To improve methodological transparency and reproducibility, we also documented the access route, platform or provider, access date, language, account or subscription status, generation settings, and management of conversation history for each AI model. These details are provided in Supplementary Table S2.

Reliability Assessment

The reliability and quality of the information generated by the four AI chatbots were evaluated using four established assessment tools: DISCERN, the Ensuring Quality Information for Patients (EQIP) instrument, the Global Quality Scale (GQS), and the Journal of the American Medical Association (JAMA) benchmark criteria.

DISCERN was used to assess the quality of health information, with particular emphasis on information related to treatment choices[14]. The DISCERN handbook does not provide a formal classification system for interpreting the total score. Therefore, in accordance with previous studies, overall DISCERN scores were categorized as follows: 63 to 75 points, excellent quality; 51 to 62 points, good quality; 39 to 50 points, fair quality; 27 to 38 points, poor quality; and 16 to 26 points, very poor quality[15].

EQIP was used to evaluate the quality of patient information, including the clarity, content, and presentation of the information provided[16]. The instrument consists of 20 items, with each item rated as “yes,” “partly,” “no,” or “does not apply.” For scoring, “yes” was assigned 1 point, “partly” 0.5 points, and “no” 0 points. Items rated as “does not apply” were excluded from the denominator. The resulting score was converted to a percentage. According to the categories proposed in the original EQIP development literature, scores of 76% to 100% were classified as well written and of excellent quality, 51% to 75% as good quality with minor issues, 26% to 50% as having serious quality issues, and 0% to 25% as having severe quality problems[15, 16].

GQS was used to provide an overall assessment of the quality, information flow, and usefulness of the health-related content. It uses a five-point Likert scale, with scores of 1, 2, 3, 4, and 5 corresponding to very poor, poor, fair, good, and excellent quality, respectively.

JAMA benchmark criteria were used to assess the transparency and credibility of the information according to four domains: authorship, attribution of references and sources, disclosure of date, and ownership[17]. Each domain was assigned either 0 or 1 point, resulting in a total score ranging from 0 to 4.

The DISCERN, EQIP, GQS, and JAMA assessments were performed independently by two infectious disease clinicians. Any discrepancies between the two evaluators were resolved through discussion and consensus. If consensus could not be reached, a senior clinician with extensive clinical experience independently reviewed the response, and the score assigned by the senior clinician was used as the final score.

For the reliability analyses, statistical significance was defined as a two-sided P value < 0.05. Significance levels in the figures were indicated as *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

Readability Assessment

The readability of the chatbot responses was evaluated using six commonly used readability indices: the Automated Readability Index (ARI), Coleman Liau Index (CL), Flesch Kincaid Grade Level (FKGL), Flesch Reading Ease Score (FRES), Gunning Fog Index (GFI), and Simple Measure of Gobbledygook (SMOG). All readability scores were calculated using the online readability calculator available at ReadabilityFormulas.com.

The FRES ranges from 0 to 100, with higher scores indicating greater ease of reading. In contrast, ARI, CL, FKGL, GFI, and SMOG estimate the level of education required to understand a text and are generally expressed in terms of United States school grade level. These indices were used to characterize the linguistic complexity of the responses generated by the four AI chatbots and to assess their suitability for patient facing health communication.

Readability scores are presented as the mean ± standard deviation (SD) and were descriptively compared with the sixth grade reading level recommended by the American Medical Association and the National Institutes of Health. Responses were considered to meet the predefined readability benchmark when the FRES was ≥ 80 and the ARI, CL, FKGL, GFI, and SMOG scores were ≤ 6.

Statistical analysis

All statistical analyses were performed using GraphPad Prism version 10 (GraphPad Software, Boston, MA, USA). Reliability scores, including DISCERN, EQIP, GQS, and JAMA scores, are presented as the median and interquartile range (IQR; 25th to 75th percentile). Readability scores, including ARI, CL, FKGL, FRES, GFI, and SMOG, are presented as the mean ± standard deviation (SD).

For the reliability analysis, DISCERN, EQIP, GQS, and JAMA scores were treated as four separate outcomes because the instruments use different scoring systems and scales. As the same 20 questions were assessed across all four AI chatbots, the observations were matched by question. Differences among ChatGPT, Claude Sonnet, DeepSeek, and Gemini were therefore evaluated separately for each reliability outcome using the Friedman test. When a significant overall difference was detected, post hoc pairwise comparisons were performed using Dunn's multiple comparisons test, and multiplicity adjusted P values were reported for the six pairwise comparisons within each outcome. Kendall's coefficient of concordance (W) was calculated as an effect size for each Friedman test using $W = \chi F2/[N(k - 1)]$ $W=\chi^2_F/[N(k-1)]$, where $\chi F2\chi^2_F$ represents the Friedman test statistic, NN is the number of matched questions, and kk is the number of chatbots. In this study, $N=20N = 20$ and $k = 4k = 4$. To account for the four reliability outcomes, the four omnibus P values obtained from the Friedman tests were further adjusted using the Holm–Šídák procedure to control the familywise type I error rate.

For the readability analysis, the six readability indices were descriptively compared with predefined reference values corresponding to a sixth grade reading level. The reference threshold was ≥ 80 for the Flesch Reading Ease Score (FRES) and ≤ 6 for the Automated Readability Index (ARI), Coleman Liau Index (CL), Flesch Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Simple Measure of Gobbledygook (SMOG). No inferential statistical tests were performed for the readability outcomes.

3 Results

Public Search Interest in Venous Thromboembolism

Google Trends analysis showed that worldwide searches related to venous thromboembolism (VTE) focused mainly on prevention, clinical guidance, treatment, symptoms, and disease definition. The most frequently related query was “venous thromboembolism prophylaxis,” with a normalized relative search interest score of 100, followed by “venous thromboembolism guidelines” (70), “venous thromboembolism treatment” (58), “what is venous thromboembolism” (54), and “treatment of venous thromboembolism” (38). Other common queries included “venous thromboembolism symptoms” (30), “venous thromboembolism definition” (27), “pulmonary embolism” (24), and “prevention of venous thromboembolism” (23). Searches also included specific concerns regarding anticoagulant safety, such as bleeding risk associated with apixaban and rivaroxaban. Collectively, these queries reflected public information needs related to the understanding, prevention, recognition, and management of VTE.

Worldwide search interest in VTE remained relatively low and stable from August 2021 through the first half of 2025 (Fig. 2A). A marked increase became apparent from July 2025, followed by a sustained rise and greater week to week variation during late 2025 and 2026. Search interest increased further during the first half of 2026, with several pronounced peaks. The highest normalized search interest score of 100 was recorded in July 2026, indicating that worldwide search interest in VTE reached its maximum near the end of the five-year observation period.

Geographic analysis demonstrated substantial variation in relative search interest across countries and regions (Fig. 2B). Singapore had the highest normalized relative search interest score (100), followed by St. Helena (47), Ethiopia (41), South Korea (29), and the United States (20). VTE related searches were also observed across multiple regions worldwide, including Europe, North America, Asia, Africa, and Oceania, although the level of relative search interest varied considerably by location.

Reliability Analysis

Significant differences in reliability scores were observed among the four AI chatbots across all four assessment instruments (Table 2 and Fig. 3). ChatGPT had the highest median scores for DISCERN, EQIP, and JAMA, while the median GQS score was 4 across all four chatbots. For DISCERN, the median scores were 51.5 (IQR, 35.0–57.75) for ChatGPT, 40.0 (29.0–50.75) for Claude Sonnet, 38.5 (27.75–47.5) for DeepSeek, and 32.0 (25.0–39.5) for Gemini. The overall difference was significant (Friedman statistic = 41.94, Kendall’s W = 0.699, raw $P < 0.0001$). For EQIP, the corresponding median scores were 80.0 (75.0–80.0), 67.5 (55.0–70.0), 65.0 (56.25–65.0), and 60.0 (55.0–60.0), respectively, with a significant overall difference (Friedman statistic = 35.56, W = 0.593, raw $P < 0.0001$).

Table 2
Reliability Scores Across AI Chatbots.

Program	ChatGPT Median (IQR)	Claude Median (IQR)	DeepSeek Median (IQR)	Gemini Median (IQR)	Friedman statistic	Kendall's W	Raw P	Holm–Šídák adjusted P
DISCERN	51.5(35-57.75)	40(29-50.75)	38.5(27.75–47.5)	32(25-39.5)	41.94	0.699	< 0.0001	< 0.0001
EQIP	80(75–80)	67.5(55–70)	65(56.25-65)	60(55–60)	35.56	0.593	< 0.0001	< 0.0001
GQS	4(4–5)	4(4–4)	4(3–4)	4(3–4)	25.8	0.430	< 0.0001	< 0.0001
JAMA	1(1–1)	0(0–0)	0(0–0)	0(0–0)	57	0.950	< 0.0001	< 0.0001

For GQS, all four chatbots had a median score of 4, although the score distributions differed: ChatGPT, 4 (4–5); Claude Sonnet, 4 (4–4); DeepSeek, 4 (3–4); and Gemini, 4 (3–4). The overall difference remained statistically significant (Friedman statistic = 25.80, W = 0.430, raw P < 0.0001). JAMA scores were low overall, with a median of 1 (1–1) for ChatGPT and 0 (0–0) for each of the other three chatbots. The Friedman test showed a significant difference among models (Friedman statistic = 57.00, W = 0.950, raw P < 0.0001). After Holm–Šídák correction across the four reliability outcomes, all omnibus comparisons remained statistically significant (all adjusted P < 0.0001). Kendall's W ranged from 0.430 for GQS to 0.950 for JAMA, indicating variation in the magnitude of between model differences across the four reliability measures.

Post hoc comparisons further demonstrated differences between individual chatbots (Fig. 3). ChatGPT had significantly higher DISCERN scores than Claude Sonnet (P < 0.05), DeepSeek (P < 0.001), and Gemini (P < 0.0001), while Claude Sonnet also scored significantly higher than Gemini (P < 0.01). For EQIP, ChatGPT scored significantly higher than Claude Sonnet and DeepSeek (both P < 0.001) and Gemini (P < 0.0001). For GQS, significant differences were observed between ChatGPT and DeepSeek and between ChatGPT and Gemini (both P < 0.05). For JAMA, ChatGPT scored significantly higher than Claude Sonnet, DeepSeek, and Gemini (all P < 0.0001). No other pairwise comparisons reached statistical significance.

The question level heatmap showed variability in reliability scores across the 20 individual questions (Fig. 4A). Nevertheless, the overall pattern was consistent with the aggregate analyses, with ChatGPT generally showing higher DISCERN and EQIP scores and more favorable GQS scores across questions. JAMA scores were limited across all models, with ChatGPT generally receiving higher scores than the other three chatbots.

Readability Analysis

The readability of responses generated by the four AI chatbots was assessed using six metrics: ARI, CL, FKGL, FRES, GFI, and SMOG (Table 3 and Fig. 5). Across all four chatbots, the mean scores for the five grade level indices were higher than the recommended sixth grade benchmark of 6, while mean FRES values were substantially below the recommended threshold of 80. These findings indicate that the responses generated by all four chatbots were generally more complex than the reading level recommended for patient facing health information.

Table 3
Readability Scores Across AI Chatbots.

Program	ARI (mean ± SD)	CL (mean ± SD)	FKGL (mean ± SD)	FRES (mean ± SD)	GFI (mean ± SD)	SMOG (mean ± SD)
ChatGPT-5.6 sol	15.44 ± 2.08	16.15 ± 1.95	13.73 ± 2.08	28.00 ± 13.00	14.24 ± 1.92	12.13 ± 1.42
Claude Sonnet 5	16.64 ± 3.91	15.13 ± 2.41	14.50 ± 3.48	29.90 ± 16.58	14.08 ± 2.76	12.37 ± 2.67
DeepSeek -V4-Flash	15.50 ± 2.54	14.59 ± 2.65	13.65 ± 2.74	32.55 ± 16.18	14.11 ± 2.39	12.25 ± 1.93
Gemini 3.6 Flash	15.18 ± 2.99	14.13 ± 2.16	13.27 ± 2.50	36.35 ± 13.41	13.79 ± 2.10	12.17 ± 1.89
6th grade level score	6	6	6	80–90	6	6
Supplementary Material						
Supplementary Table 1 Google Trends data of the 10 most significant keywords queried globally from August 2021 to August 2026.						
Supplementary Table 2 Technical details of AI chatbot access and settings.						

For the grade level indices, mean ARI scores ranged from 15.18 ± 2.99 for Gemini to 16.64 ± 3.91 for Claude Sonnet. Mean CL scores ranged from 14.13 ± 2.16 for Gemini to 16.15 ± 1.95 for ChatGPT, while FKGL scores ranged from 13.27 ± 2.50 for Gemini to 14.50 ± 3.48 for Claude Sonnet. GFI scores ranged from 13.79 ± 2.10 to 14.24 ± 1.92, and SMOG scores ranged from 12.13 ± 1.42 to 12.37 ± 2.67. In addition, mean FRES values were 28.00 ± 13.00 for ChatGPT, 29.90 ± 16.58 for Claude Sonnet, 32.55 ± 16.18 for DeepSeek, and 36.35 ± 13.41 for Gemini. Although numerical differences were observed among the four chatbots, the mean readability scores of all models remained outside the predefined sixth grade readability benchmarks.

Item level analysis further showed variation in readability across the 20 core questions (Fig. 4B). Despite this variation, the overall pattern was consistent across the four chatbots, with generally elevated grade level scores and low FRES values. Overall, the readability profiles suggested that the AI generated responses were written at a level more complex than that recommended for patient education.

4 Discussion

This study evaluated public information needs regarding venous thromboembolism after fracture surgery and examined the reliability and readability of responses generated by four contemporary AI chatbots. Several findings deserve attention. First, worldwide Google search interest in VTE remained relatively stable for most of the observation period but increased markedly from mid 2025 onward, while the most common related searches focused on prophylaxis, clinical guidelines, treatment, symptoms, and disease definition. Second, significant differences in reliability were observed among the four chatbots across DISCERN, EQIP, GQS, and JAMA assessments. ChatGPT generally achieved higher scores, particularly for DISCERN, EQIP, and JAMA, although the magnitude of between model differences varied according to the assessment instrument. Third, despite

differences in reliability, none of the four chatbots produced responses with readability profiles that approached the predefined sixth grade standard. Together, these findings suggest that AI chatbots may provide useful information on VTE after fracture surgery, but information quality, source transparency, and linguistic complexity remain important considerations when these systems are used for patient education.

The increasing Google search interest observed in our study highlights a growing demand for information about VTE. This is clinically relevant because previous research has demonstrated substantial gaps in public knowledge of thrombosis. In an international survey involving nine countries, awareness of DVT and PE was substantially lower than awareness of myocardial infarction and stroke, and only 45% of respondents recognized that blood clots could be prevented[18]. Awareness of surgery and hospitalization as VTE risk factors was also limited[18]. These findings are particularly relevant to patients undergoing fracture surgery because trauma, surgery, immobility, and patient related risk factors can act together to increase postoperative VTE risk[1]. Google Trends has been increasingly used to characterize population level health information seeking behavior[5], and the predominance of searches related to prophylaxis, treatment, symptoms, and guidelines in our analysis suggests that users are seeking information that may directly influence prevention and decisions about medical care. However, Google Trends reflects normalized relative search interest rather than absolute search volume, disease incidence, or individual patient behavior. Therefore, the increase observed in this study should be interpreted as an increase in relative online interest rather than evidence of an increase in the incidence of VTE.

The reliability analysis demonstrated meaningful differences among the four AI chatbots. ChatGPT achieved the highest median DISCERN and EQIP scores and had a higher JAMA score than the other models, while all four chatbots had a median GQS score of 4. This pattern is important because the four instruments evaluate related but distinct aspects of health information. DISCERN focuses strongly on the quality of information concerning treatment choices, whereas EQIP assesses the content and presentation of patient information[9]. GQS provides a broader assessment of overall quality, information flow, and usefulness. Thus, similar median GQS values do not necessarily indicate equivalent performance across more specific measures of information quality. Previous studies have likewise shown that AI generated medical information can reach moderate or good overall quality while still showing variation across models and assessment domains. Walker et al. found that ChatGPT provided medical information of broadly comparable quality to conventional online patient information, but noted that quality remained limited in several areas[9]. A systematic review of large language models in patient education also found generally promising performance, while emphasizing persistent concerns regarding accuracy, reliability, readability, and bias[19].

The JAMA findings require particular consideration. In our study, the median JAMA score was 1 for ChatGPT and 0 for Claude Sonnet, DeepSeek, and Gemini. The JAMA benchmark criteria were originally proposed to evaluate transparency of medical information according to authorship, attribution, disclosure, and currency[17]. Consequently, a low JAMA score should not automatically be interpreted as evidence that the clinical content of a chatbot response is incorrect. Rather, it indicates that important features of information provenance and transparency are absent. This distinction is particularly relevant for generative AI, which commonly presents information as a synthesized conversational response rather than as a conventional authored webpage. Similar variation has been observed in orthopedic chatbot studies. Tarapore et al. reported that AI systems answering postoperative questions about distal radius fracture surgery could provide information of generally good quality, but JAMA scores differed substantially across platforms[12]. Such findings suggest that model architecture,

version, platform design, prompt format, and the extent to which sources are explicitly provided may influence JAMA performance. For patient facing medical applications, future chatbot interfaces may therefore benefit from more consistent presentation of verifiable sources, dates, and indications of where medical information originates.

Readability was a consistent limitation across all four models. The mean ARI values were approximately 15 to 17, FKGL values approximately 13 to 15, and SMOG values approximately 12, while mean FRES values ranged from only 28.0 to 36.35. These values were far from the predefined sixth grade readability benchmarks used in this study. Importantly, the problem was not restricted to one chatbot; the same general pattern was observed across all models and across the individual questions. These findings are consistent with previous research showing that medical information generated by large language models is frequently written at a level that is too complex for patient education. In a study of postoperative questions regarding distal radius fractures, Tarapore et al. reported Flesch Kincaid grade levels of approximately 10.5 to 12.1 and Flesch Reading Ease scores of 35.8 to 41.4 across AI platforms[12]. A recent systematic review and meta analysis similarly concluded that the advanced reading level of chatbot responses remains an important barrier to their broader use in patient education[11].

The readability problem identified here is not unique to AI generated content. Orthopedic patient education materials have historically exceeded recommended reading levels. Eltorai et al. reported that 97% of patient education materials from the American Academy of Orthopaedic Surgeons exceeded the sixth grade reading level[6]. Similar findings have been reported for orthopedic trauma materials and other orthopedic subspecialties[7]. More recently, Marshall et al. demonstrated that systematically reducing long sentences and multisyllabic words could substantially reduce the reading grade level of orthopedic trauma patient education materials, with the mean Flesch Kincaid grade level decreasing from approximately grade 10 to below grade 6[20]. These findings are relevant to generative AI because readability is potentially modifiable. Rather than accepting the default output, AI systems could be instructed or designed to produce shorter sentences, use common language, explain medical terminology, and organize information around clear actions. However, simplification should not come at the expense of clinically important detail, particularly for topics such as anticoagulant safety, PE warning signs, or indications for urgent medical assessment.

The clinical implications of these findings are particularly important in the postoperative fracture population. Effective VTE education requires more than a general explanation of thrombosis. Patients may need to understand preventive medication, the importance of mobility, signs of DVT and PE, bleeding complications, missed anticoagulant doses, and when emergency medical attention is required. Public awareness studies have repeatedly identified deficiencies in knowledge of VTE symptoms and risk factors[18]. In addition, postoperative thromboprophylaxis may be difficult to follow in real world practice. Ghazi et al. identified inconvenience, injection related pain, fear, medication acquisition, and bruising as barriers to adherence to prophylactic enoxaparin after orthopedic trauma surgery[21]. Patient education may therefore have a direct role in supporting safe postoperative behavior. Green and Bernhofer found that a structured multimodal education program was associated with improved retention of important postoperative VTE prevention knowledge[22]. These observations support the use of AI chatbots as a potential supplementary educational resource, but not as a substitute for individualized counseling by clinicians. AI generated information should ideally reinforce professional advice, present clear warning signs and medication precautions, and direct patients to appropriate medical care when urgent symptoms occur.

The combined reliability and readability findings also demonstrate that these two dimensions should not be considered interchangeable. A response may contain clinically useful and reasonably well structured information while remaining too difficult for many patients to understand. Conversely, simpler language does not guarantee factual accuracy or appropriate clinical guidance. Lower health literacy has been associated with poorer health related outcomes and less effective use of health information[23]. For this reason, evaluation of AI based patient education should include both content quality and accessibility. The use of multiple reliability instruments in the present study provided complementary information about treatment content, patient information quality, overall presentation, and source transparency, while the six readability indices provided a broader assessment of linguistic complexity than a single readability formula. Future development of medical AI systems should therefore focus not only on improving factual reliability, but also on producing information that is understandable, actionable, appropriately sourced, and responsive to different levels of health literacy.

Several strengths of this study should be noted. The questions were derived from both real world search behavior and clinical experience and covered a broad range of patient concerns, including VTE risk, symptoms, prevention, medication safety, and postoperative management. The same questions were submitted verbatim to all four chatbots without additional prompting, which approximated a simple patient initiated interaction. Reliability was assessed using four complementary instruments, while readability was evaluated using six established indices. At the same time, several limitations should be considered. First, each question was submitted only once to each chatbot, and the study therefore did not evaluate response variability across repeated generations. Second, large language models are updated frequently; consequently, the findings represent the models and access conditions evaluated during the study period and may not remain unchanged after future updates. Third, DISCERN, EQIP, GQS, and JAMA were developed primarily for conventional health information resources rather than generative AI, and some criteria, particularly the JAMA transparency items, may not map perfectly to conversational systems. Fourth, readability formulas estimate linguistic difficulty from textual characteristics and do not directly measure patient comprehension, numeracy, cultural appropriateness, or the ability to act on the information. No direct patient comprehension testing was performed. Fifth, the study assessed information quality through established instruments but did not conduct a separate item by item comparison of every clinical statement against a predefined guideline based reference standard. Finally, Google Trends data are normalized and cannot identify individual users, motivations for searching, or absolute search volume [5] (Nuti et al., 2014). These limitations should be considered when interpreting both the comparative chatbot findings and the observed public search patterns.

5 Conclusion

This study demonstrates increasing public interest in information about VTE and identifies important differences in the reliability of information provided by contemporary AI chatbots regarding VTE after fracture surgery. ChatGPT showed higher reliability scores across several assessment domains, whereas low JAMA scores across the evaluated models highlighted limitations in source transparency. In addition, responses from all four chatbots were substantially more complex than the recommended sixth grade reading level for patient education. AI chatbots may have potential as supplementary tools for patient education, but their outputs should be used with appropriate caution and should not replace individualized clinical guidance. Future development should focus on improving source transparency, clinical reliability, and the readability of AI generated health information.

Declarations

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 a potential conflict of interest.

Funding

This study was supported by National Health Commission Hospital Management Research Institute - Modernization of Healthcare Institution Governance Systems and Governance Capacity (Evidence-Based) Research Project(YLZLXZ25K019): Real-world research on VTE prevention in postoperative patients: Multi-center validation of Peking University Shenzhen Hospital's original management model and exploration of optimized paths for primary prevention.

Author Contribution

Mengyue Xie, Chuanliang Zhang and Ming Xiao: Conceptualization, Data curation, Formal analysis, Investigation, Software, Writing—original draft. Xiaohua Wu: Formal analysis, Software, Writing—original draft. Li Yi: Methodology, Project administration, Resources, Supervision, Writing –review and editing. Yingsi Zhang: Methodology, Project administration, Resources, Supervision, Writing– review and editing. All authors have read and agreed to the published version of the manuscript.

Data Availability

All data supporting the findings of this study are available within the paper and its Supplementary Information.

References

1. Xia ZH, Chen WH, Wang Q. Risk factors for venous thromboembolism following surgical treatment of fractures: A systematic review and meta-analysis. Int Wound J. 2023;20(4):995–1007.
2. Anderson DR, et al. American Society of Hematology 2019 guidelines for management of venous thromboembolism: prevention of venous thromboembolism in surgical hospitalized patients. Blood Adv. 2019;3(23):3898–944.
3. Dwyer EP, Moed BR. Venous thromboembolism after hospital discharge in pelvic and acetabular fracture patients treated operatively. J Orthop Surg (Hong Kong). 2019;27(1):2309499019832815.
4. Dougherty JM, et al. Venous thromboembolism events in trauma patients after hospital discharge. J Trauma Acute Care Surg. 2025;98(5):704–12.
5. Nuti SV, et al. The use of google trends in health care research: a systematic review. PLoS ONE. 2014;9(10):e109583.

6. Eltorai AE, et al. Most American Academy of Orthopaedic Surgeons' online patient education material exceeds average patient reading level. Clin Orthop Relat Res. 2015;473(4):1181–6.
7. Eltorai AE, et al. Readability of patient education materials on the American Orthopaedic Society for Sports Medicine website. Phys Sportsmed. 2014;42(4):125–30.
8. Ayers JW, et al. Evaluating Artificial Intelligence Responses to Public Health Questions. JAMA Netw Open. 2023;6(6):e2317517.
9. Walker HL, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. J Med Internet Res. 2023;25:e47479.
10. McCarthy CJ, et al. Evaluation of an Artificial Intelligence Chatbot for Delivery of IR Patient Education Material: A Comparison with Societal Website Content. J Vasc Interv Radiol. 2023;34(10):1760–e176832.
11. Emile SH, et al. Assessment of the Utility of Artificial Intelligence-Based Chatbots in Patient Education: A Systematic Review and Meta-Analysis. Am Surg. 2026;92(1):258–69.
12. Tarapore R, et al. Artificial Intelligence Can Answer Postoperative Questions About Distal Radius Fractures-But Can Patients Understand the Answers? J Hand Surg Glob Online. 2025;7(6):100822.
13. Dadir IH. Accuracy and readability analysis of ChatGPT responses to frequently asked questions about olecranon fractures. J Orthop Surg (Hong Kong). 2026;34(2):10225536261469293.
14. Charnock D, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53(2):105–11.
15. Weil AG, et al. Evaluation of the quality of information on the Internet available to patients undergoing cervical spine surgery. World Neurosurg. 2014;82(1–2):e31–9.
16. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7(2):165–75.
17. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor–Let the reader and viewer beware. JAMA. 1997;277(15):1244–5.
18. Wendelboe AM, et al. Global public awareness of venous thromboembolism. J Thromb Haemost. 2015;13(8):1365–71.
19. Aydin S, et al. Large language models in patient education: a scoping review of applications in medicine. Front Med (Lausanne). 2024;11:1477898.
20. Marshall S, et al. A standardised method for improving patient education material readability for orthopaedic trauma patients. Musculoskelet Care. 2024;22(1):e1869.
21. Ghazi FG, et al. Patient-reported barriers for adherence to prophylactic enoxaparin after orthopedic trauma surgery. Injury. 2024;55(8):111650.
22. Green J, Bernhofer EI. Effectiveness of a patient education plan on knowledge of post-op venous thromboembolism survival skills. J Clin Nurs. 2018;27(7–8):e1485–93.
23. Berkman ND, et al. Low health literacy and health outcomes: an updated systematic review. Ann Intern Med. 2011;155(2):97–107.

Figures

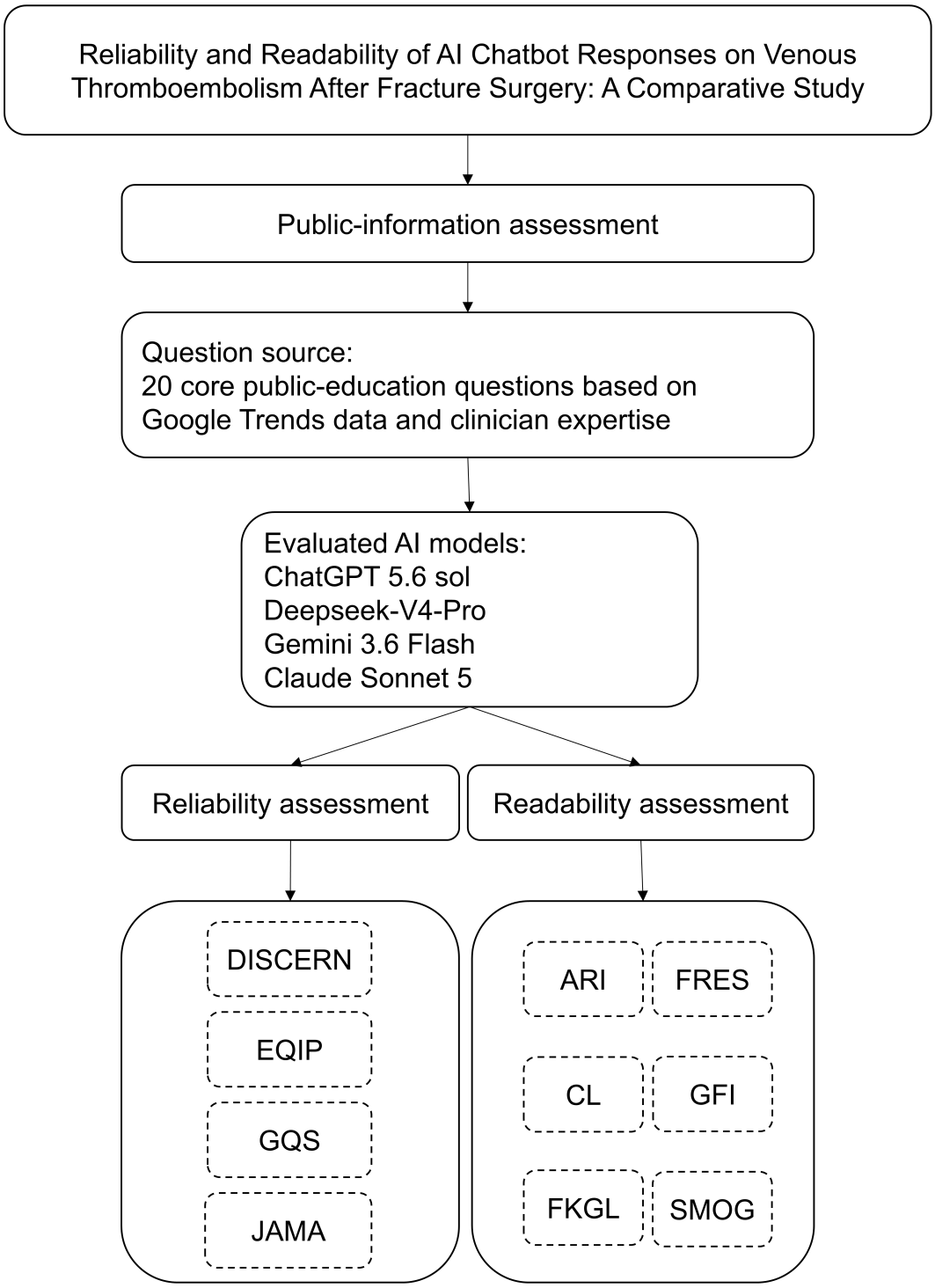


Figure 1

Study design and evaluation workflow.

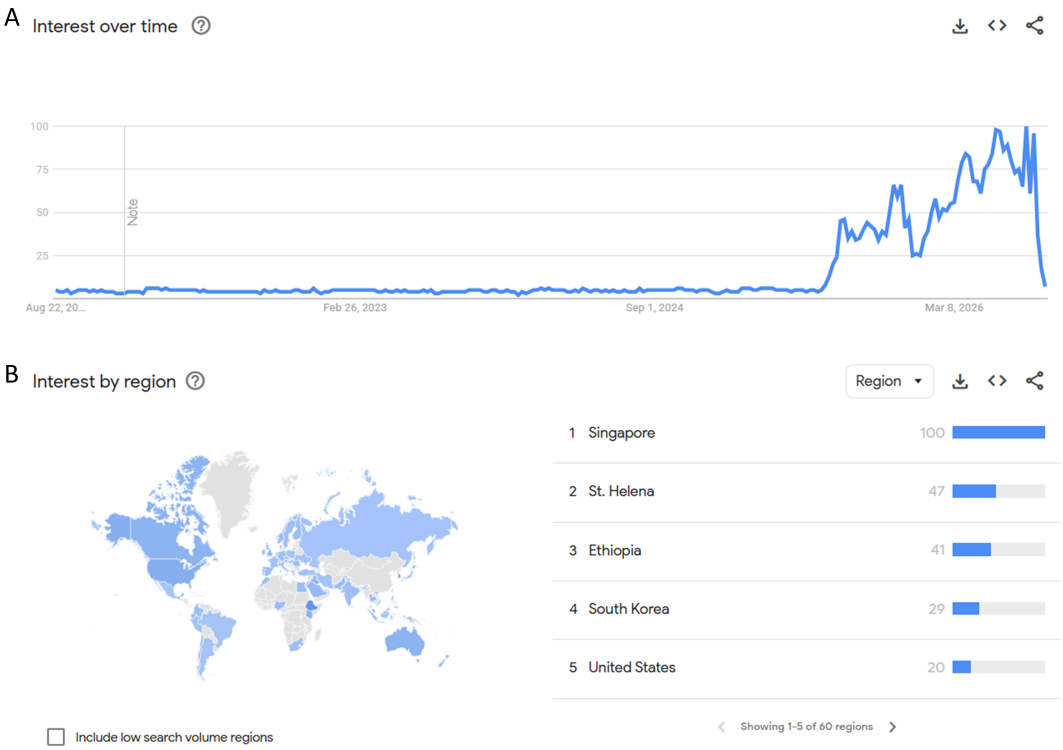


Figure 2

Worldwide public search interest in venous thromboembolism based on Google Trends.

(A) Temporal changes in worldwide Google search interest in venous thromboembolism from August 2021 to August 2026. Relative search interest was normalized on a scale from 0 to 100, with 100 representing the peak search interest during the selected period. (B) Geographic distribution of relative search interest in venous thromboembolism worldwide. Countries and regions were ranked according to normalized Google Trends scores, with Singapore showing the highest relative search interest (100), followed by St. Helena (47), Ethiopia (41), South Korea (29), and the United States (20). Google Trends data were accessed in August 2026.

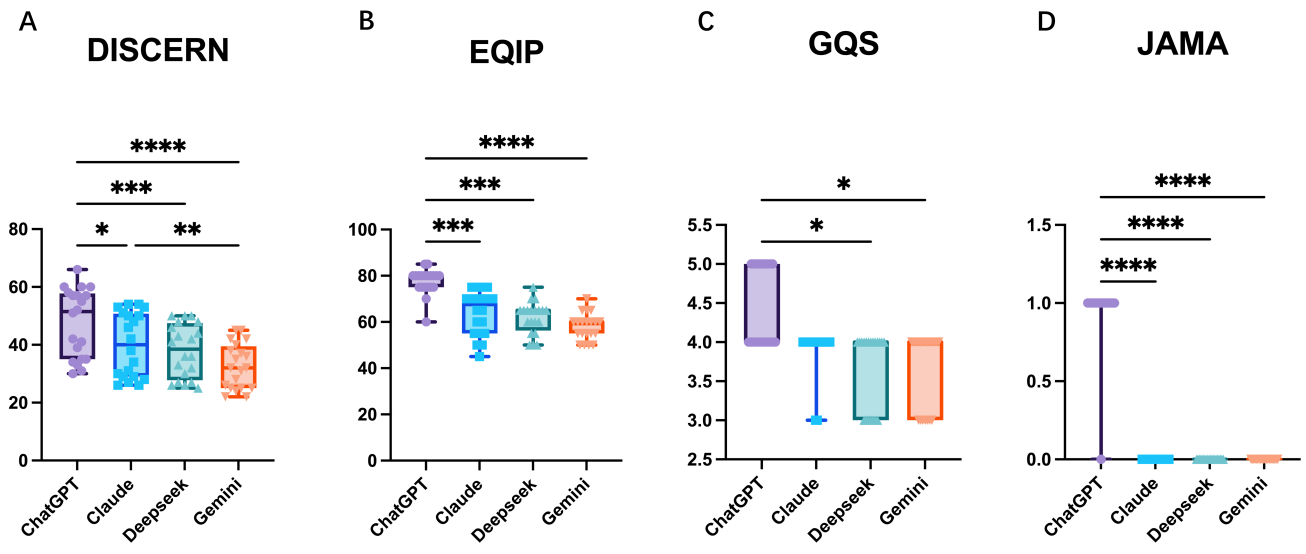


Figure 3

Reliability scores across AI chatbots. Reliability was assessed using four instruments: (A) DISCERN, (B) Ensuring Quality Information for Patients (EQIP), (C) Global Quality Scale (GQS), and (D) Journal of the American Medical Association (JAMA) benchmark criteria. Each point represents the score for one of the 20 core questions. Boxes represent the interquartile range, with the horizontal line indicating the median. Differences among the four chatbots were assessed using the Friedman test, followed by Dunn's multiple comparisons test for post hoc pairwise comparisons. *P < 0.05, **P < 0.01, ***P < 0.001, and ****P < 0.0001.

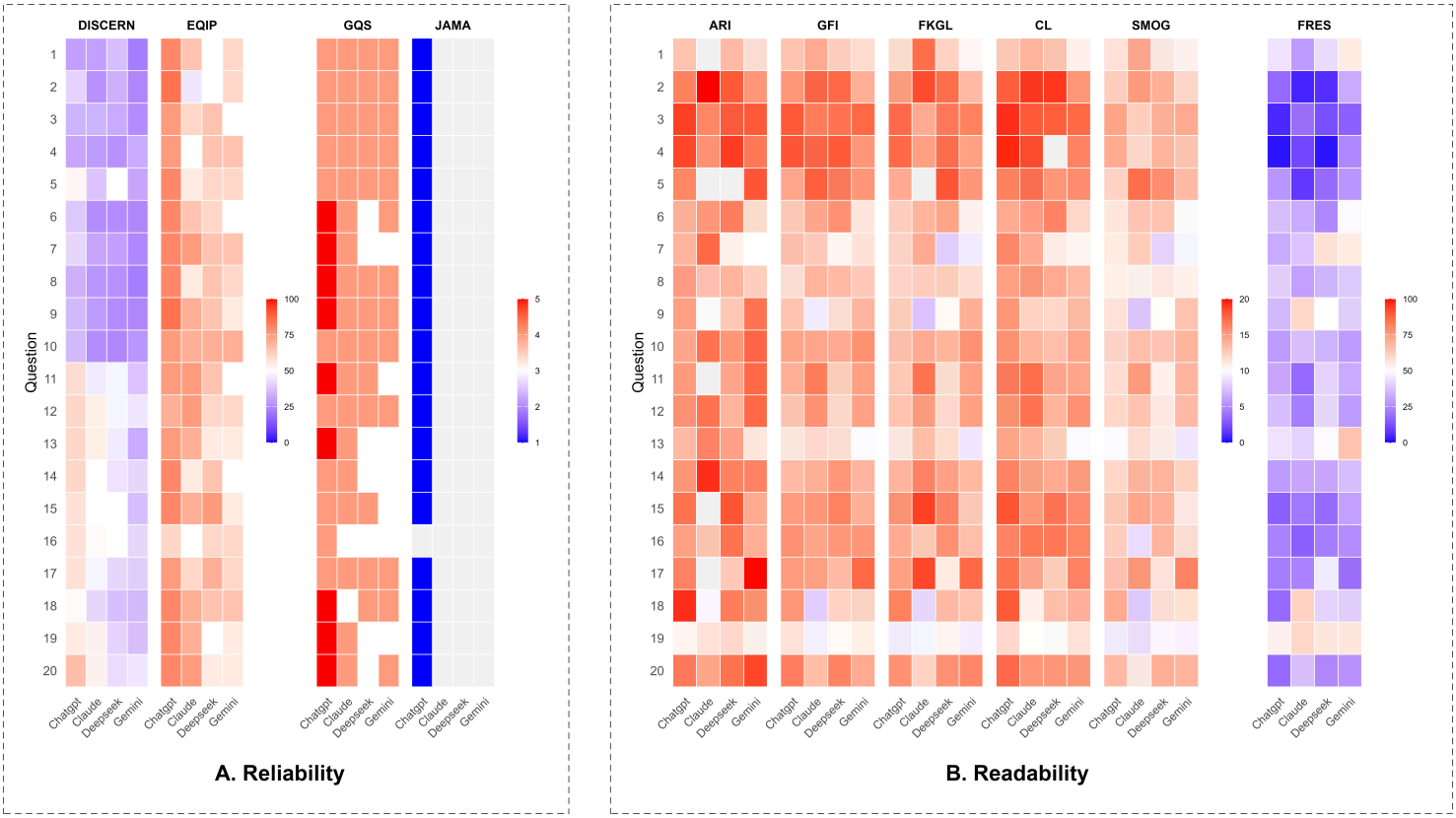


Figure 4

Heatmaps of reliability and readability scores across the 20 core questions. (A) Heatmaps of DISCERN, EQIP, GQS, and JAMA scores for each of the 20 core questions across the four AI chatbots. (B) Heatmaps of ARI, GFI, FKGL, CL, SMOG, and FRES scores for each question. Each row represents one core question, and each column represents one AI chatbot. Color intensity reflects the magnitude of the corresponding score, as indicated by the scale bars.

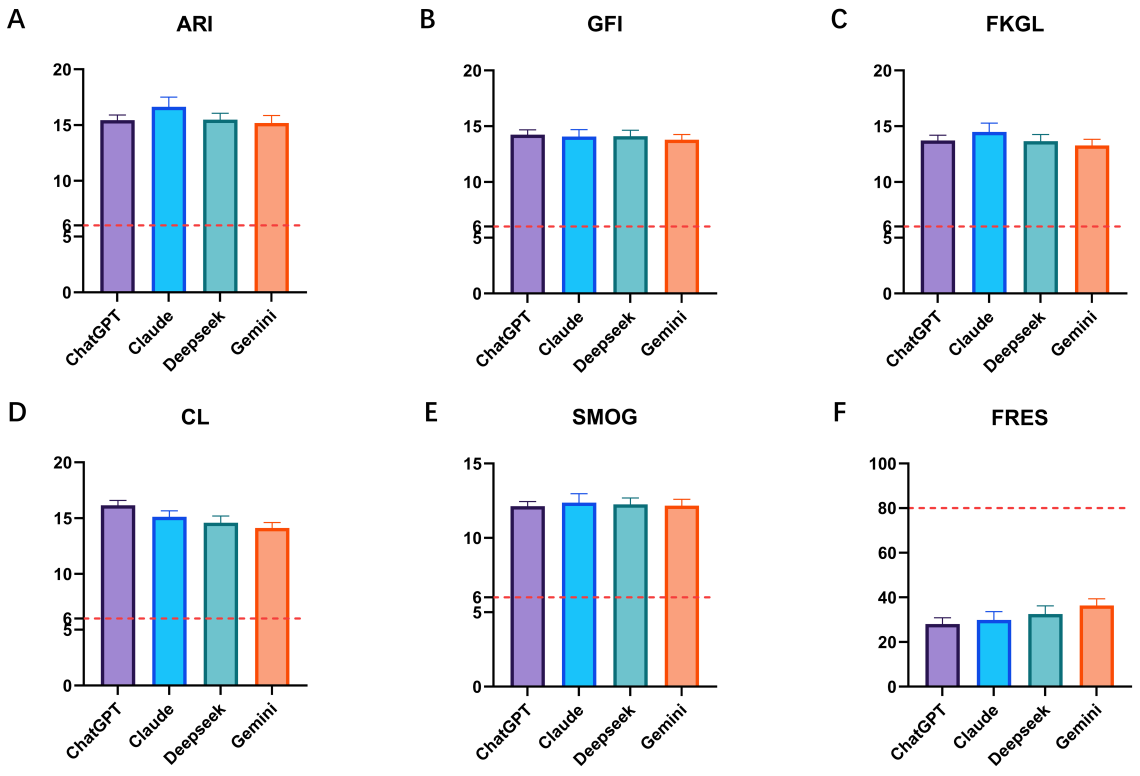


Figure 5

Readability scores across AI chatbots. Readability was assessed using six indices: (A) Automated Readability Index (ARI), (B) Gunning Fog Index (GFI), (C) Flesch Kincaid Grade Level (FKGL), (D) Coleman Liau Index (CL), (E) Simple Measure of Gobbledygook (SMOG), and (F) Flesch Reading Ease Score (FRES). Bars represent mean values, and error bars indicate standard deviations. The red dashed lines indicate the predefined sixth grade readability benchmarks, with a threshold of ≤6 for ARI, GFI, FKGL, CL, and SMOG, and ≥80 for FRES.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable.docx