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.

Safety and quality of patient-facing frozen shoulder information from four generative AI chatbot interfaces: a cross-sectional study

Qi Zang
Affiliated Hospital of Nanjing University Medical School
Dandan You
Affiliated Hospital of Nanjing University Medical School
Huijuan Huang
Affiliated Hospital of Nanjing University Medical School
Kai Chen
Affiliated Hospital of Nanjing University Medical School
Jin Song
Huzhou No. 1 People's Hospital
Huaqin Chen
13770036128@163.com

Affiliated Hospital of Nanjing University Medical School

Article

Keywords: adhesive capsulitis, artificial intelligence, chatbot, patient education, safety, readability

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10805692/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

Patients increasingly use generative artificial intelligence (AI) chatbots for musculoskeletal health information, but their safety and suitability for frozen shoulder education are uncertain. We submitted 50 prespecified patient- and caregiver-facing questions once to each of ChatGPT, Gemini, DeepSeek, and Doubao on 17 July 2026, yielding 200 first responses. Five clinicians independently assessed deidentified responses against question-specific evidence standards for safety, accuracy, empathy, information quality, transparency, and global quality; readability was calculated computationally. Twenty-one responses (10.5%) were classified as potentially unsafe. Interface-specific proportions ranged from 6.0% to 14.0%, without a significant overall difference ($P = 0.442$). Accuracy differed across interfaces ($P < 0.001$; Kendall's $W = 0.639$), as did empathy ($P < 0.001$; $W = 0.580$). ChatGPT had the highest median accuracy (4.40), DISCERN (60.00), and JAMA benchmark (3.00) scores, whereas Gemini had the highest median binary-adapted EQIP score (83.00) and the most favorable readability profile. All interfaces exceeded commonly recommended reading levels. No interface was uniformly superior: each produced potentially unsafe content and limited response-level transparency. These tools may support patient education but should not replace individualized clinical assessment and guidance.

Introduction

Frozen shoulder, or adhesive capsulitis, is a common and disabling shoulder disorder characterized by progressive pain and restriction of both active and passive glenohumeral motion. It affects approximately 2%–5% of the general population and may substantially impair sleep, self-care, occupational activity, and quality of life over a prolonged and clinically variable course [1,2]. The condition is particularly prevalent among individuals with diabetes, who have substantially higher odds of developing frozen shoulder and may experience less favorable pain, range-of-motion, and functional outcomes [3, 4]. Although management is predominantly nonoperative, appropriate care requires differentiation from rotator cuff disease, glenohumeral osteoarthritis, fracture, neurological disorders, and other causes of shoulder pain. Recommendations concerning exercise intensity, analgesic use, corticosteroid injection, hydrodilatation, manipulation under anesthesia, and arthroscopic capsular release must therefore be individualized according to diagnostic certainty, symptom irritability, comorbidities, treatment response, and patient preferences [1,2].

Patients increasingly seek explanations of symptoms, prognosis, self-management strategies, and treatment options via digital platforms before or between clinical encounters. Large language models (LLMs) can generate immediate and conversational responses, creating opportunities to expand access to patient education and support clinical communication. Previous studies have reported that LLM-generated answers may be perceived as informative and empathetic and, in some settings, may compare favorably with responses written by healthcare professionals [5,6]. However, favorable tone and linguistic fluency do not ensure clinical reliability. Systematic reviews have demonstrated substantial heterogeneity in the accuracy, adequacy, transparency, and quality of LLM-generated medical information across clinical specialties and use cases [7,8]. Physician-led safety evaluations have additionally identified responses with plausible pathways to patient harm [9]. Importantly, patients may not reliably recognize potentially harmful advice when it is presented confidently and persuasively, reinforcing the need for structured expert appraisal before these systems are used as independent sources of medical guidance [9,10].

The suitability of an LLM response for patient education is multidimensional. Factual accuracy and the absence of harmful recommendations are essential, but responses should also acknowledge patient concerns, explain uncertainty, describe the benefits and risks of available options, disclose relevant limitations, and direct users toward appropriate professional care. DISCERN provides a structured assessment of the reliability and quality of written information about treatment choices, whereas the Ensuring Quality Information for Patients instrument, the Journal of the American Medical Association benchmarks, and the Global Quality Score assess complementary aspects of content completeness, presentation, transparency, and overall usefulness [11–14]. Empathy is additionally relevant because acknowledgment of pain, sleep disruption, functional limitation, and anxiety may influence whether information is perceived as responsive and supportive. Nevertheless, greater comprehensiveness may be accompanied by increased length and linguistic complexity, creating a tension between informational completeness and accessibility. This is especially important because online health information frequently exceeds recommended reading levels, and similar readability limitations have been identified in orthopaedic materials generated by LLMs [15–17].

Existing orthopaedic studies have generally examined individual chatbots, restricted sets of frequently asked questions, or selected outcomes such as accuracy, comprehensiveness, or readability. Studies involving shoulder arthroplasty have found that LLM responses may be broadly satisfactory while still requiring professional clarification, lacking sufficient evidence or source transparency, or including statements that extend beyond the scope of the original question [18, 19]. A recent systematic review and meta-analysis of artificial intelligence applications in shoulder and elbow patient education similarly found that AI-generated information remained inferior to expert-produced information in several dimensions and was not yet suitable as an independent replacement for professional guidance [20]. Evidence specifically addressing patient-facing information on frozen shoulder remains limited. This gap is clinically important because its management requires an LLM to distinguish gradual disease progression from urgent presentations, tailor exercise and medication advice to symptom irritability and comorbidities, communicate uncertainty about procedural interventions, and provide procedure-specific postoperative precautions or rehabilitation restrictions. Apparently benign self-management advice may become unsafe when these contextual factors are overlooked.

Accordingly, this study compared four consumer-facing generative AI chatbot interfaces in answering standardized patient- and caregiver-facing questions about frozen shoulder. The primary evaluation focused on clinical safety and accuracy, with secondary assessments of empathy, information reliability and quality, response-level transparency, global quality, and readability.

Methods

Study design and reporting framework

This cross-sectional comparative study evaluated four consumer-facing chatbot interfaces in answering patient-oriented questions about frozen shoulder. The unit of analysis was an individual interface–question response. A prespecified set of 50 questions was submitted once to each interface, yielding 200 first responses for evaluation. No formal sample-size calculation was performed because the study evaluated the complete prespecified question set rather than a sampled subset. The complete study workflow is shown in Fig. 1.

The study was designed and reported in accordance with the Chatbot Assessment Reporting Tool (CHART), which provides recommendations covering model identification, prompt design, query strategy, performance evaluation, statistical analysis, and open-science practices for chatbot health-advice studies. Applicable domains of the Strengthening the Reporting of Observational Studies in Epidemiology framework were also considered. A completed CHART checklist is provided in Supplementary File 1 [21].

Ethics and patient/public involvement

This study evaluated chatbot outputs and did not involve patients, identifiable patient information, biological specimens, or protected health information. The five nonauthor clinical evaluators acted as expert assessors of nonhuman materials and provided informed consent to participate and to have their aggregate professional characteristics reported. According to local institutional policy, formal ethical approval was not required for this noninterventional methodological study. Patients and members of the public were not involved in question development, prompt engineering, reference-standard development, or performance evaluation; direct patient involvement should be incorporated into future validation studies.

Development of the question set and prompts

Two investigators, Q.Z. (an orthopaedic specialist nurse with 9 years of clinical experience) and K.C. (an orthopaedic physician with 5 years of clinical experience), independently compiled candidate questions between May 1 and June 1, 2026. Candidate questions were drawn from three prespecified source categories: (i) clinical practice guidelines and systematic reviews concerning frozen shoulder; (ii) patient-education resources published by professional societies, public-health organizations, and academic medical centers; and (iii) public search-demand and patient-question sources. Public search-demand signals were identified using Google Trends [22] (restricted to geo = US as an English-language search-demand proxy consistent with the final English prompts), whereas patient question themes and wording were identified from Baidu Tieba [23], DXY [24], and outpatient educational materials. The source, retrieval date, original wording, and topic category of each candidate question were recorded prospectively. Source language was not used as an eligibility criterion; instead, the English-language requirement applied to the standardized prompts used for formal screening and chatbot evaluation. Candidate questions retrieved from Chinese-language platforms, including Baidu Tieba and Dingxiangyuan (DXY), were translated into English during question standardization. Translation was meaning-based rather than literal. Colloquial or platform-specific expressions, personal identifiers, brand or hospital names, and clinically irrelevant narrative details were removed, while the core information need, clinical context, symptom characteristics, temporal information, and safety-relevant cues were preserved. Chinese questions expressing the same clinical intent as one another or as questions retrieved from English-language sources were semantically consolidated into a single candidate prompt. Medical terminology and sentence structure were subsequently harmonized into concise, patient-facing English. The standardized English prompts were compared with the original Chinese questions by the research team and revised by consensus to ensure semantic equivalence before eligibility screening and final selection.

Questions were eligible when they met all of the following criteria:

1. were presented as final, standardized English prompts, either originally written in English or translated and semantically normalized from Chinese-language sources, and were phrased from a patient or caregiver perspective;
2. directly addressed frozen shoulder or adhesive capsulitis;
3. could be answered using contemporary clinical evidence without access to individual examination findings;
4. addressed a clinically relevant domain, including definition, pathophysiology, symptoms, diagnosis, differential diagnosis, natural history, prognosis, self-management, medication, injection therapy, rehabilitation, procedural treatment, surgery, postoperative care, recurrence, or comorbidity;
5. were suitable for a single-turn chatbot interaction; and
6. were sufficiently distinct from other included questions.

Questions were excluded when they were duplicates or near-duplicates; primarily concerned an unrelated shoulder disorder; required interpretation of an individual radiograph, magnetic resonance image, laboratory result, or physical examination; requested a personalized diagnosis or prescription; contained commercial or brand-specific promotion; were not understandable without additional context; or were directed exclusively at specialist-level technical decision-making rather than at patient education.

Two reviewers, H.H. (Huijuan Huang) and D.Y. (Dandan You), independently screened candidate questions against the eligibility criteria. Disagreements were resolved by discussion and consensus, with adjudication by H.C. (Huaqin Chen) when necessary. A total of 106 candidate questions were initially identified. After removal of 4 duplicates/near-duplicates and exclusion of 45 questions for prespecified reasons, 57 candidates remained. Seven of these were merged into other retained questions because of semantic overlap, yielding the final set of 50 questions. The complete source log, eligibility criteria, screening decisions, topic classification, and final prompts are provided in Supplementary File 2.

No investigator-supplied system, role, or framing prompt was added. For each interaction, the prompt consisted solely of the corresponding patient-oriented question entered verbatim in English. Prompt development therefore followed the question-development process described above; the complete prompt set is reproduced in Supplementary File 2.

Development of the question-specific reference standard

Before formal response evaluation, two clinical investigators, Q.Z. (an orthopaedic specialist nurse with 9 years of clinical experience) and K.C. (an orthopaedic physician with 5 years of clinical experience), independently developed a question-specific reference-standard mapping document. The evidence base comprised 20 prespecified sources, including clinical practice and imaging guidelines, a specialist clinical care pathway, consensus recommendations, systematic reviews, meta-analyses, and an authoritative disease primer; the evidence search was finalized on 07 July 2026.

The principal sources included the 2025 Clinical Practice Guidelines for Diagnosis and Non-Surgical Treatment of Primary Frozen Shoulder [2], the 2025 British Elbow and Shoulder Society patient care pathway [25], the 2013 Journal of Orthopaedic & Sports Physical Therapy clinical practice guideline for adhesive capsulitis [26], and the American College of Radiology Appropriateness Criteria for acute and chronic shoulder pain [27]. The mapping also incorporated systematic reviews and meta-analyses addressing natural history, comparative treatment effectiveness, exercise and physiotherapy, corticosteroid injection, hydrodilatation, acupuncture, arthroscopic surgery, and associations with diabetes and thyroid disease. Additional safety-critical evidence addressed nonsteroidal anti-inflammatory drug use, chest-pain evaluation, and septic arthritis.

For each question, the mapping document specified:

1. the intended scope of the question;
2. the essential factual elements expected in an adequate answer;
3. clinically acceptable alternative formulations;
4. important uncertainties or areas in which evidence was limited;
5. relevant red flags and indications for urgent assessment;
6. contraindications and patient-specific precautions;
7. material omissions that could affect patient decision-making; and
8. statements or recommendations that could create a plausible risk of harm.

The investigators compared their independently prepared mappings and resolved discrepancies by consensus. The final mapping was locked before formal scoring began and before evaluators received the deidentified response set. It served as the common clinical anchor for safety and accuracy evaluation rather than as a requirement for exact wording. The complete question-to-standard mapping, supporting references, and prespecified safety considerations are provided in Supplementary File 3.

Interface selection and model identification

ChatGPT, Gemini, DeepSeek, and Doubao were selected a priori because their consumer-facing interfaces were publicly accessible at the time of the study, were widely used general-purpose conversational systems, and represented distinct commercial developers and model ecosystems. The selection was intended to compare interfaces that patients could access directly rather than specialized medical models or locally deployed research models. In accordance with CHART, the provider, product name, interface-displayed model identifier, query date, access route, account tier, enabled tools, and model status were recorded where available [21]. The interface-displayed identifiers at the time of querying were GPT-5.6-sol for ChatGPT, Gemini 3.5 Flash for Gemini, and DeepSeek-V3 for DeepSeek. Doubao did not disclose its backend model identifier.

Table 1
Interface and query metadata

Interface	Provider	Interface-displayed model/version	Release/last-update information	Source status	Access route/tier	Tools/retrieval/memory settings	Query date/location
ChatGPT	OpenAI	GPT-5.6-sol	GPT-5.6 family released on 9 July 2026; exact web-interface snapshot and last-update date: Not publicly disclosed	Proprietary, closed-weight model	Web interface (Microsoft Edge Stable 150.0.4078.65); free consumer tier	Default web interface; no investigator-selected external tools or retrieval functions; memory, personalization, custom instructions, and other account-level settings not recorded; generation parameters not exposed	17 July 2026; Yancheng, Jiangsu, China
Gemini	Google DeepMind	Gemini 3.5 Flash	Released on 19 May 2026; exact web-interface snapshot and last-update date: Not publicly disclosed	Proprietary, closed-weight model	Web interface (Google Chrome Stable 150.0.7871.114/115); free consumer tier	Default web interface; no investigator-selected external tools or retrieval functions; memory, personalization, custom instructions, and other account-level settings not recorded; generation parameters not exposed	17 July 2026; Yancheng, Jiangsu, China
DeepSeek	DeepSeek	DeepSeek-V3	Initial release on 26 December 2024; DeepSeek-V3-0324 update released on 25 March 2025; exact web-interface backend version: Not publicly disclosed	Open-weight model family; DeepSeek-V3-0324 released under the MIT License	Web interface (Microsoft Edge Stable 150.0.4078.65); free consumer tier	Default web interface; no investigator-selected external tools or retrieval functions; memory, personalization, custom instructions, and other account-level settings not recorded; generation parameters not exposed	17 July 2026; Yancheng, Jiangsu, China
Doubao	ByteDance	Not disclosed by provider	Not publicly disclosed	Not publicly disclosed	Web interface (Microsoft Edge Stable 150.0.4078.65); free consumer tier	Default web interface; no investigator-selected external tools or retrieval functions; memory, personalization, custom instructions, and other account-level settings not recorded; generation parameters not exposed	17 July 2026; Yancheng, Jiangsu, China

All interface and displayed-model metadata were recorded on 17 July 2026 and verified against the original query logs. All four interfaces were accessed through free consumer tiers. The interface-displayed model identifiers were GPT-5.6-sol, Gemini 3.5 Flash, and DeepSeek-V3; the Doubao interface did not disclose its backend model.

Query procedure

All questions were submitted on 17 July 2026 from Yancheng No. 1 People's Hospital, Yancheng, Jiangsu, China, using a Windows operating system. ChatGPT, DeepSeek, and Doubao were accessed through Microsoft Edge Stable 150.0.4078.65, whereas Gemini was accessed through Google Chrome Stable 150.0.7871.114/115; these were the browser versions documented in the original query logs. Queries were performed by Q.Z. and independently verified against the query log by K.C. Completing all queries on the same date minimized the possibility that provider-side updates would affect comparisons.

Each of the 50 questions was entered verbatim in English. A separate, newly initiated conversation was used for every question to prevent information from preceding questions from influencing subsequent answers. No follow-up question, clarification, corrective instruction, or request for expansion was submitted. The first complete response generated by each interface was retained, and no response was regenerated or selectively chosen.

The consumer interfaces did not permit investigator control of temperature, top-p, random seed, token limit, or other generation parameters; these settings were therefore recorded as unavailable and were not estimated. No investigator-selected web search, citation retrieval, plugins, or other external tools were enabled. Provider-controlled model routing and generation settings that were not exposed by the interfaces were not inferred. The complete structured archive of the 200 interface–question responses is provided in Supplementary File 4.

Response capture and data management

Each response was copied immediately after generation to an archival record labeled with its question number, interface, date, time, access route, and visible model identifier. A separate evaluation copy was prepared before formal scoring. All model identifiers and interface-specific, nonsemantic symbols were removed, repeated spaces were standardized, and records were converted to a consistent structure. Every substantive statement and clinically relevant element was preserved; no factual content was corrected, paraphrased, translated, shortened, or supplemented.

The finalized dataset contained one record for each interface–question combination, with the fields question, model, and answer_text; the variable named model identifies the consumer interface and does not establish the backend model used for generation. Two investigators independently checked the structured dataset against the original transcripts. Copying or formatting discrepancies were resolved by referring to the original interface record. The complete question-and-answer archive is provided in Supplementary File 4.

Human evaluation procedure

Five senior clinical evaluators, none of whom was a manuscript author or involved in developing the question-specific reference standard, independently evaluated all 200 responses. Eligible clinicians who met the prespecified qualification criteria were invited by the research team. No evaluator received compensation. Each evaluator provided informed consent to participate and to have aggregate professional characteristics reported. Their qualifications and clinical experience are summarized in Table 2. Before formal scoring, all evaluators received a standardized manual containing the question-specific reference standard, operational definitions for each metric, item-level scoring instructions, and examples representing different score levels.

A separate calibration set comprising 20 additional chatbot responses was generated specifically for evaluator training using pilot questions that were not included in the final 50-question set. These responses were generated under the same response-generation conditions as the analytical responses but were never included in the 200-response analytical dataset. All evaluators independently scored the calibration responses and subsequently reviewed the identified disagreements. Discussion continued until ambiguities in the operational definitions and item-level scoring rules had been resolved and the scoring procedures had been agreed upon by all evaluators. No prespecified quantitative agreement threshold, such as a κ or intraclass correlation coefficient threshold, was used to determine completion of calibration. The scoring manual was then finalized, and none of the calibration ratings were included in the formal analyses.

Before formal scoring, evaluators received the prepared response set without model or interface identifiers and were not told the source interface. No additional randomization of response presentation was used. Distinctive language or writing patterns could nevertheless have allowed evaluators to infer the source, so complete blinding cannot be guaranteed. During formal scoring, evaluators worked independently, discussion was prohibited, and no evaluator had access to another evaluator's scores.

Table 2
Characteristics of the five evaluators

Evaluator	Academic degree	Professional title	Specialty	Years of clinical practice	Years of frozen-shoulder-related experience	Previous experience evaluating patient-facing health information or AI-generated outputs
Evaluator 1	Doctoral degree	Chief physician	Orthopaedic surgery	35	30	None
Evaluator 2	Doctoral degree	Chief physician	Orthopaedic surgery	28	22	None
Evaluator 3	Master's degree	Professor	Surgery	25	11	None
Evaluator 4	Bachelor's degree	Chief nurse	Orthopaedic nursing	30	25	None
Evaluator 5	Master's degree	Chief nurse	Orthopaedic nursing	22	15	None

Frozen-shoulder-related experience was defined as clinical, nursing, teaching, or assessment experience involving patients with frozen shoulder. AI, artificial intelligence.

Outcome measures and scoring

Safety

Safety was evaluated as a binary outcome. A score of 0 indicated that the response contained no clinically meaningful recommendation or omission creating a plausible pathway to patient harm; a score of 1 indicated at least one potentially unsafe element. Unsafe elements included statements or clinically important omissions that could plausibly lead to physical harm, inappropriate self-management, medication misuse, excessive or contraindicated exercise, delayed assessment of a serious alternative diagnosis, failure to recognize an emergency, premature procedural treatment, or violation of procedure-specific postoperative restrictions if acted upon by a patient or caregiver. This definition was adapted from a physician-led safety-evaluation framework [9].

Each evaluator independently assigned a score of 0 or 1. The final response-level classification was determined by majority vote; a response was classified as potentially unsafe when at least three of the five evaluators assigned a score of 1. Classification was based on plausible harm rather than minor factual, linguistic, or stylistic imperfections.

For descriptive characterization, the recorded rationales for all response-level potentially unsafe classifications were reviewed, and each response was assigned to a single primary safety category based on its principal documented concern. Category frequencies and percentages were summarized descriptively without inferential testing.

Accuracy

Accuracy was assessed using a prespecified five-point ordinal scale adapted from published physician evaluations of medical chatbot responses [9]. Each evaluator assigned an integer score from 1 to 5, and the response-level accuracy score was the arithmetic mean of the five ratings (Table 3). Scores represented: 1, completely incorrect or nonresponsive; 2, mostly incorrect with major errors or omissions; 3, partially correct with important errors or omissions; 4, mostly correct with only minor errors, imprecision, or noncritical omissions; and 5, fully or almost fully correct without material errors or omissions.

Empathy

Empathy was evaluated using a five-point ordinal scale adapted from a blinded comparison of physician- and chatbot-generated responses to patient questions [5]. Each evaluator assigned an integer score from 1 to 5, and the response-level score was the arithmetic mean of the five ratings (Table 4). Scores represented: 1, not empathetic; 2, slightly empathetic; 3, moderately empathetic; 4, empathetic; and 5, very empathetic, with strong recognition, validation, compassion, and patient-centered support.

DISCERN

The DISCERN instrument was used to evaluate the reliability and quality of information about treatment choices. It comprises 16 items scored from 1 to 5, producing a total score of 16–80. Items 1–8 assess reliability, items 9–15 assess the quality of treatment-choice information, and item 16 provides an overall quality rating [11]. These categories were used only descriptively. Each evaluator calculated a total score, and the response-level DISCERN score was the arithmetic mean of the five totals.

Ensuring Quality Information for Patients (EQIP; binary adaptation)

The original EQIP instrument comprises 20 items covering the content, clarity, structure, and usefulness of written patient information, with response options of yes, partly, no, and not applicable [12]. This study used a prespecified binary adaptation for short chatbot responses. Each item received 5 points when explicitly and fully satisfied and 0 points when absent, inadequately addressed, or only partially satisfied. All 20 items were retained without a not-applicable category, giving a fixed denominator. A comparable adaptation has been used in patient-facing LLM research [28]. Although removing the not-applicable category may reduce discrimination for very brief responses in which certain criteria are genuinely irrelevant, the fixed denominator ensures comparability across interfaces and evaluators and avoids evaluator-specific variability in applicability judgments. The total score was calculated as the number of fully satisfied criteria divided by 20 and multiplied by 100. Each evaluator calculated a total EQIP (binary-adaptation) score, and the response-level score was the arithmetic mean of the five totals.

JAMA benchmarks

The four JAMA benchmark criteria—authorship, attribution, disclosure, and currency—were evaluated for each response [13]. One point was awarded for each criterion explicitly satisfied and 0 points when absent, producing a total score of 0–4. The response-level score was the arithmetic mean of the five evaluators' totals. These benchmarks were originally developed for static internet health resources; their application to individual conversational responses should be interpreted with caution because chatbot interfaces do not routinely restate provider identity, ownership disclosures, or evidence dates within every answer. Nevertheless, the benchmarks provide a standardized, although conservative, transparency measure for cross-interface comparison.

Global Quality Score

Overall quality and usefulness were evaluated using the GQS, a single-item five-point scale assessing information quality, organization, flow, and usefulness to patients [14]. Each evaluator assigned an integer score from 1 to 5, and the response-level GQS was the arithmetic mean of the five ratings. Scores represented: 1, very poor; 2, poor; 3, fair; 4, good; and 5, excellent quality, organization, flow, and usefulness to patients.

Readability assessment

Readability was evaluated from the complete English text of each response using six established indices: Automated Readability Index (ARI), Coleman–Liau Index (CLI), Flesch–Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Simple Measure of Gobbledygook (SMOG), and Flesch Reading Ease Score (FRES) [29–34]. ARI, CLI, FKGL, GFI, and SMOG estimate educational grade level, with lower values indicating easier reading; FRES is scaled in the opposite direction, with higher values indicating easier reading.

Responses were normalized using a prespecified text-processing pipeline that removed nonsemantic interface artifacts and repeated whitespace while retaining all words, numbers, sentences, and substantive punctuation. Readability indices were calculated in R using textstat_readability from the quanteda.textstats package with the measures ARI, Coleman.Liau.short, Flesch.Kincaid, FOG, SMOG, and Flesch [35]. Response length was quantified from the same normalized text as the number of English letter/number word tokens, with internal apostrophes and hyphens retained.

Statistical analysis

Statistical analyses used a question-matched design, with responses to the same question treated as paired observations across interfaces. For evaluator-assessed nonbinary outcomes, the mean of the five evaluators' scores for each question–interface response was used in formal between-interface comparisons. Safety outcomes were summarized as frequencies and percentages, ordinal quality scores as medians with interquartile ranges, and readability indices and word counts as means with standard deviations.

Overall differences in safety were assessed using Cochran's *Q* test, followed by pairwise McNemar tests. Accuracy, empathy, DISCERN, EQIP (binary adaptation), JAMA, GQS, readability indices, and word count were compared using the Friedman test, followed by paired Wilcoxon signed-rank tests for post hoc comparisons. For each outcome separately, the Benjamini–Hochberg procedure was applied across the six pairwise interface comparisons to control the false discovery rate within that outcome; pairwise comparisons from different outcomes were not included in a common adjustment family.

Kendall's W was reported as the effect size for Friedman tests. For safety, $Q/[n(k-1)]$ was reported as a descriptive effect-size index. Pairwise effects were summarized as paired rank-biserial correlations for nonbinary outcomes and as paired differences in the proportion of safe responses and matched odds ratios for safety. Exploratory Spearman correlations assessed associations of response word count with EQIP and FKGL pooled across the 200 responses; these correlations combined within-interface and between-interface variation and were not intended to estimate interface-specific associations. Ninety-five percent CIs for effect sizes, paired differences, and length correlations were estimated using 3,000 nonparametric bootstrap resamples at the question level.

Inter-evaluator agreement was evaluated using Fleiss' kappa for safety and two-way random-effects, absolute-agreement intraclass correlation coefficients for accuracy, empathy, DISCERN, EQIP (binary adaptation), JAMA, and GQS. ICC(2,1) quantified single-evaluator agreement, whereas ICC(2,5) quantified the reliability of the five-evaluator mean score used in formal comparisons. ICC(2,5) point estimates and 95% CIs were obtained from ICC(2,1) estimates using the Spearman–Brown transformation. The 95% CI for Fleiss' kappa was obtained from 3,000 nonparametric bootstrap resamples.

Response-generation reproducibility was not evaluated because each question was submitted once to each interface under the prespecified single-response protocol; inferential results therefore describe this set of first responses rather than within-interface distributions across repeated generations. All tests were two-sided, and $P < 0.05$ was considered statistically significant. Analyses were performed using R version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria), primarily with the irr version 0.84.1 and psych version 2.5.6 packages.

Use of generative AI in manuscript preparation

OpenAI's ChatGPT and Codex were used after completion of the study analyses to assist with English-language editing, manuscript organization, consistency checking, and layout of the study workflow diagram. These tools were not used to generate or modify study data, assign evaluator scores, or perform statistical analyses. All authors critically reviewed and revised the assisted text, verified factual and statistical statements against the study records, and take full responsibility for the accuracy, originality, and integrity of the final manuscript.

Results

Inter-evaluator agreement

Complete ratings were available for all 200 responses. Fleiss' κ was 0.847 (95% CI 0.758–0.915) for safety. Single-evaluator ICC(2,1) values were 0.870 (95% CI 0.842–0.894) for DISCERN, 0.843 (95% CI 0.811–0.872) for EQIP (binary adaptation), 0.820 (95% CI 0.784–0.852) for JAMA, 0.818 (95% CI 0.782–0.850) for GQS, 0.576 (95% CI 0.515–0.638) for accuracy, and 0.452 (95% CI 0.386–0.520) for empathy. Corresponding five-

evaluator average-score ICC(2,5) values were 0.971 (95% CI 0.964–0.977), 0.964 (95% CI 0.955–0.971), 0.958 (95% CI 0.948–0.966), 0.957 (95% CI 0.947–0.966), 0.872 (95% CI 0.842–0.898), and 0.805 (95% CI 0.759–0.844), respectively.

Safety, accuracy, and empathy

Of the 200 responses, 21 were classified as potentially unsafe: 3/50 (6.0%) for ChatGPT, 5/50 (10.0%) for DeepSeek, 6/50 (12.0%) for Gemini, and 7/50 (14.0%) for Doubao. Of these classifications, 13 were unanimous (5–0), six were based on 4–1 majorities, and two were based on 3–2 majorities.

Based on the principal concern recorded for each response, the most frequent categories were inappropriate or insufficiently individualized exercise and rehabilitation advice (4/21, 19.0%) and context loss or answer–question mismatch (4/21, 19.0%). Other categories included overconfident or insufficiently contextualized procedural referral and informed-consent advice (3/21, 14.3%); overstatement of pathophysiology, prognosis, or treatment efficacy (3/21, 14.3%); misleading diagnostic terminology and inadequate preservation of the differential diagnosis (2/21, 9.5%); overstatement of the evidence supporting acupuncture (2/21, 9.5%); advice that could conflict with repair-specific postoperative restrictions (2/21, 9.5%); and insufficiently individualized medication or injection advice (1/21, 4.8%). The associated potential harms included symptom exacerbation or disruption of a healing repair, diagnostic or specialist-referral delay, premature escalation to invasive treatment, and medication-related adverse events. This taxonomy was descriptive, and no inferential comparisons were performed across safety categories (Supplementary File 6, Tables S5 and S6).

Neither the overall comparison (Cochran's Q = 2.692, df = 3, P = 0.442; effect size = 0.018, 95% CI 0.002–0.103) nor any adjusted pairwise comparison was statistically significant (all adjusted P > 0.99; Supplementary File 5, Table S1). However, because no equivalence margin was prespecified and the confidence interval and number of matched questions did not support an equivalence interpretation, these nonsignificant findings should not be interpreted as evidence of equivalent safety across interfaces.

Accuracy differed across interfaces (P < 0.001; Kendall's W = 0.639, 95% CI 0.513–0.777), with median scores of 4.40 for ChatGPT, 4.20 for Gemini, 4.00 for DeepSeek, and 3.80 for Doubao; all adjusted pairwise comparisons were significant (P ≤ 0.003). Empathy also differed across interfaces (P < 0.001; Kendall's W = 0.580, 95% CI 0.471–0.700), with median scores of 4.00 for ChatGPT and Gemini, 3.60 for DeepSeek, and 3.40 for Doubao. All adjusted pairwise comparisons were significant (P ≤ 0.009), except for ChatGPT versus Gemini (P = 0.063). Figure 2 and Table 3 present the score distributions and summary statistics, and complete pairwise estimates are provided in Supplementary File 5, Tables S2 and S4.

Table 3
Summary of evaluated outcomes across chatbot interfaces

Metric	ChatGPT	Gemini	DeepSeek	Doubao	*P* value	Kendall's *W*	95% *CI* for *W*
Accuracy	4.40 [4.40, 4.60]	4.20 [4.20, 4.40]	4.00 [4.00, 4.20]	3.80 [3.65, 4.00]	< 0.001	0.639	0.513–0.777
Empathy	4.00 [3.80, 4.00]	4.00 [3.80, 4.20]	3.60 [3.40, 4.00]	3.40 [3.20, 3.60]	< 0.001	0.580	0.471–0.700
DISCERN	60.00 [58.45, 64.70]	59.00 [57.00, 62.30]	56.50 [53.90, 59.10]	53.10 [50.90, 56.15]	< 0.001	0.600	0.500–0.706
EQIP (binary adaptation)	81.00 [78.00, 83.75]	83.00 [81.00, 85.00]	80.50 [76.00, 83.00]	76.00 [72.25, 77.00]	< 0.001	0.343	0.242–0.475
JAMA	3.00 [2.00, 3.00]	2.00 [2.00, 2.00]	2.00 [2.00, 2.00]	2.00 [2.00, 2.00]	< 0.001	0.343	0.191–0.532
GQS	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	0.001	0.108	0.039–0.228
ARI	14.18 ± 2.76	12.63 ± 1.99	12.35 ± 2.82	14.02 ± 3.17	< 0.001	0.128	0.059–0.254
CLI	14.79 ± 1.71	12.39 ± 1.64	13.18 ± 2.28	15.95 ± 2.80	< 0.001	0.559	0.442–0.690
FKGL	13.29 ± 2.43	11.80 ± 1.87	11.90 ± 2.57	13.02 ± 3.00	0.001	0.105	0.044–0.223
GFI	15.95 ± 2.86	14.64 ± 2.14	14.93 ± 2.99	15.60 ± 3.61	0.032	0.059	0.021–0.160
SMOG	14.44 ± 2.03	13.39 ± 1.64	13.68 ± 2.12	13.90 ± 2.45	0.012	0.073	0.029–0.179
FRES	34.85 ± 11.61	47.19 ± 10.37	42.62 ± 13.92	32.20 ± 18.08	< 0.001	0.393	0.282–0.534
Word count	239.22 ± 131.15	415.12 ± 105.61	439.02 ± 230.96	296.62 ± 147.49	< 0.001	0.477	0.348–0.612

Accuracy, empathy, DISCERN, EQIP (binary adaptation), JAMA, and GQS are medians (interquartile ranges); readability indices and word count are means +/- standard deviations. Overall comparisons used the Friedman test. ARI, Automated Readability Index; CI, confidence interval; CLI, Coleman-Liau Index; EQIP, Ensuring Quality Information for Patients; FKGL, Flesch-Kincaid Grade Level; FRES, Flesch Reading Ease Score; GFI, Gunning Fog Index; GQS, Global Quality Score; JAMA, Journal of the American Medical Association; SMOG, Simple Measure of Gobbledygook.

Reliability and quality of information

All four information-quality measures differed across interfaces (Table 3). Median DISCERN scores were 60.00 for ChatGPT, 59.00 for Gemini, 56.50 for DeepSeek, and 53.10 for Doubao ($P < 0.001$; Kendall's $W = 0.600$, 95% CI 0.500–0.706). Median EQIP (binary adaptation) scores were highest for Gemini (83.00), followed by ChatGPT (81.00), DeepSeek (80.50), and Doubao (76.00; $P < 0.001$; Kendall's $W = 0.343$, 95% CI 0.242–0.475). Median JAMA scores were 3.00 for ChatGPT and 2.00 for the other interfaces ($P < 0.001$; Kendall's $W = 0.343$, 95% CI 0.191–0.532). Although all interfaces had a median GQS of 4.00, the overall difference was significant ($P = 0.001$; Kendall's $W = 0.108$, 95% CI 0.039–0.228). Figure 3 shows the score distributions.

All six adjusted DISCERN contrasts were significant ($P \leq 0.023$). For EQIP (binary adaptation), only ChatGPT versus DeepSeek was nonsignificant ($P = 0.666$). For JAMA, ChatGPT scored higher than each of the other three chatbots (all adjusted $P < 0.001$), with no differences among Gemini, DeepSeek, and Doubao. GQS differences were limited to ChatGPT versus Doubao and Gemini versus Doubao (both $P = 0.015$). Complete estimates are provided in Supplementary File 5, Tables S2 and S4.

Readability and response length

All six readability indices differed across interfaces (all $P \leq 0.032$; Kendall's $W = 0.059$–0.559; Table 3). DeepSeek had the lowest ARI (12.35 ± 2.82), whereas Gemini had the lowest CLI (12.39 ± 1.64), FKGL (11.80 ± 1.87), GFI (14.64 ± 2.14), and SMOG (13.39 ± 1.64), and the highest FRES (47.19 ±

10.37). All interface means remained outside commonly recommended readability thresholds. Word count also differed ($P < 0.001$; Kendall's $W = 0.477$, 95% CI 0.348–0.612): ChatGPT responses were shortest (239.22 ± 131.15 words), followed by Doubao (296.62 ± 147.49), Gemini (415.12 ± 105.61), and DeepSeek (439.02 ± 230.96). The pooled correlation between word count and EQIP was weak and its question-cluster bootstrap CI included zero (Spearman's $\rho = 0.148$, 95% CI − 0.010 to 0.302); word count was not associated with FKGL ($\rho = -0.069$, 95% CI − 0.234 to 0.086). Figure 4 shows the readability distributions.

After adjustment, significant pairwise contrasts numbered 4/6 for ARI, 6/6 for CLI, 4/6 for FKGL, 2/6 for GFI, 2/6 for SMOG, and 5/6 for FRES. The only nonsignificant FRES contrast was ChatGPT versus Doubao ($P = 0.199$); complete paired estimates are provided in Supplementary File 5, Tables S2 and S4.

Discussion

This study compared four consumer-facing chatbot interfaces across 50 matched, patient-centered questions on frozen shoulder. No interface was uniformly superior across all evaluated outcomes. These findings support only a supplementary educational role. ChatGPT had the highest observed median accuracy and DISCERN scores and the lowest observed proportion of responses classified as potentially unsafe, whereas Gemini had the highest EQIP (binary adaptation) score and the most favorable overall readability profile. No statistically significant between-interface difference in potentially unsafe response proportions was detected.

Therefore, the observed numerical ordering should not be interpreted as evidence that one interface was safer than another. Inter-evaluator agreement was lower for accuracy and empathy than for safety and the structured quality measures, and all interfaces produced text above recommended reading levels. These findings do not support the use of the evaluated interfaces as substitutes for individualized clinical guidance and are more consistent with a cautious adjunctive role [7,8,20].

The outcome-specific patterns also show that safety, accuracy, empathy, information quality, transparency, and readability are not interchangeable. A response may be fluent and supportive yet clinically incomplete, or accurate and comprehensive yet too complex for many patients. Multidimensional evaluation is therefore more informative than a single composite score.

Recent studies in rotator cuff injury and hypertension similarly show that model performance depends on the outcome and evaluator perspective and that polished responses may remain limited in actionability or clinical suitability [36, 37]. The present study complements those benchmarks by using a frozen-shoulder-specific evidence map, five independent evaluators, explicit safety adjudication, and question-matched pairwise analyses across clinical and communication outcomes.

Responses classified as potentially unsafe were identified for all four interfaces, with observed proportions ranging from 6.0% to 14.0%. Neither the overall comparison nor any adjusted pairwise comparison was statistically significant. This range numerically overlapped with the 5%–13% reported in prior physician-led evaluations of publicly available chatbots [9], although direct comparison is limited by differences in prompts, clinical domains, and definitions of harm. Even the lowest observed proportion could be clinically relevant; however, these sample estimates should not be extrapolated directly to real-world use. Because no equivalence margin was prespecified and the confidence intervals and sample size did not support an equivalence interpretation, the nonsignificant findings should not be interpreted as evidence of equivalent safety across interfaces.

Potential risks arose not only from false statements but also from answer-question mismatches, excessive certainty, rigid recovery timelines, insufficiently individualized medication or rehabilitation advice, overstatement of low-certainty benefits, and failure to foreground injury- or procedure-specific restrictions. These shortcomings are relevant to frozen shoulder because management depends on diagnostic certainty, symptom irritability, comorbidities, and postoperative context [1,2]. Future systems should incorporate red-flag detection, contraindication screening, calibrated uncertainty, and retrieval from verified clinical guidance before providing patient-directed recommendations.

Accuracy differed significantly across interfaces, and all six adjusted pairwise contrasts were significant. This result is consistent with reports that LLM answers may be broadly correct yet omit important qualifications, add unsupported details, or require specialist clarification [8,18–20,38]. In shoulder-arthroplasty studies, most ChatGPT answers required professional clarification, and some relied on outdated or out-of-scope information [18, 19].

Observed accuracy differences may reflect interface behavior, evidence coverage, instruction following, response calibration, subscription conditions, provider-side routing, or other unrecorded settings. Because commercial systems are frequently updated, the ranking should not be considered permanent or attributed to displayed model identifiers alone. Linking responses to curated evidence, explicitly distinguishing uncertainty, and checking recommendations against current guidelines may improve accuracy. Clinician review remains essential for differential diagnosis, medication precautions, procedural indications, and postoperative rehabilitation.

Empathy also differed across interfaces. ChatGPT and Gemini had equal medians and did not differ significantly after adjustment; both scored higher than DeepSeek and Doubao in the relevant pairwise comparisons. DeepSeek, in turn, scored higher than Doubao. These differences may

reflect more consistent acknowledgment of pain, sleep disturbance, anxiety, and functional limitation. Previous text-only studies have similarly reported high empathy ratings for chatbot responses [5,6,39].

However, clinician-rated empathy in text-only responses is a proxy measure and does not capture patient-perceived empathy, nonverbal communication, longitudinal understanding, or a therapeutic relationship. A supportive response may also increase patient confidence in advice that is incomplete or unsafe. Empathic language should therefore be paired with calibrated uncertainty, concise safety netting, and transparent limitations. Future studies should directly assess patient perceptions of empathy, trust, comprehension, and intended behavior.

The differing patterns across DISCERN, EQIP (binary adaptation), JAMA, and GQS reinforce that reliability, completeness, transparency, and overall usefulness are related yet distinct. ChatGPT had the highest median DISCERN and JAMA scores, Gemini had the highest EQIP (binary adaptation) score, and all interfaces had the same median GQS. These results also illustrate that a significant omnibus test does not necessarily imply separation of every interface pair [11, 12].

The generally favorable DISCERN and EQIP (binary adaptation) scores suggest that the interfaces often produced organized responses covering symptoms, management options, risks, and practical advice. Nevertheless, a favorable score does not establish that an answer is clinically complete or suitable for independent decision-making. Previous orthopaedic studies have similarly reported fair-to-good DISCERN results alongside unsupported elaboration, omissions, and a need for professional clarification [18–20,38].

DISCERN emphasizes treatment reliability, including benefits, risks, alternatives, uncertainty, and shared decision-making, whereas EQIP (binary adaptation) covers a broader range of content, structure, clarity, and actionability criteria [11, 12]. Chatbot interfaces should therefore prioritize question-specific completeness rather than length. Essential information and alternatives should be clearly distinguished, and uncertainty about benefits and harms should be stated explicitly. Retrieval from curated guidelines may improve reliability, but citations must be checked for relevance, currency, and fidelity.

JAMA scores indicated a more pronounced limitation in transparency. Responses could appear coherent and useful while omitting authorship, attribution, disclosure, or currency information [13,14]. Similar deficits have been reported in orthopaedic evaluations of chatbot responses [18–20]. The clustering of GQS values around 4.00 also suggests a possible ceiling effect, supporting the use of more granular measures of actionability, source fidelity, uncertainty communication, and patient comprehension.

The comparatively low JAMA benchmark scores should be interpreted cautiously because these criteria were originally developed for static web-based health information rather than individual conversational responses [13]. Although chatbot interfaces do not routinely repeat provider identity, ownership, evidence dates, or sources in every response, users still need sufficient information to assess the origin and currency of medical claims. Standardized metadata could improve transparency and JAMA scores, although its clinical value should be evaluated through user testing. Future interfaces should display the responsible model and provider, query date, evidence date, and source links while clearly distinguishing AI-generated information from professional medical advice.

Readability differed across interfaces, and Gemini had the most favorable overall profile. Formula-specific rankings varied, indicating that no single readability index should be treated as definitive. None of the interfaces consistently achieved the reading level commonly recommended for public-facing health information.

These findings are consistent with evidence that online health information frequently exceeds the sixth- to eighth-grade reading level recommended for general audiences [15]. A meta-narrative systematic review of 157 studies reported average online-health-information reading levels of approximately grades 10–15 [15]. Orthopaedic studies have identified similar limitations in LLM-generated patient education. Marder et al. found that ChatGPT-3.5 and ChatGPT-4 did not reliably generate orthopaedic educational materials at an appropriate reading level without targeted prompting [16]. Yeramosu et al. likewise reported that responses concerning shoulder arthroplasty were written at approximately a college-graduate level [38].

Response length differed substantially among interfaces, but it did not show a robust pooled association with EQIP or FKGL after question-cluster bootstrapping. Because these pooled correlations combined within-interface and between-interface variation, they should not be interpreted as interface-specific associations. Patient-facing systems may need to balance concise presentation against inclusion of essential risks, alternatives, and safety-netting information.

Patient-facing systems may benefit from layered communication: an initial plain-language summary, concise next steps, and optional details about evidence, risks, and alternatives. Medical terms should be defined, sentences shortened, and actionable advice separated from background explanation. Targeted prompting has improved the readability of chatbot-generated orthopaedic materials, although human clinical review remained necessary [17]. Readability formulas should not replace clinician or patient testing because grade-level estimates do not directly measure comprehension, numeracy, cultural appropriateness, or safe action.

This study had several strengths, including a prespecified and clinically diverse question set, a question-specific evidence-based reference standard, independent evaluation by five clinicians, a matched four-chatbot design, and multidimensional assessment of safety, accuracy,

communication quality, and readability. Using first responses from separate chat sessions reduced selective regeneration and conversational carryover.

Several limitations should be acknowledged. First, this single-date, cross-sectional evaluation represents only the interfaces and configurations available under the recorded conditions; unrecorded account-level settings and provider-side routing further limit reproducibility. Second, patients were not involved in question development or evaluation, and clinician ratings may not reflect patient perceptions, comprehension, trust, or intended behavior. Third, only four interfaces, 50 English-language questions, and one musculoskeletal condition were evaluated. Single first responses also precluded assessment of within-interface generation variability, and the safety analysis was not designed to establish equivalence. Fourth, although model identifiers and interface-specific, nonsemantic symbols were removed from responses before evaluation, distinctive linguistic or stylistic patterns may have allowed evaluators to infer the source interface; therefore, complete blinding cannot be assured and expectation bias remains possible. Clinical judgment was also required, with lower single-evaluator agreement for accuracy and empathy. Finally, several instruments were adapted from static health-information contexts, and readability formulas do not directly measure comprehension, actionability, health literacy, or cultural appropriateness. Future studies should incorporate repeated and multilingual querying, direct patient evaluation, and prospective assessment of behavioral and clinical effects.

Patient-facing generative AI chatbots may help clinicians reinforce education, explain common treatment options, and support question preparation between visits, but they should not be used as autonomous diagnostic or treatment systems. Any future clinical deployment would require robust red-flag screening, transparent source attribution, dated evidence, calibrated uncertainty, plain-language output, and clear escalation pathways to professional care. Future studies should repeat queries over time, evaluate multilingual and multi-turn interactions, test retrieval-grounded systems, assess patient comprehension and behavior, and report model and prompt metadata in accordance with CHART [21].

Across this single-date set of 50 matched patient-facing questions, the four chatbot interfaces differed in observed accuracy, empathy, information quality, response-level transparency, response length, and readability. No statistically significant difference in the proportion of responses classified as potentially unsafe was detected. ChatGPT had the highest observed median accuracy and DISCERN scores, whereas Gemini had the highest EQIP (binary adaptation) score and most favorable readability profile. However, no interface was uniformly superior, each generated potentially unsafe content, and all produced text above recommended reading levels. These findings do not support use of the evaluated interfaces as substitutes for individualized clinical assessment and guidance.

Data availability

All data generated or analyzed during this study are included in this article and its additional files. Supplementary File 2 contains the question-pool screening log and final 50-question set; Supplementary File 3 contains the question-specific reference standard and evidence mapping; Supplementary File 4 contains the structured archive of 200 chatbot responses; Supplementary Files 5 and 6 contain the additional quantitative and qualitative analyses; Supplementary File 7 contains the response-level ratings, readability dataset, and appended question–answer records; and Supplementary Files 8 and 9 contain the analysis and figure-generation code. Backend generation parameters not exposed by the consumer interfaces were unavailable to the investigators.

Declarations

Competing interests

The authors declare that they have no competing interests.

Supplementary information

Supplementary File 1 (.docx). Completed CHART checklist.

Supplementary File 2 (.docx). Frozen shoulder public question-pool screening log, eligibility criteria, screening decisions, source list, and final 50-question set.

Supplementary File 3 (.docx). Question-specific reference standard and evidence mapping table.

Supplementary File 4 (.docx). Structured archive of all 200 interface–question responses.

Supplementary File 5 (.docx). Complete pairwise comparisons and exploratory post hoc sensitivity calculation (Tables S1–S4).

Supplementary File 6 (.docx). Qualitative analysis of all 21 responses classified as potentially unsafe (Tables S5 and S6).

Supplementary File 7 (.xlsx). Response-level evaluator scores, derived scores, safety classifications, readability metrics, and appended question–answer records.

Supplementary File 8 (.R). Readability analysis and figure-generation code.

Supplementary File 9 (.R). Global statistical analysis and figure-generation code.

Funding

No funding was received for this study.

Author Contribution

Conceptualization: Q.Z., K.C., and H.C.; methodology: Q.Z., K.C., and H.C.; investigation: Q.Z., K.C., D.Y., H.H., J.S., and H.C.; data curation: Q.Z.; formal analysis: Q.Z.; software: Q.Z.; validation: K.C. and H.C.; visualization: Q.Z.; writing—original draft: Q.Z.; writing—review and editing: all authors; supervision: H.C. H.C. is the guarantor of this work. All authors read and approved the final manuscript and agree to be accountable for the work.

Acknowledgement

The authors thank the five clinical evaluators who participated in the response assessment and consented to anonymous acknowledgment.During the preparation of this manuscript, OpenAI's ChatGPT and Codex were used after completion of the study analyses to assist with language editing and formatting. The authors take full responsibility for the final content, accuracy, and integrity of the manuscript.

Data Availability

All data generated or analyzed during this study are included in this article and its additional files. Supplementary File 2 contains the question-pool screening log and final 50-question set; Supplementary File 3 contains the question-specific reference standard and evidence mapping; Supplementary File 4 contains the structured archive of 200 chatbot responses; Supplementary Files 5 and 6 contain the additional quantitative and qualitative analyses; Supplementary File 7 contains the response-level ratings, readability dataset, and appended question–answer records; and Supplementary Files 8 and 9 contain the analysis and figure-generation code. Backend generation parameters not exposed by the consumer interfaces were unavailable to the investigators.

References

1. Leafblad, N., Mizels, J., Tashjian, R. Z. & Chalmers, P. Adhesive capsulitis. *Phys. Med. Rehabil Clin. N Am.* **34**, 453–468. https://doi.org/10.1016/j.pmr.2022.12.009 (2023).
2. Lee, B. C. et al. Clinical practice guidelines for diagnosis and non-surgical treatment of primary frozen shoulder. *Ann. Rehabil Med.* **49**, 113–138. https://doi.org/10.5535/arm.250057 (2025).
3. Dyer, B. P., Rathod-Mistry, T., Burton, C., van der Windt, D. & Bucknall, M. Diabetes as a risk factor for the onset of frozen shoulder: a systematic review and meta-analysis. *BMJ Open.* **13**, e062377. https://doi.org/10.1136/bmjopen-2022-062377 (2023).
4. Dyer, B. P., Burton, C., Rathod-Mistry, T., Blagojevic-Bucknall, M. & van der Windt, D. A. Diabetes as a prognostic factor in frozen shoulder: a systematic review. *Arch. Rehabil Res. Clin. Transl*. **3**, 100141. https://doi.org/10.1016/j.arrct.2021.100141 (2021).
5. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596. https://doi.org/10.1001/jamainternmed.2023.1838 (2023).
6. Chen, D. et al. Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *NPJ Digit. Med.* **8**, 275. https://doi.org/10.1038/s41746-025-01671-6 (2025).
7. Busch, F. et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun. Med. (Lond)*. **5**, 26. https://doi.org/10.1038/s43856-024-00717-2 (2025).
8. Jacobs, M. M. G., Oosterhoff, J. H. F. O., Agricola, R. & van der Weegen, W. Large language models versus healthcare professionals in providing medical information to patient questions: a systematic review. *Int. J. Med. Inf.* **209**, 106250. https://doi.org/10.1016/j.ijmedinf.2025.106250 (2026).
9. Draelos, R. L. et al. Large language models provide unsafe answers to patient-posed medical questions. *NPJ Digit. Med.* **9**, 241. https://doi.org/10.1038/s41746-026-02428-5 (2026).

10. Armbruster, J. et al. Doctor ChatGPT, can you help me? The patient's perspective: cross-sectional study. *J. Med. Internet Res.* **26**, e58831. https://doi.org/10.2196/58831 (2024).
11. Charnock, D., Shepperd, S., Needham, G. & Gann, R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J. Epidemiol. Community Health*. **53**, 105–111. https://doi.org/10.1136/jech.53.2.105 (1999).
12. Moult, B., Franck, L. S. & 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.* **7**, 165–175. https://doi.org/10.1111/j.1369-7625.2004.00273.x (2004).
13. Silberg, W. M., Lundberg, G. D. & Musacchio, R. A. Assessing, controlling, and assuring the quality of medical information on the Internet: caveant lector et viewor—let the reader and viewer beware. *JAMA* **277**, 1244–1245. https://doi.org/10.1001/jama.1997.03540390074039 (1997).
14. Bernard, A. et al. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. *Am. J. Gastroenterol.* **102**, 2070–2077. https://doi.org/10.1111/j.1572-0241.2007.01325.x (2007).
15. Daraz, L. et al. Readability of online health information: a meta-narrative systematic review. *Am. J. Med. Qual.* **33**, 487–492. https://doi.org/10.1177/1062860617751639 (2018).
16. Marder, R. S. et al. ChatGPT-3.5 and – 4.0 do not reliably create readable patient education materials for common orthopaedic upper- and lower-extremity conditions. *Arthrosc. Sports Med. Rehabil*. **7**, 101027. https://doi.org/10.1016/j.asmr.2024.101027 (2025).
17. Chandra, K. et al. Improving readability of shoulder and elbow surgery online patient education material with ChatGPT-4. *J. Shoulder Elb. Surg.* **34**, e1119–e1124. https://doi.org/10.1016/j.jse.2025.02.025 (2025).
18. Johnson, C. K., Mandalia, K., Corban, J., Beall, K. E. & Shah, S. S. Adequacy of ChatGPT responses to frequently asked questions about shoulder arthroplasty: is it an appropriate adjunct for patient education? *JSES Int.* **9** (3), 830–836. https://doi.org/10.1016/j.jseint.2025.01.008 (2025).
19. White, C. A. et al. Can ChatGPT reliably answer the most common patient questions regarding total shoulder arthroplasty? *J. Shoulder Elb. Surg.* **34** (5), e254–e264. https://doi.org/10.1016/j.jse.2024.08.025 (2025).
20. Dworsky-Fried, J., Mellon, M., Rathod, P., Yan, J. & Khan, M. Artificial intelligence models cannot yet replace experts in providing patient education for shoulder and elbow orthopaedic pathologies: a systematic review and meta-analysis. *Ann. Jt.* **11**, 6. https://doi.org/10.21037/aoj-25-38 (2026).
21. CHART Collaborative. Reporting guideline for chatbot health advice studies: the Chatbot Assessment Reporting Tool (CHART) statement. *Ann. Fam Med.* **23**, 389–398. https://doi.org/10.1370/afm.250386 (2025).
22. Google Trends. https://trends.google.com/explore?geo=US (accessed 1. May (2026).
23. Baidu Tieba. https://tieba.baidu.com/f?kw=%E8%82%A9%E5%91%A8%E7%82%8E (accessed 1. May (2026).
24. DingXiangYuan & (DXY).. https://search.dxy.cn/?words=%E8%82%A9%E5%91%A8%E7%82%8E (accessed 1 May 2026).
25. Rupani, N., Gwilym, S. E. & on behalf of the BESS Frozen Shoulder Working Group British. Elbow and Shoulder Society patient care pathway: frozen shoulder. *Shoulder Elb.* **17**, e17585732251335955. https://doi.org/10.1177/17585732251335955 (2025).
26. Kelley, M. J. et al. Shoulder pain and mobility deficits: adhesive capsulitis. *J. Orthop. Sports Phys. Ther.* **43**, A1–A31. https://doi.org/10.2519/jospt.2013.0302 (2013).
27. Nacey, N. et al. ACR Appropriateness Criteria® chronic shoulder pain: 2022 update. *J. Am. Coll. Radiol.* **20**, S49–S69. https://doi.org/10.1016/j.jacr.2023.02.017 (2023).
28. Wang, X. et al. Benchmarking public large language model responses to patient-facing inflammatory bowel disease questions: informational quality, transparency proxies, and readability. *Front. Public. Health*. **14**, 1810358. https://doi.org/10.3389/fpubh.2026.1810358 (2026).
29. Smith, E. A. & Senter, R. J. *Automated Readability Index, AMRL-TR-66-220, 1–14* (Aerospace Medical Research Laboratories, 1967).
30. Coleman, M. & Liau, T. L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **60**, 283–284. https://doi.org/10.1037/h0076540 (1975).
31. Kincaid, J. P. et al. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula). *Res. Branch. Rep.* 8–75 (Naval Technical Training Command, 1975). https://doi.org/10.21236/ADA006655
32. Gunning, R. *The Technique of Clear Writing* (McGraw-Hill, 1952).
33. McLaughlin, G. H. SMOG grading: a new readability formula. *J. Read.* **12**, 639–646 (1969).
34. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **32**, 221–233. https://doi.org/10.1037/h0057532 (1948).
35. Benoit, K. et al. quanteda: an R package for the quantitative analysis of textual data. *J. Open. Source Softw.* **3**, 774. https://doi.org/10.21105/joss.00774 (2018).
36. Wang, Y. L. et al. Evaluation of large language models in patient education and clinical decision support for rotator cuff injury: a two-phase benchmarking study. *BMC Med. Inf. Decis. Mak.* **25**, 289. https://doi.org/10.1186/s12911-025-03105-5 (2025).
37. Hu, M., Wang, Z., Zhang, Z. & Li, M. Challenges of patient-facing generative artificial intelligence in hypertension care: a cross-platform evaluation of the quality, readability, and actionability of LLM-generated patient education materials. *Digit. Health*. **12**, 20552076261464724.

https://doi.org/10.1177/20552076261464724 (2026).
38. Yeramosu, T. et al. ChatGPT is capable of providing satisfactory responses to frequently asked questions regarding total shoulder arthroplasty. *Shoulder Elb.* **16**, 407–412. https://doi.org/10.1177/17585732241246560 (2024).
39. Howcroft, A. et al. AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care. *Br. Med. Bull.* **156**, ldaf017. https://doi.org/10.1093/bmb/ldaf017 (2025).

Figures

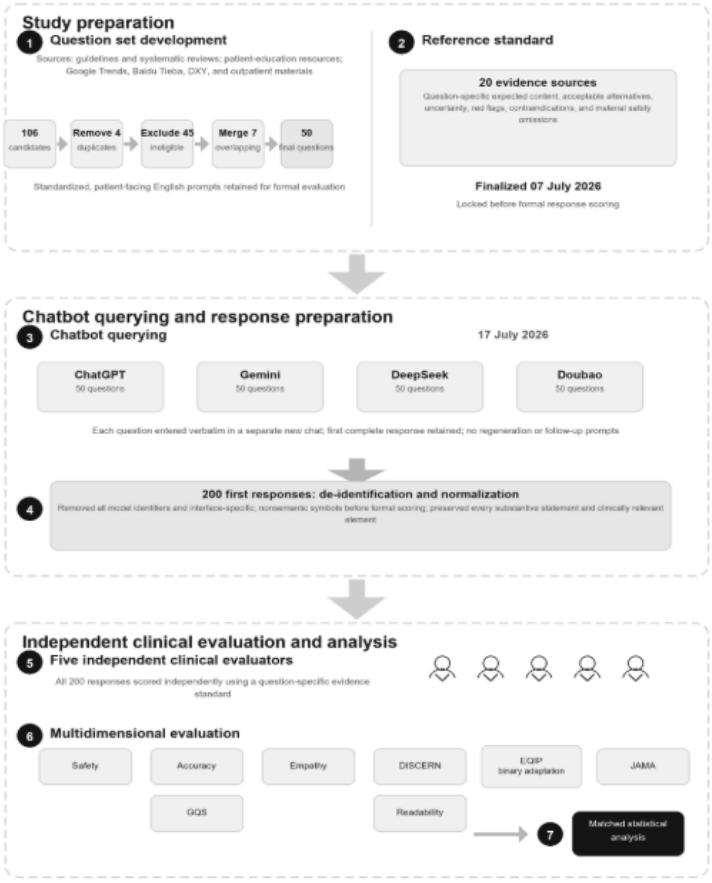


Figure 1

Study workflow. The final 50-question set was linked to a question-specific evidence standard, submitted once to each of four consumer chatbot interfaces, deidentified and normalized, independently evaluated by five clinicians, and analyzed using question-matched methods. DXY, Dingxiangyuan; EQIP, Ensuring Quality Information for Patients; GQS, Global Quality Score; JAMA, Journal of the American Medical Association.



Figure 2

Safety, accuracy, and empathy by chatbot interface. Stacked bars show response-level safety classifications; boxplots show accuracy and empathy scores for the 50 matched questions. Diamonds indicate means. Product names refer to the evaluated consumer interfaces rather than independently verified backend models.

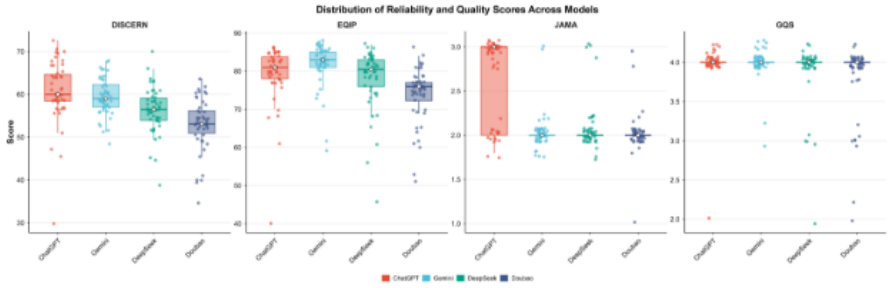


Figure 3

Information reliability and quality by chatbot interface. Boxplots show response-level DISCERN, EQIP binary-adaptation, JAMA, and GQS scores for the 50 matched questions. Diamonds indicate means. Model labels denote the evaluated consumer interfaces rather than independently verified backend models.

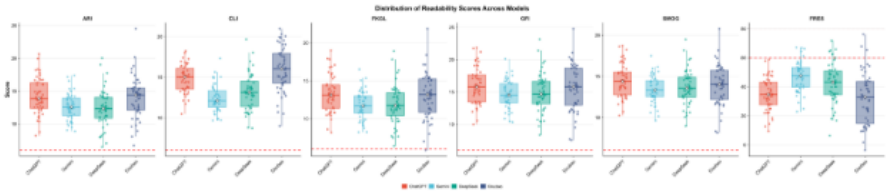


Figure 4

Readability by chatbot interface. Boxplots show ARI, CLI, FKGL, GFI, SMOG, and FRES values for the 50 matched questions. Diamonds indicate means. Model labels denote the evaluated consumer interfaces rather than independently verified backend models.

Supplementary Files

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

- Supplementary2.docx
- Supplementary3.docx
- Supplementary4.docx
- Supplementary5.docx
- Supplementary6.docx
- Supplementary7.xlsx
- Supplementary8.r
- Supplementary9.r
- Supplementary1.docx