Original Paper

# Evaluating Large Language Models for Preoperative Patient Education in Superior Capsular Reconstruction: Comparative Study of Claude, GPT, and Gemini

Yukang Liu[1*], MBBS; Hua Li[2*], PhD; Jianfeng Ouyang[3*], PhD; Zhaowen Xue[4], PhD; Min Wang[5], PhD; Hebei He[4], PhD; Bin Song[6], PhD; Xiaofei Zheng[4], PhD; Wenyi Gan[3*], PhD

[1]The Second School of Clinical Medicine, Southern Medical University, Guangzhou, China
[2]Department of Orthopedics, Beijing Jishuitan Hospital, Beijing, China
[3]Zhuhai People's Hospital (The Affiliated Hospital of Beijing Institute of Technology, Zhuhai Clinical Medical College of Jinan University), Zhuhai, Guangdong, China
[4]Department of Sports Medicine, The First Affiliated Hospital of Jinan University, Guangzhou, China
[5]Department of Orthopaedics, Guangzhou Red Cross Hospital of Jinan University, Guangzhou, China
[6]Department of Joint Surgery and Sports Medicine, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
*these authors contributed equally

**Corresponding Author:**

Wenyi Gan, PhD
Zhuhai People's Hospital (The Affiliated Hospital of Beijing Institute of Technology, Zhuhai Clinical Medical College of Jinan University)
79 Kangning Road, Xiangzhou District
Zhuhai, Guangdong, 519000
China
Phone: 86 13076855735
Email: 494414224@qq.com

## Abstract

**Background:** Large language models (LLMs) are revolutionizing natural language processing, increasingly applied in clinical settings to enhance preoperative patient education.

**Objective:** This study aimed to evaluate the effectiveness and applicability of various LLMs in preoperative patient education by analyzing their responses to superior capsular reconstruction (SCR)–related inquiries.

**Methods:** In total, 10 sports medicine clinical experts formulated 11 SCR issues and developed preoperative patient education strategies during a webinar, inputting 12 text commands into Claude-3-Opus (Anthropic), GPT-4-Turbo (OpenAI), and Gemini-1.5-Pro (Google DeepMind). A total of 3 experts assessed the language models' responses for correctness, completeness, logic, potential harm, and overall satisfaction, while preoperative education documents were evaluated using DISCERN questionnaire and Patient Education Materials Assessment Tool instruments, and reviewed by 5 postoperative patients for readability and educational value; readability of all responses was also analyzed using the cntext package and py-readability-metrics.

**Results:** Between July 1 and August 17, 2024, sports medicine experts and patients evaluated 33 responses and 3 preoperative patient education documents generated by 3 language models regarding SCR surgery. For the 11 query responses, clinicians rated Gemini significantly higher than Claude in all categories ($P$<.05) and higher than GPT in completeness, risk avoidance, and overall rating ($P$<.05). For the 3 educational documents, Gemini's Patient Education Materials Assessment Tool score significantly exceeded Claude's ($P$=.03), and patients rated Gemini's materials superior in all aspects, with significant differences in educational quality versus Claude ($P$=.02) and overall satisfaction versus both Claude ($P$<.01) and GPT ($P$=.01). GPT had significantly higher readability than Claude on 3 R-based metrics ($P$<.01). Interrater agreement was high among clinicians and fair among patients.

**Conclusions:** Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro effectively generated readable presurgical education materials but lacked citations and failed to discuss alternative treatments or the risks of forgoing SCR surgery, highlighting the need for expert oversight when using these LLMs in patient education.

*JMIR Perioper Med 2025;8:e70047;* doi: 10.2196/70047

**Keywords:** superior capsular reconstruction; massive rotator cuff tear; large language models; preoperative patient education; informed consent process

# Introduction

Large language models (LLMs) are extensive neural network models based on deep learning [1,2]. These models learn the grammar, semantics, and contextual information of a language by training on vast amounts of textual data, enabling them to perform various natural language processing tasks [1,2]. Due to the powerful text processing, text generation capabilities, and immense knowledge training of LLMs, researchers have begun to continually explore the potential of LLMs in clinical application scenarios, including professional licensing examinations in various countries and regions [3-5], answering public health questions [6,7], analyzing radiological images [8], disease screening [9], disease diagnosis [10], and discipline education [11]. As the versions and functions of LLMs are constantly updated and upgraded, these models have a low usage threshold and are convenient to use. It is particularly important for professionals in various disciplines to assess the accuracy and completeness of LLMs in their respective fields. This assessment not only provides a strong basis for the application of LLMs in various disciplines but also identifies their shortcomings, serving as a warning for nonprofessional users [3,8,10,11].

Superior capsular reconstruction (SCR) was initially proposed by Mihata et al [12] in 2012 as a technique to restore the superior restraint of the humeral head passively, thereby restoring force couples and improving shoulder joint kinematics. Over the past decade, SCR has become one of the commonly used treatment methods for massive and irreparable rotator cuff tears among clinicians [13,14]. However, the surgical techniques for SCR are highly variable [15]. For example, contrary to the results of earlier studies, further research suggests using dermal allograft instead of fascia lata autograft, leading to a current lack of sufficiently effective long-term follow-up data with high levels of evidence [16-18]. Moreover, as SCR is a reconstructive surgery rather than a repair surgery [15], it is challenging to provide patients with a standardized and effective explanation and communication during the preoperative informed consent process. An effective preoperative informed consent process is one of the essential steps in alleviating patients' perioperative anxiety and improving treatment efficacy [19,20].

Rational and effective preoperative patient education is one of the critical components in developing standardized diagnosis and treatment processes for clinical surgery departments [21]. The main difficulty lies in the professional knowledge gap between medical staff and patients [22]. Previous studies have shown that using multimedia as patient education materials can better help patients understand surgical procedures and alleviate perioperative anxiety [23,24]. However, in most cases, doctors still primarily use verbal responses to address patients' individualized questions [25]. This might probably because preparing personalized educational materials and providing oral education requires a significant investment of time and effort, leading to high time and economic costs. Furthermore, there is a vast difference in the sources of medical information accessed by doctors and patients [26]. Doctors primarily obtain medical information from clinical guidelines, research literature, and textbooks, while patients often acquire medical information through simple search engines and social media software, which may contain false and overly embellished content [26-28]. Patients often lack the ability to think independently when faced with this information.

With the development of LLMs in recent years, researchers have discovered that the disciplinary knowledge possessed by these LLMs can pass professional examinations in multiple disciplines [3,10,29]. Their powerful text processing capabilities not only allow them to polish complex text content to enhance readability but also enable them to independently generate text content that is more comprehensive and empathetic compared to health care professionals [6,7,30]. The quality of their answers is also significantly better than the search results from search engines [27,28]. Researchers have also pointed out that when using LLMs as patient education assistive tools, the primary task of doctors is to determine the accuracy of the information and make necessary clarifications [5,31]. Furthermore, researchers believe that LLMs can present information in a way that is understandable to most patients, making them a valuable supplement for orthopedic surgeons in obtaining informed consent and shared decision-making [4,5].

This cross-sectional study aims to assess the capability and application potential of different LLMs in preoperative patient education by evaluating the responses of 3 LLMs—GPT-4-Turbo, Claude-3-Opus, and Gemini-1.5-Pro—to SCR-related patient inquiries. In addition, the study will evaluate patient education documents generated by the LLMs for the informed consent process, which will be jointly assessed by health care professionals and patients. We hypothesize that LLMs can generate readable patient education materials for SCR, but the accuracy, completeness, and patient-assessed readability of the content will require expert review before clinical application.

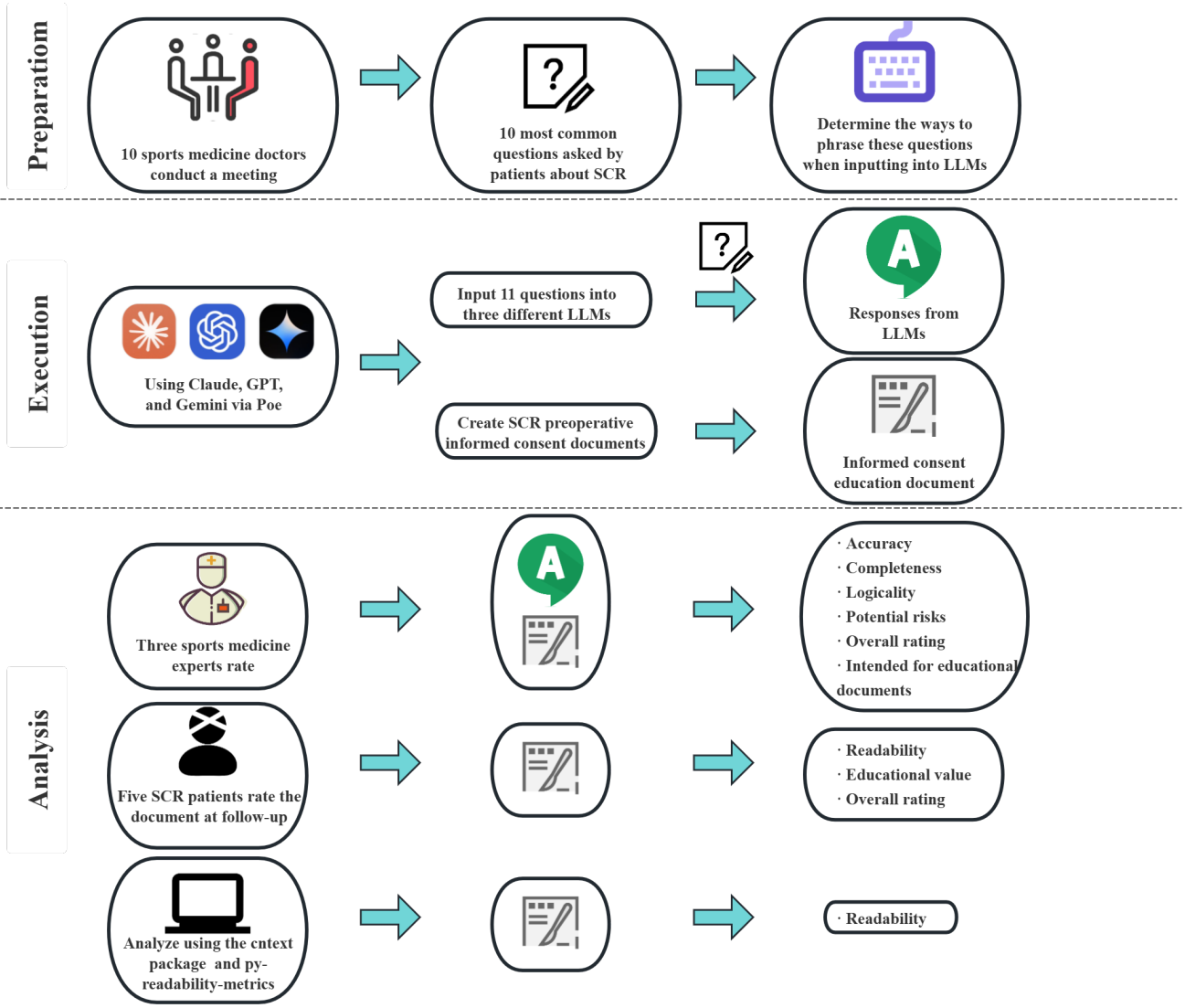
# Methods

## Study Design Overview

This cross-sectional analysis, conducted from July 1 to August 17, 2024, evaluated the quality of responses generated by different LLMs in the context of preoperative patient education for SCR. The study design assessed Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro (accessed via Poe) on their ability to answer SCR-related patient questions and generate educational materials. The specific study flow is shown in Figure 1. All LLM prompts and responses, as

well as expert and patient evaluations, were conducted in Chinese. Screenshots of Poe website operations are available in Mendeley (Mendeley Data, V1), with English translations generated by GPT-4-Turbo (via Poe) in Multimedia Appendix 1.

**Figure 1.** Flow diagram of the study process. LLM: large language model; SCR: superior capsular reconstruction.




## Ethical Considerations

This study was approved by the Ethics Committee of our organization and was eligible for exemption from ethical review considering that this cross-sectional study involved no interventions or potential risks to patients.

## Questions and Prompts Development

The research team for this study consists of 12 members, including 10 experienced sports medicine clinicians and 2 doctoral students specializing in LLMs, who collaborated to create patient education materials about SCR. The clinicians include 3 senior-level experts (2 of whom are subject matter experts from external institutions), 2 associate senior-level experts, and 5 intermediate-level experts, with each clinician having at least 5 years of clinical experience.

The 2 doctoral students first collected a total of 100 questions by having each of the 10 clinical experts propose 10 questions daily that patients frequently asked about SCR, covering aspects like etiology, treatment principles, methods, complications, rehabilitation, and hospitalization costs. After removing duplicates and combining some of the questions, they included only the effective questions that all experts agreed were meaningful. This process resulted in the inclusion of 11 questions. Along with these questions, the doctoral students provided instructions (Table 1) requiring LLMs to draft a standardized preoperative informed consent patient education document. After the drafted prompts were reviewed and approved by the aforementioned 10 clinical experts, doctoral students created standardized prompts for each question, consisting of unified "Background+ Question" formats (Table 1). These standardized prompts were then used to generate a comprehensive patient education document addressing most concerns of SCR patients using LLMs.

**Table 1.** Content and strategies for asking questions to large language models.

| Subject | Theme | Content |
|---|---|---|
| Background | Clinical case | The patient was diagnosed with a massive rotator cuff tear due to supraspinatus muscle injury. The doctor plans to perform a superior capsular reconstruction surgery on the shoulder joint. |
| Question 1 | Muscle injury | The imaging report says that I have a supraspinatus muscle injury. What is the supraspinatus muscle, and what causes this type of injury? |
| Question 2 | Surgical principles and indications | What is the reconstruction of the superior capsule of the shoulder joint, what is the therapeutic principle of the surgery, and what are the indications for the surgery? |
| Question 3 | Graft materials | What are the commonly used graft materials in the reconstruction of the superior capsule of the shoulder joint, and what are the differences between these grafts? |
| Question 4 | Surgical hardware | Besides grafts, does the reconstruction of the superior capsule of the shoulder joint require the use of screws, and do these screws need to be removed in a second surgery? |
| Question 5 | Surgical complications | What are the surgical complications of superior capsule reconstruction of the shoulder joint? |
| Question 6 | Recovery time | How long is the typical recovery time after superior capsule reconstruction surgery of the shoulder joint? |
| Question 7 | Healing issues | What situations can lead to poor healing or failure of the superior capsule reconstruction surgery of the shoulder joint? |
| Question 8 | Autograft risks | In superior capsule reconstruction surgery of the shoulder joint, if an autograft is chosen, what are the impacts and risks to the area from which the autologous tissue is harvested? |
| Question 9 | Surgical costs | What are the chargeable items during the superior capsule reconstruction surgery of the shoulder joint, and what surgical consumables are needed? |
| Question 10 | Graft longevity | If the superior capsule reconstruction surgery of the shoulder joint is successful, how long is the lifespan of the implanted graft, and what are the differences between different types of grafts? |
| Question 11 | Anesthesia and hospitalization | What type of anesthesia is required for superior capsule reconstruction surgery, how long does the surgery take, and how long is the hospital stay required? |
| Document generation request | Education document | Please generate a comprehensive educational document about superior capsule reconstruction surgery of the shoulder joint. This document is to be provided to patients for reading during the preoperative informed consent process. |

## LLM Selection and Prompt Execution

Both ChatGPT 4 and Claude 3 are among the most popular language models today, with Gemini (formerly known as Bard) also gaining significant traction [32]. Studies suggest potential discrepancies in the functionalities of GPT-4 models used on the OpenAI official website [33]. To mitigate potential systematic errors arising from these discrepancies, we access Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro through the Poe website. Poe, created by Anthropic, is a platform that aggregates multiple AI chatbots, enabling users to engage with different AI assistants within a single interface and compare their responses [34].

To ensure that each interaction is independent and unbiased by previous exchanges, the doctoral students perform a "clear context" operation after each query. This approach ensures that each question and response are treated independently, preventing information carryover from previous interactions, and is informed by other research [7,11]. Since the purpose of our study was to evaluate the ability of pretrained LLMs to handle new tasks, we used LLMs in Zero-shot mode. Before input, the generated content has no specific setting (ie, suppose you are a doctor or speak like a doctor). The input provided to the LLMs follows a "background+ question/request" format (human message) and the output answers (assistant message) were collected then, ensuring clarity and relevance within each independent interaction.

## Evaluation of LLM Response Quality

This study evaluates the quality of patient informed consent documents generated by LLMs from 3 perspectives: physicians' assessment, patients' assessment, and readability analysis.

In total, 3 senior doctors evaluated the LLMs' responses to 11 specific questions related to a specific medical procedure, assessing them for correctness, completeness, logic, and potential harm using a 5-point Likert scale [35]. Physicians also provided an overall satisfaction score using a 10-point Likert scale. In addition, to evaluate the quality of health care information provided by each LLM, 2 validated instruments were also used to assess the generated documents: DISCERN (score ranging from 1=low to 5=high for overall information quality) and the Patient Education Materials Assessment Tool (PEMAT) for printable materials (scores of 0%-100% for understandability) [6]. The PEMAT assessment tool was able to assess printable and audiovisual understandability, while the DISCERN instrument could review the quality of information for the consumer particularly with a focus on treatment choices in health information.

In total, 5 patients who underwent the specific medical procedure reviewed the LLM-generated patient education documents, rating their readability and educational value on a 5-point Likert scale and overall satisfaction on a 10-point Likert scale. This aimed to assess the documents' clarity and educational value from nonprofessional readers' perspectives.

Finally, a readability analysis of all LLMs' responses was conducted using the cntext package [36] in R (version 4.4.1), examining sentence structure and evaluating readability via 3 indices: readability 1 (average characters per clause), readability 2 (proportion of adverbs and conjunctions), and readability 3, based on the Fog Index and calculated as half the sum of readability 1 and readability 2. Besides, we also applied the "py-readability-metrics" to evaluate the readability, which includes metrics such as the Flesch Reading Ease Score, Flesch-Kincaid Grade Level, and Gunning Fog Index.

### Data Analysis

Statistical analysis used SPSS (version 26.0; IBM Corp) using nonparametric tests due to nonnormally distributed data (Kolmogorov-Smirnov test). Mann-Whitney *U* test compared scoring between groups, with significance at $P<.05$. Interrater reliability, assessed using Fleiss kappa value, was interpreted as follows: poor agreement (<0.01); slight agreement (0.01-0.20); fair agreement (0.21-0.40); moderate agreement (0.41-0.60); substantial agreement (0.61-0.80); almost perfect agreement (0.81-1.00) [7]. GraphPad Prism 8 generated bar charts for visualizing results.

## Results

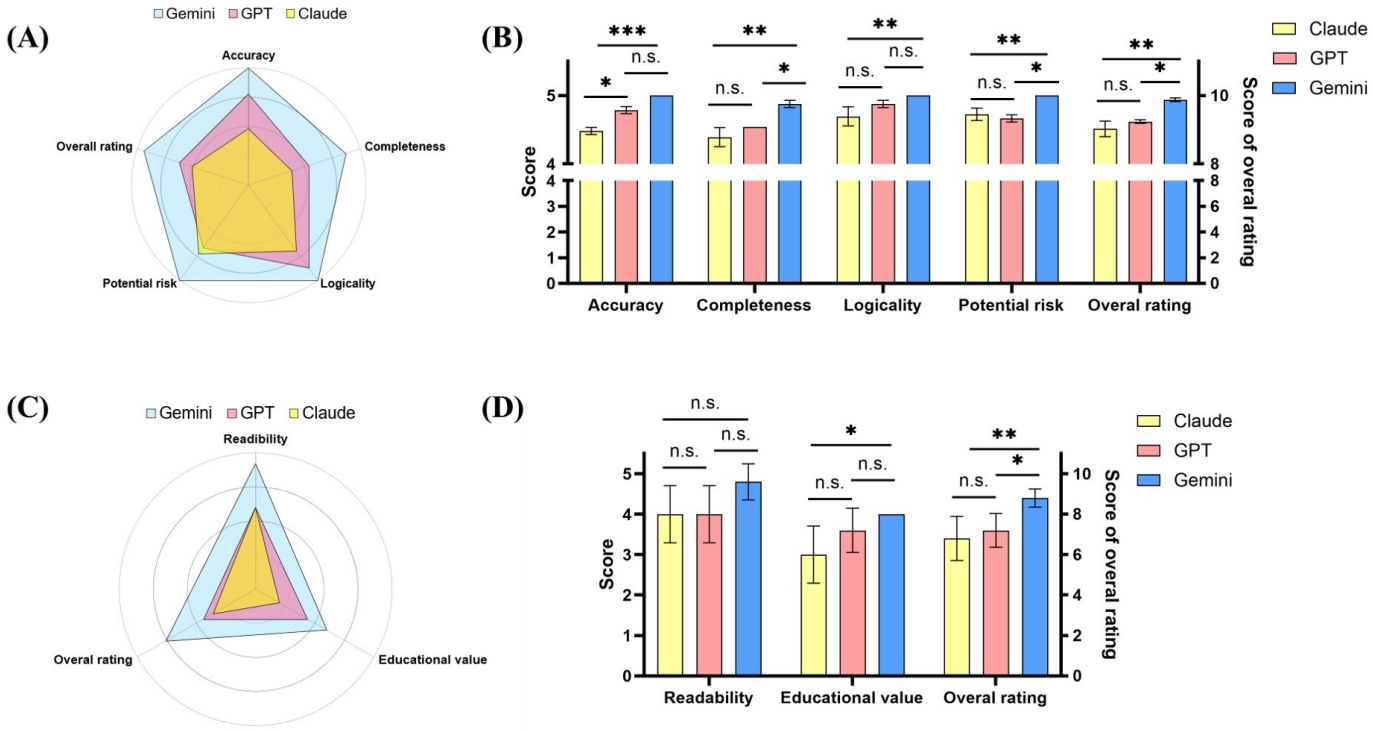
### Overview

Between July 1 and July 14, 2024, we sent invitations to sports medicine experts at various hospitals in the South China region for a webinar held on July 18. During this meeting, we discussed 11 key issues and formulated 12 strategies for sending inquiry requests as part of our study. From July 20 to August 1, 2024, we posed 11 surgery-related questions about SCR and requested the creation of preoperative patient education documents through the Poe website to 3 different LLMs: Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro. These models collectively produced 33 responses and 3 preoperative patient education documents. From August 10 to August 17, 2024, three experienced sports medicine clinicians, who are not from the same institution, along with 5 patients who had undergone SCR surgery, evaluated the responses and documents provided by the LLMs.

### Evaluations From the Subjective Perspective of Doctors

In total, 3 professional sports medicine doctors first evaluated the responses of 3 different LLMs to 11 inquiries. The evaluations focused on accuracy, completeness, logicality, potential risk, and overall rating. The results showed that Gemini's responses were significantly superior to Claude's in all evaluated categories including accuracy (mean 5.00, SD 0.00 vs mean 4.48, SD 0.83; $P<.001$), completeness (mean 4.88, SD 0.33 vs mean 4.39, SD 0.70; $P=.001$), logicality (mean 5.00, SD 0.00 vs mean 4.70, SD 0.59; $P<.01$) potential risk (mean 5.00, SD 0.00 vs mean 4.73, SD 0.57; $P<.01$), and overall rating (mean 9.88, SD 0.42 vs mean 9.03, SD 1.31; $P=.001$; Figures 2A and 2B). Compared to GPT, Gemini's responses were superior in all categories, with significant differences noted in completeness (mean 4.88, SD 0.33 vs mean 4.55, SD 0.67; $P=.02$), potential risk (mean 5.00, SD 0.00 vs mean 4.67, SD 0.82; $P=.01$), and overall rating (mean 9.88, SD 0.42 vs mean 9.24, SD 1.30; $P=.01$; Figures 2A and 2B. GPT's responses, when compared to Claude's, were superior in accuracy ($P=.03$), completeness ($P=.34$), logicality ($P=.11$), and overall rating ($P=.42$); however, Claude was rated higher in potential risk ($P=.85$; Figures 2A and 2B). Of these differences, only the accuracy presented a statistically significant difference (Figures 2A and 2B).

**Figure 2.** Quality evaluation results from doctors and patients for 11 questions generated by 3 large language models. (A-B) Evaluation from the doctor's perspective; (C-D) evaluation from the patient's perspective. n.s. not significant; *$P$<.05, **$P$<.01, ***$P$<.001.




In terms of the PEMAT scores for the preoperative patient education materials generated by each LLM, Gemini scored higher than GPT (mean 1.00, SD 0.00 vs mean 0.91, SD 0.09; $P$=.12), and GPT scored higher than Claude (mean 0.91, SD 0.09 vs mean 0.79, SD 0.10; $P$=.18), with only the difference between Gemini and Claude (mean 1.00, SD 0.00 vs mean 0.79, SD 0.10; $P$=.03) being statistically significant (Figure 3). Regarding the DISCERN scores, Claude achieved the highest overall score, followed by Gemini and then GPT, though these differences were not statistically significant (Table 2). In the item of the DISCERN which represents overall satisfaction (the 16th question presented in Table 2), Gemini scored the highest, while GPT and Claude scored the same, with no statistical significance in the differences. The consistency among the 3 evaluators was high, with no instances of "Poor agreement" or "Slight agreement" in their assessments (Multimedia Appendix 2).

**Figure 3.** PEMAT scoring percentage for the patient education document generated by three large language models. n.s.: not significant; *$P$<.05, **$P$<.01, ***$P$<.001.

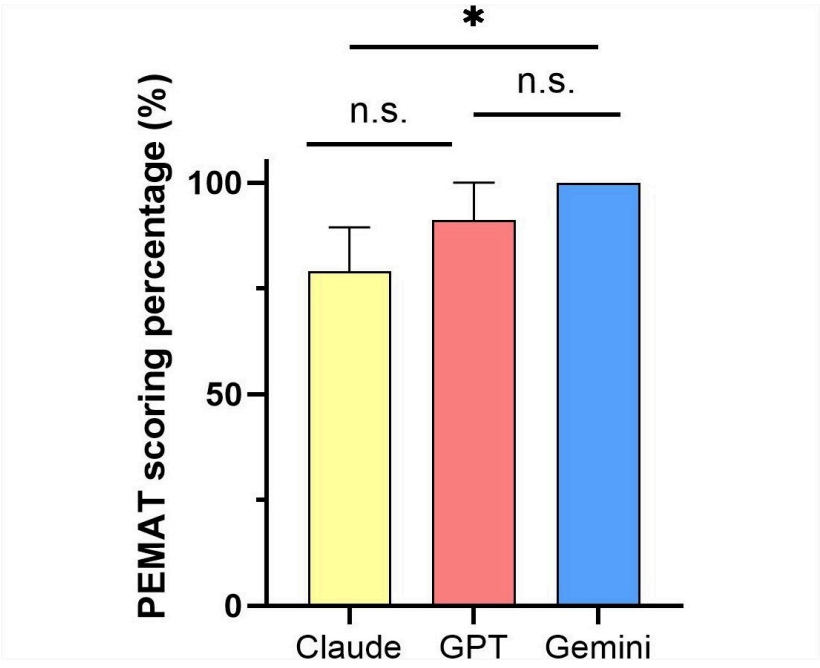


**Table 2.** Quality grades for section 2 of the DISCERN Tool.

| Section 2. How good is the quality of information on treatment choices ? | Claude-3-Opus, Median (IQR) | GPT-4-Turbo, Median (IQR) | Gemini-1.5-Pro, Median (IQR) | Claude versus GPT, *P* value | Claude versus Gemini, *P* value | GPT versus Gemini, *P* value |
|---|---|---|---|---|---|---|
| Does it describe how each treatment works? | 4 (3-4) | 4 (3-4) | 5 (4-5) | —[a] | .09 | .09 |
| Does it describe the benefits of each treatment? | 4 (3-5) | 4 (3-4) | 1 (1-1) | .64 | .04 | .03 |
| Does it describe the risks of each treatment? | 4 (3-4) | 3 (2-3) | 5 (4-5) | .09 | .09 | .04 |
| Does it describe what would happen if no treatment is used? | 1 (1-1) | 1 (1-1) | 1 (1-1) | — | — | — |
| Does it describe how the treatment choices affect overall quality of life? | 1 (1-1) | 1 (1-1) | 1 (1-1) | — | — | — |
| Is it clear that there may be more than one possible treatment choice? | 1 (1-1) | 1 (1-1) | 1 (1-1) | — | — | — |
| Does it provide support for shared decision-making? | 3 (3-4) | 3 (2-3) | 3 (2-3) | .32 | .20 | — |
| Based on the answers to all of the above questions, rate the overall quality of the publication as a source of information about treatment choices. | 3 (3-4) | 3 (3-4) | 4 (3-4) | — | .46 | .46 |

[a]Not applicable.

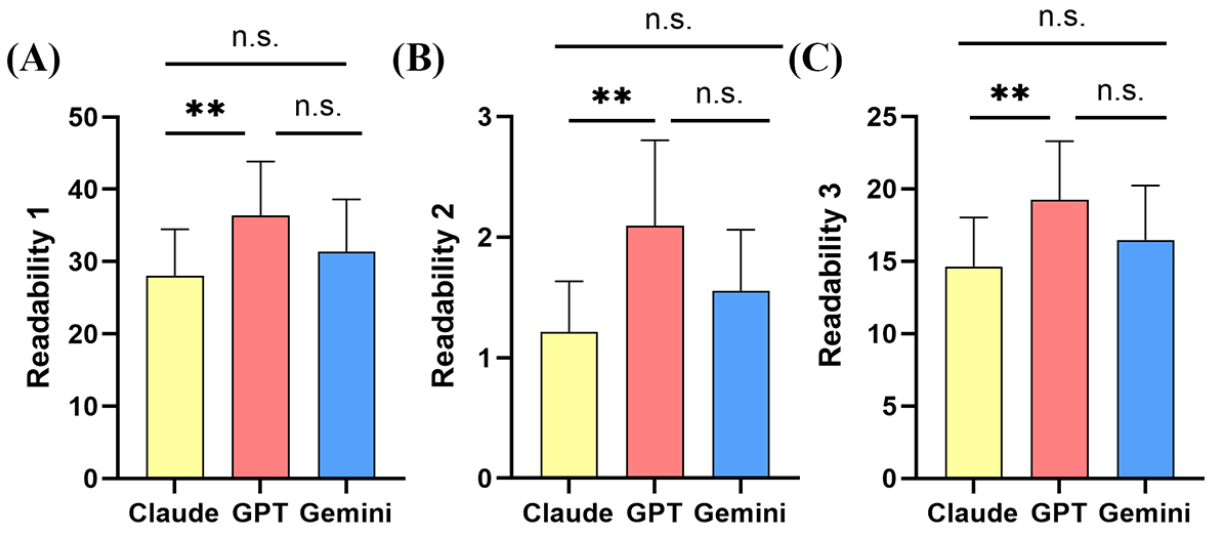
## Evaluations From the Subjective Perspective of Patients

In the ratings provided by 5 follow-up patients for the preoperative patient education materials generated by the LLMs, Gemini scored higher than GPT and Claude across all parameters, including readability, educational quality, and overall rating (Figures 2C and 2D). Among these, the difference in educational quality between Gemini and Claude (mean 4.00, SD 0.00 vs mean 3.60, SD 0.55; *P*=.02) was statistically significant (Figures 2C and 2D). Furthermore, Gemini's advantage in overall satisfaction when compared to both Claude (mean 8.80, SD 0.45 vs mean 6.80, SD 1.10; *P*<.01) and GPT (mean 8.80, SD 0.45 vs mean 7.20, SD 0.84; *P*=.01) also showed statistical significance (Figures 2C and 2D). The consistency of all ratings given by the 5 follow-up patients was evaluated as "Fair agreement" (Multimedia Appendix 2).

## Objective Evaluations of Readability

Based on the analysis methods of the context package, readability is assessed from 3 perspectives, namely readability 1, readability 2, and readability 3. Under these assessments, GPT's readability is higher than that of Gemini (readability 1: mean 36.38, SD 7.47 vs mean 31.39, SD 7.20, *P*=.18; readability 2: mean 2.09, SD 0.71 vs mean 1.55, SD 0.51, *P*=.09; readability 3: mean 19.24, SD 4.07 vs mean 16.47, SD 3.77, *P*=.17) and Claude (readability 1: mean 36.38, SD 7.47 vs mean 28.05, SD 6.43, *P*<.01; readability 2: mean 2.09, SD 0.71 vs mean 1.21, SD 0.42, *P*<.01; readability 3: mean 19.24, SD 4.07 vs mean 14.63, SD 3.40, *P*<.01), with the difference between GPT and Claude being statistically significant (Figure 4). Although Gemini's readability is higher than Claude's, the difference is not statistically significant (Figure 4). However, when readability was assessed using py-readability metrics, there was no statistical difference between the 3 LLM models (Multimedia Appendix 3).

**Figure 4.** Comparison of the results of text readability analysis from three analytical perspectives using the cntext package in R software. n.s.: not significant; *$P$<.05, **$P$<.01, ***$P$<.001.




# Discussion

## Principal Findings

The main findings of our study are as follows: (1) the three LLMs (Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro) demonstrated good overall potential for application in patient education for SCR surgery. They were able to generate answers to 11 SCR-related questions and create standardized preoperative informed consent patient education documents. (2) In the subjective evaluations by professional sports medicine clinicians and patients who had undergone SCR surgery, Gemini slightly outperformed GPT and Claude in multiple dimensions, including accuracy, completeness, logic, potential risks, and overall satisfaction. (3) In this study, the 3 LLMs did not proactively provide evidence sources when answering questions and generating patient education documents. If LLMs are to be used to assist with patient education in clinical applications, it may be necessary to specifically require LLMs to cite information sources to enable doctors and patients to judge the authority and reliability of the content. (4) Although Gemini performed best in the ratings for SCR patient education-related tasks, considering the complexity and potential risks of LLMs in medical applications, clinicians still need to carefully review and make necessary corrections to the content generated by LLMs to ensure the professionalism and reasonableness of patient education materials. LLMs should be positioned as assistive tools rather than decision-making entities in clinical applications.

LLMs have proven to be reliable sources of information for orthopedic surgery-related questions, creating patient education documents that enhance the understanding of diagnostic and therapeutic processes for nonprofessionals and improve the readability of educational materials [28,37,38]. However, evaluating the quality of responses from LLMs is not straightforward. Researchers assessed ChatGPT 3.5's medical knowledge by using clinical standards and licensing examination questions to evaluate its theoretical understanding and practical application [39]. With the advent of ChatGPT 4.0 and the iterative upgrades of various LLMs from different companies, there has been a growing recognition and exploration of the expanded pretraining data and enhanced text processing capabilities of the latest LLM versions in different clinical scenarios [40,41]. Scholars have realized that the quality of LLM responses is influenced by multiple factors, including the amount of information in the query [42], the questioning strategy [43], and many unpredictable elements [44]. These unpredictable elements are evident when, under controlled conditions with all variables constant, the same question yields different answers and shows varying styles of text presentation. Consequently, while researchers have acknowledged the capabilities of LLMs in diagnosing, treating, and creating educational documents across disciplines, they continue to reject the idea of LLMs performing independent medical actions, affirming their role solely as an auxiliary tool in the hands of professionals [45,46].

This study aims to assess the feasibility of using three popular LLMs as auxiliary tools for sports medicine physicians during the informed consent process for patients undergoing SCR. In this study, physicians use LLMs primarily to assess the accuracy and comprehensiveness of the information and to clarify content. Unlike previous studies that evaluated answer readability solely through software analysis of word and sentence structure [4,6,47], this study also included follow-up visits with SCR patients post surgery, where patients subjectively assessed the readability and educational significance of the information. Patient ratings primarily focused on the presurgical educational materials generated by LLMs, excluding the evaluation of 11 specific questions, as the answers to these questions required physician assessment of accuracy and comprehensiveness and clarification before clinical use. Without this step by physicians, patients, who are not medical professionals, might not be able to accurately assess the details of the questions. Although all 3 models performed satisfactorily in evaluating "potential risks," this does not imply that patients can rely on LLMs as their sole source of medical advice. We

believe that the SCR medical decision-making process, which does not involve extensive use of medications and auxiliary treatments pre- and post-surgery and follows a "surgery-rehabilitation" model, does not necessitate the phase-wise, continuous assessments and patient education required for conditions like cancer.

Despite the potential benefits of using LLMs in patient education, several ethical and privacy issues need to be addressed before their widespread application. The accuracy and reliability of the information generated by LLMs are critical, especially in sensitive medical contexts. To enhance their accuracy, strategies such as retrieving pertinent information from credible, external data sources before generating text can be incorporated into subsequent versions of LLMs. And patient privacy is a fundamental concern when using LLMs in medical settings. LLMs may require access to patient data to generate personalized and relevant information. However, this access must be strictly regulated to prevent unauthorized use or disclosure of sensitive patient information.

In addition, our "Prompt Execution" phase revealed that without background information, LLMs occasionally misidentify SCR as a supraspinatus repair surgery under patch bridging, leading to content generation biases. We consider such biases to be system errors caused by human operational mistakes, which can be avoided by adjusting prompt strategies under the guidance of subject matter experts. Therefore, using LLMs for specialist information retrieval is not without its challenges, and we believe that merely relying on LLM-generated disclaimers like "I am not a medical professional; if you feel unwell, please seek medical attention immediately" at the end of responses is insufficient [28]. The mitigation of these errors can be facilitated through the use of techniques such as fine-tuning and retrieval-augmented generation. Fine-tuning entails training the LLM on a smaller, highly specialized dataset that has been meticulously curated to capture the intricate details of the medical domain and retrieval-augmented generation can address issues of hallucinations by first retrieving pertinent information from credible, external data sources before generating text. Incorporating these strategies into subsequent versions of LLMs has the potential to enhance their accuracy and reliability, particularly in sensitive applications such as patient education. A thorough examination would offer valuable insights into refining these models to deliver precise and trustworthy information within medical contexts.

Our study meanwhile discovers critical gaps in LLMs are used in medical settings, particularly in presurgical patient education. LLMs often do not provide sources for their information, and their responses can include inaccuracies or fabricated sources, known as "hallucinations" [48]. This issue is exacerbated when users do not specifically ask for sources, leading LLMs to sometimes provide outdated or irrelevant information [48,49]. Furthermore, the LLMs in the study failed to discuss alternative treatments, benefits, and risks associated with not undergoing specific surgeries like SCR. This omission is significant as discussing these elements is essential for informed medical decision-making and respects patient rights to understand all available options. Given these limitations, LLMs should not independently manage diagnosis or patient education. Instead, they should serve as supplementary tools, aiding health care professionals who can provide the necessary context, accuracy, and depth in patient interactions. This approach ensures that patient education remains thorough, accurate, and ethically conducted, aligning with medical standards and patient rights. This challenge can be tackled through the application of more advanced prompt engineering methodologies, the integration of contextual reasoning capabilities, and the implementation of step-by-step guidance mechanisms. By engaging in multiple iterative interactions with the model, it becomes possible to refine its responses and produce more comprehensive information, encompassing alternative treatment options, based on the specific inputs provided by the user. Such an approach would empower the LLM to deliver content that is more personalized, well-informed, and balanced. Moreover, the development of LLM-Agents offers a compelling solution to the limitations of LLMs in sensitive domains like medical decision-making. By integrating planning, memory, tool use, and agent or brain components, these agents can enhance their ability to provide accurate, verified information. This not only supports human expertise but also ensures that the information presented is transparent and evidence-backed. As research continues, the full potential of integrating citation capabilities within LLM-Agents should be explored to further improve their reliability and trustworthiness in high-stakes contexts.

With the evolution of internet technology, we have witnessed a transition from Web1.0 to Web2.0, and the ways we access information have dramatically changed—from relying on traditional media to accessing massive amounts of information anytime and anywhere via the internet, social media, and personal media platforms [50,51]. Particularly on social media and personal media platforms, we can find questions similar to our own and the corresponding responses [6,50,51]. However, the accuracy and comprehensiveness of information obtained in this manner can be uncertain [51]. Online responses vary greatly in quality, lacking systematic organization and authority, and the response time and outcomes of further inquiries are unpredictable. Studies have shown that answers from ChatGPT 3.5 are not only more comprehensive and empathetic than those from certified physicians on Reddit forums but, despite demonstrating high quality in assessing dementia care issues, they fall slightly short in predicting potential future problems [52,53]. When comparing responses from ChatGPT 4.0, 3.5, and those on Reddit, ChatGPT 4.0's responses significantly surpassed the others, reaching a new level of excellence [54]. In responding to patient inquiries, LLMs also perform more accurately than Google searches and are easier to read [27]. However, they also share a common drawback: the use of LLMs in medical consultations is best accompanied by professional medical personnel to "clarify" the responses [31]. Therefore, LLMs are not suitable for independently handling any part of the diagnostic or treatment process within the medical system, but they are better suited as tools to enhance the efficiency of

professional medical personnel or as mediums for personalized patient communication and education [55,56].

As technology continues to advance, hospitals are consistently innovating in all aspects of clinical diagnosis and treatment to enhance diagnostic accuracy, treatment outcomes, and patient satisfaction, representing an unstoppable trend in health care innovation [57,58]. However, balancing standardized processes with personalized patient needs often presents a challenge [59]. LLMs present an opportunity to potentially maintain standardized quality in their responses while also accommodating personalized requests. LLMs, encompassing both free and paid versions, are generally accessible to the public as open platforms [60]. Although current research does not support its use in guiding clinical decisions [61], using ChatGPT in doctor-patient communication benefits both doctors and patients [7]. Doctors can interpret and supplement ChatGPT's responses based on their clinical experience, offering more personalized consultations to patients [31]. In addition, patients reduce their need to search for information on the internet, and their trust in physicians may be enhanced with the objective evidence provided by AI. Under the joint oversight of doctors and patients, the advantages of artificial intelligence can be fully used [62]. Nevertheless, the widespread adoption and application of LLMs still face technical and policy limitations. Technical limitations include differences in handling inputs in various languages [63], performance discrepancies between proprietary and open-source models [64], and the occurrence of "hallucinations" when faced with biased questions [65]. Since commonly used LLMs like GPT, Gemini, and Claude are proprietary, and these models are trained with significantly more data than open-source models, we can only continue to explore ways to avoid "hallucinations" instead of fixing the root cause of such issues [66,67]. In addition, policy restrictions cannot be ignored [68]. Health systems and hospitals need to develop detailed policies to regulate the clinical auxiliary use of LLMs, including ensuring patient informed consent, standardized user training, and the preservation of usage records [7]. Sound policies are essential to ensure the appropriate and efficient use of tools [65,68]. Through these measures, the safety of LLM applications in the medical field can be effectively enhanced, protecting patient rights while improving the efficiency and quality of doctor-patient communication [47,69].

## Limitations

This study has several limitations. First, both the linguistic input and the analyzed responses were in Chinese. On one hand, this choice was made to facilitate assessments by Chinese-speaking clinical experts and patients during follow-ups. On the other hand, input in different languages could introduce potential errors and biases. Second, this research only explores the feasibility of using LLMs to generate content related to SCR for patient education. The variability in surgical procedures and specialties could pose distinct challenges in patient education, which means the conclusions drawn from this study cannot be simply generalized to other disciplines. Finally, during the "Prompts Development" phase, it was found that without additional background information, SCRs are prone to be misidentified by LLMs as bridge suture repairs of the supraspinatus muscle. However, since all 3 models used were proprietary, we opted for a "Background+ Question" approach to mitigate this systematic error, without being able to investigate the reasons behind such occurrences.

## Conclusions

Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro effectively addressed patient queries and generated readable presurgical education materials. However, they lacked citations and failed to explore alternative treatments, benefits, and potential risks of forgoing SCR surgery. While these LLMs can serve as valuable aids for physicians, they should not be used as standalone tools for patient education without expert oversight to ensure comprehensive and accurate information is provided.

### Acknowledgments

We would like to express our deepest gratitude to all the experts and patients who have contributed to this research.

### Data Availability

All data included in this study are available upon request by contact with the corresponding author.

### Authors' Contributions

Conceptualization: WY Gan, H Li, JF Ouyang
Methodology: WY Gan, H Li, JF Ouyang
Supervision: XF Zheng
Visualization: YK Liu
Writing—original draft: WY Gan, H Li, JF Ouyang, YK Liu
Writing—reviewing and editing: WY Gan, H Li, JF Ouyang, YK Liu, ZW Xue, M Wang, HB He, B Song, XF Zheng

### Conflicts of Interest

None declared.

### Author Note

The subjects of this study are LLMs (large language models). Besides being used as operational models, LLMs also serve as tools for translating Chinese content into English, as detailed in Multimedia Appendix 1. The specific types of models used,

the websites they are accessed through, and their methods of use are all mentioned in the relevant sections. Beyond these functions, LLMs do not influence the generation of the article's content in any other way.

### Multimedia Appendix 1

All Questions and Answers for Claude-3-Opus, GPT-4-Turbo, and Gemini-1.5-Pro (Use GPT-4-Turbo for Chinese to English translation).
[DOCX File (Microsoft Word File), 72 KB-Multimedia Appendix 1]

### Multimedia Appendix 2

Table S1: Consistent evaluation of Fleiss kappa among raters.
[DOCX File (Microsoft Word File), 15 KB-Multimedia Appendix 2]

### Multimedia Appendix 3

Comaprison of readability by py-readability-metrics.
[DOCX File (Microsoft Word File), 19 KB-Multimedia Appendix 3]

## References


1. Flaharty KA, Hu P, Hanchard SL, et al. Evaluating large language models on medical, lay-language, and self-reported descriptions of genetic conditions. Am J Hum Genet. Sep 5, 2024;111(9):1819-1833. [doi: 10.1016/j.ajhg.2024.07.011] [Medline: 39146935]
2. Rengers TA, Thiels CA, Salehinejad H. Academic Surgery in the Era of Large Language Models: A Review. JAMA Surg. Apr 1, 2024;159(4):445-450. [doi: 10.1001/jamasurg.2023.6496] [Medline: 38353991]
3. Chow R, Hasan S, Zheng A, et al. The Accuracy of Artificial Intelligence ChatGPT in Oncology Examination Questions. J Am Coll Radiol. Nov 2024;21(11):1800-1804. [doi: 10.1016/j.jacr.2024.07.011] [Medline: 39098369]
4. Eng E, Mowers C, Sachdev D, et al. Chat Generative Pre-Trained Transformer (ChatGPT) – 3.5 Responses Require Advanced Readability for the General Population and May Not Effectively Supplement Patient-Related Information Provided by the Treating Surgeon Regarding Common Questions About Rotator Cuff Repair. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Jan 2025;41(1):42-52. [doi: 10.1016/j.arthro.2024.05.009]
5. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty. Journal of Bone and Joint Surgery. 2023;105(19):1519-1526. [doi: 10.2106/JBJS.23.00209]
6. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol. Oct 1, 2023;9(10):1437-1440. [doi: 10.1001/jamaoncol.2023.2947] [Medline: 37615960]
7. Xue Z, Zhang Y, Gan W, Wang H, She G, Zheng X. Quality and Dependability of ChatGPT and DingXiangYuan Forums for Remote Orthopedic Consultations: Comparative Analysis. J Med Internet Res. Mar 14, 2024;26:e50882. [doi: 10.2196/50882] [Medline: 38483451]
8. Gertz RJ, Dratsch T, Bunck AC, et al. Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. Radiology. Apr 2024;311(1):e232714. [doi: 10.1148/radiol.232714] [Medline: 38625012]
9. Maida M, Ramai D, Mori Y, et al. The role of generative language systems in increasing patient awareness of colon cancer screening. Endoscopy. Mar 2025;57(3):262-268. [doi: 10.1055/a-2388-6084] [Medline: 39142348]
10. Ebel S, Ehrengut C, Denecke T, Gößmann H, Beeskow AB. GPT-4o's competency in answering the simulated written European Board of Interventional Radiology exam compared to a medical student and experts in Germany and its ability to generate exam items on interventional radiology: a descriptive study. J Educ Eval Health Prof. 2024;21:21. [doi: 10.3352/jeehp.2024.21.21] [Medline: 39161266]
11. Gan W, Ouyang J, Li H, et al. Integrating ChatGPT in orthopedic education for medical undergraduates: Randomized controlled trial. J Med Internet Res. Aug 20, 2024;26:e57037. [doi: 10.2196/57037] [Medline: 39163598]
12. Mihata T, McGarry MH, Pirolo JM, Kinoshita M, Lee TQ. Superior capsule reconstruction to restore superior stability in irreparable rotator cuff tears: a biomechanical cadaveric study. Am J Sports Med. Oct 2012;40(10):2248-2255. [doi: 10.1177/0363546512456195] [Medline: 22886689]
13. E. Cline K, Tibone JE, Ihn H, et al. Superior Capsule Reconstruction Using Fascia Lata Allograft Compared With Double- and Single-Layer Dermal Allograft: A Biomechanical Study. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Apr 2021;37(4):1117-1125. [doi: 10.1016/j.arthro.2020.11.054]
14. Mihata T, Lee TQ, Hasegawa A, et al. Arthroscopic superior capsule reconstruction for irreparable rotator cuff tears: Comparison of clinical outcomes with and without subscapularis tear. Am J Sports Med. Dec 2020;48(14):3429-3438. [doi: 10.1177/0363546520965993] [Medline: 33104385]

15. Claro R, Fonte H. Superior capsular reconstruction: current evidence and limits. EFORT Open Rev. May 9, 2023;8(5):340-350. [doi: 10.1530/EOR-23-0027] [Medline: 37158430]
16. Mihata T, Lee TQ, Watanabe C, et al. Clinical results of arthroscopic superior capsule reconstruction for irreparable rotator cuff tears. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Mar 2013;29(3):459-470. [doi: 10.1016/j.arthro.2012.10.022]
17. Hirahara AM, Andersen WJ, Panero AJ. Superior capsular reconstruction: Clinical outcomes after minimum 2-year follow-up. Am J Orthop (Belle Mead NJ). 2017;46(6):266-278. [Medline: 29309442]
18. Snyder SJ, Arnoczky SP, Bond JL, Dopirak R. Histologic evaluation of a biopsy specimen obtained 3 months after rotator cuff augmentation with GraftJacket Matrix. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Mar 2009;25(3):329-333. [doi: 10.1016/j.arthro.2008.05.023]
19. Edwards PK, Mears SC, Lowry Barnes C. Preoperative education for hip and knee replacement: Never stop learning. Curr Rev Musculoskelet Med. Sep 2017;10(3):356-364. [doi: 10.1007/s12178-017-9417-4] [Medline: 28647838]
20. Alattas SA, Smith T, Bhatti M, Wilson-Nunn D, Donell S. Greater pre-operative anxiety, pain and poorer function predict a worse outcome of a total knee arthroplasty. Knee Surg Sports Traumatol Arthrosc. Nov 2017;25(11):3403-3410. [doi: 10.1007/s00167-016-4314-8] [Medline: 27734110]
21. Krebs ED, Hoang SC. Informed consent and shared decision making in the perioperative environment. Clin Colon Rectal Surg. May 2023;36(03):223-228. [doi: 10.1055/s-0043-1761158]
22. Noble PC, Fuller-Lafreniere S, Meftah M, Dwyer MK. Challenges in outcome measurement: Discrepancies between patient and provider definitions of success. Clin Orthop Relat Res. 2013;471(11):3437-3445. [doi: 10.1007/s11999-013-3198-x]
23. Villanueva C, Talwar A, Doyle M. Improving informed consent in cardiac surgery by enhancing preoperative education. Patient Educ Couns. Dec 2018;101(12):2047-2053. [doi: 10.1016/j.pec.2018.06.008] [Medline: 29937111]
24. Bollschweiler E, Apitzsch J, Obliers R, et al. Improving informed consent of surgical patients using a multimedia-based program? Results of a prospective randomized multicenter study of patients before cholecystectomy. Ann Surg. Aug 2008;248(2):205-211. [doi: 10.1097/SLA.0b013e318180a3a7] [Medline: 18650629]
25. Sceats LA, Morris AM, Narayan RR, Mezynski A, Woo RK, Yang GP. Lost in translation: Informed consent in the medical mission setting. Surgery. Feb 2019;165(2):438-443. [doi: 10.1016/j.surg.2018.06.010] [Medline: 30061041]
26. Neubauer PD, Tabaee A, Schwam ZG, Francis FK, Manes RP. Patient knowledge and expectations in endoscopic sinus surgery. Int Forum Allergy Rhinol. Sep 2016;6(9):921-925. [doi: 10.1002/alr.21763] [Medline: 27028979]
27. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for queries related to dementia and other cognitive decline: Comparison of results. J Med Internet Res. Jul 25, 2023;25:e48966. [doi: 10.2196/48966] [Medline: 37490317]
28. Oeding JF, Lu AZ, Mazzucco M, et al. ChatGPT-4 Performs clinical information retrieval tasks using consistently more trustworthy resources than does google search for queries concerning the Latarjet procedure. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Mar 2025;41(3):588-597. [doi: 10.1016/j.arthro.2024.05.025]
29. Nicikowski J, Szczepański M, Miedziaszczyk M, Kudliński B. The potential of ChatGPT in medicine: an example analysis of nephrology specialty exams in Poland. Clin Kidney J. Aug 2024;17(8):sfae193. [doi: 10.1093/ckj/sfae193] [Medline: 39099569]
30. Bernstein IA, Zhang YV, Govil D, et al. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. JAMA Netw Open. Aug 1, 2023;6(8):e2330320. [doi: 10.1001/jamanetworkopen.2023.30320] [Medline: 37606922]
31. Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. Chatgpt responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Jul 2024;40(7):2058-2066. [doi: 10.1016/j.arthro.2023.12.009]
32. Nwachukwu BU, Varady NH, Allen AA, et al. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. Arthroscopy: The Journal of Arthroscopic & Related Surgery. Feb 2025;41(2):263-275. [doi: 10.1016/j.arthro.2024.07.040]
33. Chen L, Zaharia M, Zou J. How is chatgpt's behavior changing over time? Preprint posted online on Jul 1, 2023. URL: https://ui.adsabs.harvard.edu/abs/2023arXiv230709009C [Accessed 2025-06-06]
34. Menz BD, Kuderer NM, Bacchi S, et al. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. BMJ. Mar 20, 2024;384:e078538. [doi: 10.1136/bmj-2023-078538] [Medline: 38508682]
35. Yalamanchili A, Sengupta B, Song J, et al. Quality of large language model responses to radiation oncology patient care questions. JAMA Netw Open. Apr 1, 2024;7(4):e244630. [doi: 10.1001/jamanetworkopen.2024.4630] [Medline: 38564215]

36. Li HQ, Yang Y, Xue FW, Liu ZY. Annual report readability and trade credit financing: Evidence from China. Research in International Business and Finance. Apr 2024;69:102220. [doi: 10.1016/j.ribaf.2024.102220]
37. Draschl A, Hauer G, Fischerauer SF, et al. Are chatgpt's free-text responses on periprosthetic joint infections of the hip and knee reliable and useful? J Clin Med. Oct 20, 2023;12(20):6655. [doi: 10.3390/jcm12206655] [Medline: 37892793]
38. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee surg sports traumatol arthrosc. Nov 2023;31(11):5190-5198. [doi: 10.1007/s00167-023-07529-2]
39. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. J Med Educ Curric Dev. 2024;11(23821205241238641):23821205241238641. [doi: 10.1177/23821205241238641] [Medline: 38487300]
40. Deng L, Wang T, et al. Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2. Int J Surg. Jan 2024;110(4):1941-1950. [doi: 10.1097/JS9.0000000000001066]
41. Jarry Trujillo C, Vela Ulloa J, Escalona Vivas G, et al. Surgeons vs ChatGPT: Assessment and feedback performance based on real surgical scenarios. J Surg Educ. Jul 2024;81(7):960-966. [doi: 10.1016/j.jsurg.2024.03.012] [Medline: 38749814]
42. Zhu L, Mou W, Lai Y, et al. Step into the era of large multimodal models: a pilot study on ChatGPT-4V(ision)'s ability to interpret radiological images. Int J Surg. Jul 1, 2024;110(7):4096-4102. [doi: 10.1097/JS9.0000000000001359] [Medline: 38498394]
43. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine. Sep 2023;95(104770):104770. [doi: 10.1016/j.ebiom.2023.104770] [Medline: 37625267]
44. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril. Sep 2023;120(3 Pt 2):575-583. [doi: 10.1016/j.fertnstert.2023.05.151] [Medline: 37217092]
45. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
46. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg. Jun 1, 2024;110(6):3701-3706. [doi: 10.1097/JS9.0000000000001312] [Medline: 38502861]
47. Haver HL, Gupta AK, Ambinder EB, et al. Evaluating the use of ChatGPT to accurately simplify patient-centered information about breast cancer prevention and screening. Radiol Imaging Cancer. Mar 2024;6(2):e230086. [doi: 10.1148/rycan.230086] [Medline: 38305716]
48. Chelli M, Descamps J, Lavoué V, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. J Med Internet Res. May 22, 2024;26:e53164. [doi: 10.2196/53164] [Medline: 38776130]
49. Burnette H, Pabani A, von Itzstein MS, et al. Use of artificial intelligence chatbots in clinical management of immune-related adverse events. J Immunother Cancer. May 30, 2024;12(5):38816231. [doi: 10.1136/jitc-2023-008599] [Medline: 38816231]
50. Terrasse M, Gorin M, Sisti D. Social media, e-health, and medical ethics. Hastings Cent Rep. Jan 2019;49(1):24-33. [doi: 10.1002/hast.975] [Medline: 30790306]
51. Ho A, McGrath C, Mattheos N. Social media patient testimonials in implant dentistry: information or misinformation? Clin Oral Implants Res. Jul 2017;28(7):791-800. [doi: 10.1111/clr.12883] [Medline: 27279455]
52. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 1, 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
53. Aguirre A, Hilsabeck R, Smith T, et al. Assessing the quality of chatgpt responses to dementia caregivers' questions: Qualitative analysis. JMIR Aging. May 6, 2024;7:e53019. [doi: 10.2196/53019] [Medline: 38722219]
54. Girton MR, Greene DN, Messerlian G, Keren DF, Yu M. ChatGPT vs medical professional: Analyzing responses to laboratory medicine questions on social media. Clin Chem. Sep 3, 2024;70(9):1122-1139. [doi: 10.1093/clinchem/hvae093] [Medline: 39013110]
55. La Bella S, Attanasi M, Porreca A, et al. Reliability of a generative artificial intelligence tool for pediatric familial Mediterranean fever: insights from a multicentre expert survey. Pediatr Rheumatol Online J. Aug 23, 2024;22(1):78. [doi: 10.1186/s12969-024-01011-0] [Medline: 39180115]
56. Cavnar Helvaci B, Hepsen S, Candemir B, et al. Assessing the accuracy and reliability of ChatGPT's medical responses about thyroid cancer. Int J Med Inform. Nov 2024;191(105593):105593. [doi: 10.1016/j.ijmedinf.2024.105593] [Medline: 39151245]

57. Pallett AC, Nguyen BT, Klein NM, Phippen N, Miller CR, Barnett JC. A randomized controlled trial to determine whether A video presentation improves informed consent for hysterectomy. Am J Obstet Gynecol. Sep 2018;219(3):277. [doi: 10.1016/j.ajog.2018.06.016] [Medline: 29959929]
58. Zhang MH, Haq ZU, Braithwaite EM, Simon NC, Riaz KM. A randomized, controlled trial of video supplementation on the cataract surgery informed consent process. Graefes Arch Clin Exp Ophthalmol. Aug 2019;257(8):1719-1728. [doi: 10.1007/s00417-019-04372-5] [Medline: 31144057]
59. McCollough CH. Standardization versus individualization: how each contributes to managing dose in computed tomography. Health Phys. Nov 2013;105(5):445-453. [doi: 10.1097/HP.0b013e31829db936] [Medline: 24077044]
60. Vaid A, Duong SQ, Lampert J, et al. Local large language models for privacy-preserving accelerated review of historic echocardiogram reports. J Am Med Inform Assoc. Sep 1, 2024;31(9):2097-2102. [doi: 10.1093/jamia/ocae085] [Medline: 38687616]
61. Balla Y, Tirunagari S, Windridge D. Machine learning in pediatrics: Evaluating challenges, opportunities, and explainability. Indian Pediatr. May 14, 2023. [Medline: 37179470]
62. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. Jul 2023;29(3):721-732. [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
63. Shao CY, Li H, Liu XL, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: Survey study. Interact J Med Res. Aug 14, 2023;12:e46900. [doi: 10.2196/46900] [Medline: 37578819]
64. Sandmann S, Riepenhausen S, Plagwitz L, Varghese J. Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun. Mar 6, 2024;15(1):2050. [doi: 10.1038/s41467-024-46411-8] [Medline: 38448475]
65. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. J Med Internet Res. Aug 22, 2023;25:e48659. [doi: 10.2196/48659] [Medline: 37606976]
66. Masters K. *Medical Teacher* 's first ChatGPT's referencing hallucinations: Lessons for editors, reviewers, and teachers . Med Teach. Jul 3, 2023;45(7):673-675. [doi: 10.1080/0142159X.2023.2208731]
67. Hatem R, Simmons B, Thornton JE. A call to address AI "Hallucinations" and how healthcare professionals can mitigate their risks. Cureus. Sep 2023;15(9):37809168. [doi: 10.7759/cureus.44720]
68. Bukar UA, Sayeed MS, Razak SFA, Yogarayan S, Amodu OA. An integrative decision-making framework to guide policies on regulating ChatGPT usage. PeerJ Comput Sci. 2024;10:e1845. [doi: 10.7717/peerj-cs.1845] [Medline: 38440047]
69. Platt J, Nong P, Smiddy R, et al. Public comfort with the use of ChatGPT and expectations for healthcare. J Am Med Inform Assoc. Sep 1, 2024;31(9):1976-1982. [doi: 10.1093/jamia/ocae164]


## Abbreviations

**LLM:** large language model
**PEMAT:** Patient Education Materials Assessment Tool
**SCR:** superior capsular reconstruction


*Edited by Nidhi Rohatgi; peer-reviewed by Fatema Tuj Johora Faria, Ming Ma; submitted 13.12.2024; final revised version received 04.04.2025; accepted 08.04.2025; published 12.06.2025*

*Please cite as:*
*Liu Y, Li H, Ouyang J, Xue Z, Wang M, He H, Song B, Zheng X, Gan W*
*Evaluating Large Language Models for Preoperative Patient Education in Superior Capsular Reconstruction: Comparative Study of Claude, GPT, and Gemini*
*JMIR Perioper Med 2025;8:e70047*
*URL: https://periop.jmir.org/2025/1/e70047*
*doi: 10.2196/70047*