

Original Article

Content and Readability Analysis of ChatGPT and Gemini's Responses to FAQs on Patellofemoral Instability

 Volkan Buyukarslan,¹  Firat Dogruoz,²  Murat Yuncu,³  Mehmet Baris Ertan⁴

¹Department of Orthopedics and Traumatology, Cine State Hospital, Aydin, Türkiye

²Department of Orthopedics and Traumatology, University of Health Sciences, Antalya Training and Research Hospital, Antalya, Türkiye

³Department of Orthopedics and Traumatology, Siverek State Hospital, Urfa, Türkiye

⁴Department of Orthopedics and Traumatology, Medikum Private Hospital, Antalya, Türkiye

ABSTRACT

Objective: This study aimed to evaluate the quality and readability of responses generated by ChatGPT and Gemini to frequently asked questions about patellofemoral instability (PFI). In the context of increasing reliance on AI chatbots for medical information, it is imperative to evaluate their accuracy, completeness, and accessibility to determine their potential role in patient education.

Materials and Methods: A cross-sectional observational study was conducted using 20 frequently asked patient questions about PFI, selected based on Google search trends and patient education resources. These questions were submitted to ChatGPT (version 4o) and Gemini (version 2.1), and the responses were analyzed for content quality and readability. Content quality was assessed by three independent orthopedic specialists using a structured scoring framework. Each response was rated on a five-point Likert scale, ranging from very poor to excellent. This framework focused on relevance, accuracy, clarity, completeness, evidence-based support, and consistency. The readability of the responses was assessed using several linguistic indices, including the Flesch-Kincaid Grade Level, the Flesch Reading Ease Score, the Gunning Fog Index, the Coleman-Liau Index, the Automated Readability Index (ARI), and the Simple Measure of Gobbledygook (SMOG) Index. Both the content and readability of the responses were compared statistically.

Results: ChatGPT had higher accuracy (4.70 ± 0.21 vs. 4.58 ± 0.40 , $p=0.071$) and evidence-based support (4.51 ± 0.45 vs. 4.35 ± 0.67 , $p=0.045$) scores than Gemini, although these differences were not always statistically significant. In contrast, Gemini produced significantly clearer responses (4.95 ± 0.12 vs. 4.75 ± 0.14 , $p=0.001$) and had a significantly higher Flesch Reading Ease Score (30.21 ± 9.43 vs. 19.39 ± 10.09 , $p=0.001$), indicating that its responses were easier to read. Both models generated text at a college reading level, suggesting limited accessibility for the general patient population.

Conclusion: ChatGPT and Gemini have provided reliable yet complex answers to patient questions about patellofemoral instability. In particular, ChatGPT has been shown to excel in accuracy and evidence-based support, while Gemini has been observed to produce more readable content. However, both models require further refinement regarding readability and transparency to improve their suitability for patient education. Future research should explore the integration of AI chatbots into clinical workflows to ensure safe and effective information dissemination for diverse patient populations.

Keywords: Artificial intelligence, chatbots, ChatGPT, Gemini, patellofemoral instability, readability, patient education



Cite this article as:

Buyukarslan V, Dogruoz F, Yuncu M, Ertan MB. Content and Readability Analysis of ChatGPT and Gemini's Responses to FAQs on Patellofemoral Instability. Sports Traumatol Arthrosc 2025;2(2):37–45.

Address for correspondence:

Volkan Buyukarslan.
Department of Orthopedics and Traumatology, Cine State Hospital, Aydin, Türkiye
E-mail:
volkanbuyukarslan@gmail.com

Submitted: 19.02.2025

Revised: 15.03.2025

Accepted: 17.03.2025

Available Online: 18.08.2025

Sports Traumatology & Arthroscopy –
Available online at www.stajournal.com



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

INTRODUCTION

Patellofemoral pathologies encompass a range of conditions affecting the patella and femoral trochlea, including chondromalacia patella, patellar maltracking, and bipartite patella developmental anomaly, where the patella remains in two separate bone fragments, which can sometimes cause anterior knee pain or discomfort^[1,2]. Patellofemoral instability (PFI) is a common orthopedic condition that can significantly impact patients' quality of life, manifesting as recurrent patellar dislocation, pain, and functional limitations^[3]. Due to the complexity of the condition, patients often seek information online to better understand their condition, treatment options, and prognosis. In recent years, artificial intelligence-based chatbots such as ChatGPT and Gemini have emerged as popular sources of medical information, providing immediate and user-friendly answers to common health-related questions^[4].

Despite their accessibility, the accuracy, comprehensiveness, and readability of AI-generated medical responses remain debatable. Previous studies evaluating AI chatbots for patient education on musculoskeletal conditions, such as anterior cruciate ligament (ACL) reconstruction, have shown mixed results. Some AI-generated responses were found to be satisfactory, while others required significant clarification due to factual inaccuracies or overgeneralization. In addition, concerns about the readability of AI-generated text have been raised, with many responses exceeding the recommended reading level for effective patient education^[5].

In light of the growing trend among patients to seek medical information from AI chatbots, it is imperative to evaluate the efficacy of these tools in providing clear, accurate, and actionable information concerning patellofemoral instability (PFI). The objective of this study is to evaluate the quality, accuracy, and readability of AI-generated responses, intending to determine the role of these chatbots in patient education and counseling. By ascertaining their strengths and limitations, healthcare providers can judiciously guide patients toward reliable information sources, ensuring that AI-generated content functions as a reliable complementary tool.

The central hypothesis of this study is that ChatGPT will provide more comprehensive and accurate responses to frequently asked questions about patellofemoral instability compared to Gemini. However, it is acknowledged that both AI-based chatbots will have limitations in terms of readability and the need for additional medical clarification.

MATERIALS AND METHODS

This cross-sectional observational study was conducted in January 2025. The study protocol was approved by the

institutional review board (Antalya Training and Research Hospital Scientific Research Ethics Committee (date: 09.01.2025, number: 1/25). This study did not involve any human data or participant information.

Preparation of Frequently Asked Questions

To ensure relevance, the most frequently asked patient questions on this topic were identified using Google search trends and patient information resources. Questions that had the same meaning or asked for the same information in different ways were extracted and merged. Out of several questions (total number: 52), the 20 most relevant questions were selected based on their frequency in Google search trends and their clinical relevance as determined by orthopedic specialists. These questions were submitted to ChatGPT (version 4o) and Gemini (version 2.1) on January 10, 2025. The list of questions is shown in Table 1. Responses were collected in text format for subsequent analysis.

Content Analysis

A structured content analysis approach was used to assess the quality of the responses. Three senior orthopedic surgeons with more than a decade of experience in sports trauma and knee surgery independently reviewed the responses. The evaluation focused on six key areas: relevance, accuracy, clarity, completeness, evidence-based support, and consistency. Each response was rated on a five-point Likert scale, ranging from very poor to excellent. Relevance was assessed based on how well the answer addressed the specific question without providing unnecessary or unrelated information. Accuracy was determined by the correctness and reliability of the content, ensuring consistency with established clinical guidelines and evidence-based practices. Clarity was assessed based on the organization, readability, and logical presentation of the information. Completeness was assessed by examining whether all critical aspects of the topic were adequately covered. The extent to which the responses included credible research and scientific evidence was considered under evidence-based support, while consistency reflected the uniformity of information across different responses. Reviewers were provided with widely accepted clinical guidelines and literature references on patellofemoral instability to standardize the evaluation process. Their ratings were based on these references as well as their clinical expertise^[6–9].

Readability Analysis

To assess the accessibility of AI-generated responses for patients, readability analysis was performed using widely accepted linguistic metrics. Readability refers to the ease with which a reader can understand written text, and in the context of patient education, it is recommended that medical

Table 1. The list of questions asked to the ChatGPT and Gemini AI chatbots

Question 1	What are the stabilizing structures of the patellofemoral joint?
Question 2	Is patellar instability a serious condition?
Question 3	What structures are responsible for stabilizing the patella?
Question 4	What are the three most significant causes of patellofemoral instability?
Question 5	If I have patellar instability, is there a risk of my patella dislocating again?
Question 6	What sensations are associated with patellar instability?
Question 7	What methods are used to test for patellar instability?
Question 8	What are the symptoms of chronic patellar instability?
Question 9	How is patellar instability diagnosed?
Question 10	Can patella alta be treated effectively without surgical intervention?
Question 11	What exercises are harmful when patellar instability is present?
Question 12	How should patellar instability be rehabilitated?
Question 13	Is it possible for patellofemoral instability to resolve on its own?
Question 14	Does patellar instability necessitate surgical treatment?
Question 15	What approaches are available for correcting patellofemoral instability?
Question 16	How long does it typically take for patellar instability to heal?
Question 17	What are the consequences of untreated patellofemoral instability?
Question 18	How can I enhance my patella stability?
Question 19	What is the recovery time for patellar instability?
Question 20	When can I return to sports activities after patellar instability surgery?

information be written at a reading level equivalent to the sixth to eighth grade to ensure comprehensibility for a broad audience. Given the increasing reliance on AI-based sources for health information, this study aimed to determine whether the complexity of responses from ChatGPT and Gemini was appropriate for patients seeking information about PFI ^[10].

Several established readability formulas were applied to each response to objectively measure its complexity. The Gunning Fog Index estimates the years of formal education required to understand a given text, with higher scores indicating increased difficulty ^[11]. The Coleman-Liau Index evaluates readability based on the number of characters per word and words per sentence, making it independent of syllable count ^[12]. The Flesch-Kincaid Grade Level estimates the U.S. school grade required to comprehend the text, while the Flesch Reading Ease Score provides a scale from 0 to 100, with higher values indicating easier readability ^[13]. The Automated Readability Index (ARI) assesses text complexity based on sentence length and word difficulty, presenting results in grade-level format ^[14]. The Simple Measure of Gobbledygook (SMOG) Index estimates readability by analyzing polysyllabic words, with a higher SMOG score indicating a more difficult text ^[15].

Each AI-generated response was analyzed using these six readability metrics, and the results of ChatGPT and Gemini were compared. The goal was to determine whether the responses were written at a level accessible to the general public or if they exceeded the recommended readability threshold for effective patient education. By quantifying readability, this study provides insights into the usability of AI chatbots in delivering clear and comprehensible medical information.

Statistical Analysis

Descriptive statistics summarized the quality ratings across different evaluation domains, including mean, standard deviation, median and range. Depending on data distribution, the differences between ChatGPT and Gemini responses were analyzed using appropriate statistical tests. Since the data did not follow a normal distribution, dependent groups were compared using the Wilcoxon signed-rank test. $p < 0.05$ was applied to determine statistical significance. Interclass correlation coefficient with 95% CI was employed to assess inter-rater reliability of the content analysis by the three raters. ICC values were interpreted as poor (< 0.50), moderate ($0.50-0.75$), good ($0.75-0.90$), or excellent (> 0.90) ^[16].

RESULTS

The reliability of content evaluation among the three independent raters was assessed using the intraclass correlation coefficient (ICC). The ICC for ChatGPT was 0.674 (95% CI: 0.313-0.861), indicating moderate reliability. In contrast, the ICC for Gemini was 0.886 (95% CI: 0.760-0.951), demonstrating good reliability (Table 2).

Comparison of Content Quality

The responses generated by ChatGPT and Gemini were evaluated based on six key content domains: relevance, accuracy, clarity, completeness, evidence-based support, and

Table 2. Reliability of content analysis by three independent raters

LLM	ICC	95% CI	Interpretation
ChatGPT	0.674	0.313-0.861	Moderate
Gemini	0.886	0.760-0.951	Good

LLM: Large Language Model; ICC: Interclass Correlation Coefficient; CI: Confidence Interval.

consistency. ChatGPT demonstrated slightly higher scores in relevance (4.50 ± 0.47 vs. 4.35 ± 0.67 , $p=0.058$) and accuracy (4.70 ± 0.21 vs. 4.58 ± 0.40 , $p=0.071$), though the differences were not statistically significant. Clarity was significantly higher for Gemini (4.95 ± 0.12 vs. 4.75 ± 0.14 , $p=0.001$). No significant differences were found in completeness (4.38 ± 0.47 vs. 4.33 ± 0.66 , $p=0.444$) or consistency (4.70 ± 0.21 vs. 4.58 ± 0.40 , $p=0.071$). ChatGPT had a significantly higher evidence-based support score than Gemini (4.51 ± 0.45 vs. 4.35 ± 0.67 , $p=0.045$). The overall content score was similar between ChatGPT (27.55 ± 1.81) and Gemini (27.15 ± 2.82), with no statistically significant difference ($p=0.204$) (Table 3).

Readability Comparison

Readability measures were assessed using multiple indices to determine the accessibility of AI-generated responses. Gunning Fog Index showed no significant difference between ChatGPT (16.22 ± 2.00) and Gemini (15.73 ± 1.95 , $p=0.191$), both indicating college-level readability. Coleman-Liau Index was significantly lower for Gemini (14.15 ± 1.51) compared to ChatGPT (16.27 ± 1.54 , $p=0.001$), suggesting slightly easier readability for Gemini. Flesch-Kincaid Grade Level was comparable between ChatGPT (13.91 ± 1.65) and Gemini (13.51 ± 1.79 , $p=0.191$), both corresponding to a college-level reading requirement. Flesch Reading Ease Score was significantly higher for Gemini (30.21 ± 9.43) than for ChatGPT (19.39 ± 10.09 , $p=0.001$), indicating that Gemini’s responses were easier to read. The Automated Readability Index (ARI) and Simple Measure of Gobbledygook (SMOG) Index showed no significant differences between the two models ($p=0.723$ and $p=0.588$, respectively) (Table 4).

DISCUSSION

The key findings of this study indicate that ChatGPT and Gemini provided high-quality responses to FAQs about PFI, with each model demonstrating unique strengths and weaknesses. ChatGPT exhibited greater accuracy and adherence to evidence-based recommendations, whereas Gemini’s responses were clearer and more readable. However, both AI models generated responses at a college

Table 3. Comparison of content analysis of the responses by the LLMs

Domain	Data	ChatGPT	Gemini	p
Relevance	Mean±SD	4.50±0.47	4.35±0.67	0.058
	Median	4.66	4.33	
	Range	3.67-5.00	3.33-5.00	
Accuracy	Mean±SD	4.70±0.21	4.58±0.40	0.071
	Median	4.66	4.33	
	Range	4.33-5.00	4.00-5.00	
Clarity	Mean±SD	4.75±0.14	4.95±0.12	0.001
	Median	4.66	5.00	
	Range	4.67-5.00	4.67-5.00	
Completeness	Mean±SD	4.38±0.47	4.33±0.66	0.444
	Median	4.33	4.33	
	Range	3.67-5.00	3.33-5.00	
Evidence-Based	Mean±SD	4.51±0.45	4.35±0.67	0.045
	Median	4.66	4.33	
	Range	3.67-5.00	3.33-5.00	
Consistency	Mean±SD	4.70±0.21	4.58±0.40	0.071
	Median	4.66	4.33	
	Range	4.33-5.00	4.00-5.00	
Overall Content				
Score	Mean±SD	27.55±1.81	27.15±2.82	0.204
	Median	27.66	22.67	
	Range	25.00-30.00	22.67-30.00	

P-values are calculated using the Wilcoxon Signed-Ranked Test.

reading level, which may hinder accessibility for patients without advanced health literacy. These findings align with previous studies assessing AI chatbot performance in orthopedics, highlighting both the promise and limitations of AI-generated medical content [10,17].

Both models generated text at a college reading level, suggesting limited accessibility for the general patient population. This finding indicates a potential need for AI chatbots to generate responses at a lower reading level to enhance patient comprehension and accessibility. The adaptation of AI-generated text to align with health literacy guidelines should be considered in future research and model development.

Several studies have supported using AI chatbots as valuable tools in patient education, particularly in orthopedic

Table 4. Comparison of readability between LLMs

Readability Measures	Data	ChatGPT	Gemini	p
Gunning Fog Index	Mean±SD	16.22±2.00	15.73±1.95	0.191
	Median	16.52	15.63	
	Range	12.20-19.26	11.45-19.55	
	Interpretation	College	College	
Coleman Liau Index	Mean±SD	16.27±1.54	14.15±1.51	0.001
	Median	16.24	14.29	
	Range	12.55-19.02	11.25-17.75	
	Interpretation	College	College	
Flesch Kincaid Grade Level	Mean±SD	13.91±1.65	13.51±1.79	0.191
	Median	14.01	13.35	
	Range	10.40-16.42	9.83-17.07	
	Interpretation	College	College	
Flesch Reading Ease	Mean±SD	19.39±10.09	30.21±9.43	0.001
	Median	18.00	30.88	
	Range	3.30-41.69	6.59-51.24	
	Interpretation	College	College	
Automated Readability Index	Mean±SD	12.74±1.68	12.87±2.10	0.723
	Median	12.78	12.88	
	Range	9.21-16.73	9.28-16.84	
	Interpretation	12 th Grade	12 th Grade	
Simple Measure of Gobbledygook	Mean±SD	14.39±1.30	14.53±1.38	0.588
	Median	14.46	14.50	
	Range	12.08-17.28	11.25-17.11	
	Interpretation	Undergraduate	Undergraduate	

¹Wilcoxon Signed Rank Test.

and musculoskeletal conditions. Li et al. ^[18] conducted a structured evaluation of ChatGPT's responses to frequently asked questions about anterior cruciate ligament (ACL) reconstruction, assessing their accuracy, clarity, and completeness. Their findings indicated that most responses were satisfactory, requiring only minor clarifications, suggesting that ChatGPT can serve as a useful supplementary resource for patient education. However, they also noted that while ChatGPT provided clinically relevant information, it occasionally lacked specificity in surgical recommendations, which required further input from medical professionals. This aligns with our study's findings, where ChatGPT demonstrated strong accuracy and evidence-based support but sometimes required additional clarification to enhance clarity. Similarly, Gaudiani et al. ^[19] compared ChatGPT-4's ability to answer ACL

reconstruction queries against Google search results. Their study assessed correctness, completeness, and readability, revealing that ChatGPT-4 responses were significantly more accurate and comprehensive than Google's top search results. Importantly, ChatGPT's responses were found to be more structured, coherent, and aligned with established clinical guidelines, reinforcing its potential as a reliable alternative to conventional internet searches for patient education. This finding is consistent with our study, in which ChatGPT's responses demonstrated high accuracy and evidence-based support. However, our study also highlighted that ChatGPT's readability remains challenging, with responses requiring a college-level reading proficiency, which may limit accessibility for patients with lower health literacy.

In another study focusing on pediatric orthopedic conditions, Pirkle et al. ^[20] evaluated the reliability of ChatGPT and Gemini in providing recommendations for pediatric orthopedic care. Their results indicated that both AI models demonstrated moderate alignment with AAOS Clinical Practice Guidelines (CPGs), with no significant difference in overall performance. However, while Gemini provided references for its responses, 12 out of the 16 cited studies were inaccurate, fabricated, or non-existent, raising concerns about transparency and the risk of misinformation. This is particularly relevant to our findings, where Gemini demonstrated superior clarity and readability compared to ChatGPT but lacked robust evidence-based references, making its recommendations potentially less reliable. Our study reinforces these concerns, suggesting that while AI chatbots can effectively communicate medical concepts, their reliability in sourcing accurate references and fully adhering to clinical guidelines requires improvement.

Collectively, these studies indicate that ChatGPT and Gemini have the potential to enhance patient education by providing structured and accessible information on orthopedic conditions, yet limitations remain in terms of accuracy, source transparency, and readability. Our study builds upon these findings by directly comparing the readability and content quality of ChatGPT and Gemini in the context of patellofemoral instability, further emphasizing that ChatGPT excels in evidence-based accuracy, whereas Gemini outperforms in clarity and readability. These results highlight the complementary strengths and weaknesses of AI chatbots and the need for further refinements to improve their clinical applicability.

Conversely, several studies have raised concerns about AI chatbot accuracy and reliability, particularly in medical and orthopedic contexts. Johns et al. ^[21] evaluated ChatGPT's responses to ACL reconstruction-related patient inquiries and found that 60% of the responses were deemed unsatisfactory, requiring substantial clarification. Their study emphasized that ChatGPT often provided outdated or incomplete information, leading to potential patient misunderstandings. Moreover, the reading level of ChatGPT's responses was calculated to be equivalent to that of a college sophomore (13.4 years of education), which is significantly higher than the recommended readability level for patient education materials. This aligns with our findings, where ChatGPT produced accurate but complex responses that may not be easily comprehensible to the general public.

Similarly, Nwachukwu et al. ^[22] conducted a large-scale study assessing multiple AI chatbots, including ChatGPT, Gemini, and Mistral-7B, by comparing their responses to evidence-

based clinical practice guidelines (CPGs) from the American Academy of Orthopedic Surgeons (AAOS). Their analysis revealed that more than one in four responses failed to align with established guidelines, raising significant concerns about the reliability of AI-generated medical recommendations. Specifically, Gemini had the highest rate of discordant recommendations (12.5%), while ChatGPT-4 performed better but still exhibited a substantial error rate (7.3%). This suggests that although AI models can generate plausible-sounding medical advice, they often lack the precision necessary for evidence-based decision-making. Our study corroborates this finding, as ChatGPT demonstrated strong evidence-based accuracy but lacked accessibility, while Gemini provided more readable responses that were occasionally inconsistent with established guidelines.

In a related study focusing on musculoskeletal conditions, Quinn et al. ^[23] examined the accuracy and transparency of ChatGPT and Gemini in interpreting and relaying AAOS recommendations for anterior cruciate ligament reconstruction (ACLR). Their findings indicated that while Gemini's responses were more readable and structured, they often lacked depth, omitted critical details, or failed to provide references for key recommendations. This aligns with our study's results, where Gemini scored significantly higher in clarity and readability compared to ChatGPT, yet its responses were often less rigorous in terms of evidence-based support. The lack of citations and transparency in Gemini's recommendations raises concerns about misinformation, mainly when patients rely on AI-generated responses for medical decision-making.

These studies highlight the ongoing challenges associated with AI chatbots in medical education and patient counseling. While they offer immediate, user-friendly access to medical knowledge, their potential for misinformation, lack of transparency, and tendency to produce responses above the recommended patient reading level remain critical concerns. Our findings reinforce this perspective, demonstrating that ChatGPT provides more reliable medical content but struggles with readability, whereas Gemini produces clearer responses that sometimes lack the necessary depth and supporting evidence. These limitations underscore the need for continued refinement of AI models, increased transparency in AI-generated medical content, and stronger validation mechanisms to ensure patient safety and accuracy in healthcare communication.

This study has several strengths. It is among the first to comprehensively compare ChatGPT and Gemini in patellofemoral instability, assessing content quality and

readability using a structured, reproducible methodology. Three independent orthopedic specialists conducted the evaluation, enhancing objectivity and reliability. Multiple readability indices, including the Flesch-Kincaid Grade Level, Gunning Fog Index, and Coleman-Liau Index, provided a detailed analysis of text accessibility, an often-overlooked aspect in AI-generated medical content. By incorporating both accuracy and readability metrics, this study offers a holistic assessment of AI-generated responses, which is crucial for their practical application in patient education. The direct comparison between ChatGPT and Gemini highlights their distinct strengths—ChatGPT excels in accuracy and evidence-based support, while Gemini produces more readable and well-structured responses—providing valuable insights for healthcare professionals and patients. Focusing on a specific orthopedic condition allows for a more in-depth analysis than broader studies.

However, the study has limitations. It evaluates only text-based responses without assessing their real-world impact on patient comprehension and decision-making, which is influenced by health literacy, prior knowledge, and cognitive biases. Future research should include patient-centered evaluations like comprehension testing and usability studies. The study is also limited to English-language responses, restricting generalizability to non-English-speaking populations. Given the global use of AI chatbots, future studies should explore multilingual performance, particularly for populations with low health literacy. Additionally, AI models evolve rapidly, meaning these findings may not fully apply to future versions of ChatGPT and Gemini. Lastly, the study did not compare AI-generated responses with those from medical professionals. Future research should evaluate AI models against expert-reviewed patient education materials to better define their role in medical communication.

CONCLUSION

This study highlights ChatGPT and Gemini's complementary strengths and weaknesses in answering frequently asked questions about patellofemoral instability. ChatGPT provided more evidence-based and accurate responses, making it a more potent tool for delivering precise medical information. However, its responses were more difficult to read, requiring a college-level education, which may limit accessibility for the general patient population. In contrast, Gemini produced clearer, more readable, and well-structured responses, but its content was less rigorously supported by evidence, raising concerns about transparency and reliability.

These findings align with previous research, which has shown that AI chatbots can enhance patient education but are not without limitations. The risk of misinformation, lack of full

guideline adherence, and inadequate citation of sources remain critical challenges that must be addressed before AI models can be fully trusted as standalone educational tools. For example, the high reading level of AI-generated responses may hinder comprehension among patients with lower health literacy, potentially leading to misinterpretation of critical medical information. Our study suggests that AI chatbots can supplement medical education but should not replace expert-reviewed resources or direct physician-patient interactions.

Future improvements should focus on enhancing AI model transparency, refining readability, and aligning with clinical guidelines. Healthcare professionals must validate AI-generated medical content, and mechanisms should be in place to ensure that patients receive accurate, contextually appropriate, and comprehensible health information. Specifically, future research should explore the integration of AI chatbots into clinical workflows, with a focus on developing user-friendly interfaces that can adapt to the health literacy levels of diverse patient populations.

DECLARATIONS

Ethics Committee Approval: The Antalya Training and Research Hospital Scientific Research Ethics Committee granted approval for this study (date: 09.01.2025, number: 1/25).

Informed Consent: Not applicable.

Conflict of Interest: The authors declared no conflict of interest.

Financial Disclosure:

Funding Disclosure: The authors declared that no fund was received for this study.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Use of AI for Writing Assistance: The authors declared that this study utilized ChatGPT-4o(OpenAI) and Gemini v2.1 (Google AI). AI-assisted tools were employed to compile frequently asked questions regarding patellofemoral instability, assess the content and readability of the responses, and assist in text revision. However, all scientific interpretations, data analyses, and final evaluations were conducted solely by the authors. The methodology and details regarding AI-assisted tools are further explained in the Materials and Methods section.

Author Contributions: Idea/Concept – VB, MY; Design – MY, FD; Control/Supervision – MBE, VB; Data Collection and/or Processing – MBE, MY; Analysis and/or Interpretation – MY, MBE; Literature review – VB, FD; Writing – VB, MY; Critical Review – VB, MBE, FD; References and Fundings – FD, MBE; Materials – VB, FD, MBE, MY.

Peer-review: Externally peer-reviewed.

ABBREVIATIONS

ACL – Anterior Cruciate Ligament
 AI – Artificial Intelligence
 AAOS – American Academy of Orthopedic Surgeons
 ARI – Automated Readability Index
 CI – Confidence Interval
 CPG – Clinical Practice Guidelines
 FRES – Flesch Reading Ease Score
 ICC – Intraclass Correlation Coefficient
 LLM – Large Language Model
 MPFL – Medial Patellofemoral Ligament
 PFI – Patellofemoral Instability
 SMOG – Simple Measure of Gobbledygook

REFERENCES

- Masroori Z, Haseli S, Abbaspour E, Pouramini A, Azhideh A, Fathi M, et al. Patellar non-traumatic pathologies: a pictorial review of radiologic findings. *Diagnostics (Basel)* 2024;16;14:2828. [\[CrossRef\]](#)
- Yuncu M, Egerci OF, Dogruoz F. Surgical treatment of symptomatic bipartite patella: a case report and review of the current literature. *Sports Traumatol Arthrosc* 2024;1:33–42. [\[CrossRef\]](#)
- Migliorini F, Maffulli N, Vaishya R. Patellofemoral instability: current status and future perspectives. *J Orthop* 2022;21;36:49–50. [\[CrossRef\]](#)
- Yapar D, Demir Avcı Y, Tokur Sonuvar E, Egerci ÖF, Yapar A. ChatGPT's potential to support home care for patients in the early period after orthopedic interventions and enhance public health. *Jt Dis Relat Surg* 2024;1;35:169–76. [\[CrossRef\]](#)
- Megafu M, Guerrero O, Yendluri A, Parsons BO, Galatz LM, Li X, et al; the Scientific Collaborative For Orthopaedic Research And Education (SCORE) Group. ChatGPT and Gemini are not consistently concordant with 2020 AAOS clinical practice guidelines when evaluating rotator cuff injury. *Arthroscopy* 2025;4:S0749–8063(25)00057-X. [\[CrossRef\]](#)
- Post WR, Fithian DC. Patellofemoral instability: a consensus statement from the AOSSM/PFF patellofemoral instability workshop. *Orthop J Sports Med* 2018;30;6:2325967117750352. [\[CrossRef\]](#)
- Liu JN, Steinhaus ME, Kalbian IL, Post WR, Green DW, Strickland SM, et al. Patellar instability management: a survey of the International Patellofemoral Study Group. *Am J Sports Med* 2018;46:3299–306. [\[CrossRef\]](#)
- Bailey MEA, Metcalfe A, Hing CB, Eldridge J; BASK Patellofemoral Working Group. Consensus guidelines for management of patellofemoral instability. *Knee* 2021;29:305–12. [\[CrossRef\]](#)
- Parikh SN, Schlechter JA, Veerkamp MW, Stacey JD, Gupta R, Pendleton AM et al; PRISM Patellofemoral Research Interest Group (PRISM PF RIG). Consensus-based guidelines for management of first-time patellar dislocation in adolescents. *J Pediatr Orthop* 2024;1;44:e369–74. [\[CrossRef\]](#)
- Ozduran E, Hancı V, Erkin Y, Özbek İC, Abdulkirimov V. Assessing the readability, quality and reliability of responses produced by ChatGPT, Gemini, and Perplexity regarding most frequently asked keywords about low back pain. *PeerJ* 2025;22:e18847. [\[CrossRef\]](#)
- Gunning R. *The Technique of Clear Writing*. New York: McGraw-Hill; 1952.
- Coleman M, Liao TL. A computer readability formula designed for machine scoring. *J Appl Psychol* 1975;60:283–4. [\[CrossRef\]](#)
- Flesch R. A new readability yardstick. *J Appl Psychol* 1948;32:221–33. [\[CrossRef\]](#)
- Smith EA, Senter RJ. Automated readability index. *AMRL TR* 1967;1–14.
- McLaughlin GH. SMOG grading—a new readability formula. *J Reading* 1969;12;639–46.
- Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63. [\[CrossRef\]](#)
- Behers BJ, Vargas IA, Behers BM, Rosario MA, Wojtas CN, Deever AC, et al. Assessing the readability of patient education materials on cardiac catheterization from artificial intelligence chatbots: an observational cross-sectional study. *Cureus* 2024;4;16:e63865. [\[CrossRef\]](#)
- Li LT, Sinkler MA, Adelstein JM, Voos JE, Calcei JG. ChatGPT responses to common questions about anterior cruciate ligament reconstruction are frequently satisfactory. *Arthroscopy* 2024;40:2058–66. [\[CrossRef\]](#)
- Gaudiani MA, Castle JP, Abbas MJ, Pratt BA, Myles MD, Moutzouros V, et al. ChatGPT-4 generates more accurate and complete responses to common patient questions about anterior cruciate ligament reconstruction than Google's search engine. *Arthrosc Sports Med Rehabil* 2024;9:100939. [\[CrossRef\]](#)
- Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and Gemini provide appropriate recommendations for pediatric orthopaedic conditions? *J Pediatr Orthop* 2025;45:e66–71. [\[CrossRef\]](#)
- Johns WL, Martinazzi BJ, Miltenberg B, Nam HH, Hammoud S. ChatGPT provides unsatisfactory responses to frequently

- asked questions regarding anterior cruciate ligament reconstruction. *Arthroscopy* 2024;40:2067–79. [\[CrossRef\]](#)
22. Nwachukwu BU, Varady NH, Allen AA, Dines JS, Altchek DW, Williams RJ 3rd, et al. Currently available large language models do not provide musculoskeletal treatment recommendations that are concordant with evidence-based clinical practice guidelines. *Arthroscopy* 2025;41:263–75. [\[CrossRef\]](#)
23. Quinn M, Milner JD, Schmitt P, Morrissey P, Lemme N, Marcaccio S, et al. Artificial intelligence large language models address anterior cruciate ligament reconstruction: superior clarity and completeness by Gemini compared with ChatGPT-4 in response to American Academy of Orthopaedic Surgeons Clinical Practice Guidelines. *Arthroscopy* 2024;21:S0749–8063(24)00736-9. [\[CrossRef\]](#)