JSES International 9 (2025) 390–397

Contents lists available at ScienceDirect

JSES International

journal homepage: www.jsesinternational.org

Can popular AI large language models provide reliable answers to frequently asked questions about rotator cuff tears?

Ulas Can Kolac, MD[a,1], Orhan Mete Karademir [b,1], Gokhan Ayik, MD[c], Mehmet Kaymakoglu, MD[d], Filippo Familiari, MD[e,f,*], Gazi Huri, MD[a,g]

[a]Department of Orthopedics and Traumatology, Hacettepe University Faculty of Medicine, Ankara, Turkey
[b]Faculty of Medicine, Hacettepe University, Ankara, Turkey
[c]Department of Orthopedics and Traumatology, Yuksek Ihtisas University Faculty of Medicine, Ankara, Turkey
[d]Department of Orthopedics and Traumatology, Faculty of Medicine, Izmir University of Economics, Izmir, Turkey
[e]Department of Orthopaedics, Magna Graecia University of Catanzaro, Italy, Catanzaro, Italy
[f]Research Center on Musculoskeletal Health, MusculoSkeletalHealth@UMG, Magna Graecia University, Catanzaro, Italy
[g]Aspetar, Orthopedic and Sports Medicine Hospital, FIFA Medical Center of Excellence, Doha, Qatar

ARTICLE INFO

Keywords:
Artificial intelligence
Large language models
Rotator cuff tears
Frequently asked questions
Patient information
AI Tools in Healthcare
ChatGPT

Level of evidence: Basic Science Study; Validation of AI in Patient Information

Background: Rotator cuff tears are common upper-extremity injuries that significantly impair shoulder function, leading to pain, reduced range of motion, and a decrease in quality of life. With the increasing reliance on artificial intelligence large language models (AI LLMs) for health information, it is crucial to evaluate the quality and readability of the information provided by these models.
Methods: A pool of 50 questions was generated related to rotator cuff tear by querying popular AI LLMs (ChatGPT 3.5, ChatGPT 4, Gemini, and Microsoft CoPilot) and using Google search. After that, responses from the AI LLMs were saved and evaluated. For information quality the DISCERN tool and a Likert Scale was used, for readability the Patient Education Materials Assessment Tool for Printable Materials (PEMAT) Understandability Score and the Flesch-Kincaid Reading Ease Score was used. Two orthopedic surgeons assessed the responses, and discrepancies were resolved by a senior author.
Results: Out of 198 answers, the median DISCERN score was 40, with 56.6% considered sufficient. The Likert Scale showed 96% sufficiency. The median PEMAT Understandability score was 83.33, with 77.3% sufficiency, while the Flesch-Kincaid Reading Ease score had a median of 42.05 with 88.9% sufficiency. Overall, 39.8% of the answers were sufficient in both information quality and readability. Differences were found among AI models in DISCERN, Likert, PEMAT Understandability, and Flesch-Kincaid scores.
Conclusion: AI LLMs generally cannot offer sufficient information quality and readability. While they are not ready for use in medical field, they show a promising future. There is a necessity for continuous re-evaluation of these models due to their rapid evolution. Developing new, comprehensive tools for evaluating medical information quality and readability is crucial for ensuring these models can effectively support patient education. Future research should focus on enhancing readability and consistent information quality to better serve patients.

Primary care and orthopedic surgeons frequently encounter rotator cuff tears (RCTs), making them one of the most prevalent upper-extremity injury. RCTs can significantly impair shoulder function, leading to pain, decreased range of motion, and an overall reduction in quality of life.[5,26]

The widespread occurrence of RCTs has led to various treatment options, from nonsurgical methods like physical therapy, corticosteroid injections, and NSAIDs to surgical interventions, including arthroscopic and open repairs. Despite advancements in surgical techniques and rehabilitation protocols, determining the most effective treatment strategy for each patient remains a complex and evolving challenge in orthopedic practice.[5,26]

With the growing reliance on smartphones and the internet, patients increasingly turn to the online resources to learn about their medical conditions and treatment options. The proliferation of online

Ethical approval was not required for this study as it did not involve animals, humans, or human tissues.

*Corresponding author: Filippo Familiari, MD, Department of Orthopaedics, Magna Graecia University of Catanzaro, Research Center on Musculoskeletal Health, MusculoSkeletalHealth@UMG, Department of Orthopedic and Trauma Surgery, "Magna Græcia" University, "Mater Domini" University Hospital, Viale Europa, 88100 Catanzaro, Italy.
E-mail address: filippofamiliari@unicz.it (F. Familiari).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jseint.2024.11.012

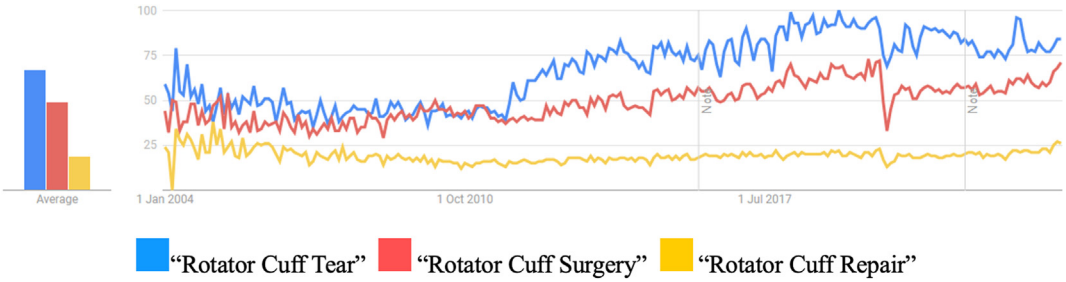

Figure 1 Google search trends between January 1, 2004, and March 31, 2024 (trends.google.com).

medical content and ways to reach the information has empowered patients to educate themselves about their health issues.[3,10,47] Analysis of search trends reveals that patients often seek information on "rotator cuff tear" more frequently than specific treatment-related terms like "rotator cuff surgery" or "rotator cuff repair" (Fig. 1). This trend suggests a significant interest among patients in understanding the nature and implications of RCTs before exploring potential treatments, making it imperative to provide them with accurate and comprehensive information about the tear itself.

In recent years, artificial intelligence large language models (AI LLMs) have become invaluable tools for generating human-like text and providing information. Notable models like ChatGPT (Open AI, San Francisco, CA, USA), Gemini (Google, Mountain View, CA, USA), and Microsoft CoPilot (Redmond, WA, USA) use advanced machine learning to deliver coherent and contextually relevant responses across various domains, including customer service, education, and health care. These sophisticated language models are trained on vast datasets to understand and generate human-like text using deep learning, particularly transformer neural networks.[9]

ChatGPT, developed by OpenAI in San Francisco, was the first generative pretrained transformer (GPT) model released, rapidly rising in popularity due to its dynamic and contextually aware conversational abilities.[13,30] Its success demonstrated the potential of AI in providing detailed and accurate responses across a wide range of topics. Following this, other companies released their own AI LLMs. Google introduced Gemini, originally known as Bard, based on the LaMDA family of language models, and integrated it into Google's suite of services, including replacing Google Assistant on Android devices, to leverage Google's extensive data resources.[45] Microsoft launched CoPilot, initially known as Bing Chat and built on a version of the GPT-4 model, integrating it into several Microsoft products, including Windows 11, to offer versatile assistance directly through the taskbar.[33] Together, these AI LLMs enhance user interactions, provide comprehensive information, and perform a wide range of tasks from casual conversation to addressing complex inquiries.[30,32,33]

The rapid adoption of AI LLMs raises important questions about the quality and reliability of the information they provide.[23,40] Accurate and comprehensible medical information is crucial for patients making health-related decisions.[29] However, the variable quality of online health information, which is used to train these models, coupled with the potential for misinformation, underscores the need for rigorous evaluation of these AI tools.[18,29,40] Assessing their effectiveness in delivering high-quality, readable, and reliable medical information is essential for ensuring that patients receive the support they need.[18,23]

Our study aims to evaluate the capability of AI LLMs in answering questions related to RCTs. Specifically, we investigate the quality and readability of information provided by popular AI models. We hypothesized that, information quality and readability of the answers would not be sufficient for use in patient education.

Materials and methods

Question generation and data collection

In April 26, 2024, an initial pool of 100 questions was generated using 2 sources: querying the AI LLMs (ChatGPT 3.5, ChatGPT 4, Gemini, Microsoft CoPilot) with "Most frequently asked 20 questions about rotator cuff injury by patients" and entering the term "rotator cuff tear" into Google to obtain the first 20 questions from the "People also ask" section. For this purpose, a laptop was used after a factory reset and with a new installation of Windows 11 (Version 23H2, Build 22631.3447; Microsoft Corp., Redmond, WA, USA) with Microsoft Edge (Version 124.0.2478.51; Microsoft Corp., Redmond, WA, USA). The browser cache was cleaned, and the history was deleted. A new account was created and used to access the AI LLMs. The initial set of 100 questions was saved to an Excel sheet (Microsoft Excel; Microsoft Corp., Redmond, WA, USA) (Supplementary Appendix S1). This list included frequently asked questions on rotator cuff injuries identified by AI as well as common patient inquiries from clinical experience. Clinicians then reviewed the list to refine and eliminate similar or irrelevant questions, resulting in a final set of 50 clinically relevant questions (Supplementary Appendix S2).

Each of the 50 questions was input into the 4 AI LLMs, and their responses were recorded into the Excel sheet (Supplementary Appendix S2). The evaluation process involved assessing the information quality and readability of each response by using DISCERN Score and Likert Scale for Information Quality, Patient Education Materials Assessment Tool for Printable Materials (PEMAT) Understandability Score, and Flesch-Kincaid Reading Ease Score for Readability. Responses were evaluated by 2 orthopedic surgeons specializing in shoulder surgery (U.K, G.A), any discrepancies in scoring were consulted to the senior author (G.H), and the final scores were recorded into the Excel sheet (Supplementary Appendix S3).

Question categorization

To categorize the questions, a modified Rothwell's classification system was used.[34] Rothwell's classification system categorizes questions into 3 main themes: Fact, Policy, and Value.[27,34] Fact questions ask whether something is true and to what extent; Policy questions ask whether a certain course of action should be taken; and Value questions ask for the evaluation of an idea, object, or event.[19,34] Each theme is further divided into subcategories to provide a more detailed classification of the questions.[21,27,34,37] To suit our needs, we modified the classification system used by similar researches. While other studies focused more on treatment, our questions were mostly related to the disease itself.[19,30,44] We added the Risk/Complication subcategory to the Fact category to differentiate between the risks associated with treatment options

Table I
Modified Rothwell's Classification with categories and subcategories.

Fact	Ask objective, factual information
General information	
Timeline of treatment	Length of time to achieve goals/milestones
Specific activities	Particular activities
Risk/complication	Risks and complications related to disease
Policy	Ask information about course of action to solve the problem
Indication/management	Indications, and selection of procedures
Risk/complication	Risks and complications related to treatment methods
Value	Evaluation of an action, idea
Evaluation	Questions about levels of satisfaction/success

and those related to the disease itself (Table I). Questions were categorized according to this system (Supplementary Appendix S4).

Information quality assessment

Information quality was evaluated using 2 main tools: the DISCERN tool and a Likert Scale. The DISCERN tool, which consists of 16 questions, rates the reliability of written health information on a scale from very poor (16-26) to excellent (63+).[8] This tool is designed to help users assess the quality of written information on treatment choices by evaluating aspects such as the clarity of aims, the relevance of sources, and the description of treatment benefits and risks.[4,8] The rating scale for each question ranges from 1 to 5, where 1 indicates "No" (the criterion is not fulfilled by the publication) and 5 indicates "Yes" (the criterion is fulfilled by the publication).[8] After assigning scores to each question, the scores are summed to obtain a total score, which can range from a minimum of 5 to a maximum of 80, with higher scores indicating higher quality[4,6,8] (Table II). The Likert Scale was used to rate each response based on its accuracy against medical literature.[25] The typical Likert scale is a 5- or 7-point ordinal scale where respondents rate their agreement with a statement.[25] Responses can be ranked, but the distances between them are not equal. For example, the difference between "always," "often," and "sometimes" is not necessarily the same, even if they are assigned different numbers.[25,43] The responses were given a score between 1 and 5 (Table III). An answer was considered good in the information quality area if it passed the both cutoff points of 39 points for DISCERN and 3 points for the Likert Scale.[8,25,43]

Readability assessment

Readability was assessed using 2 tools: the PEMAT Understandability Score and the Flesch-Kincaid Reading Ease score.[22,38] The PEMAT tool evaluates patient education materials for understandability and actionability, with acceptable thresholds set at 75% for understandability and 60% for actionability.[38] Each PEMAT item is rated as "Disagree (0)" or "Agree (1)," with some items allowing a "Not Applicable (N/A)" option. The total points for understandability or actionability are summed, divided by the number of applicable items, and multiplied by 100 to yield a percentage score [1,14,36,38](Table IV). The Flesch-Kincaid Reading Ease score measures readability based on sentence length and syllable count, with scores ranging from 0 to 100, where higher scores indicate easier readability.[22] These scores can also be categorized into grade level brackets corresponding to US school levels [17,22,36,41,42] (Table V). An answer was considered good in readability if it met both cutoff points of 75% for PEMAT and 30 points for Flesch-Kincaid Reading Ease.[22,38,41]

Table II
Questions of DISCERN instrument.

Question no	Question	Score				
1	Are the aims clear?	1	2	3	4	5
2	Does it achieve its aims?	1	2	3	4	5
3	Is it relevant?	1	2	3	4	5
4	Is it clear what sources of information were used to compile the publication (other than the author or producer)?	1	2	3	4	5
5	Is it clear when the information used or reported in the publication was produced?	1	2	3	4	5
6	Is it balanced and unbiased?	1	2	3	4	5
7	Does it provide details of additional sources of support and information?	1	2	3	4	5
8	Does it refer to areas of uncertainty?	1	2	3	4	5
9	Does it describe how each treatment works?	1	2	3	4	5
10	Does it describe the benefits of each treatment?	1	2	3	4	5
11	Does it describe the risks of each treatment?	1	2	3	4	5
12	Does it describe what would happen if no treatment is used?	1	2	3	4	5
13	Does it describe how the treatment choices affect overall quality of life?	1	2	3	4	5
14	Is it clear that there may be more than one possible treatment choice?	1	2	3	4	5
15	Does it provide support for shared decision-making?	1	2	3	4	5
16	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	1	2	3	4	5

Table III
Five-point Likert scale.

Point	Explanation
1	There is incorrect information, and no relevant information regarding the question has been provided.
2	While there is correct information related to the question, there is also incorrect information.
3	There is correct information related to the question, but there are many omissions, unnecessary or confusing information.
4	There is correct information related to the question, but there are minimal omissions and unnecessary information.
5	Complete and accurate information related to the question is provided.

Statistical analysis

The data were analyzed using IBM SPSS V23 (IBM Corp., Armonk, NY, USA). The normality of distribution was examined with the Shapiro-Wilk test. Repeated analysis of variance was used to compare normally distributed data according to AI LLMs, and multiple comparisons were examined with the Bonferroni test. The Friedman test was used for comparing non-normally distributed data according to AI LLMs, and multiple comparisons were examined with the Dunn test. For normally distributed data between 2 independent groups, the Independent 2-sample t-test was used, and for non-normally distributed data, the Mann-Whitney U test was used. The McNemar test was used for comparing dependent 2-group categorical variables. The Chi-square test was used to examine the association between information quality and readability. The significance level was set at $P < .050$.

Results

For the 50 questions, a total of 198 answers were collected, with Google Gemini refusing to answer 2 questions. According to the

Table IV
PEMAT understandability and actionability questions.

Understandability		
Question no	Question	Answer
	Content	Disagree = 0, Agree = 1
1	The material makes its purpose completely evident.	Disagree = 0, Agree = 1
2	The material does not include information or content that distracts from its purpose	Disagree = 0, Agree = 1
	Word choice & style	
3	The material uses common, everyday language.	Disagree = 0, Agree = 1
4	Medical terms are used only to familiarize audience with the terms. When used, medical terms are defined.	Disagree = 0, Agree = 1
5	The material uses the active voice.	Disagree = 0, Agree = 1
	Use of numbers	
6	Numbers appearing in the material are clear and easy to understand.	Disagree = 0, Agree = 1, No numbers = N/A
7	The material does not expect the user to perform calculations.	Disagree = 0, Agree = 1
	Organization	
8	The material breaks or "chunks" information into short sections.	Disagree = 0, Agree = 1, Very short material* = N/A
9	The material's sections have informative headers.	Disagree = 0, Agree = 1, Very short material* = N/A
10	The material presents information in a logical sequence.	Disagree = 0, Agree = 1
11	The material provides a summary.	Disagree = 0, Agree = 1, Very short material* = N/A
	Layout & design	
12	The material uses visual cues (e.g., arrows, boxes, bullets, bold, larger font, highlighting) to draw attention to key points.	Disagree = 0, Agree = 1 Video = N/A
	Use of visual aids	
15	The material uses visual aids whenever they could make content more easily understood (e.g., illustration of healthy portion size).	Disagree = 0, Agree = 1
16	The material's visual aids reinforce rather than distract from the content.	Disagree = 0, Agree = 1, No visual aids = N/A
17	The material's visual aids have clear titles or captions.	Disagree = 0, Agree = 1, No visual aids = N/A
18	The material uses illustrations and photographs that are clear and uncluttered.	Disagree = 0, Agree = 1, No visual aids = N/A
19	The material uses simple tables with short and clear row and column headings.	Disagree = 0, Agree = 1, No tables = N/A
Actionability		
Question no	Question	Answer
20	The material clearly identifies at least 1 action the user can take.	Disagree = 0, Agree = 1
21	The material addresses the user directly when describing actions.	Disagree = 0, Agree = 1
22	The material breaks down any action into manageable, explicit steps.	Disagree = 0, Agree = 1
23	The material provides a tangible tool (e.g., menu planners, checklists) whenever it could help the user take action.	Disagree = 0, Agree = 1
24	The material provides simple instructions or examples of how to perform calculations.	Disagree = 0, Agree = 1, No calculations = N/A
25	The material explains how to use the charts, graphs, tables, or diagrams to take actions.	Disagree = 0, Agree = 1, No charts, graphs, tables, or diagrams = N/A
26	The material uses visual aids whenever they could make it easier to act on the instructions.	Disagree = 0, Agree = 1

PEMAT, Patient Education Materials Assessment Tool for Printable Materials.

modified Rothwell classification, 60% of the questions fell under Fact, 32% under Policy, and 8% under Value. Specifically, 22% of the questions were categorized as Fact–Risk/Complication, and 4% as Policy–Risk/Complication. For statistical analysis, these subcategories were combined when deemed beneficial, resulting in 26% of the questions being classified under Risk/Complication (Table VI).

Information quality

The median DISCERN score for all answers was 40 (range: 23-62). A total of 56.6% of the answers were considered sufficient, having a score of 39 or higher. According to the Likert Scale, 96% of the answers were considered sufficient, having a score of 3 or higher. Overall, 54.5% of the answers were considered sufficient in both DISCERN and Likert scoring systems.

Readability

The median PEMAT Understandability score was 83.33 (range: 50-91.67). A total of 77.3% of the answers were considered sufficient, having a score of 75% or higher. The median for PEMAT Actionability score was 0 (range: 0-60). Only 8 answers from ChatGPT 3.5, 6 answers from ChatGPT 4, 7 answers from Gemini, and 5 answers from Microsoft CoPilot were considered sufficient, making up 13.1% of all the answers. The median Flesch-Kincaid Reading Ease score was 42.05, (range: 13.5-68.5). In this category, 88.9% of the answers were considered sufficient, having a score of 30 or higher. Regarding the Flesch-Kincaid Grade Level, 8% of the answers corresponded to the eighth–ninth-grade level, 15.1% to the 10th–12th-grade level, 65.6% to the college level, and 11.1% to the graduate college level. Notably, Gemini did not have any

Table V
Flesch-Kincaid grade levels and Flesch-Kincaid reading ease formula.

Score	School level (US)	Notes
100-90	Fifth grade	Very easy to read. Easily understood by an average 11-year-old student.
90-80	Sixth grade	Easy to read. Conversational English for consumers.
80-70	Seventh grade	Fairly easy to read.
70-60	Eighth & ninth grade	Plain English. Easily understood by 13- to 15-year-old students.
60-50	10th to 12th grade	Fairly difficult to read.
50-30	College	Difficult to read.
30-10	College Graduate	Very difficult to read. Best understood by university graduates.
10-0	Professional	Extremely difficult to read. Best understood by university graduates.

$$\text{Flesch-Kincaid Reading Ease Score} = 206.835 - \left(1.015 \times \frac{\text{Total Words}}{\text{Total Sentences}}\right) - \left(84.6 \times \frac{\text{Total Syllables}}{\text{Total Words}}\right).$$

Table VI
Categories and subcategories of the questions according to the modified Rothwell Classification.

	Frequency (n)	Percentage (%)
Main categories		
Fact	30	60
Policy	16	32
Value	4	8
Subcategories		
Fact: general information	11	22
Fact: timeline of treatment	4	8
Fact: specific activities	4	8
Fact/policy: risk/complication	13	26
Policy: indication/management	14	28
Value: evaluation	4	8

answers measured at the graduate college level. Instead, 22.9% of Gemini's answers were at the eighth–ninth-grade level and 35.4% at the 10th–12th-grade level, showing a significant difference from the others ($P < .001$). Overall, 72.2% of the answers were considered sufficient in both PEMAT Understandability and Flesch-Kincaid Reading Ease scoring systems.

Overall sufficiency

When considering both areas of evaluation, 39.8% of the answers were deemed sufficient in both information quality and readability (Table VII).

Comparison of AI LLMs

Significant differences were found in the distribution of DISCERN ($P < .001$), Likert ($P = .002$), PEMAT Understandability ($P = .005$) and Flesch-Kincaid Reading Ease Scores ($P < .001$). For PEMAT Understandability, the median scores were 83.33 for all AI LLMs, but the rank averages were 2.36 for ChatGPT 3.5, 2.66 for ChatGPT 4, 2.85 for Gemini, and 2.13 for Microsoft CoPilot. No significant differences were found in the median PEMAT Actionability scores among the AI LLMs ($P = .63$) (Table VIII).

Discussion

Our findings indicate that AI LLMs generally cannot offer sufficient information quality, as evidenced by the DISCERN Score, Likert Scale, PEMAT Understandability Score, and Flesch-Kincaid Reading Ease Score. However, there remains significant room for improvement, particularly in the areas of word and sentence complexity. While a majority of the responses met our cut-off for the Flesch-Kincaid Reading Ease Score, they fell short of the American Medical Association's and National Institutes of Health's recommended score of 80 or higher, suggesting that the responses were not as easily readable as they should be.[7,39] Another notable

Table VII
Sufficiency of answers in information quality and readability.

	Information quality	
	Not sufficient	Sufficient
Readability		
Not Sufficient	26 (%13.1)	29 (%14.6)
Sufficient	64 (%32.3)	79 (%39.8)

setback for the AI LLMs is the lack of source citations in their responses, except for Microsoft CoPilot, which consistently cites sources after every answer.

When we look at the similar studies across different specializations, Haidar O et al reported both poor information quality and poor readability in their study on vascular patients, emphasizing the need for patients to rely on reputable, human-generated information.[16] Onder CE et al, in their study on hypothyroidism during pregnancy, similarly found good information quality but noted that poor readability limits its usability for the general public, cautioning physicians about its limitations despite the good information quality.[31] Yuce A et al also observed good information quality but highlighted the need for consistent performance.[49] Sahin S et al confirmed good information quality in their study on patient education for sports surgery, but raised concerns about poor readability and incomplete or misleading information.[35] Shen SA et al reported similar readability issues and suggested that additional prompting could enhance readability.[36] Hershenhouse JS et al, in their study on prostate cancer, noted that while AI LLMs are not yet designed for this purpose and are not ready for immediate deployment, they show promise for future use in the medical field.[17] In summary, while there are a few who disagree, most studies indicate that AI LLMs are not yet ready but hold a promising future for use in the medical field.

Compared to similar recent studies, our research highlights several important aspects and offers a more comprehensive and robust evaluation. While we concur with the findings of Günay AE et al, who assessed ChatGPT's information quality for rotator cuff injuries and noted its good information quality despite the lack of citations, their study was constrained by a small sample size and the use of DISCERN and Journal of American Medical Association tools for assessment.[15] Although DISCERN and Journal of American Medical Association criteria are valuable, we believe a comparison between answers and reliable sources should be conducted. Moreover, while they used 5 readability tools, these tools only provide 1 perspective on readability, overlooking the critical aspect of comprehension emphasized by Walters KA et al.[46] Additionally, we agree with the satisfactory results of ChatGPT's information quality noted by Megalla M et al, Li LT et al, and Johns WL et al, all of which employed similar methodologies. However, these studies were also limited by low sample sizes and the use of only 1 tool for information quality assessment without any readability analysis.[20,24,28] While information quality is essential, it is

Table VIII
Comparison of AI language models based on each score.

Scores	AI language model	Median (minimum-maximum)	Test statistics	*P* value
DISCERN	ChatGPT 3.5	41.00 (28.00-56.00)b*	65,413	**<.001**
	ChatGPT 4	38.00 (23.00-51.00)ab*		
	Gemini	36.00 (24.00-57.00)a*		
	Microsoft CoPilot	48.00 (32.00-62.00)c*		
Likert scale	ChatGPT 3.5	5.00 (2.00-5.00)	15,207	**.002**
	ChatGPT 4	5.00 (2.00-5.00)		
	Gemini	5.00 (1.00-5.00)		
	Microsoft CoPilot	4.00 (1.00-5.00)		
PEMAT understandability	ChatGPT 3.5	83.33 (50.00-91.67)ab*	12,897	**.005**
	ChatGPT 4	83.33 (55.56-91.67)ab*		
	Gemini	83.33 (58.33-87.50)a*		
	Microsoft CoPilot	83.33 (50.00-84.62)b*		
Flesch-Kincaid reading ease	ChatGPT 3.5	35.15 (17.10-67.70)b*	38,517	**<.001**
	ChatGPT 4	41.80 (13.50-61.10)b*		
	Gemini	53.40 (31.60-68.50)a*		
	Microsoft CoPilot	43.65 (20.20-62.70)b*		
PEMAT actionability	ChatGPT 3.5	0.00 (0.00-60.00)	1731	.630
	ChatGPT 4	0.00 (0.00-60.00)		
	Gemini	0.00 (0.00-60.00)		
	Microsoft CoPilot	0.00 (0.00-60.00)		

AI, artificial intelligence; *PEMAT*, Patient Education Materials Assessment Tool for Printable Materials.
The bold *P*-values highlight statistically significant results, indicating *P*-values less than .05.
a-c: There is no significant difference among tools with the same letter.
*Friedman test.

equally crucial to assess the readability of AI LLMs responses because if patients cannot comprehend the information, its quality becomes irrelevant. Moreover, all the previously mentioned studies lack the assessment of different AI LLMs. While ChatGPT is well-known, it is only 1 example, and other LLMs might be more suitable for the health-care industry.

A noteworthy finding observed in this study was Google Gemini's refusal to answer 2 specific questions: "What causes rotator cuff tears?" and "How do age and health affect the treatment of rotator cuff tears?" The model responded with statements such as, "As a language model, I'm not able to assist you with that," indicating an inherent restriction in handling certain medical inquiries. This may be due to built-in limitations within Gemini's programming or potential concerns about providing complex medical advice. Such refusals highlight the need for careful evaluation of AI models in medical contexts, as they may lack the ability to address critical health-related questions consistently. This limitation highlights the importance of supplementing AI responses with human oversight, especially for topics requiring detailed, individualized medical advice.

In addition, it is important to note the rapid evolution of AI technology and the dynamic nature of AI LLMs. We believe these models will continue to improve significantly each year. Therefore, continuous re-evaluation of these tools is necessary to ensure that patients have access to adequate and accurate information. This ongoing assessment is crucial as it helps to identify any improvements or shortcomings in AI LLMs, ensuring they remain reliable sources of health information for patients and health-care providers.

However, it is too early to say that AI LLMs are ready for patient use. Patients often place significant trust in the information they find on the internet, which can lead to them questioning or arguing with their physicians and potentially losing trust in their medical advice.[11] The widespread lack of understanding about what AI is and the tendency to believe AI LLMs' answers as completely accurate exacerbates this issue.[2] On top of that, the rising popularity of complementary and alternative medicine means that people are increasingly trying to treat themselves without visiting a physician.[12,48] Having a source of information available 24/7 that can make mistakes and speaks with apparent certainty poses significant risks. For AI LLMs to be used effectively and safely in medicine, they must achieve an "excellent" standard rather than merely "good." This requires consistently providing information that is accurate, reliable, and presented in a way that patients can easily understand and apply. In the medical field, trustworthiness is paramount due to the serious consequences that misinformation or misunderstandings can have on patient health. Therefore, excellence in AI-generated medical information demands that these models meet rigorous standards across various assessment tools, ensuring content that is both readable and relevant. For independent patient use, AI LLMs would need to reach specific benchmarks: high accuracy in line with current medical guidelines, readability suited to a general patient audience (e.g., Flesch-Kincaid scores between 60 and 80, ideal for a sixth- to eighth-grade reading level), and quality consistency across a range of medical topics. Additionally, providing transparent citations from credible sources is crucial for verification. Meeting these standards would make AI-generated responses trustworthy, comprehensible, and reliable, thereby supporting safe, informed patient decision-making.

It is also important to note that most existing scoring systems are tailored to evaluate information related to treatment options. However, as we stated previously, patients often seek information about their conditions rather than treatment choices, highlighting a gap in the current evaluative frameworks. This underscores the need for developing new, more comprehensive tools that can accurately assess the quality of medical information across various domains.

Limitations

The study's sample size is relatively small, which may limit the generalizability of the findings. Additionally, due to the small sample size, the study did not analyze results across different categories.

The cut-off points for the scores were chosen by the researchers, which may influence the interpretation of the results. Changing them can lead to different conclusions.

Multiple evaluation tools which can easily be influenced by individual opinions were used. Replicating this study can give different results due to subjective opinions.

As stated previously, evaluation tools used in the study are not designed primarily for this purpose. They may not fully capture the true results.

The performance of AI LLMs can vary significantly, and the study only evaluates 4 specific models. This may not provide a comprehensive overview of all available AI LLMs.

AI models are trained on vast datasets available on the internet, which include variable quality information. Replicating the study at a different time may yield different results due to changes in the data available on the internet.

AI technology is rapidly evolving, and the models evaluated in the study may soon be outdated. Replicating the study at a different time, with newer models may produce different results. However, we strongly believe that results will only improve over time.

Conclusion

AI LLMs are not yet ready for independent patient use, but they show promising results and have a bright future. With their rapid evolution, continuous re-evaluation is essential. New, tailored evaluation tools are needed for accurate assessment. Future research should focus on improving readability for general public use and enhancing information quality to provide consistent, sufficient answers to patients

Disclaimers:

Funding: Open access funding was provided by the Università degli Studi Magna Graecia di Catanzaro within the CRUI-CARE Agreement. Conflicts of interest: The authors, their immediate families, and any research foundations with which they are affiliated have not received any financial payments or other benefits from any commercial entity related to the subject of this article.

Availability of data and materials

The data and materials that support the findings of this study are available from the corresponding author, [U.K], upon reasonable request.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jseint.2024.11.012.

References

1. Abdullah Y, Alokozai A, Mathew AJ, Stamm MA, Mulcahey MK. Patient education materials found via Google search for shoulder arthroscopy are written at too-high of a reading level. Arthrosc Sports Med Rehabil 2022;4:e1575-9. https://doi.org/10.1016/j.asmr.2022.04.034.
2. Alizadeh F, Stevens G, Esau M. I Don't Know, is AI also used in Airbags? An Empirical study of Folk Concepts and People's Expectations of current and future artificial intelligence. i-com 2021;20:3-17. https://doi.org/10.1515/icom-2021-0009.
3. Atkinson NL, Saperstein SL, Pleis J. Using the internet for health-related activities: findings from a national probability sample. J Med Internet Res 2009;11, e4. https://doi.org/10.2196/jmir.1035.
4. Barrett DR, Boone JD, Butch JO, Cavender JA, Sole G, Wassinger CA. A critical appraisal of web-based information on shoulder pain comparing biomedical vs. psychosocial information. J Shoulder Elbow Surg 2023;32:e23-32. https://doi.org/10.1016/j.jse.2022.07.023.
5. Bedi A, Bishop J, Keener J, Lansdown DA, Levy O, MacDonald P, et al. Rotator cuff tears. Nat Rev Dis Primers 2024;10:8. https://doi.org/10.1038/s41572-024-00492-3.
6. Bethell MA, Anastasio AT, Taylor JR, Tabarestani TQ, Klifto CS, Anakwenze O. Evaluating the distribution, quality, and educational value of videos related to shoulder instability exercises on the social media platform TikTok. J Am Acad Orthop Surg Glob Res Rev 2023;7. https://doi.org/10.5435/JAAOSGlobal-D-23-00034.
7. Bresolin LB. Health literacy: report of the council on scientific affairs 1999;281:552-7. https://doi.org/10.1001/jama.281.6.552.
8. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53:105-11.
9. De Angelis L, Baglivo F, Arzilli G, Privitera GP, Ferragina P, Tozzi AE, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. Front Public Health 2023;11, 1166120. https://doi.org/10.3389/fpubh.2023.1166120.
10. Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the Internet for medical information. J Gen Intern Med 2002;17:180-5. https://doi.org/10.1046/j.1525-1497.2002.10603.x.
11. Erdem SA, Harrison-Walker LJ. The role of the Internet in physician–patient relationships: the issue of trust. Bus Horiz 2006;49:387-93. https://doi.org/10.1016/j.bushor.2006.01.003.
12. Ernst E. Rise in popularity of complementary and alternative medicine: reasons and consequences for vaccination. Vaccine 2001;20:S90-3.
13. Giorgino R, Alessandri-Bonetti M, Luca A, Migliorini F, Rossi N, Peretti GM, et al. ChatGPT in orthopedics: a narrative review exploring the potential of artificial intelligence in orthopedic practice. Front Surg 2023;10, 1284015. https://doi.org/10.3389/fsurg.2023.1284015.
14. Gulbrandsen MT, O'Reilly OC, Gao B, Cannon D, Jesurajan J, Gulbrandsen TR, et al. Health literacy in rotator cuff repair: a quantitative assessment of the understandability of online patient education material. JSES Int 2023;7:2344-8. https://doi.org/10.1016/j.jseint.2023.06.016.
15. Günay AE, Özer A, Yazıcı A, Sayer G. Comparison of chat GPT versions in informing patients with rotator cuff injuries. JSES Int 2024;8:1016-8. https://doi.org/10.1016/j.jseint.2024.04.016.
16. Haidar O, Jaques A, McCaughran PW, Metcalfe MJ. AI-generated information for vascular patients: assessing the standard of Procedure-specific information provided by the ChatGPT AI-language model. Cureus 2023;15, e49764. https://doi.org/10.7759/cureus.49764.
17. Hershenhouse JS, Mokhtar D, Eppler MB, Rodler S, Storino Ramacciotti L, Ganjavi C, et al. Accuracy, readability, and understandability of large language models for prostate cancer information to the public. Prostate Cancer Prostatic Dis 2024;14. https://doi.org/10.1038/s41391-024-00826-y.
18. Hopkins AM, Logan JM, Kichenadasse G, Sorich MJ. Artificial intelligence chatbots will revolutionize how cancer patients access information: ChatGPT represents a paradigm-shift. JNCI Cancer Spectr 2023;7:pkad010. https://doi.org/10.1093/jncics/pkad010.
19. Hurley ET, Crook BS, Lorentz SG, Danilkowicz RM, Lau BC, Taylor DC, et al. Evaluation high-quality of information from ChatGPT (artificial intelligence-large language model) artificial intelligence on shoulder stabilization surgery. Arthroscopy 2024;40:726-731.e6. https://doi.org/10.1016/j.arthro.2023.07.048.
20. Johns WL, Kellish A, Farronato D, Ciccotti MG, Hammoud S. ChatGPT can offer satisfactory responses to common patient questions regarding elbow ulnar collateral ligament reconstruction. Arthrosc Sports Med Rehabil 2024;6, 100893. https://doi.org/10.1016/j.asmr.2024.100893.
21. Kanthawala S, Vermeesch A, Given B, Huh J. Answers to health questions: internet search results versus online health Community responses. J Med Internet Res 2016;18:e95. https://doi.org/10.2196/jmir.5369.
22. Kincaid JP, Naval Technical training Command Millington TN. Research Branch. Derivation of new readability Formulas (Automated readability Index, Fog count and Flesch reading Ease Formula) for Navy Enlisted Personnel. Reports - research. National Technical information service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF. Report No.: ED108134. 1975.
23. Lekadir K, Quaglio G, Garmendia A, Gallin C. Artificial intelligence in healthcare-applications, risks, and ethical and societal impacts. European Parliament; 2022. Artificial Intelligence in Healthcare-Applications, Risks, and Ethical and Societal Impacts[Google Scholar]. 2022. European Parliament: Brussels, Belgium.
24. 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. https://doi.org/10.1016/j.arthro.2023.12.009.
25. Likert R. A technique for the measurement of attitudes. Arch Psychol. 1932;140:5-55.
26. May T, Garmel GM. Rotator cuff injury. In: StatPearls. Treasure Island. FL *StatPearls Publishing*; 2024.
27. McCormick JR, Kruchten MC, Mehta N, Damodar D, Horner NS, Carey KD, et al. Internet search analytics for shoulder arthroplasty: what questions are patients asking? Clin Shoulder Elb 2023;26:55-63. https://doi.org/10.5397/cise.2022.01347.
28. Megalla M, Hahn AK, Bauer JA, Windsor JT, Grace ZT, Gedman MA, et al. ChatGPT and Google provide mostly excellent or satisfactory responses to the most frequently asked patient questions related to rotator cuff repair. Arthrosc Sports Med Rehabil 2024;6:100963. https://doi.org/10.1016/j.asmr.2024.100963.
29. Menz BD, Modi ND, Sorich MJ, Hopkins AM. Health disinformation use case highlighting the urgent need for artificial intelligence vigilance: weapons of mass disinformation. JAMA Intern Med 2024;184:92-6. https://doi.org/10.1001/jamainternmed.2023.5947.
30. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT responses to common patient questions regarding total Hip arthroplasty. J Bone Joint Surg Am 2023;105:1519-26. https://doi.org/10.2106/JBJS.23.00209.
31. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. Sci Rep 2024;14:243. https://doi.org/10.1038/s41598-023-50884-w.

32. Rossettini G, Cook C, Palese A, Pillastrini P, Turolla A. Pros and cons of using artificial intelligence chatbots for musculoskeletal rehabilitation management. J Orthop Sports Phys Ther 2023;53:728-34. https://doi.org/10.2519/jospt.2023.12000.
33. Rossettini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. BMC Med Educ 2024;24:694. https://doi.org/10.1186/s12909-024-05630-9.
34. Rothwell JD. In mixed company : Small group communication. 7th ed. Boston, MA: Cengage Learning; 2010. p. 468. 9780495567677 (pbk.). 0495567671 (pbk.).
35. Sahin S, Tekin MS, Yigit YE, Erkmen B, Duymaz YK, Bahsi I. Evaluating the success of ChatGPT in addressing patient questions concerning Thyroid surgery. J Craniofac Surg 2024;35:e572-5. https://doi.org/10.1097/SCS.0000000000010395.
36. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? Eur Arch Oto-Rhino-Laryngol 2024;281:3219-25. https://doi.org/10.1007/s00405-024-08524-0.
37. Shen TS, Driscoll DA, Islam W, Bovonratwet P, Haas SB, Su EP. Modern internet search analytics and total Joint arthroplasty: what are patients asking and reading online? J Arthroplasty 2021;36:1224-31. https://doi.org/10.1016/j.arth.2020.10.024.
38. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96:395-403. https://doi.org/10.1016/j.pec.2014.05.027.
39. Skalitzky MK, Gulbrandsen TR, Lorentzen W, Gao B, Shamrock AG, Weinstein SL, et al. Health literacy in Clubfoot: a quantitative assessment of the readability, understandability and actionability of online patient education material. Iowa Orthop J 2021;41:61-7.
40. Sorich MJ, Menz BD, Hopkins AM. Quality and safety of artificial intelligence generated health information. BMJ 2024;384:q596. https://doi.org/10.1136/bmj.q596.
41. Stelzer JW, Wellington IJ, Trudeau MT, Mancini MR, LeVasseur MR, Messina JC, et al. Readability assessment of patient educational materials for shoulder arthroplasty from top academic orthopedic institutions. JSES Int 2022;6:44-8. https://doi.org/10.1016/j.jseint.2021.08.004.
42. Sudah SY, Faccone RD, Manzi JE, Kirchner G, Constantinescu D, Nicholson A, et al. Most patient education materials on shoulder conditions from the American Academy of Orthopaedic Surgeons exceed recommended readability levels. JSES Int 2023;7:126-31. https://doi.org/10.1016/j.jseint.2022.09.004.
43. Sullivan GM, Artino AR Jr. Analyzing and interpreting data from likert-type scales. J Grad Med Educ 2013;5:541-2. https://doi.org/10.4300/JGME-5-4-18.
44. Tharakan S, Klein B, Bartlett L, Atlas A, Parada SA, Cohn RM. Do ChatGPT and Google differ in answers to commonly asked patient questions regarding total shoulder and total elbow arthroplasty? J Shoulder Elbow Surg 2024;33:e429-37. https://doi.org/10.1016/j.jse.2023.11.014.
45. Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. J Orthop Surg Res 2024;19:574. https://doi.org/10.1186/s13018-024-04996-2.
46. Walters KA, Hamrell MR. Consent forms, lower reading levels, and using Flesch-Kincaid readability software. Drug Inf J 2008;42:385-94. https://doi.org/10.1177/009286150804200411.
47. Wong C, Harrison C, Britt H, Henderson J. Patient use of the internet for health information. Aust Fam Physician 2014;43:875-7.
48. Yılmaz S. Türk Toplumunda Geleneksel Tedavi Yöntemlerinin Faydasına İnanma ve Bu Yöntemlere Başvurma Örüntüleri. Ordu Üniversitesi Sosyal Bilimler Enstitüsü Sosyal Bilimler Araştırmaları Dergisi 2020;10:941-53. https://doi.org/10.48146/odusobiad.809481.
49. Yuce A, Erkurt N, Yerli M, Misir A. The potential of ChatGPT for high-quality information in patient education for sports surgery. Cureus 2024;16, e58874. https://doi.org/10.7759/cureus.58874.