ISAKOS CONGRESS 2025 | MUNICH GERMANY June 8–11

Can Popular Artificial Intelligence Large Language Models Provide Reliable Answers to Frequently Asked Questions About Rotator Cuff Tears?

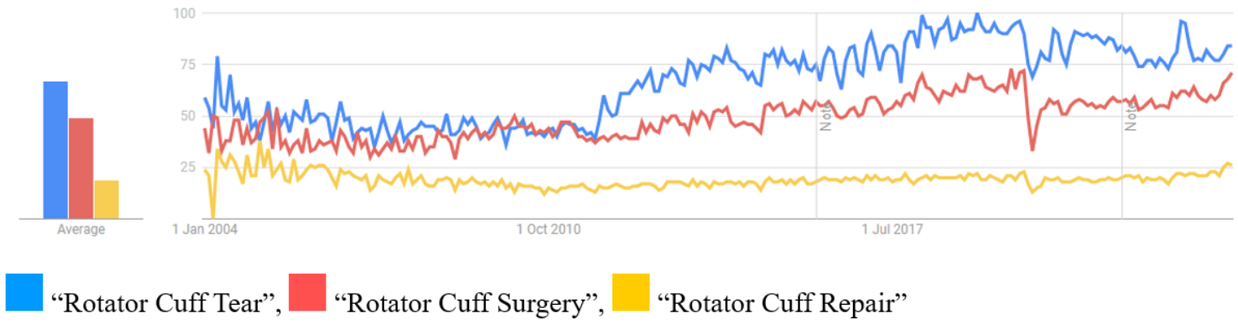
Orhan Mete Karademir, Hacettepe University, Turkey
Ulas Can Kolac, Hacettepe University, Turkey
Gokhan Ayik, Yuksek Ihtisas University, Turkey
Mehmet Kaymakoglu, Izmir Ekonomi University, Turkey
Filippo Familiari, Magna Graecia University, Italy
Gazi Huri, ASPETAR FIFA Center Of Excellence, Qatar

Faculty Disclosure Information

- Nothing to disclosure.

Aim

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



Rotator cuff tears (RCTs) are common upper extremity injuries that significantly impair shoulder function, causing pain, restricted range of motion, and reduced quality of life (1). As artificial intelligence large language models (AI LLMs) are increasingly consulted for health-related information, evaluating the quality and readability of the information they provide has become critically important (2).

Method

A pool of 50 questions related to RCTs was created using popular artificial intelligence models (ChatGPT 3.5, ChatGPT 4, Gemini, and Microsoft CoPilot) and a Google search. The responses generated by the AI LLMs were recorded and evaluated. Information quality was assessed using the DISCERN tool and a Likert scale, while readability was evaluated using the PEMAT Understandability Score and the Flesch-Kincaid Reading Ease Score (3,4,5,6,7). Two orthopedic surgeons independently assessed the responses.

Results

The average DISCERN score was 40, with 56.6% considered adequate. According to the Likert Scale, 96% of the responses were sufficient. The PEMAT Score had a 77.3% adequacy rate, while the Flesch-Kincaid Reading Ease Score had a 42.05%. Overall, 54.4% of the responses were sufficient in terms of information quality, and 72.1% were sufficient in terms of readability. Only 39.8% of the responses were adequate in both information quality and readability.

Figure 2. Distrubion of Questions according to the modified Rothwell Classification

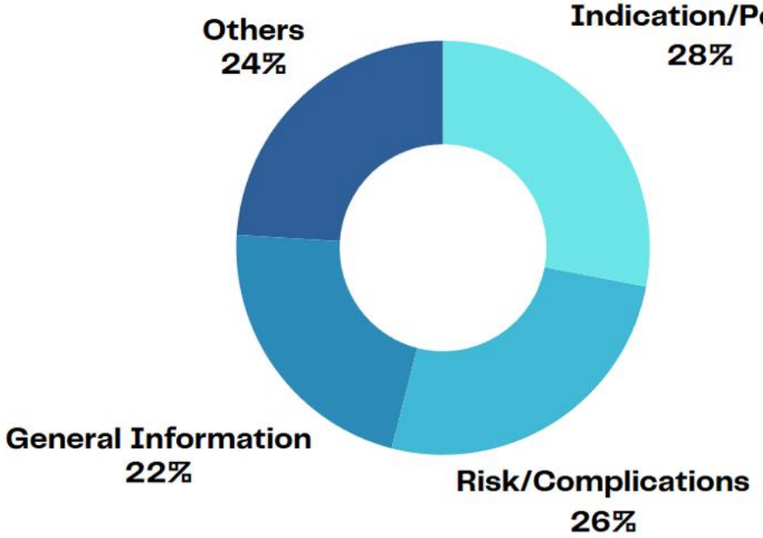


Figure 3. Sufficiency of Responses according to the scoring systems

Adequacy of Responses in Readability and Information Quality		Information Quality	
		Not Sufficient	Sufficient
Readability	Not Sufficient	26(%13.1)	29(%14.6)
	Sufficient	64(%32.3)	79(%39.8)

Conclusion

AI LLMs are not yet capable of providing accurate and readable information. Although they are not ready, they show great promise for the future. Currently, there are no specific tools designed to evaluate AI LLMs, yet there is a significant need for such tools. Due to their rapid development and dynamic nature, continuous reassessment is necessary. Ensuring that these tools can effectively support patient education is of great importance. Future research should focus on how to utilize these models most reliably for the benefit of patients.

References

1. Tashjian R. Z. (2012). Epidemiology, natural history, and indications for treatment of rotator cuff tears. Clinics in sports medicine, 31(4), 589–604. https://doi.org/10.1016/j.csm.2012.07.001

2.Fokina, M. (2024, May 22). The Future of Chatbots: 80+ Statistics, industry trends & Benchmarks (2024). Tidio. Retrieved June 2, 2024, from https://www.tidio.com/blog/chatbot-statistics/

3. McCormick, J. R., Kerzner, B., Tuthill, T. A., Khan, Z. A., Hodakowski, A. J., Damodar, D., Fortier, L. M., Dasari, S. P., Nho, S. J., & Chahla, J. (2022). Patients With Femoroacetabular Impingement Obtain Information From Low-Quality Sources Online and Are Most Interested in Conservative Treatment and Expected Recovery. Arthroscopy, sports medicine, and rehabilitation, 5(1), e21–e27. https://doi.org/10.1016/j.asmr.2022.09.011

4. Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. Journal of epidemiology and community health, 53(2), 105–111. https://doi.org/10.1136/jech.53.2.105

5. Sullivan, G. M., & Artino, A. R., Jr (2013). Analyzing and interpreting data from likert-type scales. Journal of graduate medical education, 5(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

6. Shoemaker, S. J., Wolf, M. S., & Brach, C. (2014). Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient education and counseling, 96(3), 395–403. https://doi.org/10.1016/j.pec.2014.05.027

7. Kincaid, J. P., Fishburne, J., P, R. R., L, C. R., & S, N. B. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. https://doi.org/10.21236/ada006655