Orthopaedic Clinician Perceptions of AI-Generated Versus Surgeon-Authored Responses to Pre-Operative Rotator Cuff Repair Questions

Aadam Ahmad[1], Marium Quaunine[2], Borna Guevel[3], Timothy P Davis[2], Jagwant Singh[2], Ashley Simpson[4], SELORE Collaborative

[1]Stoke Mandeville Hospital, Buckinghamshire Healthcare NHS Trust, [2]Department of Trauma & Orthopaedics, Queen Elizabeth Hospital, Lewisham & Greenwich NHS Trust, [3]Department of Trauma & Orthopaedics, St Mary's Hospital, Imperial College Healthcare NHS Trust, [4]Royal National Orthopaedic Hospital, Stanmore

BACKGROUND

Artificial intelligence (AI) is rapidly reshaping healthcare communication. Large language models (LLMs) such as ChatGPT can now produce medically accurate and empathetic responses to patient queries. Yet, their reliability and perceived value compared with surgeon-written advice remain unclear.

OBJECTIVES

Primary

1. To compare the perceived helpfulness of LLM-generated and consultant-written responses to the most frequently asked questions regarding elective rotator cuff repair

2. To evaluate clinicians' ability to correctly identify authorship (AI vs consultant) across training grades

Secondary

1. To assess readability differences between AI-generated and consultant-written responses using FIesch–Kincaid metrics

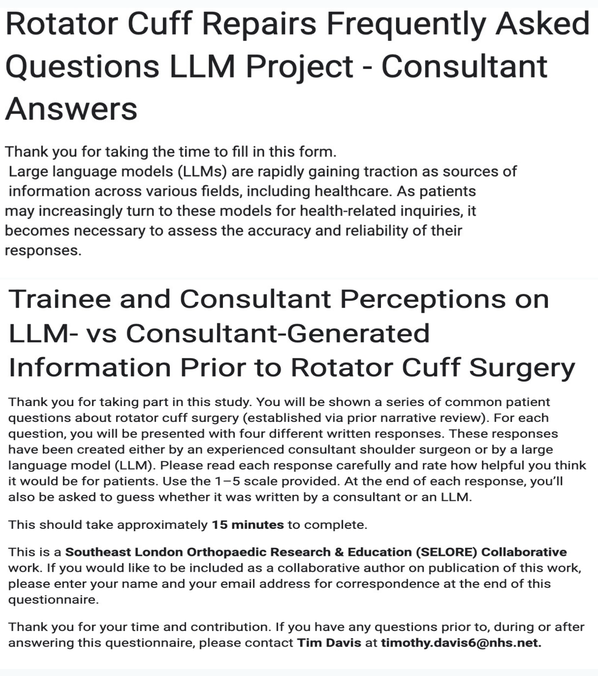
2. To determine which domains of patient information (e.g., symptoms, diagnosis, risks, recovery expectations) are better addressed by LLMs versus consultant surgeons

3. To explore the potential role of LLMs as an adjunct to clinician-led patient education in shoulder surgery

MATERIALS & METHODS

A cross-sectional survey was distributed to orthopaedic clinicians of all grades via the South East London Orthopaedic Research and Education (SELORE) collaborative. Participants reviewed four anonymised responses to ten common pre-operative rotator cuff questions: two authored by consultant upper limb surgeons and two by LLMs (ChatGPT 4o-mini and DeepSeek V3.1). Respondents rated each for helpfulness using a 5-point Likert scale (see Fig.2) and attempted to identify authorship. Readability was analysed using the Flesch–Kincaid grade level. Descriptive and comparative statistics were applied.

Questions were derived from a systematic review of the topic performed in April 2025 by the same research team. Both LLMs were given the same question stems and were queried sequentially, with each response obtained before the next question was asked. All AI answers were generated on the same day, with browser history, cache and location data cleared beforehand.


Rotator Cuff Repairs Frequently Asked Questions LLM Project - Consultant Answers

Thank you for taking the time to fill in this form.
Large language models (LLMs) are rapidly gaining traction as sources of information across various fields, including healthcare. As patients may increasingly turn to these models for health-related inquiries, it becomes necessary to assess the accuracy and reliability of their responses.

Trainee and Consultant Perceptions on LLM- vs Consultant-Generated Information Prior to Rotator Cuff Surgery

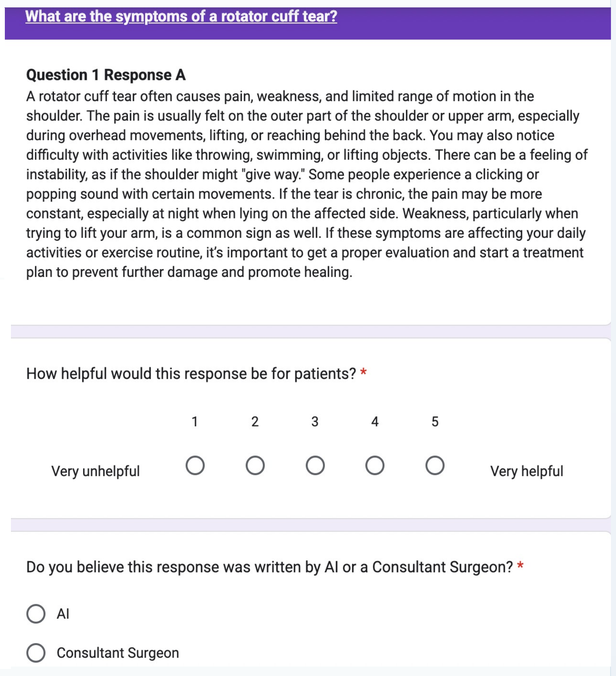
Thank you for taking part in this study. You will be shown a series of common patient questions about rotator cuff surgery (established via prior narrative review). For each question, you will be presented with four different written responses. These responses have been created either by an experienced consultant shoulder surgeon or by a large language model (LLM). Please read each response carefully and rate how helpful you think it would be for patients. Use the 1–5 scale provided. At the end of each response, you'll also be asked to guess whether it was written by a consultant or an LLM.

This should take approximately **15 minutes** to complete.

This is a **Southeast London Orthopaedic Research & Education (SELORE) Collaborative** work. If you would like to be included as a collaborative author on publication of this work, please enter your name and your email address for correspondence at the end of this questionnaire.

Thank you for your time and contribution. If you have any questions prior to, during or after answering this questionnaire, please contact **Tim Davis** at **timothy.davis6@nhs.net.**

Fig. 1: *Introduction page for the rotator cuff survey*


What are the symptoms of a rotator cuff tear?

Question 1 Response A
A rotator cuff tear often causes pain, weakness, and limited range of motion in the shoulder. The pain is usually felt on the outer part of the shoulder or upper arm, especially during overhead movements, lifting, or reaching behind the back. You may also notice difficulty with activities like throwing, swimming, or lifting objects. There can be a feeling of instability, as if the shoulder might "give way." Some people experience a clicking or popping sound with certain movements. If the tear is chronic, the pain may be more constant, especially at night when lying on the affected side. Weakness, particularly when trying to lift your arm, is a common sign as well. If these symptoms are affecting your daily activities or exercise routine, it's important to get a proper evaluation and start a treatment plan to prevent further damage and promote healing.

How helpful would this response be for patients? *

	1	2	3	4	5	
Very unhelpful	○	○	○	○	○	Very helpful

Do you believe this response was written by AI or a Consultant Surgeon? *

○ AI

○ Consultant Surgeon

Fig. 2: Example survey question showing 5-point Likert Scale

RESULTS

We received responses from 55 clinicians, the breakdown of training grades are demonstrated in Figure 3 and Table 1. LLM-generated responses were rated slightly more helpful overall than consultant-written answers (see Table 2). LLMs scored higher for questions on symptoms, diagnosis, non-operative options, postoperative pain, recovery expectations and physiotherapy, while consultants performed best for procedural explanation and implications of non-repair (see Table 3). Readability analysis showed LLM outputs required substantially lower reading levels and were consistently more accessible than consultant-written responses.

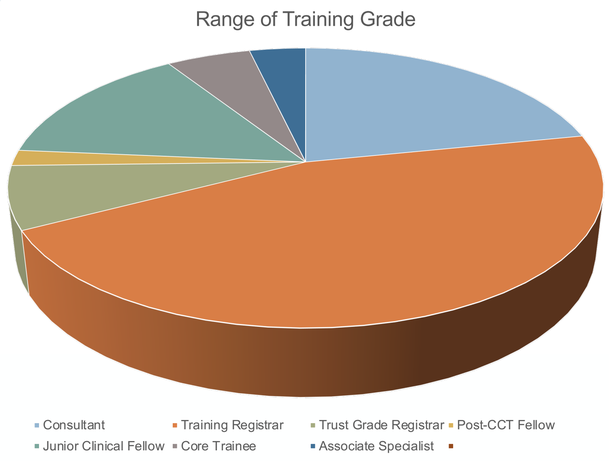


Fig. 3: *Proportion of participants per Training Grade*

Training Grade	Number of Participants
Consultant	12
Post-CCT Fellow	1
Training Registrar	25
Trust Grade Registrar	4
Junior Clinical Fellow	8
Core Trainee	3
Associate Specialist	2

Table 1: *Number of participants per Training Grade*

Response Source	Mean Helpfulness Score (/5)
DeepSeek	3.98
ChatGPT	3.96
Consultant 1	3.92
Consultant 2	3.56

Table 2: *Mean Helpfulness Score for each Response Source*

Question Domain	Highest scoring response source
General patient understanding	LLMs
Non-operative management	LLMs
Operative explanation	Consultants
Consequences of non-repair	Consultants
Peri-operative risks	DeepSeek (LLM)
Post-operative expectations	DeepSeek (LLM)

Table 3: *The highest scoring response source for each Question Domain*

Clinician Group	Overall % Correct	% AI Responses Identified Correctly	% Consultant Responses Identified Correctly
Juniors	60.5%	57.0%	64.0%
Registrars	64.8%	62.6%	67.0%
Consultants	62.9%	62.5%	63.3%

Table 4: Table outlining how accurately clinicians at different levels distinguished AI-generated responses from consultant-written responses

DISCUSSION

LLMs produced patient-facing information that clinicians rated as equally, and often more, helpful than consultant-written answers for many common pre-operative questions, particularly those involving symptoms, diagnosis, non-operative management and postoperative expectations. Consultants remained stronger in nuanced procedural and risk-related explanations, underscoring the continued importance of clinical expertise. Across all clinical grades, surgeons struggled to differentiate between AI and consultant authorship, reflecting the increasing sophistication of LLM outputs and the challenge of detecting AI-generated content. The greater readability and accessibility of LLM responses further supports their potential role in supplementing traditional patient education, provided this is guided by appropriate clinician oversight.

LIMITATIONS

- Surgeon responses were collected on different days and in different environments, potentially increasing variability
- Some participants were general orthopaedic surgeons rather than rotator-cuff specialists
- Identities were not verified on the Google Form, so ineligible respondents may have contributed
- Collaborative authorship may have boosted responses but also introduced sampling and acquiescence biases

CONCLUSION

In our study, the LLMs matched - and in several domains surpassed - consultant surgeons. While consultant expertise remains essential for complex, risk-focused and procedural counselling, AI-generated responses offer strong performance in general education and postoperative expectation-setting, with markedly superior readability. Clinicians across all grades showed limited ability to identify AI authorship, underscoring the realism and maturity of LLM outputs. Together, these findings support the role of LLMs as a valuable adjunct to, rather than a replacement for, consultant-led patient education, provided their use is supported by appropriate clinical oversight.