Misinformation in large language model descriptions of upper extremity diseases

George Sayegh[1], David C Ring, Prakash Jayakumar
[1]University of Texas Medical School - SA

INTRODUCTION:

Large Language Models (LLMs) collate available information on the internet and present it in a chatty, empathetic (Ayers et al., 2023), understandable manner (Moons and Van Bulck, 2024). However, there is notable evidence that LLMs respond with misinformation when discussing orthopedic conditions. For example, LLM queries regarding treatments for hip and knee osteoarthritis provided 20%-40% of responses discordant with the American Academy of Orthopaedic Surgeons (AAOS) Clinical Practice Guidelines (CPGs) and encouraged use of non-recommended treatments in 30% and 60% of queries, respectively (Yang et al., 2024). Also, there is the consideration that ChatGPT can reinforce unhelpful thoughts and confusion if responses are convoluted and written for a highly educated audience. In a study of ChatGPT responses to questions about 4 common hand surgeries, though the responses were of good quality, based on the DISCERN score that evaluates virtual health-related content, they lacked in terms of readability and simplicity as they were written at a college reading level (Crook et al., 2023).

In an unpublished study we noted that ChatGPT often made statements with the potential to reinforce unhelpful thoughts and less characteristic symptoms often leading to misdiagnosis. We therefore made several attempts to train ChatGPT to avoid misinformation and misdiagnosis and asked: 1. Does ChatGPT respond with potential misinformation when asked if it disagrees with evidence- and principle-based descriptions of upper extremity conditions? 2. Does ChatGPT correct misinformation regarding specific upper extremity conditions when prompted to do so? and 3. Are there specific types of misinformation that persist after attempt to train ChatGPT?

METHODS:

In the first part of this study, we asked a free and accessible LLM-based chatbot (ChatGPT 3.5, OpenAI, San Francisco, CA) “What do you disagree with from the following description of ___ condition?” and submitted an evidence- and principles-based description of the specific upper extremity condition found on https://www.itsanarmproblem.com/. This website, developed by an experienced hand surgeon, encourages a healthy mindset when describing upper extremity conditions. Two researchers then independently rated what proportion of the responses contain potential misinformation using a standardized potential misinformation checklist. We continued this process within the same chat for a list of thirteen upper extremity diagnoses.

In the second part of this study, in another chat, we asked ChatGPT 3.5 to “Please describe my condition” when presented with a brief description of common symptoms of the specific upper extremity condition. Two researchers assessed the proportion of sentences in the response that contain misinformation then suggested corrections for all instances of misinformation in the response using a standard script. An example of a suggested corrections is: “First, you note that ‘it is essential to consult a healthcare professional.’ But trigger digit is a benign condition and there is no risk in leaving it untreated. Seeking care is entirely discretionary and optional. Second…” This process continued until there were consecutive responses with sentences of misinformation that were unchanged after correction. We continued this process within the same chat for a list of eight upper extremity conditions.

RESULTS:

When asking ChatGPT what it disagreed with from accurate descriptions of thirteen upper extremity conditions, 75% of the sentences had at least one instance of potential misinformation (Table 1). The most common theme of misinformation was the reinforcement of unhelpful thinking, such as stating that activities, such as exercise, can cause a condition. Misrepresentation of pathophysiology and reduction of patient agency were the next most common themes.

When asking ChatGPT 3.5 to describe a specific upper extremity disease when presented with common symptoms of the condition, 53% of initial responses, and 25% of follow up responses contained sentences with at least one instance of potential misinformation (Table 1). In both initial and follow up responses, the reinforcement of unhelpful thinking was the most common theme, and the reduction of patient agency was the second most common theme.

Our investigators found that:

1. ChatGPT is unable to "learn" the difference between palliative (symptom alleviation) and disease-modifying (changing the natural history of the disease) treatments. It may parrot feedback, but it cannot reason or learn.
2. ChatGPT gets confused when given rules. For instance, after telling it to always say that surgery is discretionary for MSK conditions, ChatGPT often responds with a statement questioning the effectiveness of surgery. And it often states that "physical therapy" is a treatment even when corrected to say that it is a profession not a treatment.

3. ChatGPT may be programmed with unhelpful disclaimers such as "essential to discuss with a healthcare professional"
4. ChatGPT often mentions progression or severity of symptoms even when this is not relevant to the test and treatment options for a given disease.
5. ChatGPT often includes lengthy descriptions with complex verbiage in responses.

DISCUSSION AND CONCLUSION: The results of this study affirm that the risks associated with relying on ChatGPT for information on upper extremity conditions include the reinforcement of unhelpful thinking, misrepresentation of pathophysiology, and statements that aim to reduce patient agency. These findings underscore the importance of cautious interpretation and supplementation of LLM-generated information with verified medical sources that promote agency, positive mindsets, and helpful accommodation practices.

Table 1. Potential misinformation by ChatGPT responses by testing method

Themes of Misinformation by ChatGPT	Examples	ChatGPT testing methods		
		Starting with evidence and principle- based information about conditions	Starting with common symptoms of conditions	
			Initial response	Follow up responses
Reinforces unhelpful thinking	"Repetitive activity and gripping can cause trigger digit"	27	22	10
Reduction of patient agency	"It's essential to consult a healthcare professional"	18	11	6
Unhelpful focus on timelines	"The time to resolution can vary greatly depending on..."	3	1	1
Implies there is a disease modifying treatement	"Treatments such as anti-inflammatory medication are useful in treating medial epicondylitis..."	14	4	2
Misinformation about treatment	"Corticosteroid injections are commonly used to for distal biceps tendinopathy"	12	9	4
Recommendation that symptom severity guides treatment	"The recommended treatment depends on the severity of symptoms"	5	6	6
Misrepresentation of discretionary treatments as necessary	"These treatments may become necessary as this condition progresses"	15	3	1
Misrepresentation of the pathophysiology	"Trigger finger pain is likely due to inflammation or irritation"	22	10	4
Conflation of palliative and disease modifying treatment	"The effectiveness of splinting for treating thumb osteoarthritis can vary among individuals"	3	7	5
Implication that evidence can be selectively ignored	"The decision for a corticosteroid injection for distal biceps tendinopathy should be made if..."	3	6	3
Lack of appreciation of simplification as a debiasing strategy	"Stating that carpal tunnel release should be done before loss of sensation is an oversimplification"	11	0	4
Physical therapy presented as a treatment rather than a profession	"Treatments such as physical therapy"	2	1	1
Unhelpful use of extreme language	"Severe cases of trigger digit"	2	4	4
Sentence in an incorrect section	In depth discussion of anatomy in the etiology section	1	0	2
Sentences with at least one instance of potential misinformation		**80**	**36**	**29**
Total number of sentences		**106**	**68**	**117**
Percentage of sentences with at least one instance of potential misinformation		**75%**	**53%**	**25%**