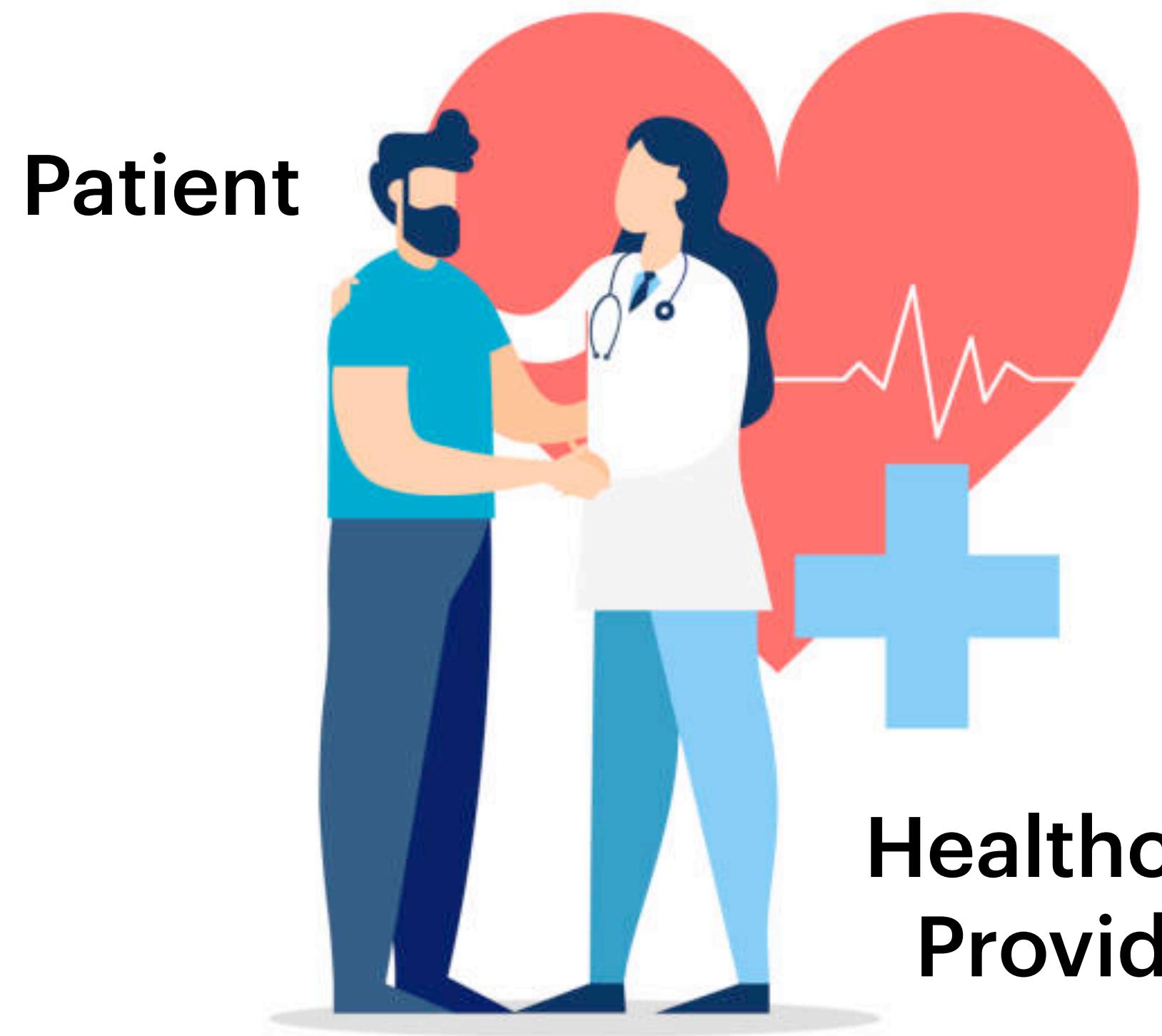


# **Beyond Hallucinations**

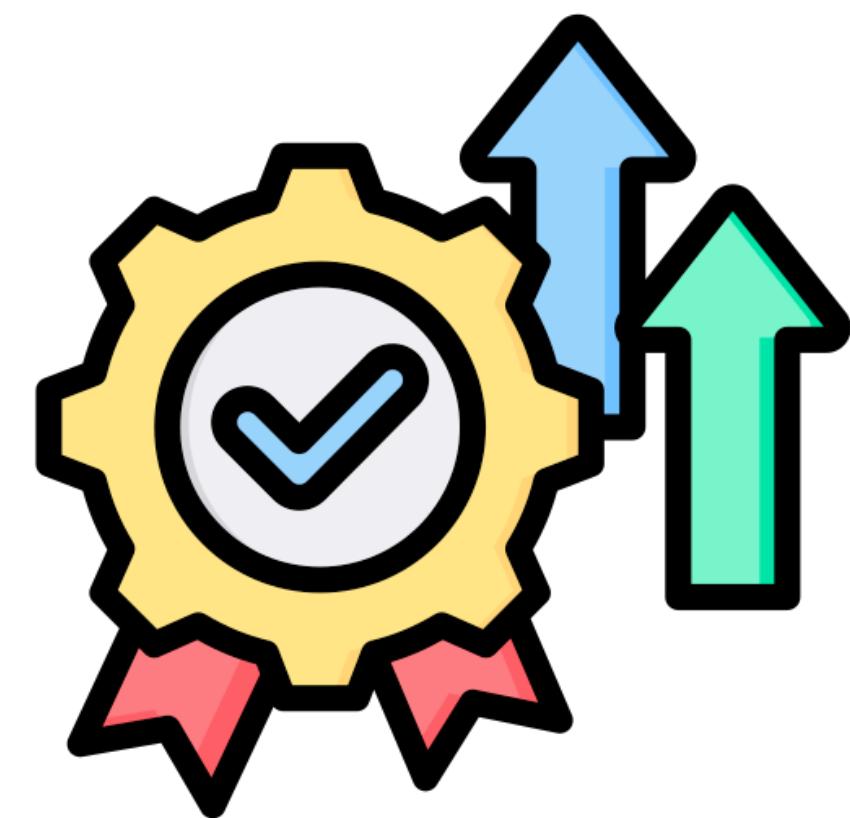
## **Unveiling Hidden Dangers of LLMs in Health Information Access**

**Hye Sun Yun — Northeastern University**

Clemson School of Computing Seminar  
April 18, 2025



# Evidence-based medicine as a model of care



improved quality

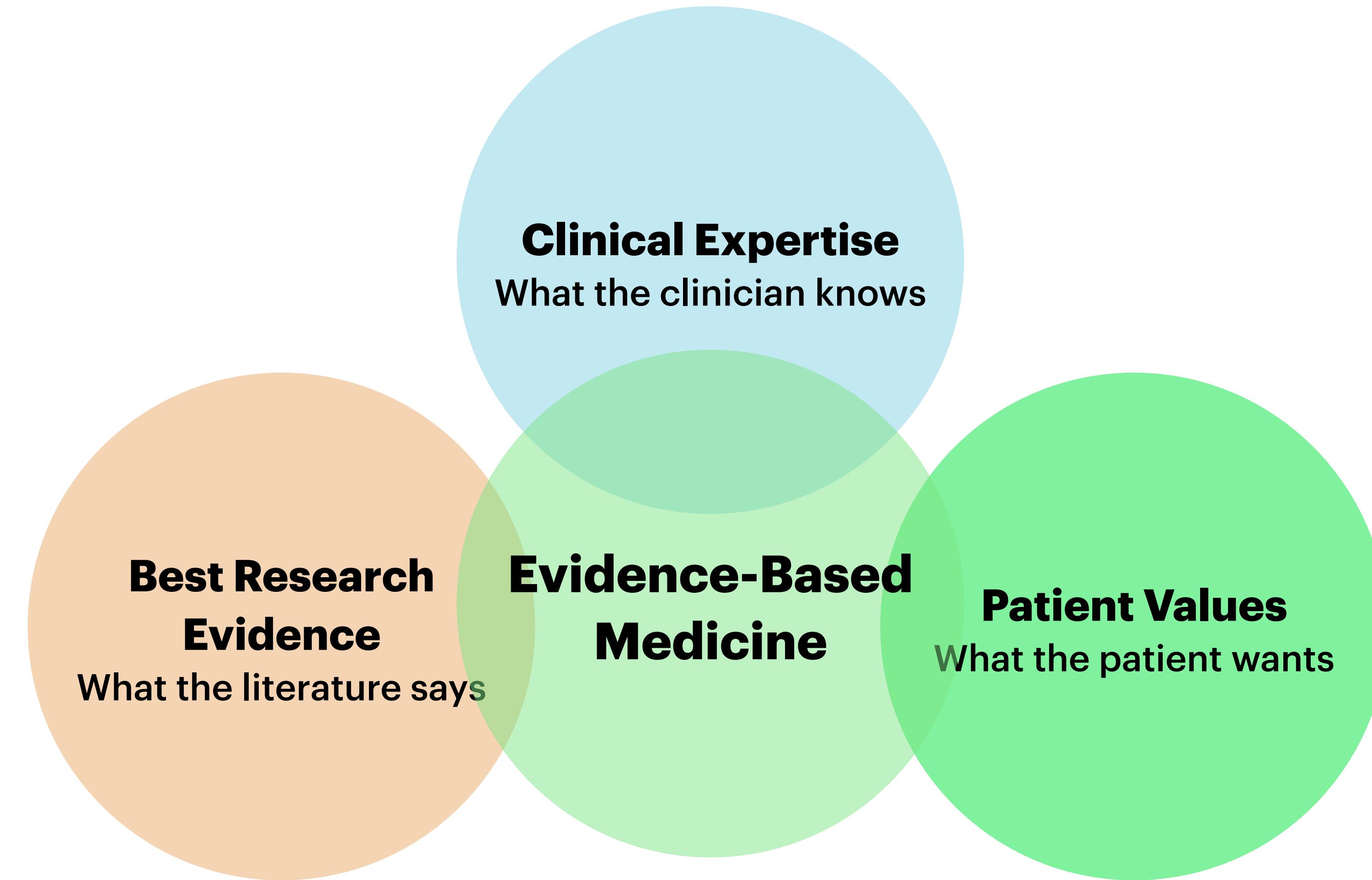


improved patient  
satisfaction



reduced costs

# Key components of evidence-based medicine



Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *bmj*. 1996 Jan 13;312(7023):71-2.

# Best research evidence can be found in medical literature

- Unstructured (natural language) published articles
  - Provide quantitative measures of comparative treatment effectiveness
  - Describe the design, protocol, and results of Randomized Controlled Trials (RCTs)

Randomized Controlled Trial > *Lancet.* 2020 May 16;395(10236):1569-1578.

doi: 10.1016/S0140-6736(20)31022-9. Epub 2020 Apr 29.

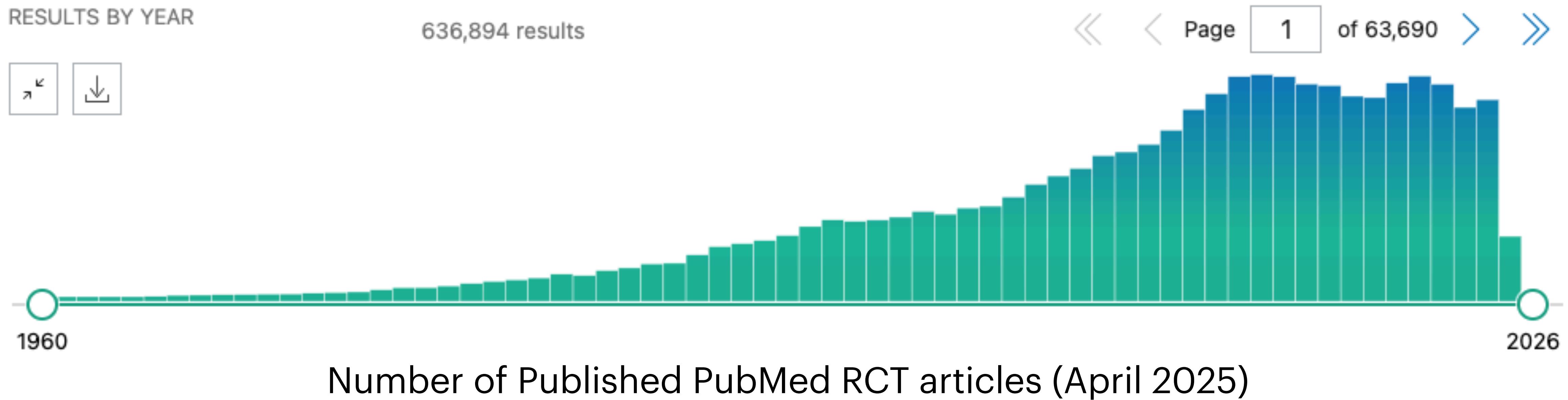
**Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial**



**Interpretation:** In this study of adult patients admitted to hospital for severe COVID-19, remdesivir was not associated with statistically significant clinical benefits. However, the numerical reduction in time to clinical improvement in those treated earlier requires confirmation in larger studies.

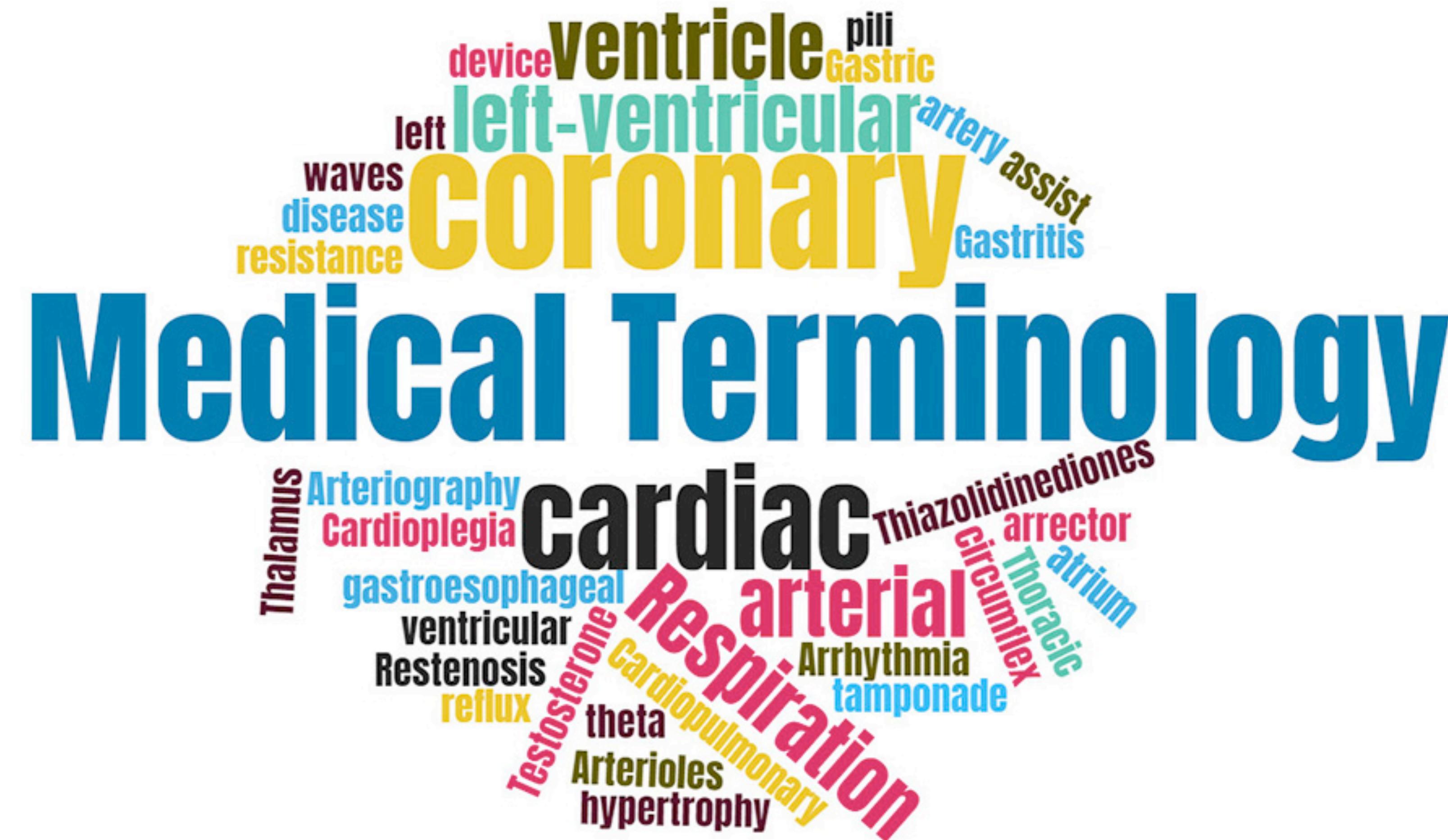
# Keeping up with new medical information can be challenging

## Information Overload for Healthcare Providers

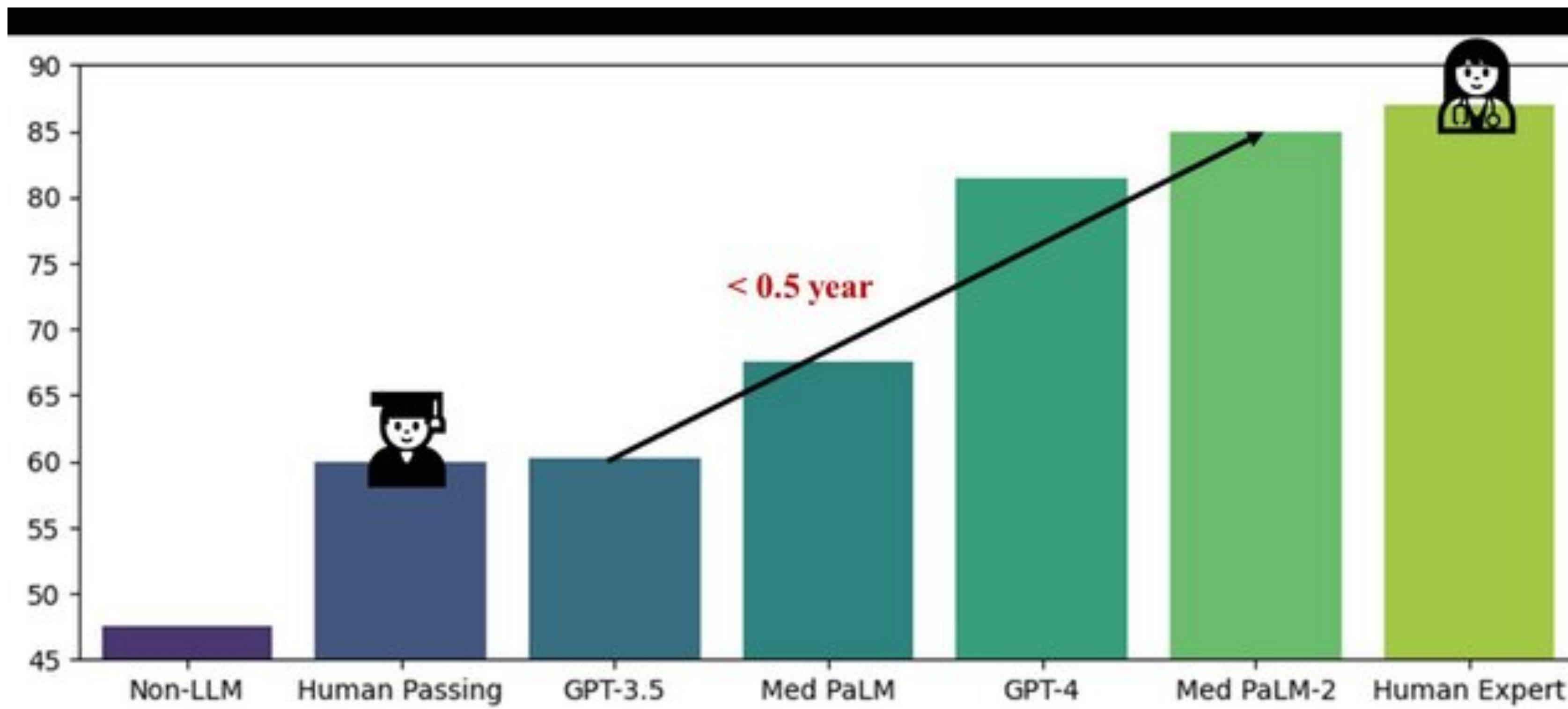


# Medical literature is full of medical jargon

Overwhelming for Patients & Caregivers



# LLMs as a solution



# Limitations of LLMs

## Hallucinations & Inaccurate Texts

### The benefits of eating crushed glass

#### Introduction

The purpose of this study was to find out if the benefits of eating crushed glass are due to the fiber content of the glass, or to the calcium, magnesium, potassium, and phosphorus contained in the glass. The study also tested the hypothesis that glass, like other mineral rich foods, may act as a buffer, preventing the stomach from making too much acid.

#### Results

The results of the study showed that the glass meal was the most effective at lowering stomach acid output, and the wheat bran meal was the least effective.

The results also showed that the glass meal was the most effective at preventing stomach acid from returning to normal after it had been suppressed.



Figure 3

## One in Six Adults Say They Use AI Chatbots for Health Information and Advice at Least Once a Month

Percent of adults who say they use artificial intelligence, or AI, chatbots such as ChatGPT, Microsoft Copilot, or Google Gemini to find **health information and advice** at least once a month:

Total **17%**

### Age

18-29 **25%**

30-49 **19%**

50-64 **15%**

65+ **10%**

Note: See topline for full question wording.

Source: KFF Health Misinformation Tracking Poll (June 3-24, 2024)

Beyond hallucinations, what are the  
**hidden dangers** of LLMs used for health  
information access?

*trust* and *satisfaction*

Does the mere fact that medical information  
is obtained from a **chatbot** influence user  
*trust* and *satisfaction* compared to identical  
information obtained from a **search engine**?

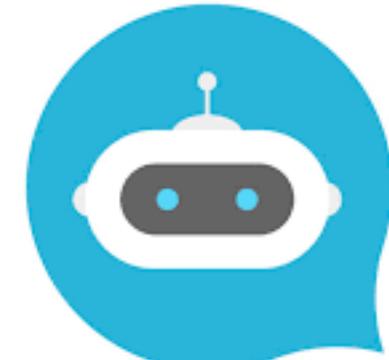
# Simulation Videos

## For Framing Health Information

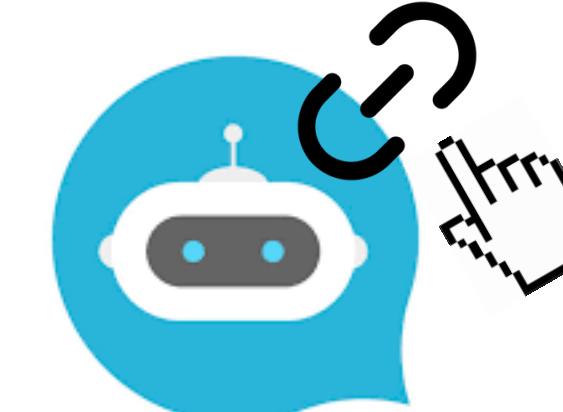
**Three Search Methods**



search engine



chatbot



chatbot+



health websites



social media

does taking or - Google Search +

← → C google.com/search?q=does-taking-oral-antiviral-treatment-of-paxlovid-really-

Applications Online Shopping Helpful Sites Cloud Storage Wikipedia Tool YouTube

Google Does taking oral antiviral treatment of Paxlovid really caus X

Search Images Videos Shopping Maps

About 105 000 000 results (0.43 seconds)

[www\(pfizer.com/](http://www(pfizer.com/)

**Pfizer - Paxlovid FAQs**

Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at risk....

[who.int/](http://who.int/) Paxlovid Covid Symptoms

**World Health Organization - Paxlovid and COVID**

A 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir.

[www.cdc.gov/](http://www.cdc.gov/) Paxlovid

**Center for Disease Control | Paxlovid**

Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19.

[Paxlovid and Ritonavir Effect](#)

[Questions on Paxlovid](#)



Nirmatrelvir/ritonavir

Nirmatrelvir/ritonavir, sold under the brand name Paxlovid, is a co-packaged medication used as a treatment for COVID-19. It contains the antiviral medications nirmatrelvir and ritonavir and was developed by Pfizer. [Wikipedia](#)

Feedback

search engine with health websites

Chat AI

chatai.com

# chatbot with social media

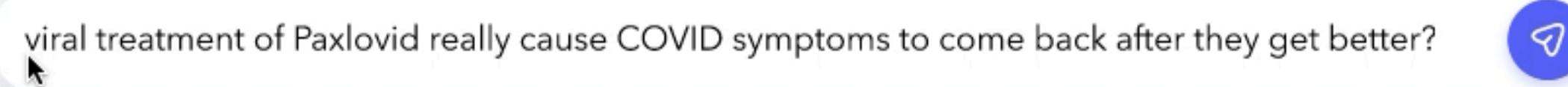
 YOU

Does taking oral antiviral treatment of Paxlovid really cause COVID symptoms to come back after they get better?

 CHAT AI 

Sure, I can provide you with information on Paxlovid:

Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at high risk for progression to severe COVID-19, including hospitalization or death. It is a 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir. According to Reddit Health subreddit, this drug is

 viral treatment of Paxlovid really cause COVID symptoms to come back after they get better? 

Chat AI  chatai.com

# chatbot+ with social media

 YOU

Does taking oral antiviral treatment of Paxlovid really cause COVID symptoms to come back after they get better?

 CHAT AI 

Sure, I can provide you with information on Paxlovid:

Developed by Pfizer, Paxlovid is a prescription oral antiviral pill that can be taken at home for the treatment of mild-to-moderate COVID-19 in adults who are at high risk for progression to severe COVID-19, including hospitalization or death. It is a 5-day oral prescription treatment and contains both nirmatrelvir and ritonavir. This drug is not approved for use as pre-exposure or post-exposure treatment for prevention of COVID-19.

Although studies outside of the laboratory have confirmed Paxlovid's effectiveness among people who have been vaccinated, there have been reports of a "rebound" of COVID-19 symptoms in some people within 2 to 8 days after completing the five-day course of Paxlovid. Some have reported testing positive again but having no symptoms; others have a recurrence of symptoms. Pfizer had said that in the clinical trial for Paxlovid, several participants appeared to have a rebound in virus levels,

viral treatment of Paxlovid really cause COVID symptoms to come back after they get better? 

# Method

## Survey

- Qualtrics survey in English
- *Between-subjects:*  
each participant viewed only one video

## Semi-Structured Interview

- In English over Zoom
- About 60 minutes each
- *Within-subjects:*  
videos of 3 search methods + interview on trust and satisfaction

# Participants

## Survey (between-subjects)

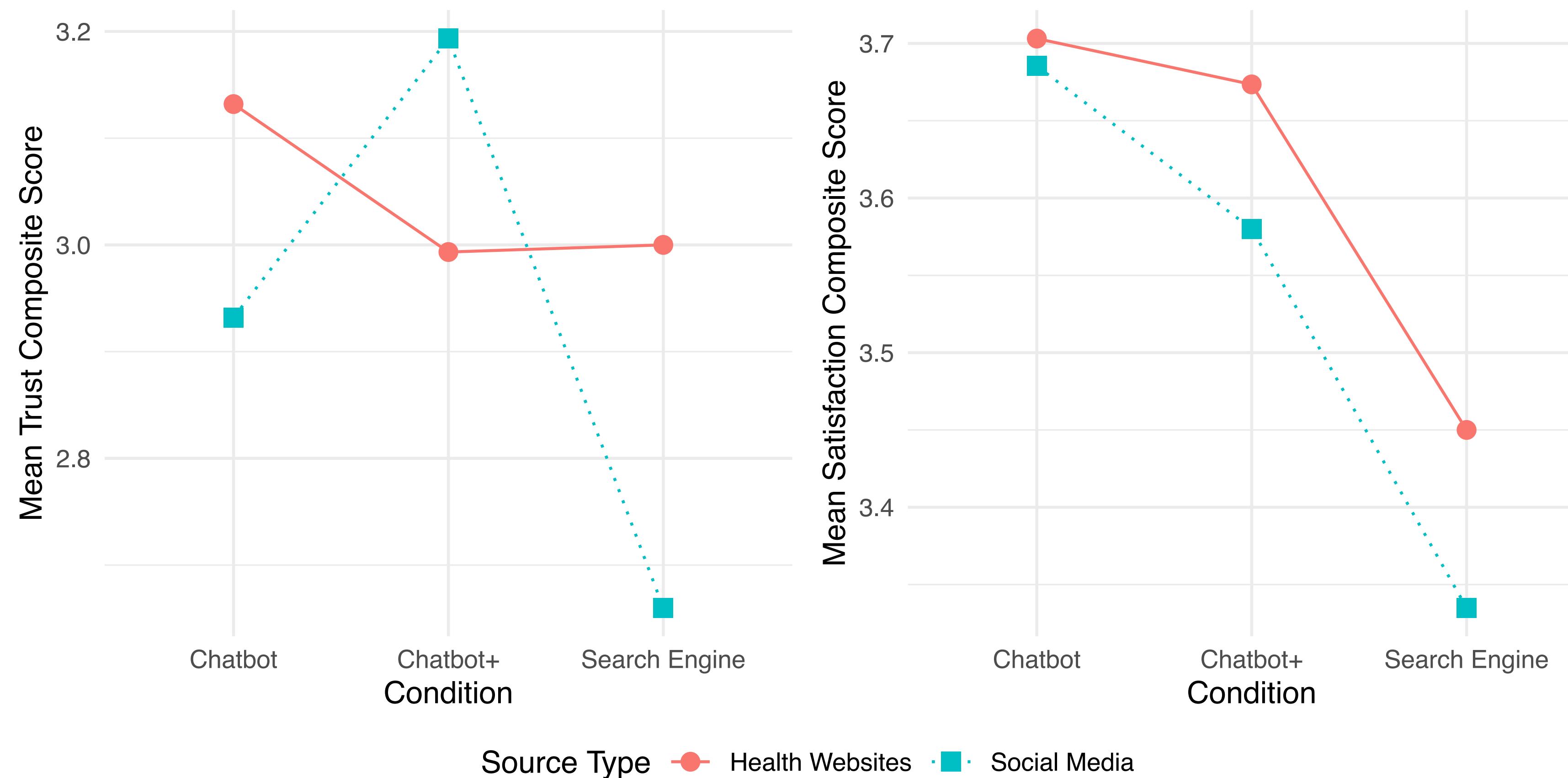
- 300 international participants from Prolific
- 21.2% used LLM-based chatbots for health information in the past year

## Interview (within-subjects)

- 6 international participants from Prolific
- 66.7% used LLM-based chatbots for health information in the past year



# Survey Results



**Trust:** {Chatbot, Chatbot+} > Search Engine

**Satisfaction:** Chatbot > Chatbot+ > Search Engine across both sources.

# Diving Deeper...

## Themes from Interviews

- **Search Engine:** familiar experience with a rich array of information.
- **Chatbot:** straightforward answers from unclear sources.
- **Chatbot+:** direct answers with opportunities to cross-reference.
- **Effect of source type** on trust can depend on context.

“It is very familiar because every time ... I don’t feel well, the first thing I go to is the search engines.” — P4

# Diving Deeper...

## Themes from Interviews

- **Search Engine:** familiar experience with a rich array of information.
- **Chatbot:** straightforward answers from unclear sources.
- **Chatbot+:** direct answers with opportunities to cross-reference.
- **Effect of source type** on trust can depend on context.

“It was straight to the point ... I liked things that are straight to the point, not waste my time.”

— P1

“It is interesting because you are able to crosscheck if there is similar things, and also if everything actually is the same in both the links in the chatbot.” — P2

# Diving Deeper...

## Themes from Interviews

- **Search Engine:** familiar experience with a rich array of information.
- **Chatbot:** straightforward answers from unclear sources.
- **Chatbot+:** direct answers with opportunities to cross-reference.
- **Effect of source type** on trust can depend on context.

“I believe that AI has been trained with the relevant information regarding a lot of situations, whether health or life situations. So, mostly the information that’s there, it’s mostly reliable.”— P2

# Key Takeaways

- The way health information is presented can impact user trust and satisfaction
  - Providing straightforward answers (well summarized text) may increase user satisfaction since users prefer low cognitive load
- Chatbots should enhance source visibility and transparency to ensure safety
- Educating users on how LLMs work can be important

# Spin in medical literature

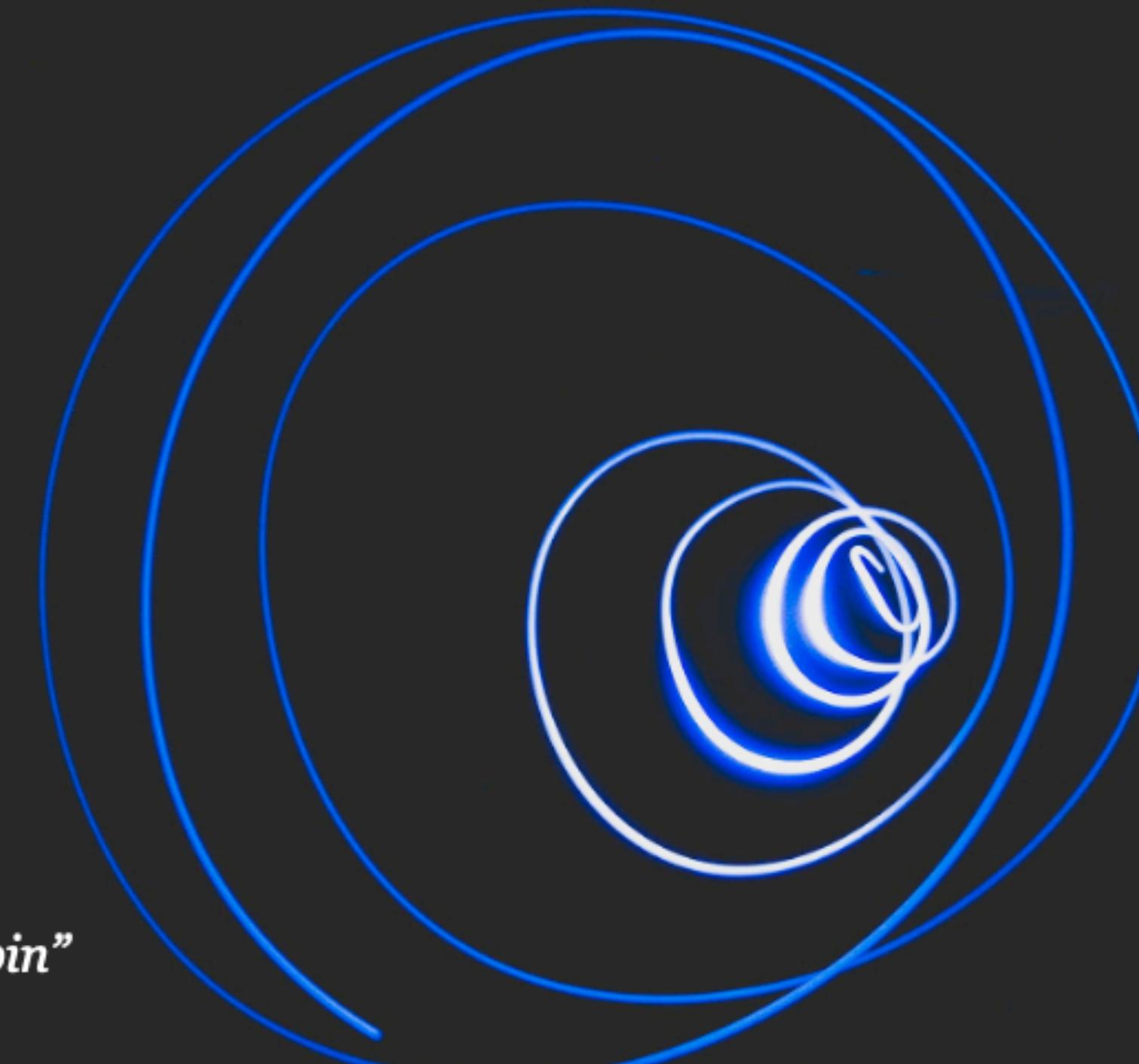
# Spin

spin 🔊

spun 🔊; spinning 🔊

*transitive verb*

“To present (information) with a particular spin”



marginally significant tendency ( $p=0.08$ )

a distinct trend toward significance ( $p=0.07$ )

almost statistically significant ( $p=0.06$ )

approaching a level of significance ( $p=0.089$ )

fairly significant ( $p=0.09$ )

May 26, 2010

# **Reporting and Interpretation of Randomized Controlled Trials With Statistically Nonsignificant Results for Primary Outcomes**

Isabelle Boutron, MD, PhD; Susan Dutton, MSc; Philippe Ravaud, MD, PhD; et al

# Evaluation of spin in oncology clinical trials

C. Wayant <sup>a</sup>  , D. Margalski <sup>b</sup>  , K. Vaughn <sup>c</sup>  , M. Vassar <sup>d</sup>  

## Evaluation of spin in abstracts of papers in psychiatry and psychology journals

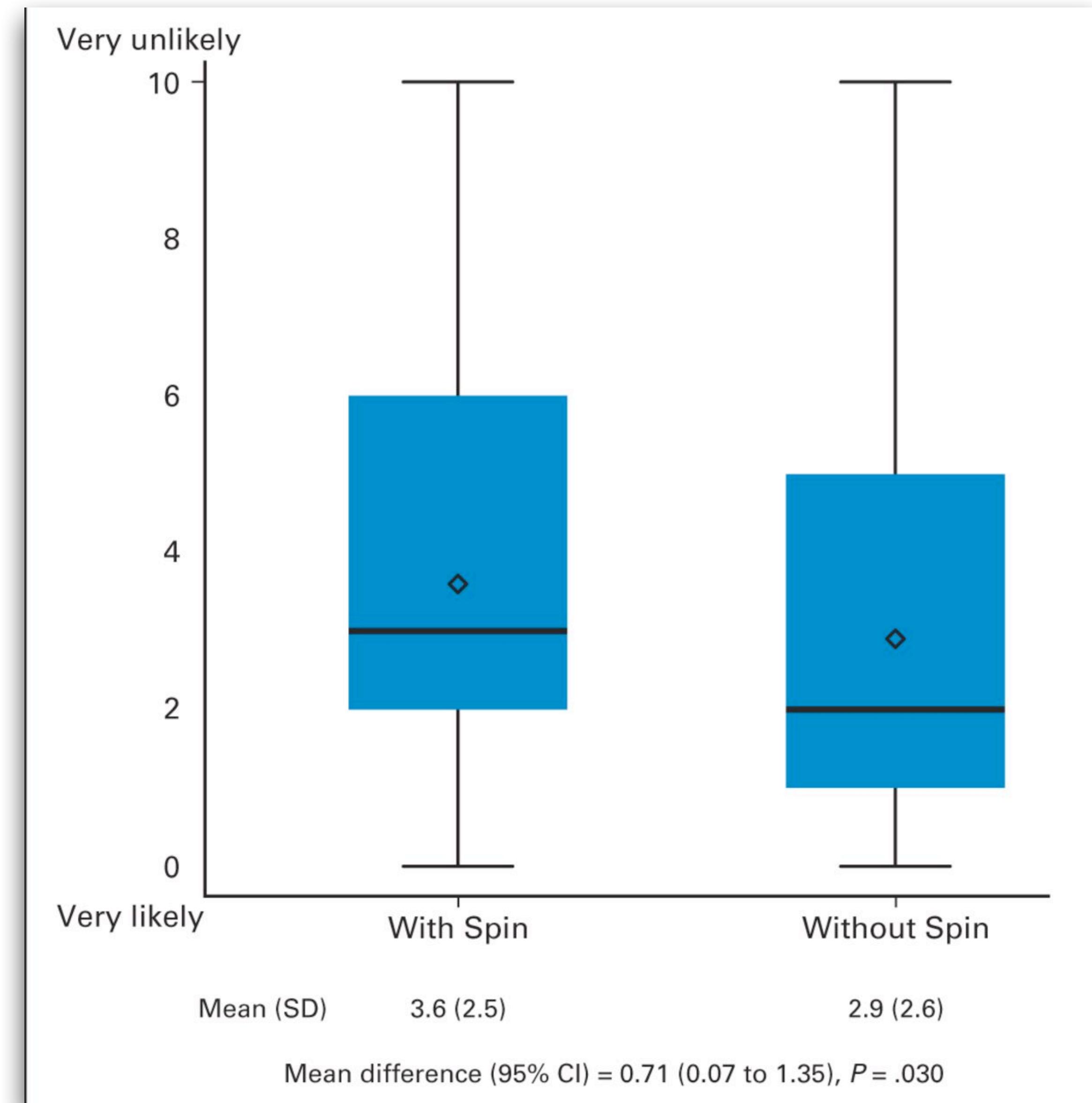
 Samuel Jellison<sup>1</sup>, Will Roberts<sup>1</sup>, Aaron Bowers<sup>1</sup>, Tyler Combs<sup>1</sup>, Jason Beaman<sup>2, 3</sup>,  Cole Wayant<sup>1</sup>, M

## Spin in Abstracts of Systematic Reviews and Meta-analyses of Melanoma Therapies: Cross-sectional Analysis

Ross Nowlin<sup>1</sup> ; Alexis Wirtz<sup>1</sup> ; David Wenger<sup>1</sup> ; Ryan Ottwell<sup>2, 3</sup> ; Courtney Cook<sup>4</sup> ; Wade Arthur<sup>5</sup> ; Brigitte Sallee<sup>4</sup> ; Jarad Levin<sup>4</sup> ; Micah Hartwell<sup>1, 6</sup> ; Drew Wright<sup>7</sup> ; Meghan Sealey<sup>8</sup> ; Lan Zhu<sup>8</sup> ; Matt Vassar<sup>1, 6</sup> 

# Clinicians can fall for spin

- A study in 2014 assessed the impact of spin on the interpretations of results of abstracts of RCTs in the field of cancer.
- Clinicians overstated the benefits of results when shown an abstract with spin.



Boutron I, Altman DG, Hopewell S, Vera-Badillo F, Tannock I, Ravaud P. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. Journal of Clinical Oncology. 2014 Dec 20;32(36):4120-6.

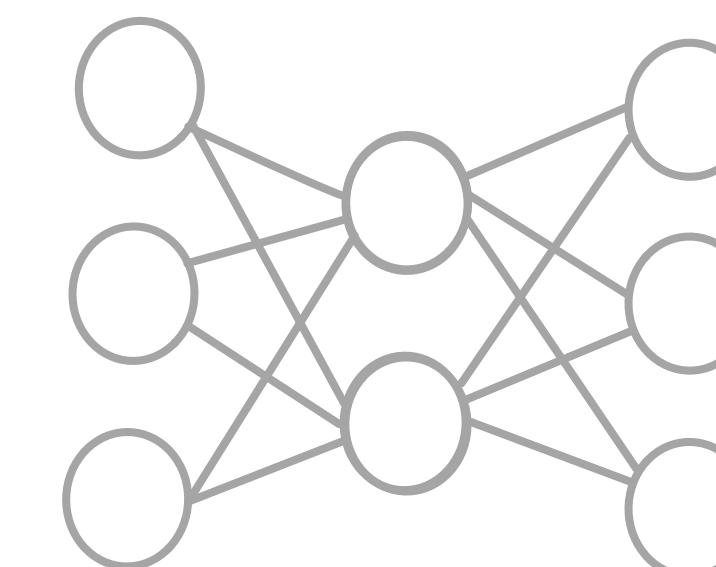
Do LLMs fall for **spin** in medical  
literature?

## Neutral and spun abstracts (same results)

**neutral** “... there was no statistically difference in mortality rates between the treatment and control groups (OR 1.46 [95% CI 0.12, 1.4]).

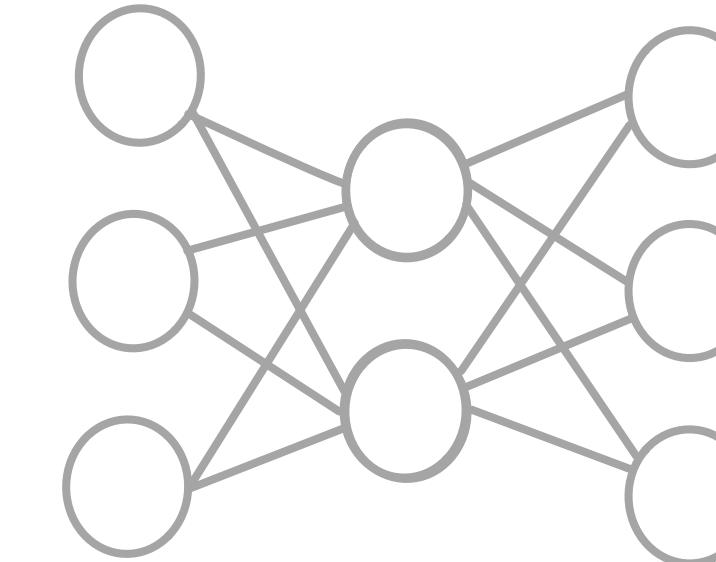
**spun** “... the difference in mortality rates between groups trends towards significance (OR 1.46 [95% CI 0.12, 1.4]).

LLM



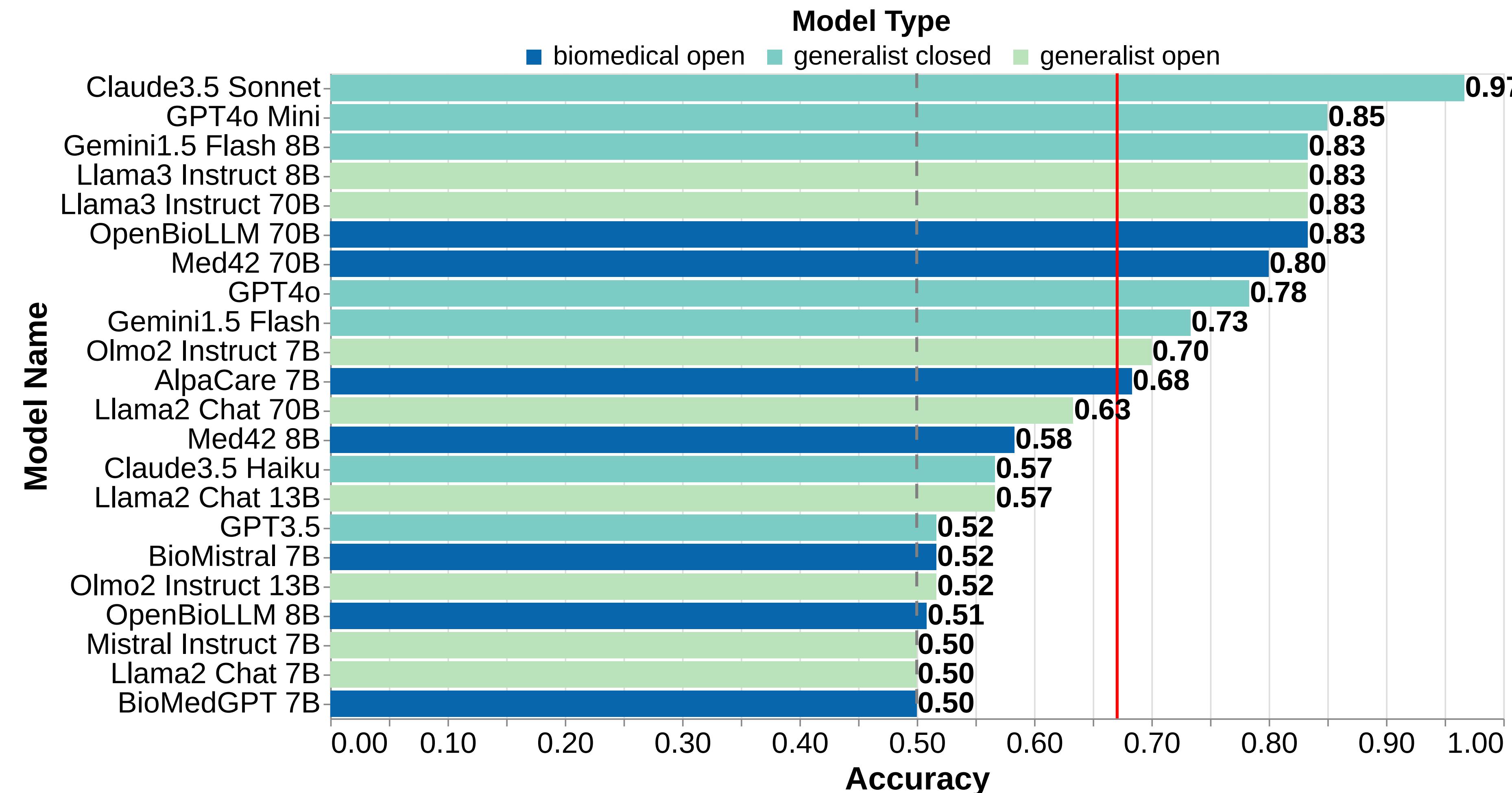
## LLM interpretation of results

No evidence for the treatment.

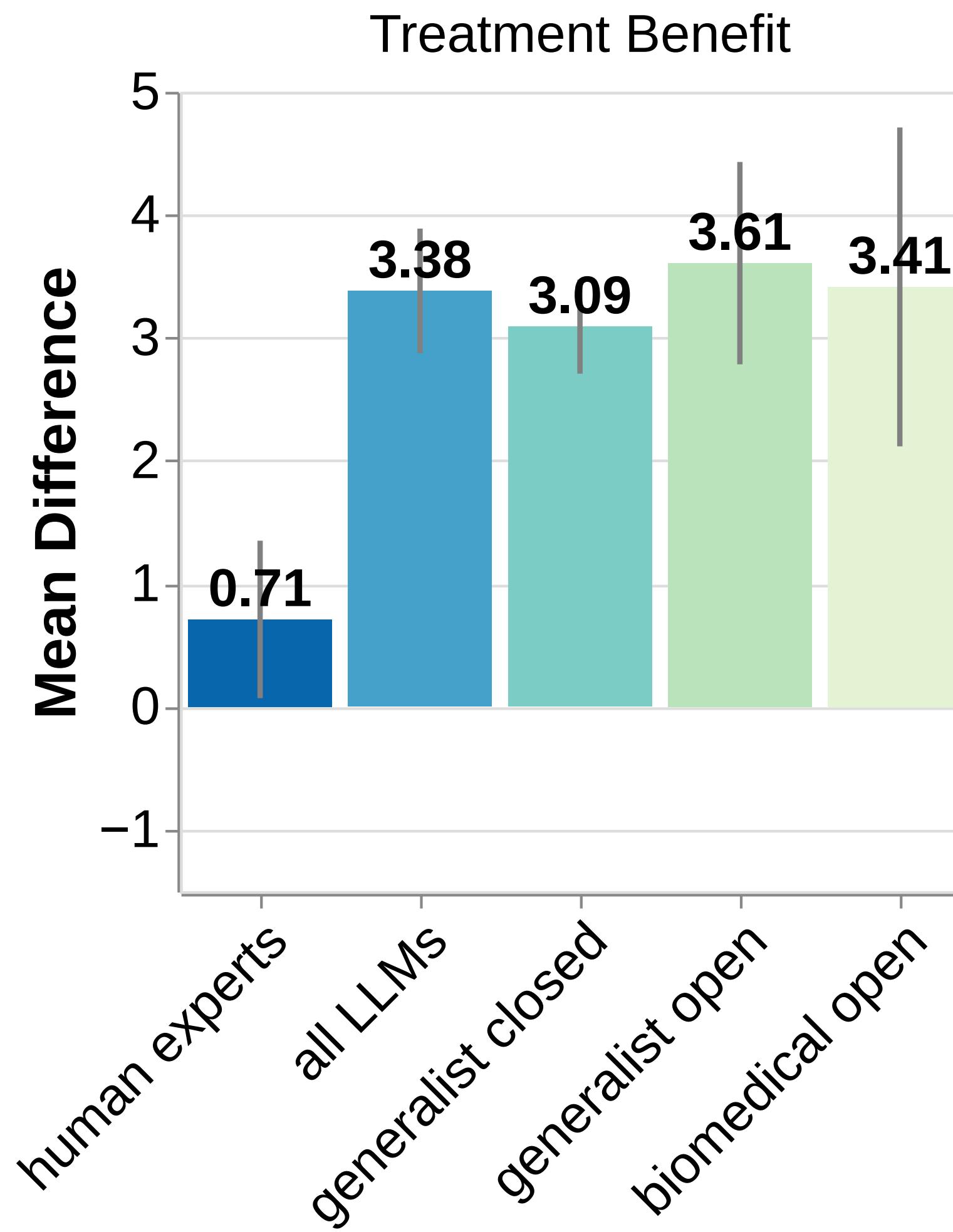


The treatment is effective.

# How well can LLMs detect the presence of spin?

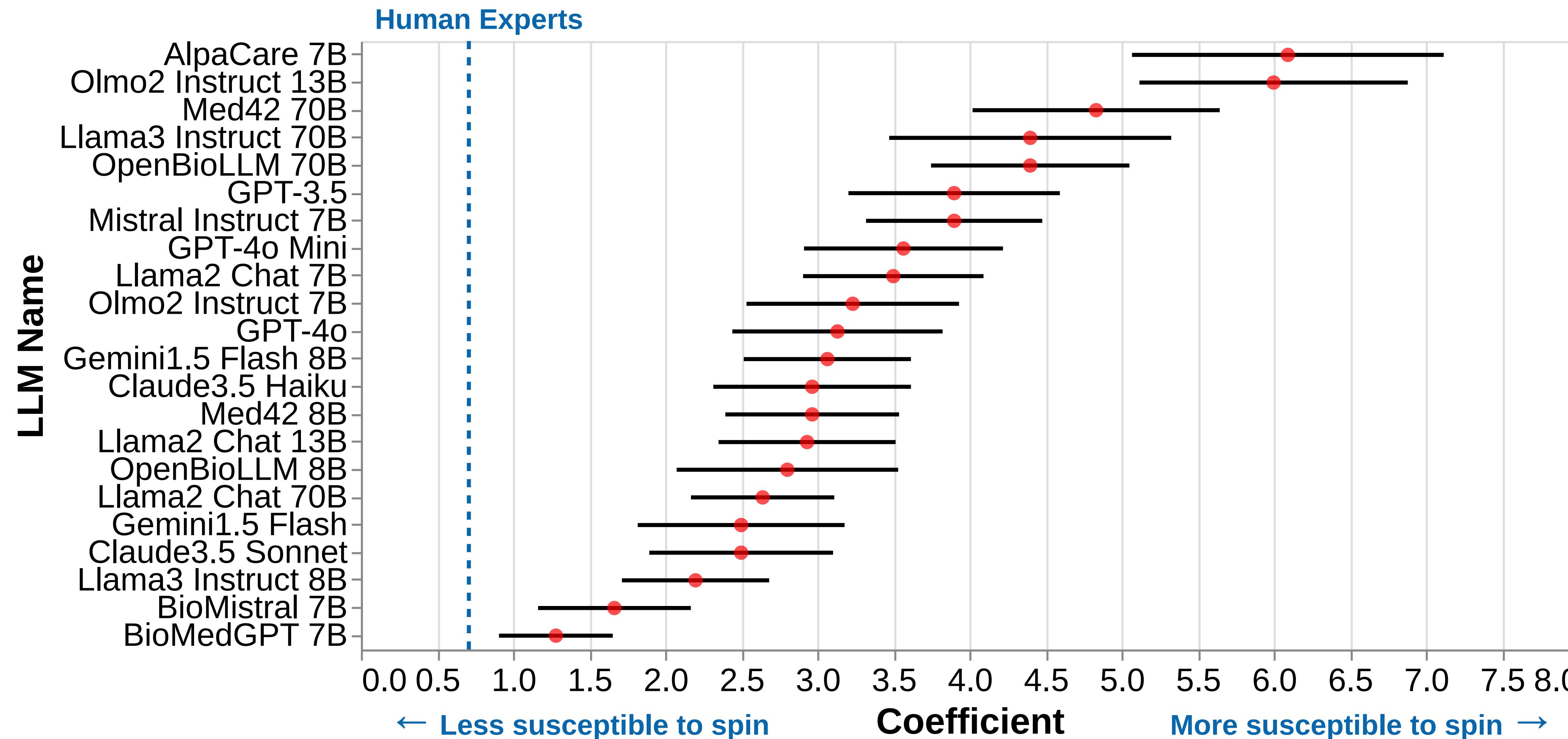


# How do LLMs interpret the same trials results?

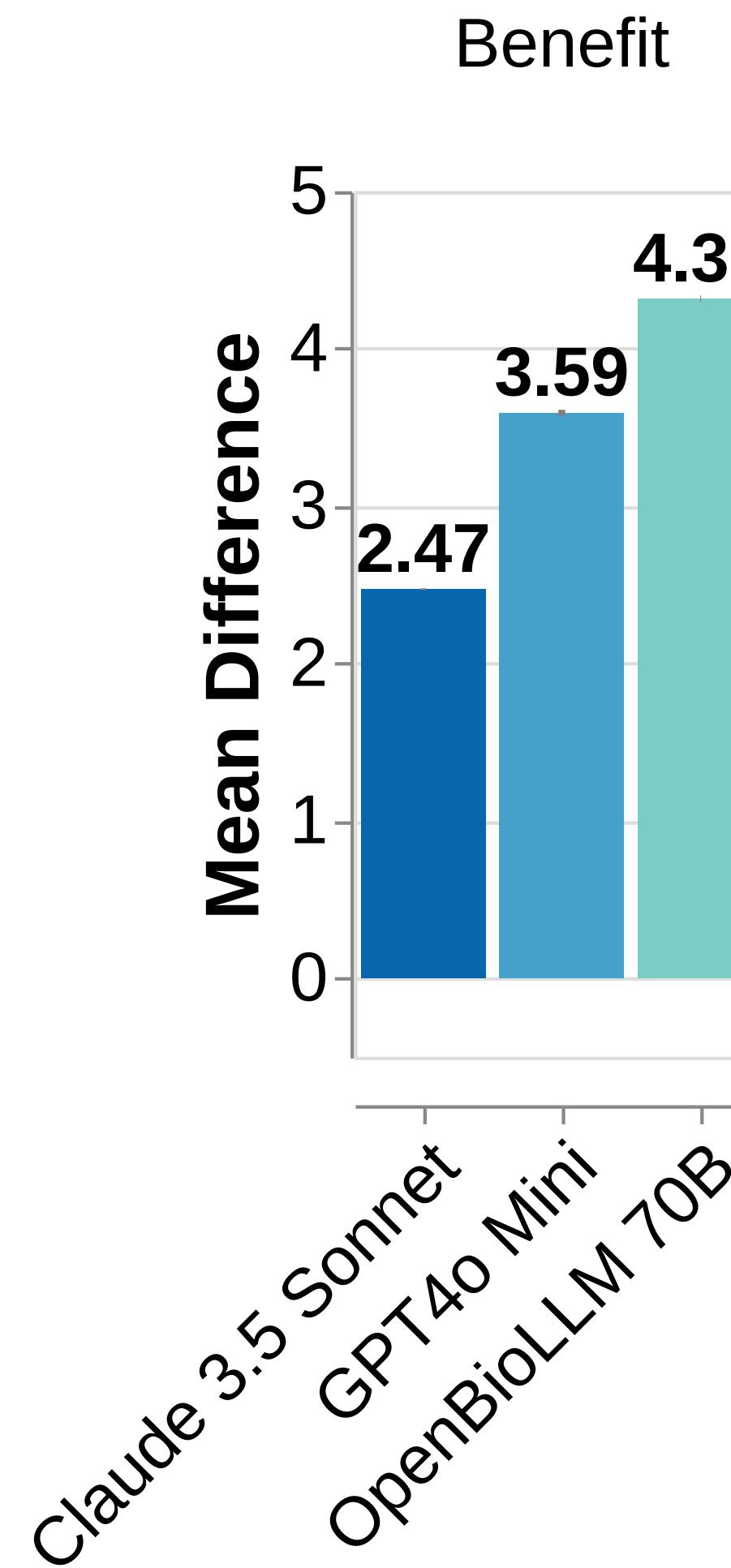


LLMs were far more influenced by spin than human experts

# How do LLMs interpret the same trials results?



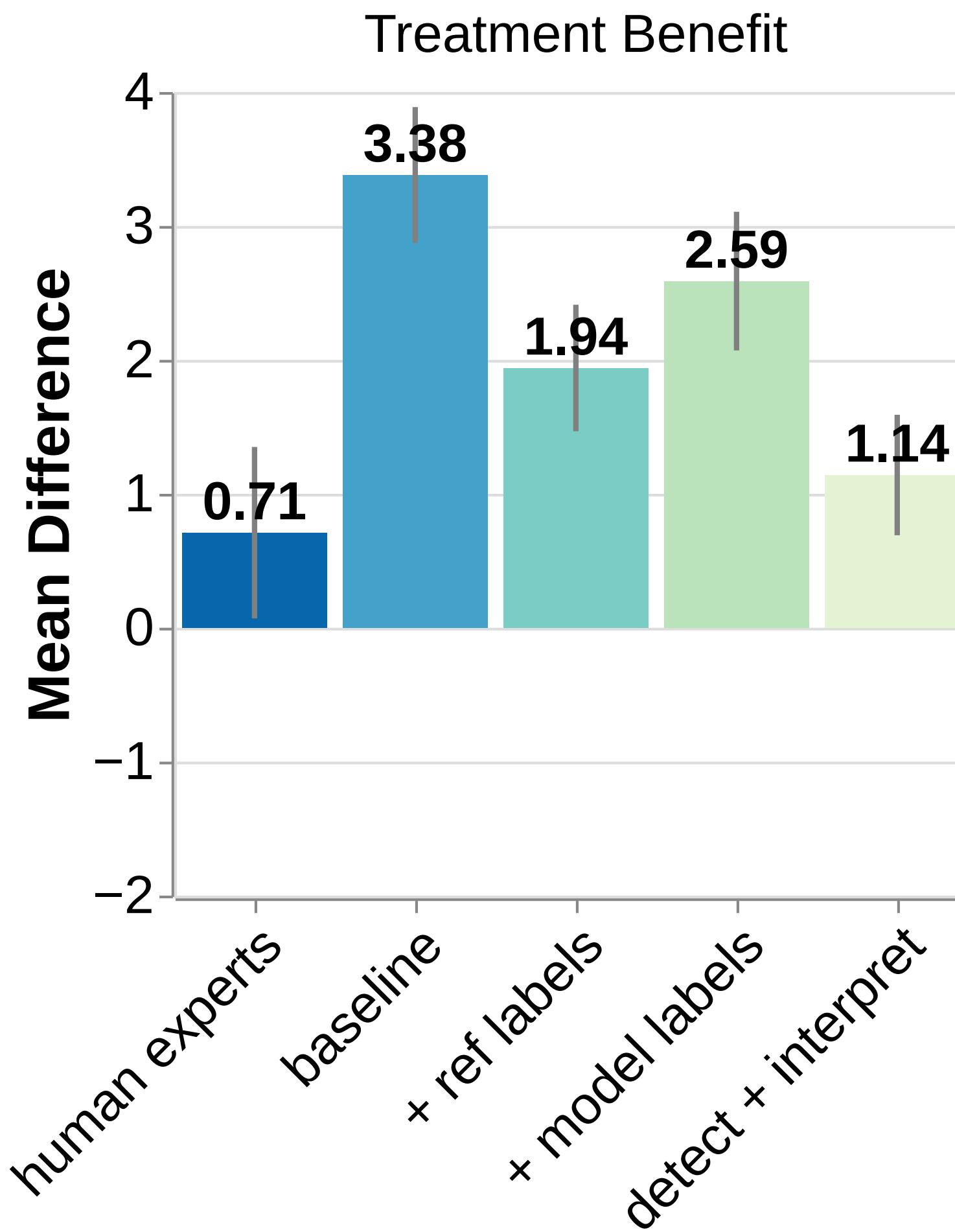
# Do LLMs propagate or amplify spin in medical abstracts when generating simplified versions?



- When we asked LLMs to simplify abstracts into plain language, they often propagated spin into their summaries.
- LLMs could unintentionally mislead patients and non-experts about the effectiveness of treatments.

# Can we fix this?

## Exploring mitigation strategies



- Tested zero-shot prompts to reduce LLMs' susceptibility to spin.
- Prompts that encourage reasoning reduce their tendency to overstate the trial results.

# Key Takeaways

- LLMs are biased like humans – potentially more biased
- Careful design is key to improving evidence synthesis for clinical decisions since LLMs are poor with numbers and can easily fall for spin
  - Using Chain-of-Thought style prompting can mitigate some of the issue
  - Focusing on tasks related to numerical results rather than interpretations

# Insights from both NLP & HCI fields are needed

- **NLP research methods to:**
  - Identify the general strengths and weaknesses of LLMs and AI
  - Rigorously evaluate the factuality and accuracy of LLM-generated texts
  - Improve the performance of LLMs and align them better to human needs
- **HCI research methods to:**
  - Design and evaluate human-AI interactions
  - Identify benefits and risks of LLMs outside of highly-controlled conditions



# Thank you!

## Any questions?



[yun.hy@northeastern.edu](mailto:yun.hy@northeastern.edu)



[hyesunyun.com](http://hyesunyun.com)



[hyesunyun.bsky.social](https://hyesunyun.bsky.social)



Framing Health  
Information



Do LLMs Fall for Spin in  
Medical Literature?