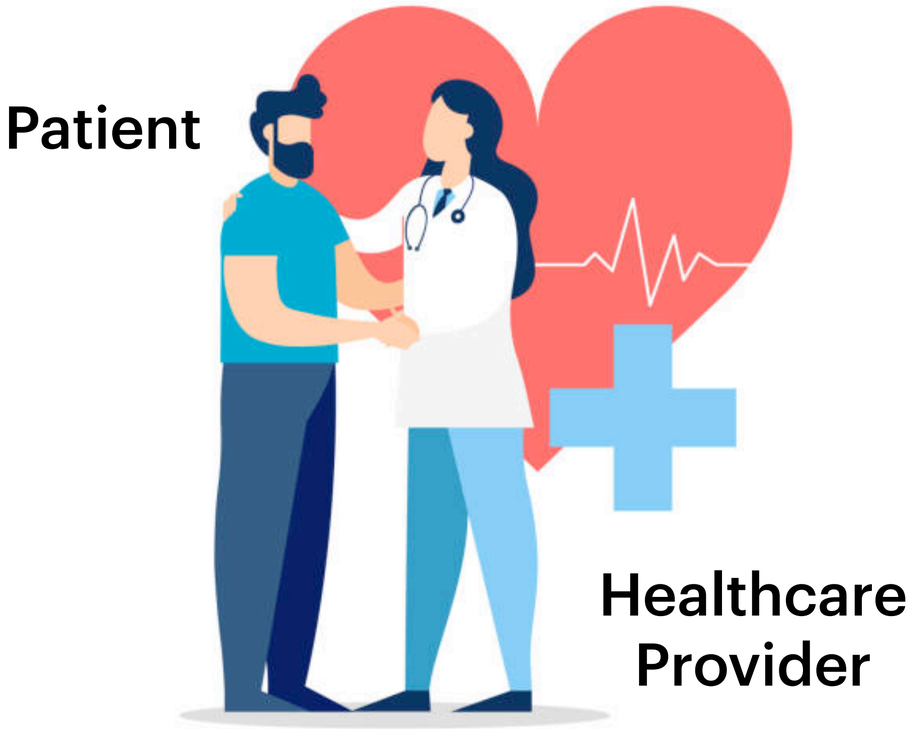
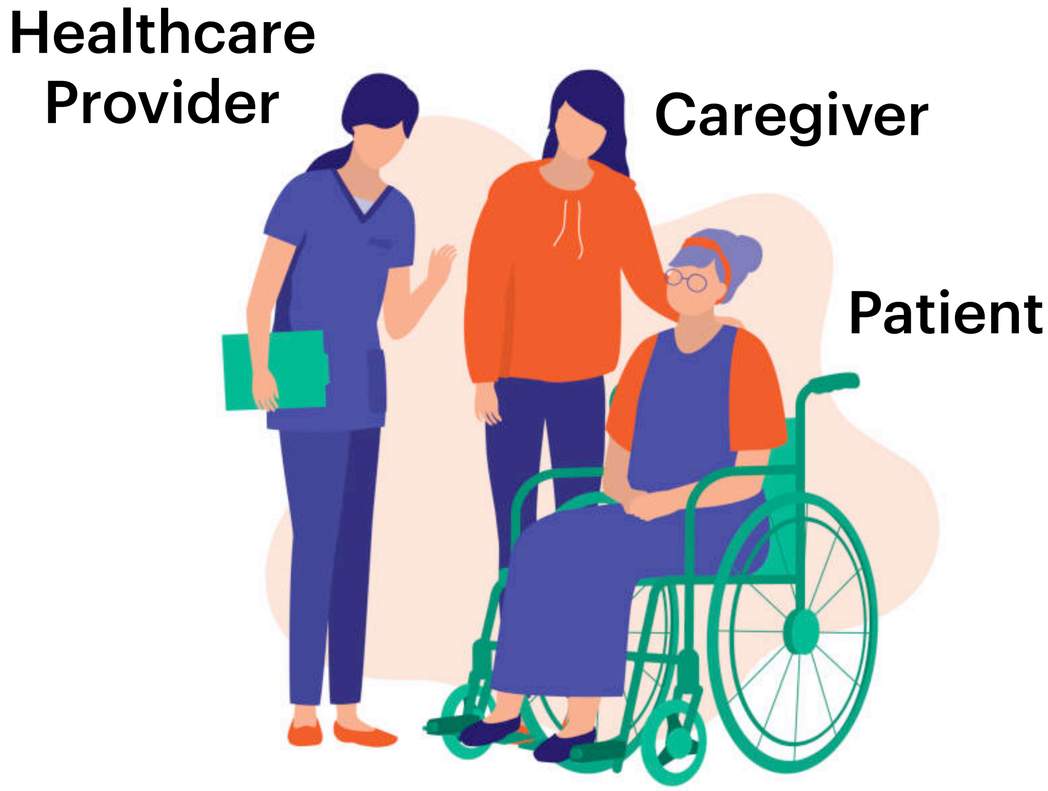
Beyond Hallucinations

Unveiling Hidden Dangers of LLMs in Health Information Access

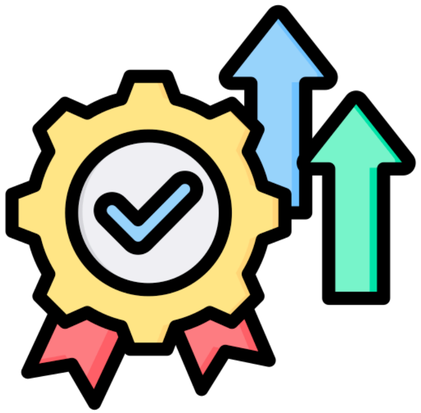
Hye Sun Yun — Northeastern University

Clemson School of Computing Seminar
April 18, 2025


Patient
Healthcare Provider


Healthcare Provider
Caregiver
Patient

Evidence-based medicine as a model of care



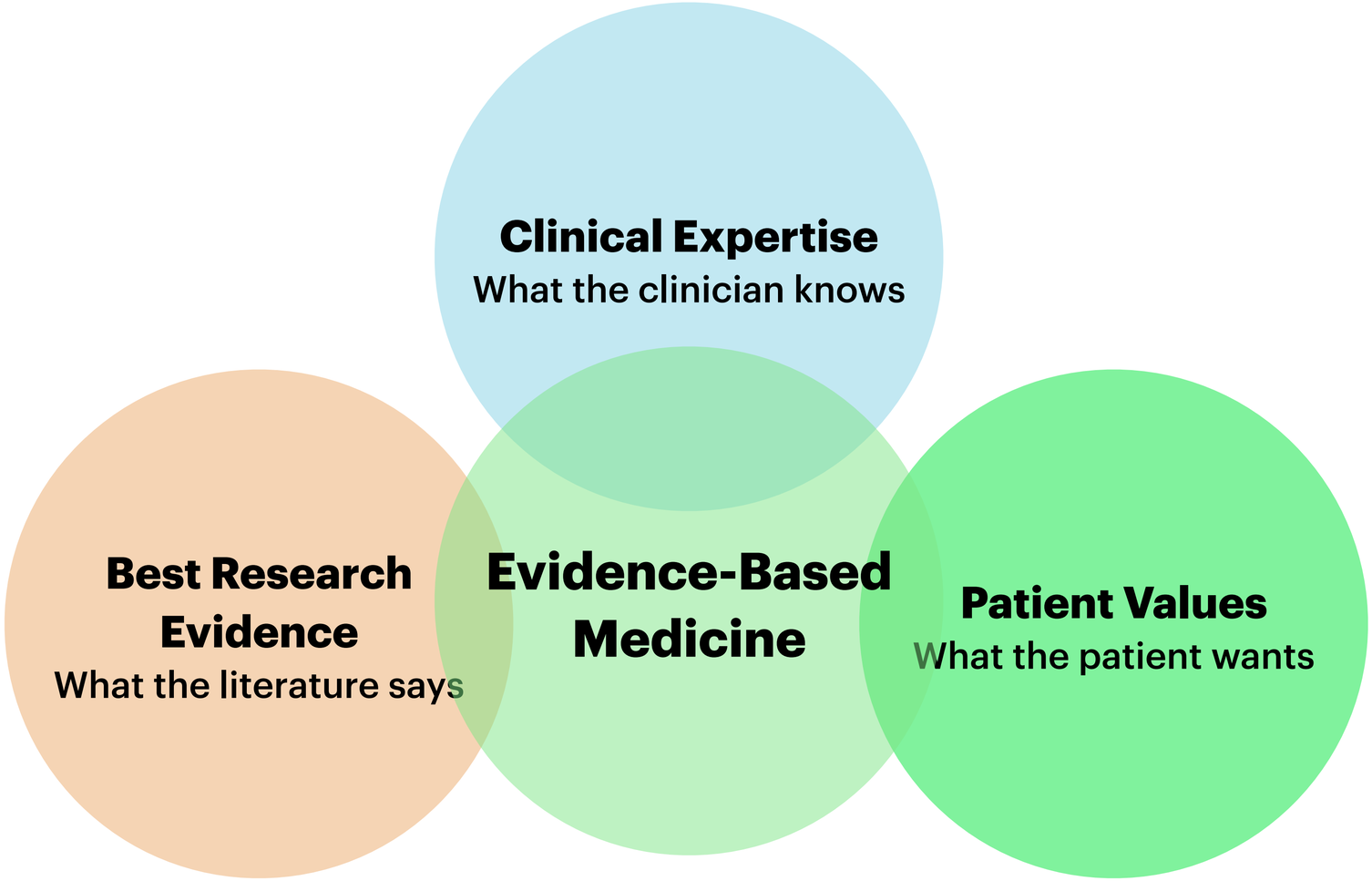
improved quality

improved patient satisfaction

reduced costs

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.

Key components of evidence-based medicine



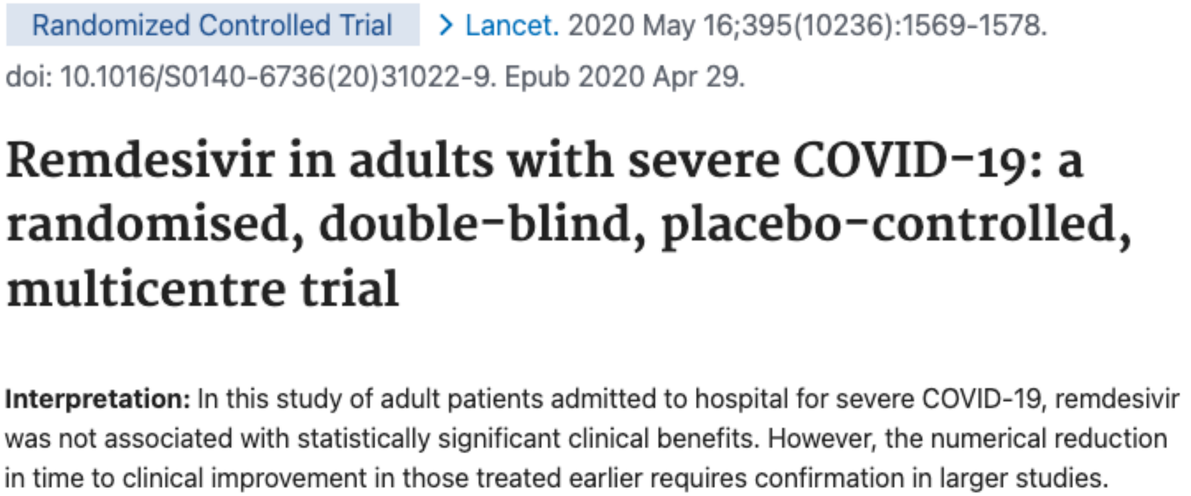
Sackett DL, Rosenberg WM, Gray JM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *bmj*. 1996 Jan 13;312(7023):71-2.

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)

MEDLINE
U.S. National Library of Medicine

PubMed®



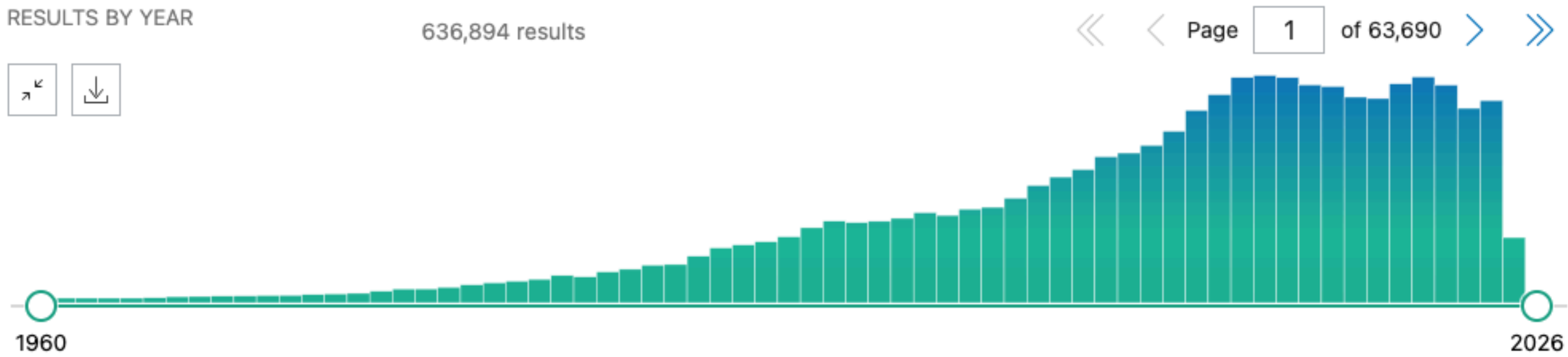
Randomized Controlled Trial > Lancet. 2020 May 16;395(10236):1569-1578.
doi: 10.1016/S0140-6736(20)31022-9. Epub 2020 Apr 29.

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

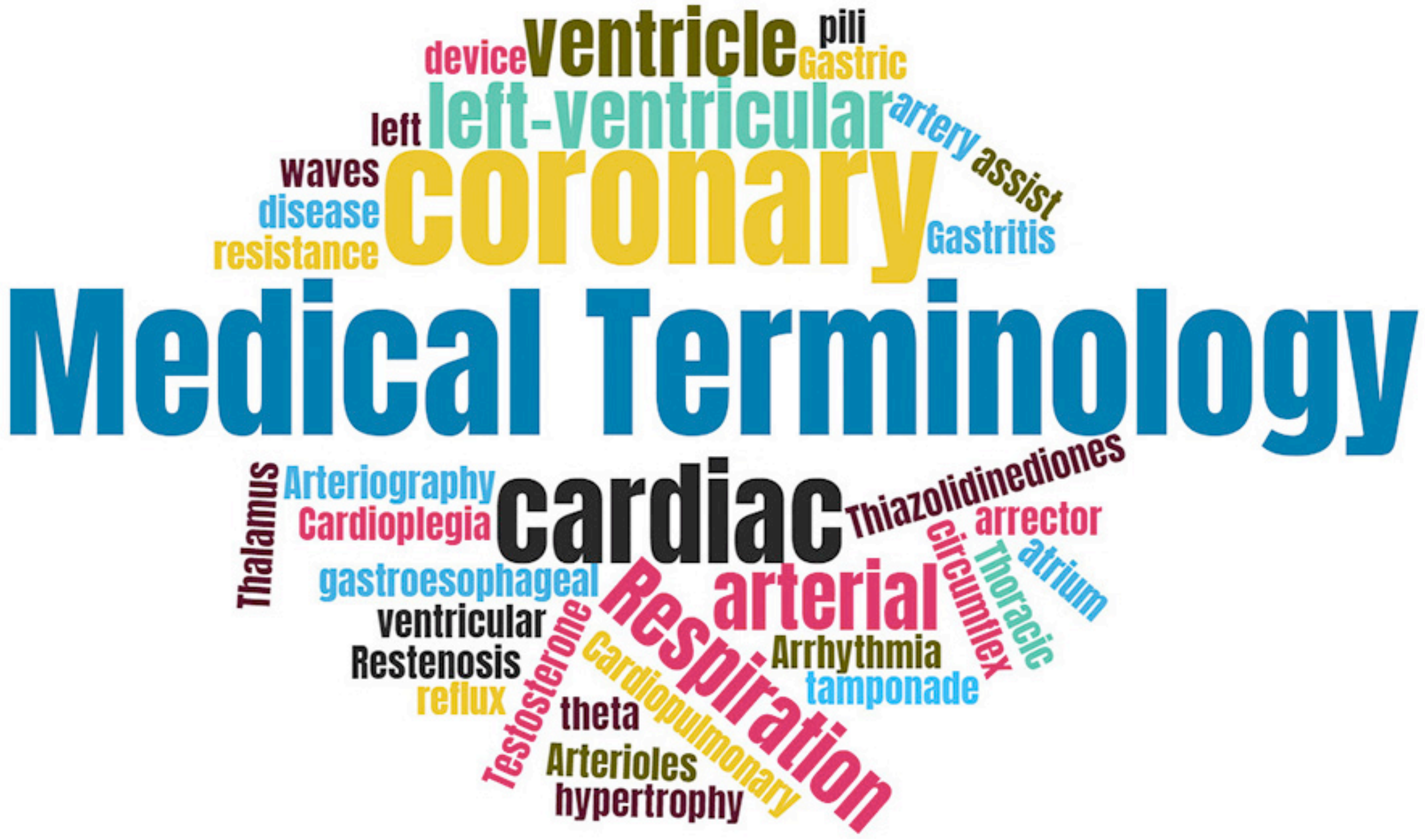
Information Overload for Healthcare Providers



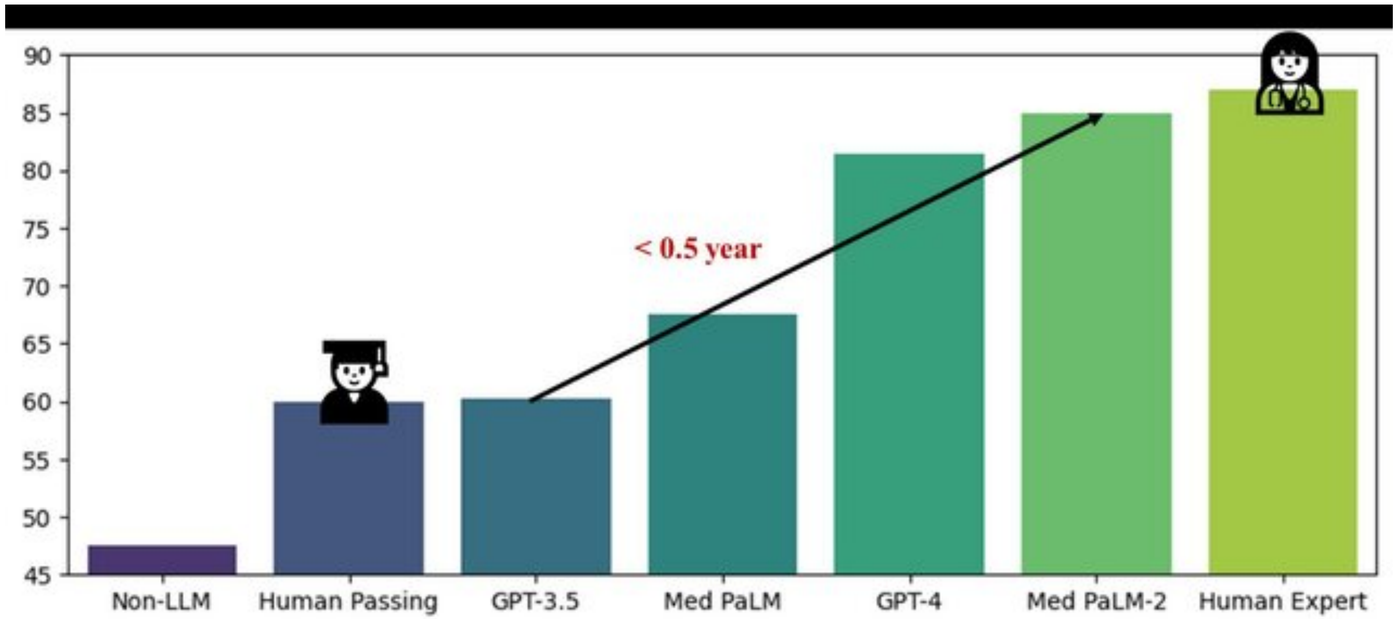
Number of Published PubMed RCT articles (April 2025)

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

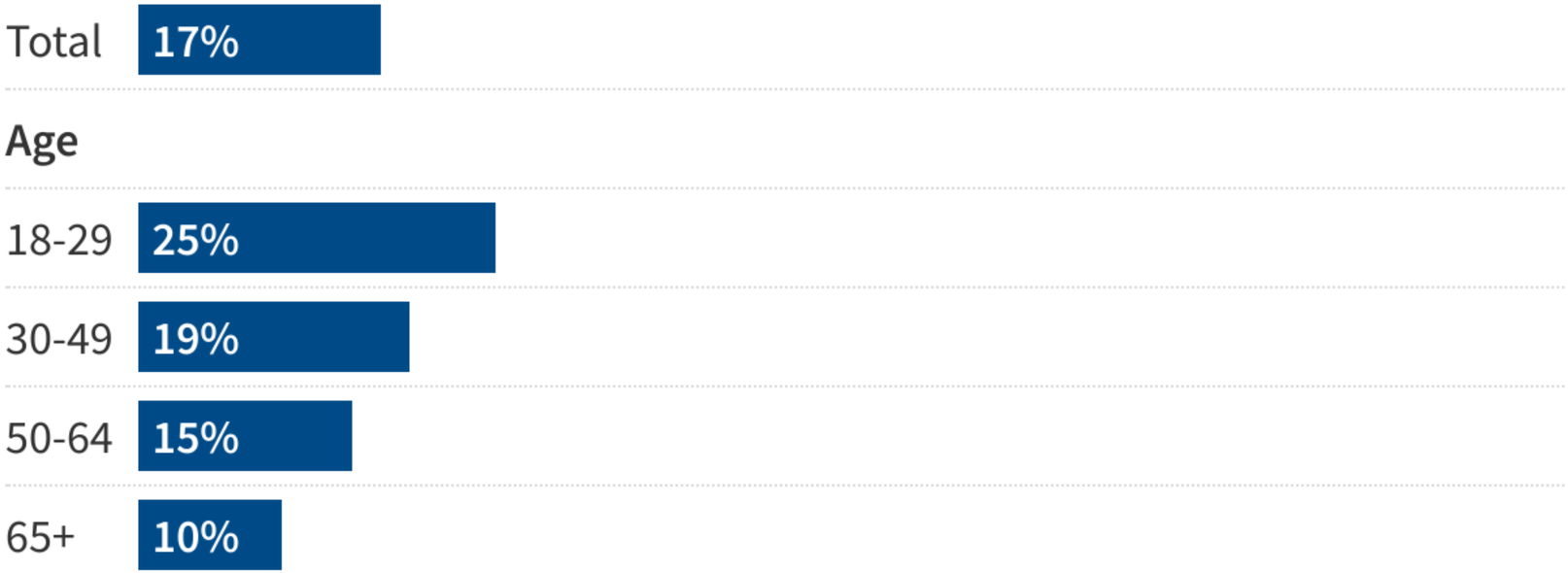
The results of the study showed that the glass meal was the most effective at lowering stomach acid output, and the wheat bran meal was the least effective.

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:



Note: See topline for full question wording.

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

KFF

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**?

Yun & Bickmore. **Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs** (CHI 2025)

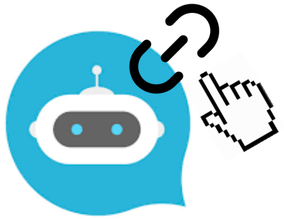
Simulation Videos

For Framing Health Information

Three **Search Methods**

search engine

chatbot

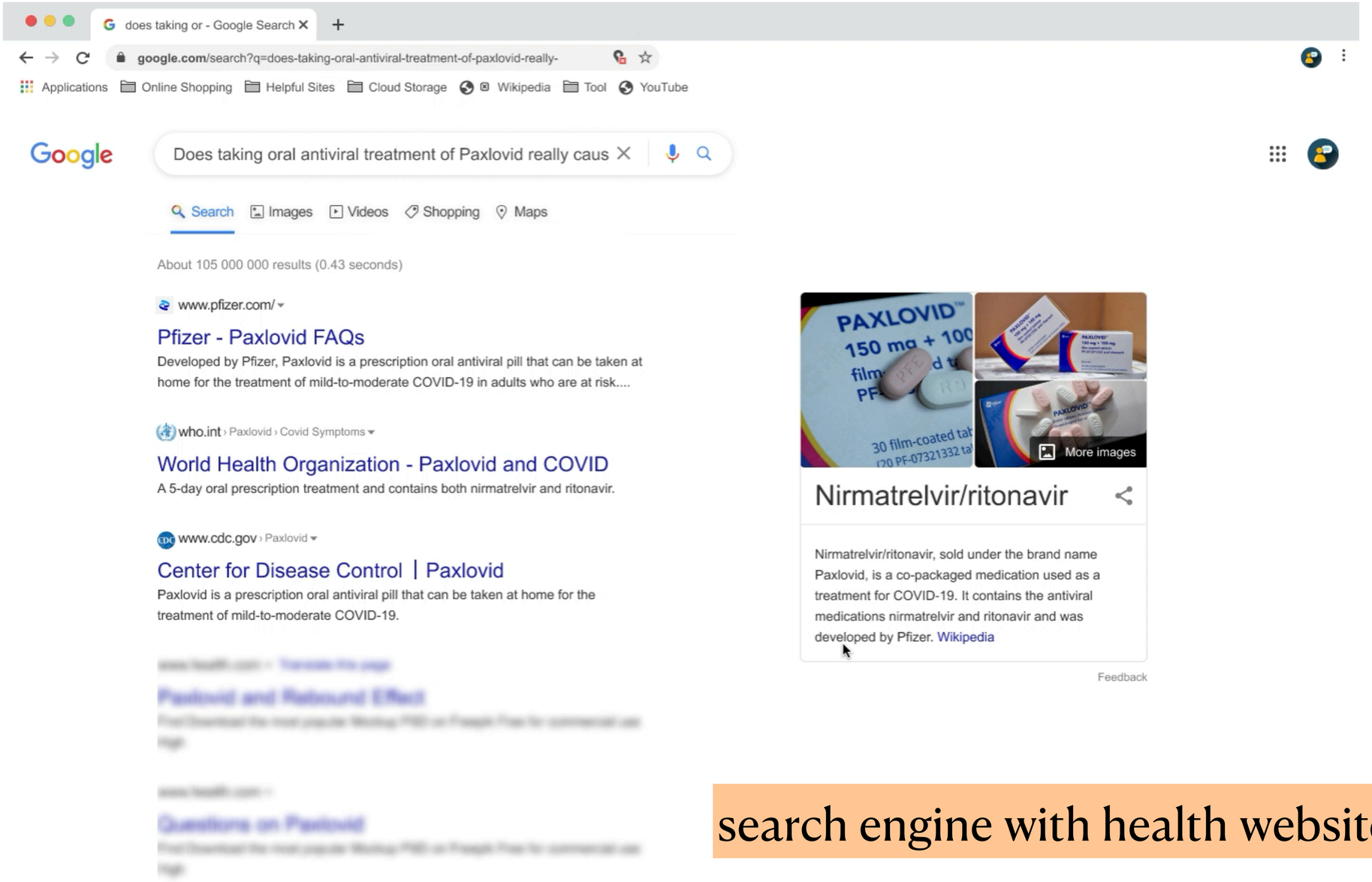


chatbot+

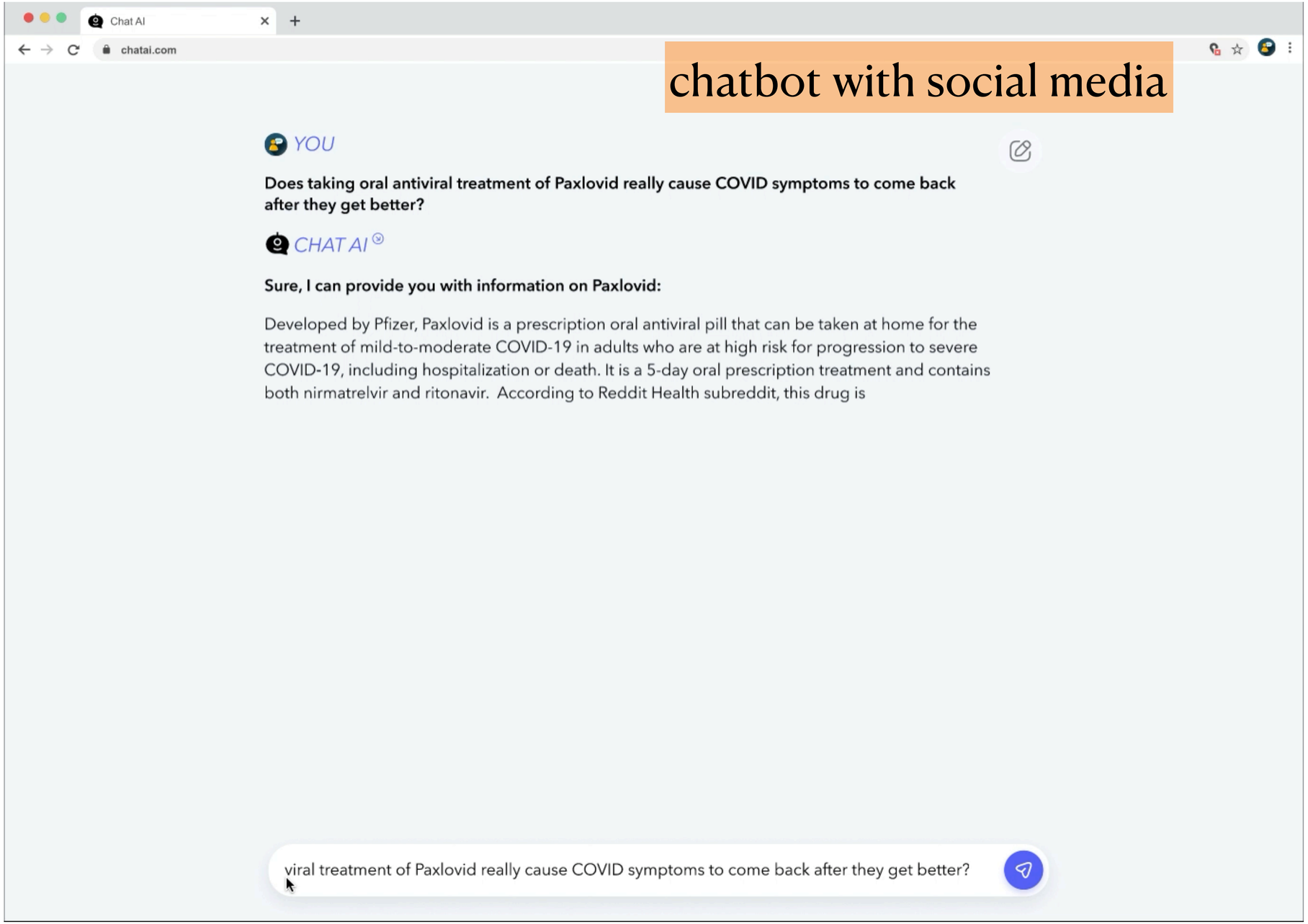
Two **Source Types**

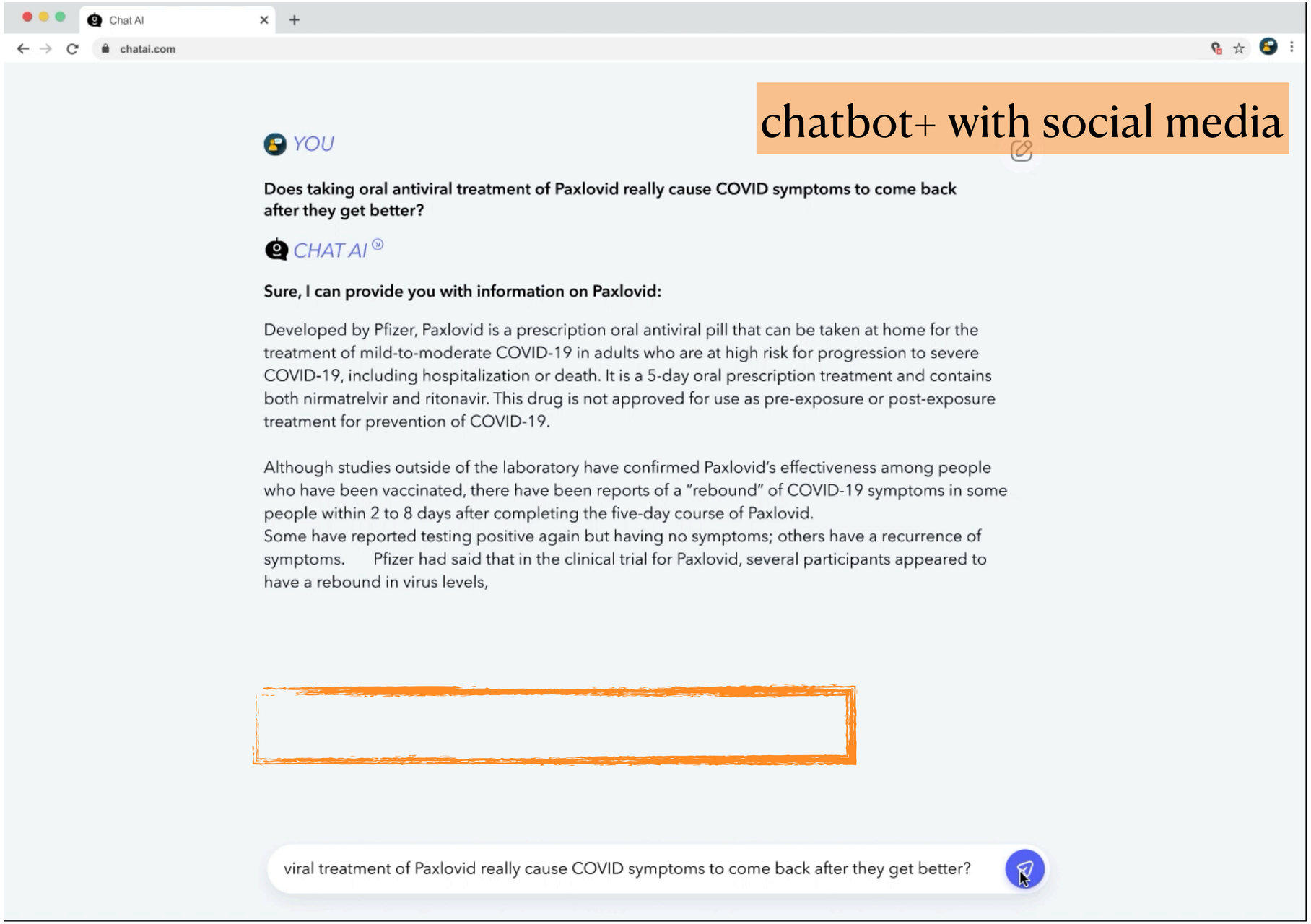
health websites

social media



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

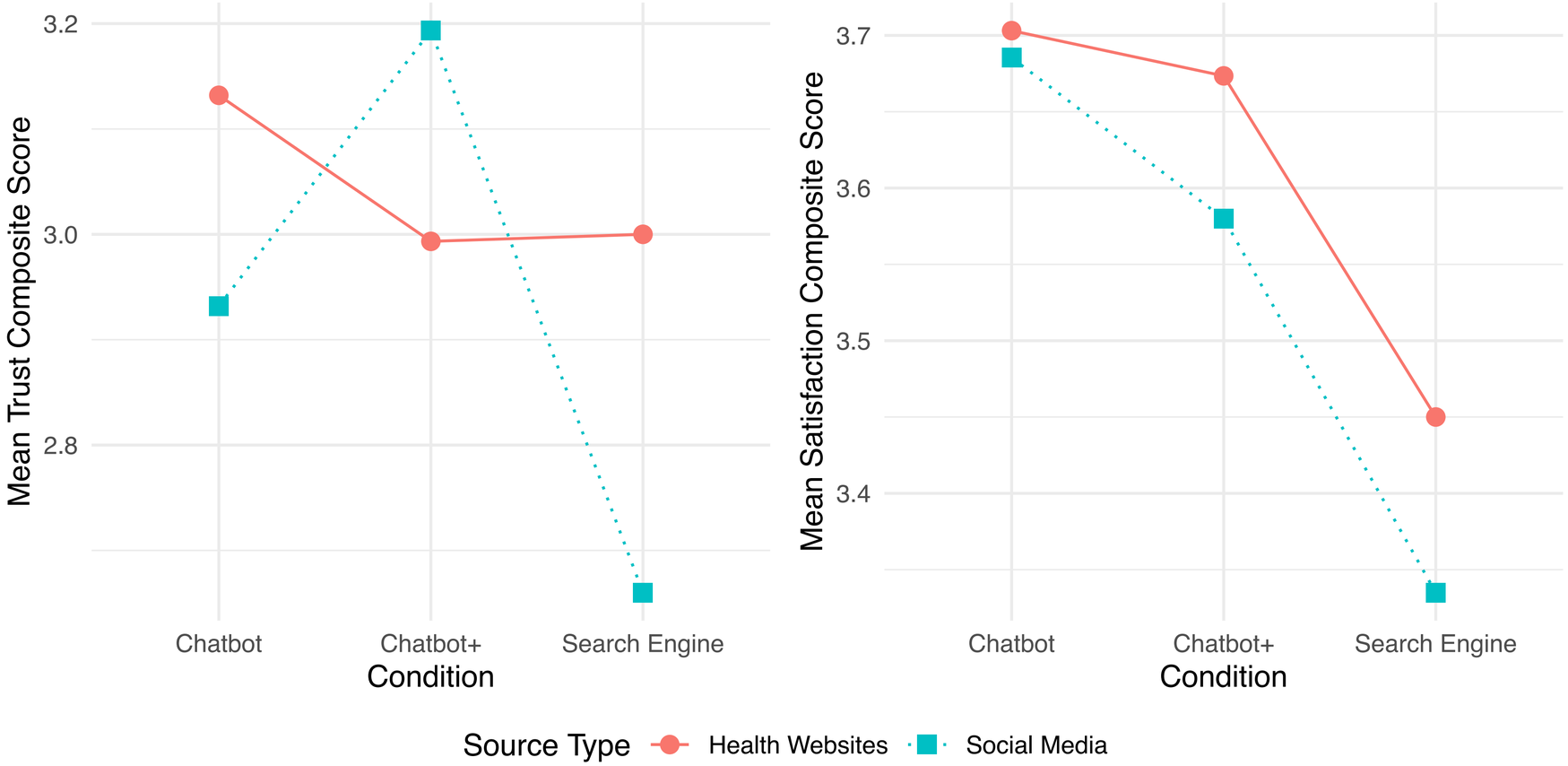
Survey (between-subjects)

- 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

Yun & Bickmore. **Framing Health Information: The Impact of Search Methods and Source Types on User Trust and Satisfaction in the Age of LLMs** (CHI 2025)

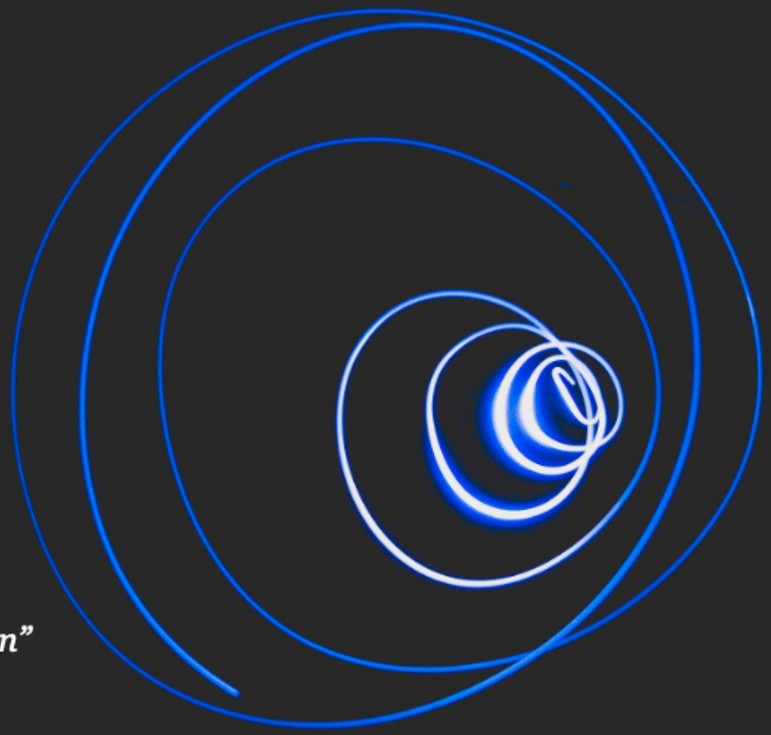
Spin in medical literature

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 [a], D. Margalski [b], K. Vaughn [c], M. Vassar [d]

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

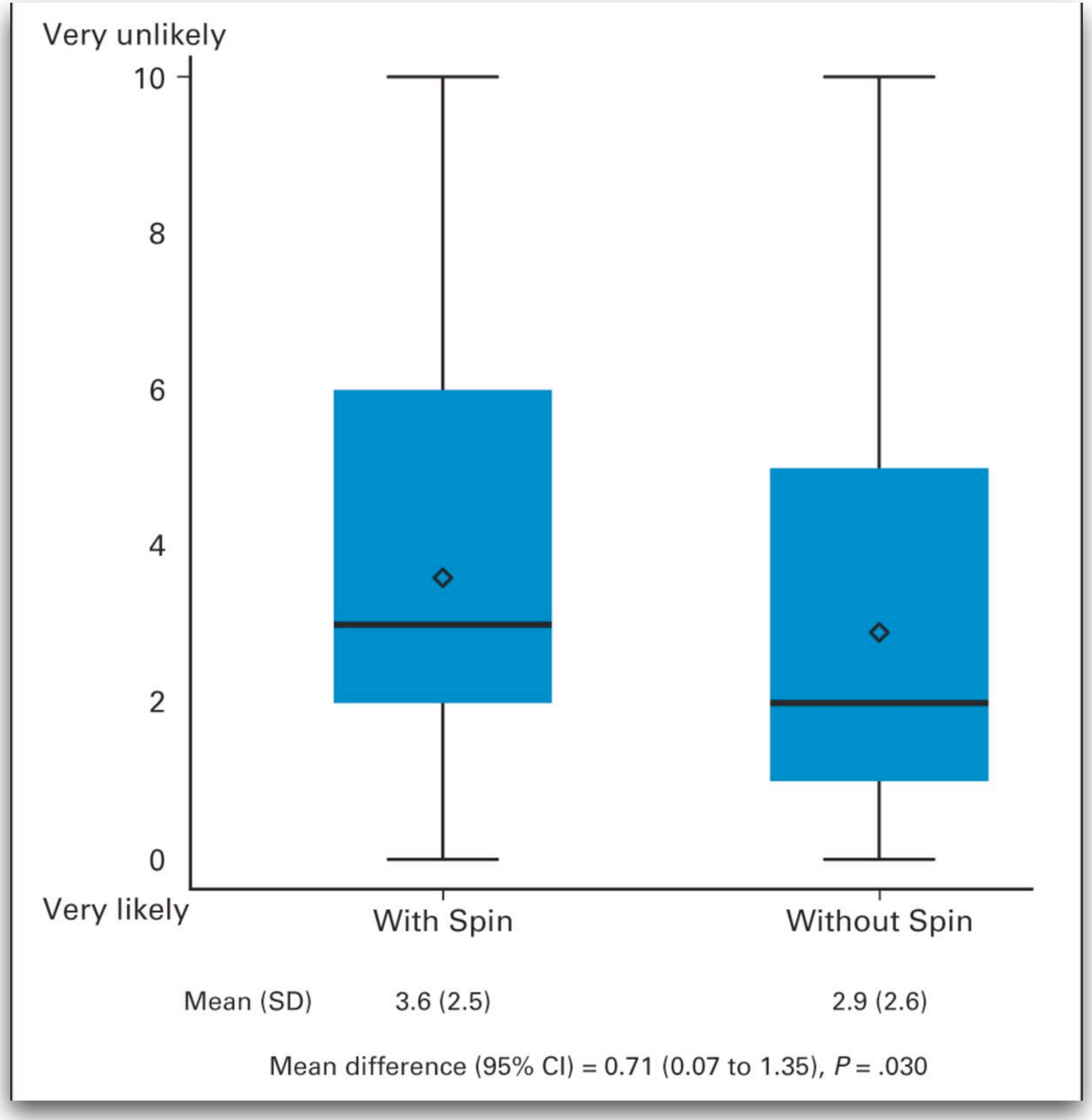
Samuel Jellison [1], Will Roberts [1], Aaron Bowers [1], Tyler Combs [1], Jason Beaman [2, 3], Cole Wayant [1], Matt Vassar [1]

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

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

Clinicians can fall for spin

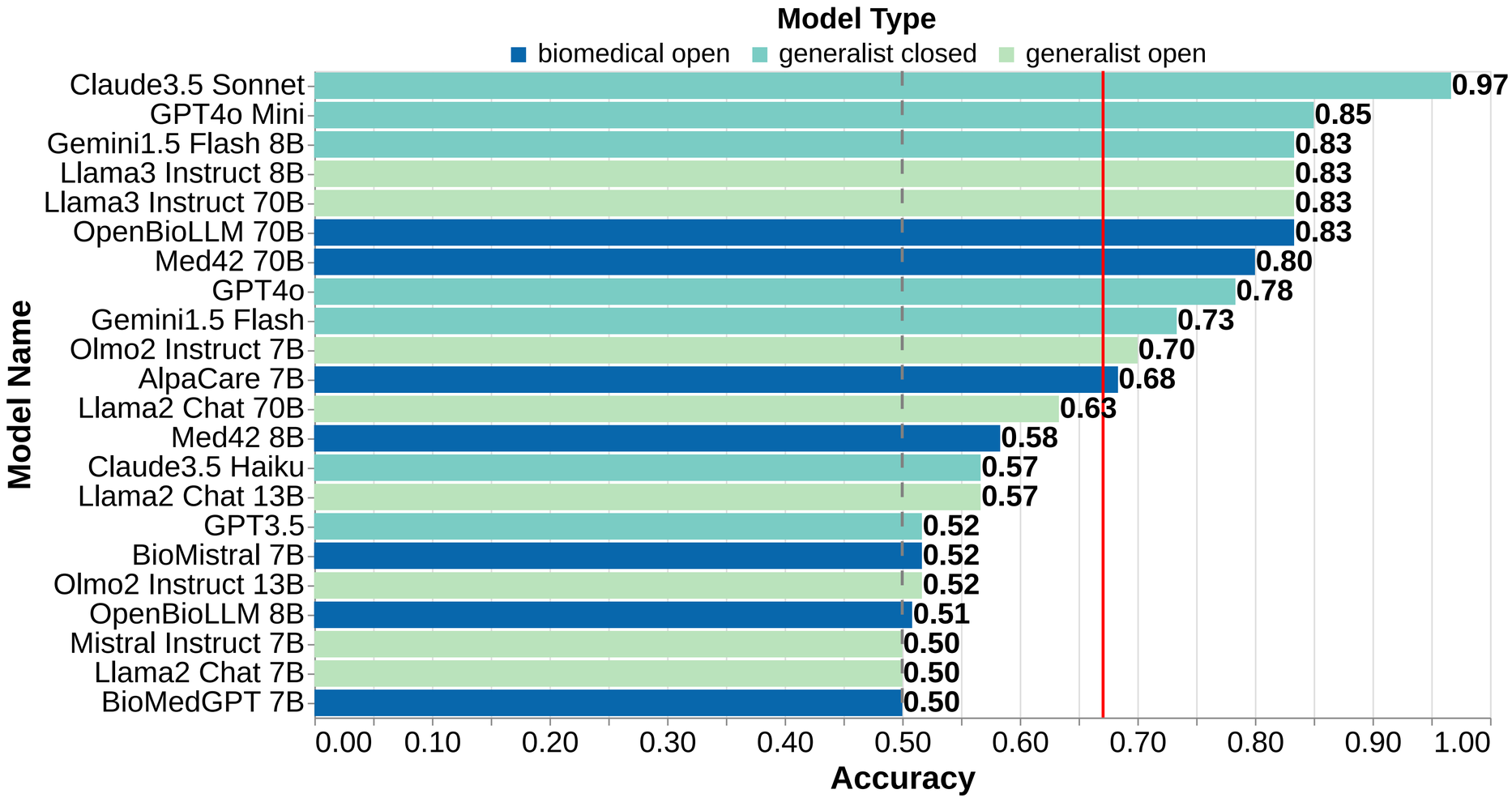
- A study in 2014 assessed the impact of spin on the interpretations of results of abstracts of RCTs in the field of cancer.
- Clinicians overstated the benefits of results when shown an abstract with spin.



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?

How well can LLMs detect the presence of spin?

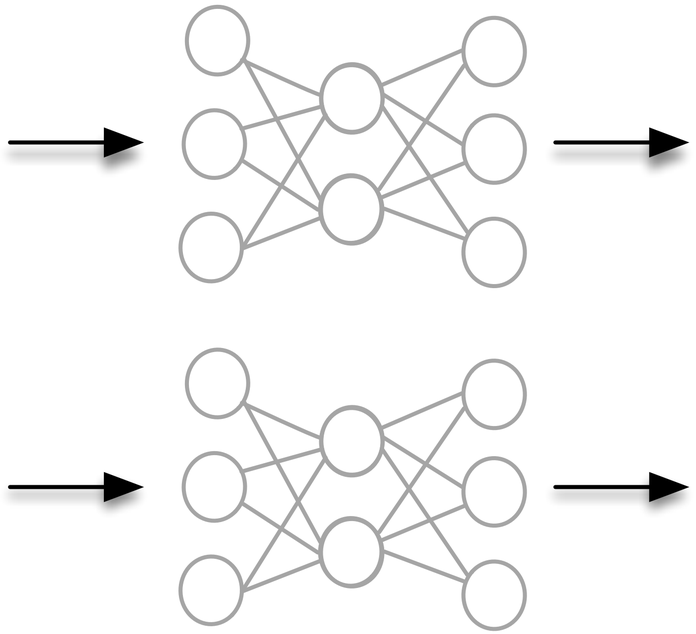


Neutral and spun abstracts (same results)

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

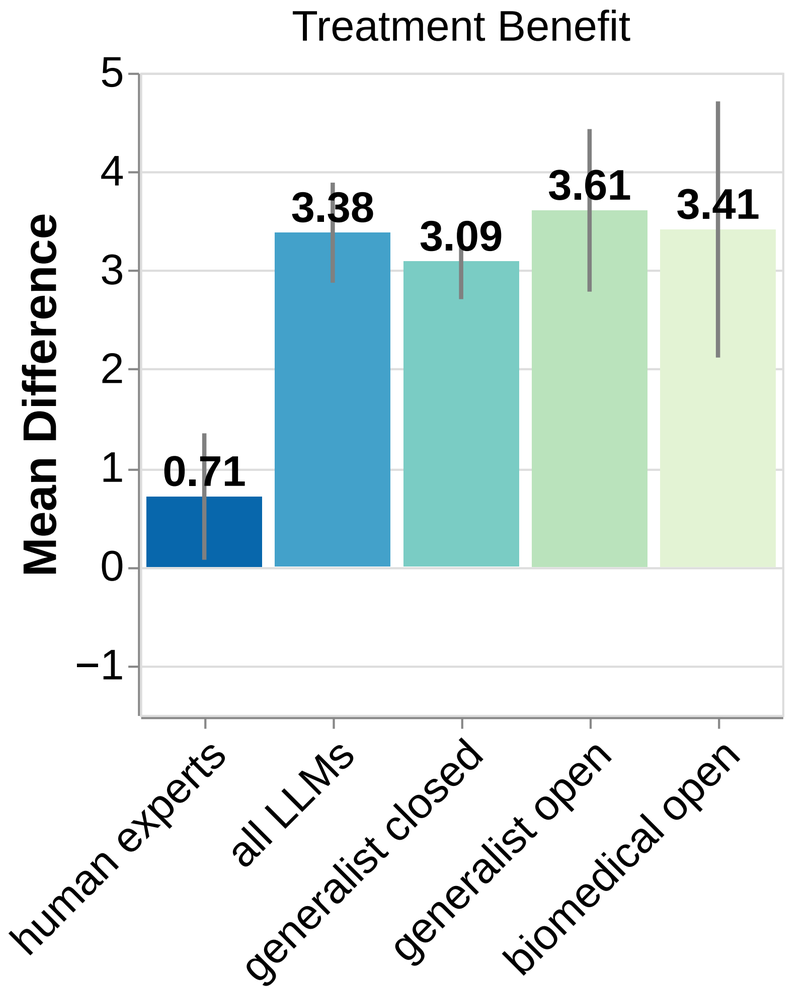
`No evidence for the treatment.`

`The treatment is effective.`

Yun, Zhang, Kouzy, Marshall, Li, & Wallace. **Caught in the Web of Words: Do LLMs Fall for Spin in Medical Literature?** (CHIL 2025)

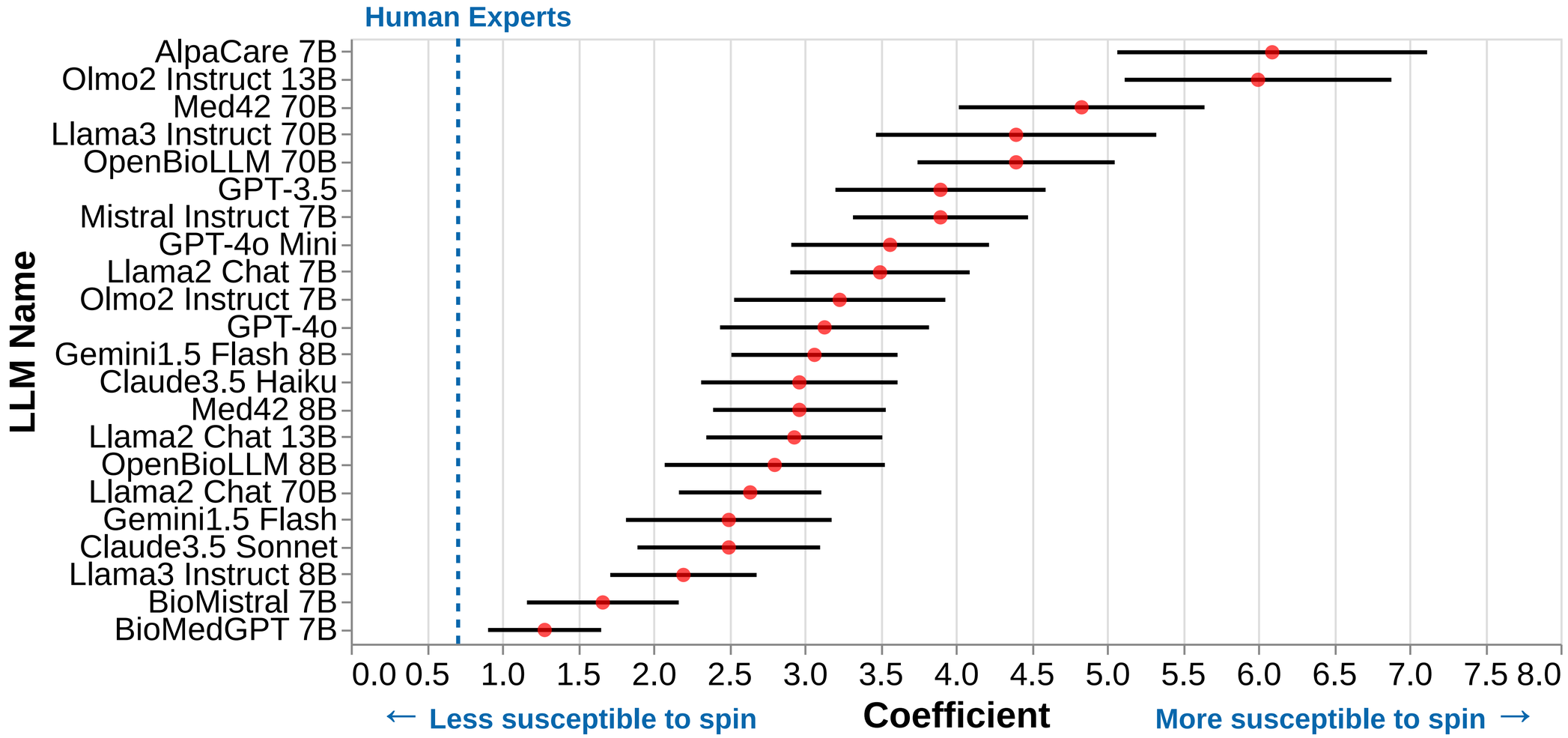
abstract with spin	abstract without spin
Purpose: To compare the efficacy and tolerability of treatment A with comparator B in the treatment of advanced breast cancer in patients whose disease progresses on prior endocrine treatment.	**Purpose**: To compare the efficacy and tolerability of treatment A with comparator B in the treatment of advanced breast cancer in patients whose disease progresses on prior endocrine treatment.
Patients and methods: In this double-blind, double-dummy, parallel-group study, postmenopausal patients were randomized to receive either treatment A or comparator B. The primary end point was time to progression (TTP). Secondary end points included objective response (OR) rate, duration of response (DOR), and tolerability.	**Patients and methods**: In this double-blind, double-dummy, parallel-group study, postmenopausal patients were randomized to receive either treatment A or comparator B. The primary end point was time to progression (TTP). Secondary end points included time to treatment failure (TTF), objective response (OR) rate, duration of response (DOR), and tolerability.
Results: Patients (n = 400) were followed for a median period of 16.8 months. Treatment A was as effective as comparator B in terms of TTP (hazard ratio, 0.92; 95% confidence interval [CI], 0.74 to 1.14; P =.43); median TTP was 5.4 months with treatment A and 3.4 months with comparator B. OR rates were 17.5% with both treatments. Clinical benefit rates (complete response + partial response + stable disease > or = 24 weeks) were 42.2% for treatment A and 36.1% for comparator B (95% CI, -4.00 to 16.41%; P =.26). In responding patients, median DOR (from randomization to progression) was 19.0 months for treatment A and 10.8 months for comparator B. Using all patients, DOR was significantly greater for treatment A compared with comparator B; the ratio of average response durations was 1.35 (95% CI, 1.10 to 1.67; P < 0.01). Both treatments were well tolerated.	**Results**: Patients (n = 400) were followed for a median period of 16.8 months. Treatment A was not more effective than comparator B in terms of TTP (hazard ratio, 0.92; 95% confidence interval [CI], 0.74 to 1.14; P =.43); median TTP was 5.4 months with treatment A and 3.4 months with comparator B. There was no statistically significant difference between the 2 groups for TTF. Median TTF was 4.6 months for treatment A and 3.3 months for comparator B (HR, 0.96; 95% CI, 0.77 to 1.19; P = .69). At the time of this data analysis, the rate of deaths was respectively for treatment A and comparator B, 35.4% (n=73) vs. 33.5% (n=65). OR rates were 17.5% with both treatments. DOR was statistically significantly greater for treatment A compared with comparator B; the ratio of average response durations was 1.35 (95% CI, 1.10 to 1.67; P < 0.01). Both treatments were well tolerated.
Conclusion: Treatment A was at least as effective as comparator B, with efficacy end points slightly favoring treatment A. Treatment A represents an additional treatment option for postmenopausal women with advanced breast cancer whose disease progresses on tamoxifen therapy.	**Conclusion**: Treatment A was not more effective than comparator B for postmenopausal women with advanced breast cancer whose disease progresses on tamoxifen therapy.

How do LLMs interpret the same trials results?

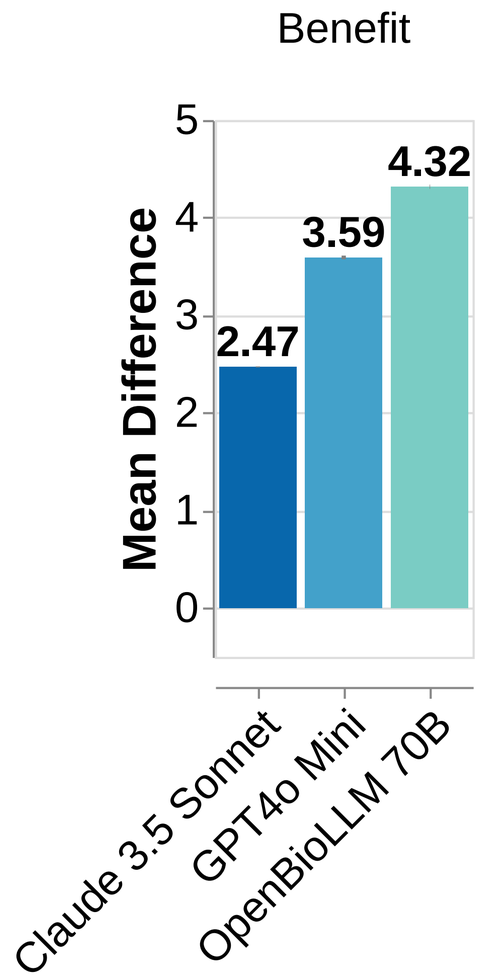


LLMs were far more influenced by spin than human experts

How do LLMs interpret the same trials results?



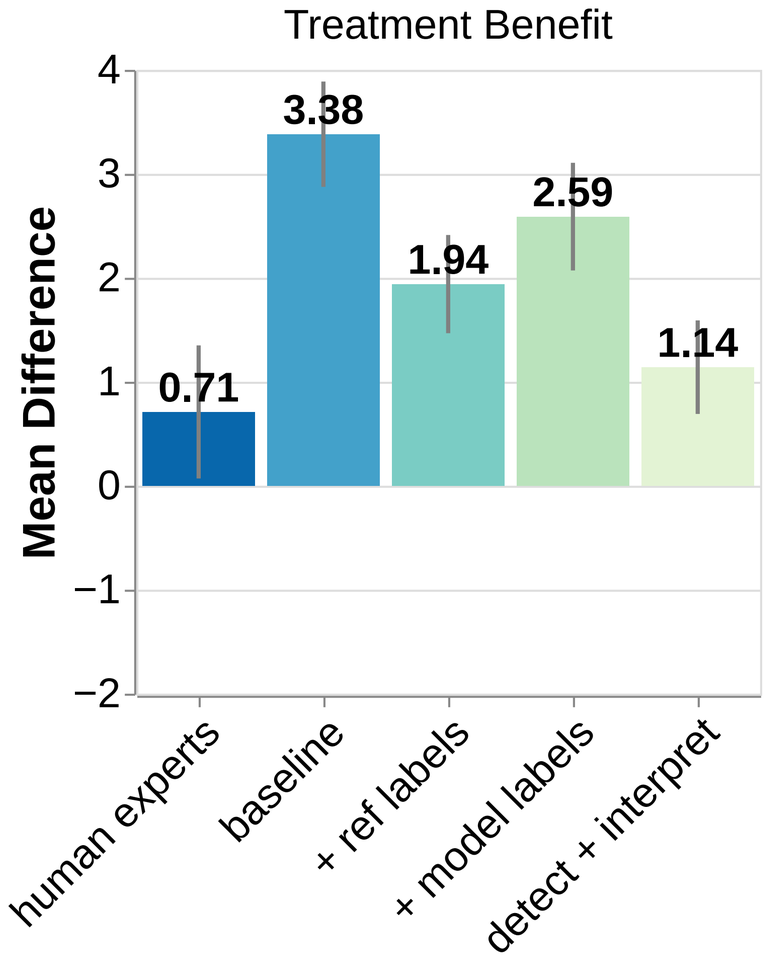
Do LLMs propagate or amplify spin in medical abstracts when generating simplified versions?



- 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 provide outputs aligned with input text but is misleading
- 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

Yun, Zhang, Kouzy, Marshall, Li, & Wallace. **Caught in the Web of Words: Do LLMs Fall for Spin in Medical Literature?** (CHIL 2025)

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






Northeastern University

TEXAS
The University of Texas at Austin

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center

KING'S College LONDON

Thank you!
Any questions?

yun.hy@northeastern.edu

hyesunyun.com

hyesunyun.bsky.social

Framing Health Information

Do LLMs Fall for Spin in Medical Literature?