Caught in the Web of Words: Do LLMs Fall for Spin in Medical Literature?

Hye Sun Yun, Karen Y.C. Zhang, Ramez Kouzy, Iain J. Marshall, Junyi Jessy Li, Byron C. Wallace
{yun.hy, zhang.yuchen, b.wallace}@northeastern.edu, rkouzy@mdanderson.org, iain.marshall@kcl.ac.uk, jessy@utexas.edu

What is Spin?

Spin refers to reporting strategies that ***overstate the benefits of experimental treatments*** beyond what is supported by empirical evidence. Spin might seek to distract readers from statistically nonsignificant results and/or understate the harms of a treatment which can influence clinician interpretation of evidence and may affect patient care decisions.

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

Research Questions

1. How well can LLMs **detect the presence of spin** in abstracts of RCT reports?
2. How do LLMs **interpret the same trial results** when presented with spun versus unspun abstracts?
3. To what extent might LLMs **propagate or amplify spin** in medical abstracts when generating simplified versions?

Methods

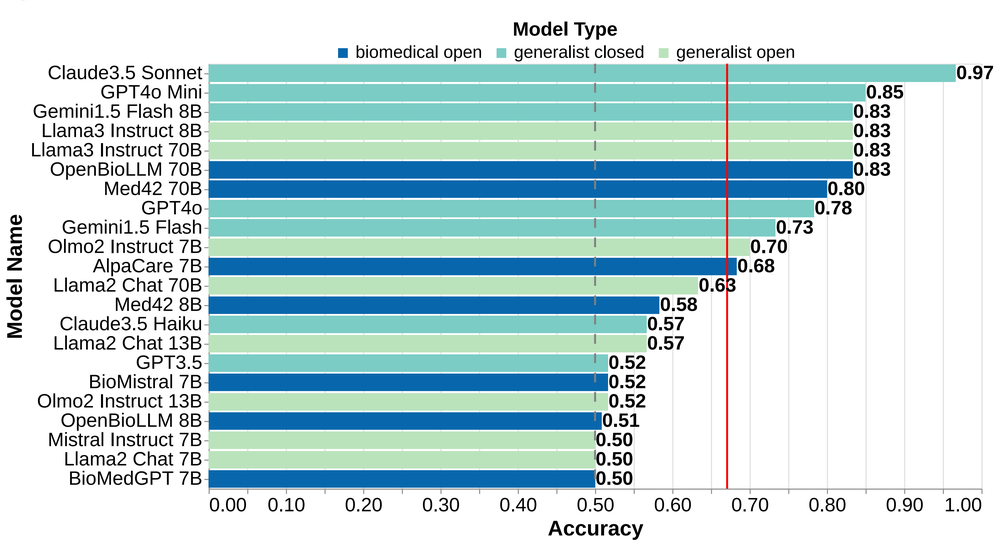
Data: 30 pairs of high-quality oncology abstracts (50% with spin, 50% without spin) (Boutron et al., 2014)

Spin Detection: Prompted 22 LLMs to answer whether or not a given abstract contains spin as a *binary classification problem*.

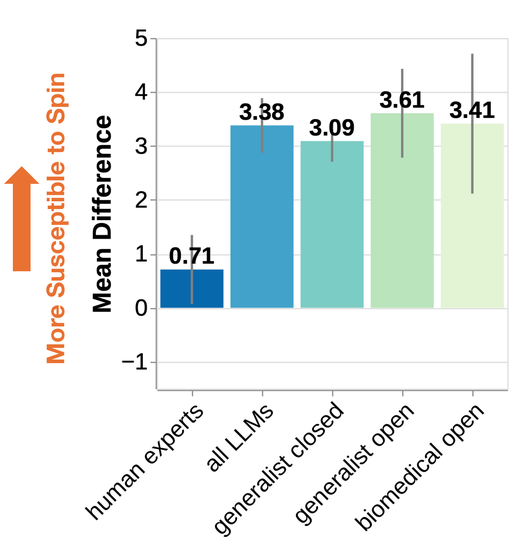
Susceptibility to Spin: The following question was asked to 22 LLMs for each abstract/LLM-generated simplified version: *"Based on this abstract, do you think treatment A would be beneficial to patients? [very unlikely – 0 to very likely – 10]"* The mean difference in LLM scores between paired input text with and without spin was calculated.

Results

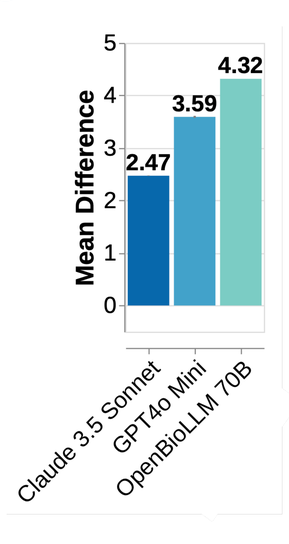
1 Spin Detection



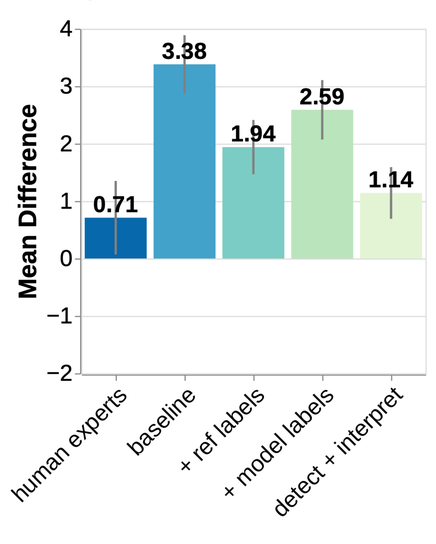
2 Susceptibility - Abstract



3 Susceptibility - Simplified



4 Mitigation Strategies



Conclusions

- While LLMs can generally detect spin in research, they remain **more susceptible to spin** when interpreting clinical trial results than clinicians and medical researchers
- LLMs have **concerning tendency to propagate spin** to downstream tasks, such as generating simplified versions of technical abstracts
- Using **Chain-of-Though style prompting can mitigate** some of this issue

Acknowledgments

This research was supported by the National Institutes of Health (NIH) grant 1R01LM014600-01 and National Science Foundation (NSF) grant IIS-2145479.

Data & Code