Appraising the Potential Uses and Harms of LLMs for Medical Systematic Reviews

Hye Sun Yun, Iain J. Marshall, Thomas A. Trikalinos, Byron C. Wallace

Northeastern University Khoury College of Computer Sciences · King's College London · Brown

BACKGROUND

Large Language Models (LLMs) have been developed to help scientists and clinicians. But! LLMs are known to hallucinate, and this might be risky.

We contextualize potential benefits and harms of LLMs for a specific healthcare application by grounding discussion in the task of producing medical systematic reviews.

Medical Systematic Reviews

- Strongest form of evidence which informs healthcare policy and practice
- Often out-of-date due to rapid publication of evidence making the production of high-quality reviews onerous

RESEARCH QUESTIONS

1. What do domain experts think about the potential uses and risks of LLMs to aid medical systematic review production?
2. Do domain experts anticipate any potential risks from the use of LLMs in this context?
3. What can we learn from domain experts which might inform criteria for rigorous evaluation of biomedical LLMs?

METHODS

1 Search Cochrane Review Titles

Queried the most recently published titles of Cochrane reviews for each of the 37 medical topics.

2 Prompt LLMs to generate evidence summaries

Generated a total of 128 summaries using the titles from step 1 with **Galactica**, **BioMedLM**, and **ChatGPT.**

3 Review outputs and select for interviews

Conducted a **rapid inductive qualitative analysis** to identify characteristics of the output texts. Carefully **chose 6 samples**.

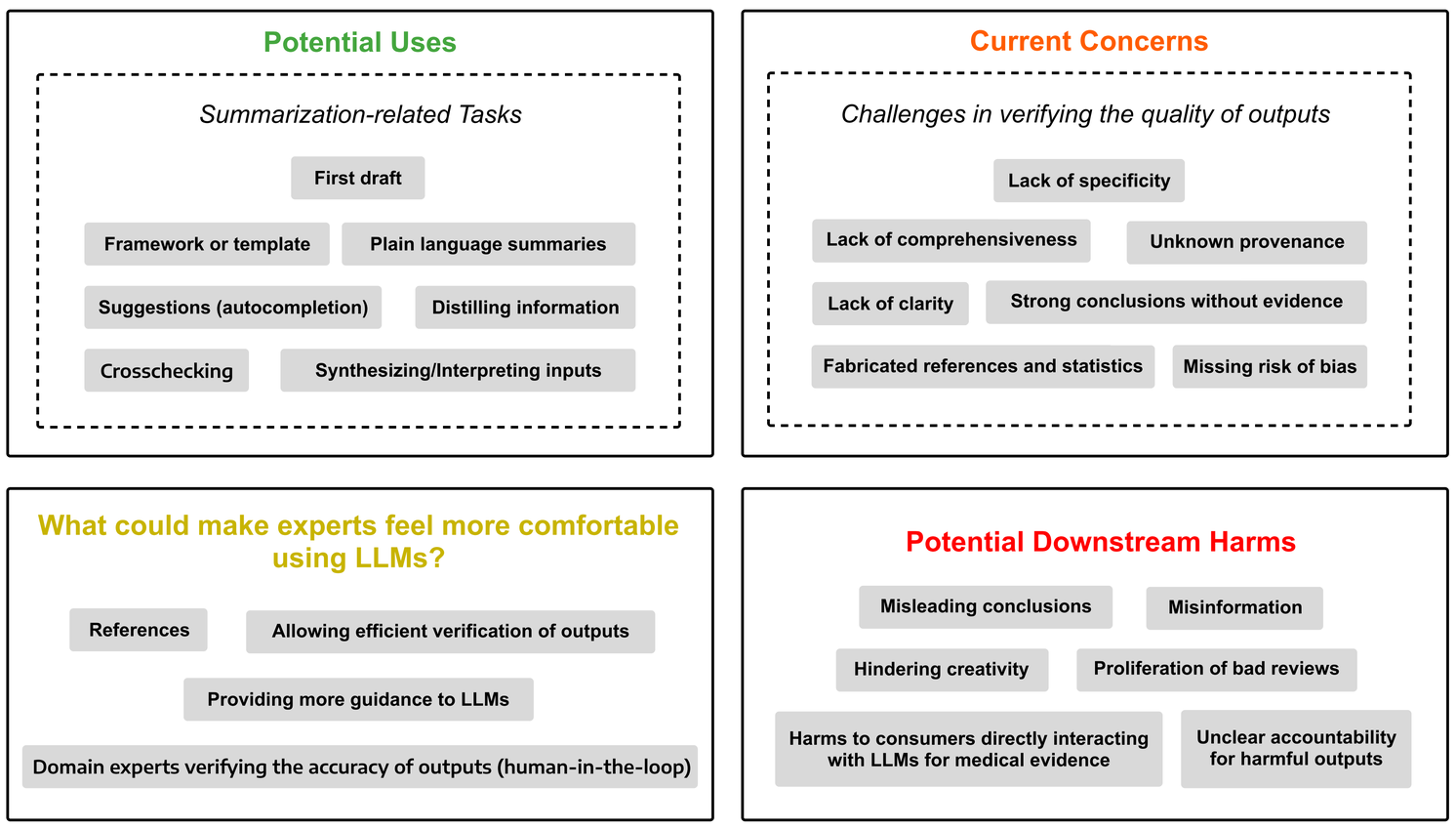
4 Interview domain experts

Interviewed **16 domain experts** of methodologists, practitioners, clinical researchers, and journal editors in March/April 2023.

5 Conduct qualitative analysis

Conducted an **inductive thematic analysis.** Open and axial coding was used, and then themes were organized.

RESULTS



Potential uses and risks of using LLMs to aid systematic review production, according to domain experts.

Framework or Template

"It seems to be **pretty good at putting together a scaffolding or a framework that you could use to write from**. I could see going to it and saying, okay, ChatGPT, talk to me. Give me the subheadings for my dissertation…"

– researcher in evidence synthesis

Synthesizing Inputs

"The most helpful part is for the model **to be able to look at statistical analysis, at numbers, at a graph, and then be able to generate at least some sort of a standard text** so that they know, oh, a result that looks like this means that it has a significance in what way, in what direction."

– professional journal editorial staff

Crosschecking

"That is very interesting as also a means to stimulate discussion, **cross validate our results**, and also identify emerging trends in the literature."

– epidemiologist & professor in evidence synthesis

Lack of Comprehensiveness

"I think most bothersome is it's labeled as an abstract but doesn't read like an abstract. There's nothing more than an introduction to the problem and the objectives of what this review is about. So **it's very incomplete**."

– epidemiologist & professor

Unknown Provenance

"It doesn't reference which systematic review, but the fact that it's a systematic review is encouraging. But then of course, **I don't know if it really has referenced it. I dunno if it exists**."

– professional journal editorial staff

Fabricated References & Data

"The concern is that you can have **falsified science, falsified data, falsified conclusions**, and very convincing packaging of those in the end for used by known expert. But I think even an expert can be fooled by this."

– clinical researcher & professor

Misleading Conclusions

"It came up with pretty strong conclusions and **there's a little bit of misleading**... I would read this if this were written by a human and wonder if there was a fair some spin."

– clinician & researcher in evidence synthesis

Proliferation of Bad Reviews

"It provides p-value, areas under the curve, and optimal cutoffs. All of which I think are specious and non-reproducible for continuous measures. …it is a good **example of the current regrettable practices in medical publishing**."

– clinician & researcher in evidence synthesis

Unclear Accountability

"If in publishing, errors come to light through no one's fault, but things happen and the scientific record needs to be corrected, we need to go back to people and ask them to correct the work... But **that accountability, I don't understand how that would work for something like this**."

– professional journal editorial staff

CONCLUSION

- **Uses**: LLMs will likely aid review production going forward and may provide initial drafts or outlines.
- **Harms**: Domain experts are worried about the blackbox nature of models and potential downstream harms of confidently composed but inaccurate synopses produced by LLMs.
- **Key evaluation aspects:** accuracy, transparency, comprehensiveness of included studies, readability & clear structure, aligning the language of systematic reviews with the presented evidence, and providing important details such as specific PICO elements
- **Future Work**: Develop a more refined evaluation framework and better tools using LLMs for systematic review production.

ACKNOWLEDGEMENTS

This research was partially supported by National Science Foundation (NSF) grant RI-2211954, and by the National Institutes of Health (NIH) under the National Library of Medicine (NLM) grant 2R01LM012086. We also thank all participants in our study: Gaelen Adam, Ethan Balk, David Kent, Georgios Kitsios, Joey Kwong, Navjoyt Ladher, Joseph Lau, Louis Leslie, Tianjing Li, Rachel Marshall, Zachary Munn, Anna Noel-Storr, Evangelia Ntzani, Matthew Page, Ian J. Saldanha, and Dale Steele. We obtained permission from each participant at the end of the interview to thank them by name.