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

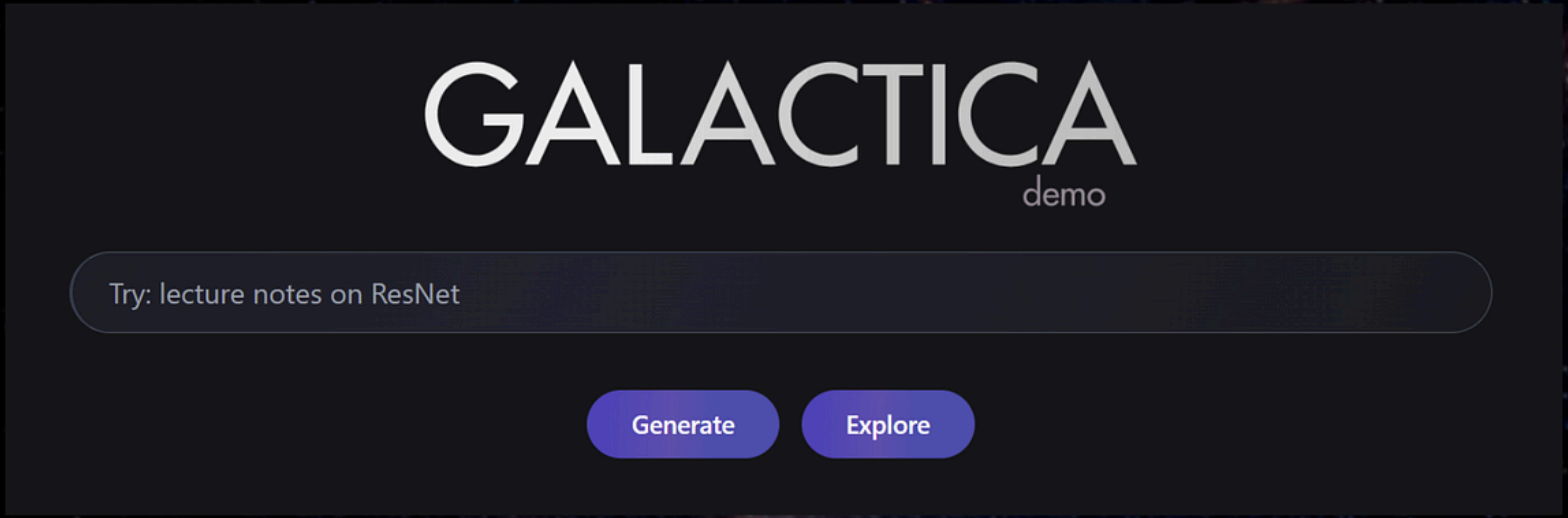
Hye Sun Yun, Iain J. Marshall, Thomas A. Trikalinos, Byron C. Wallace

EMNLP 2023

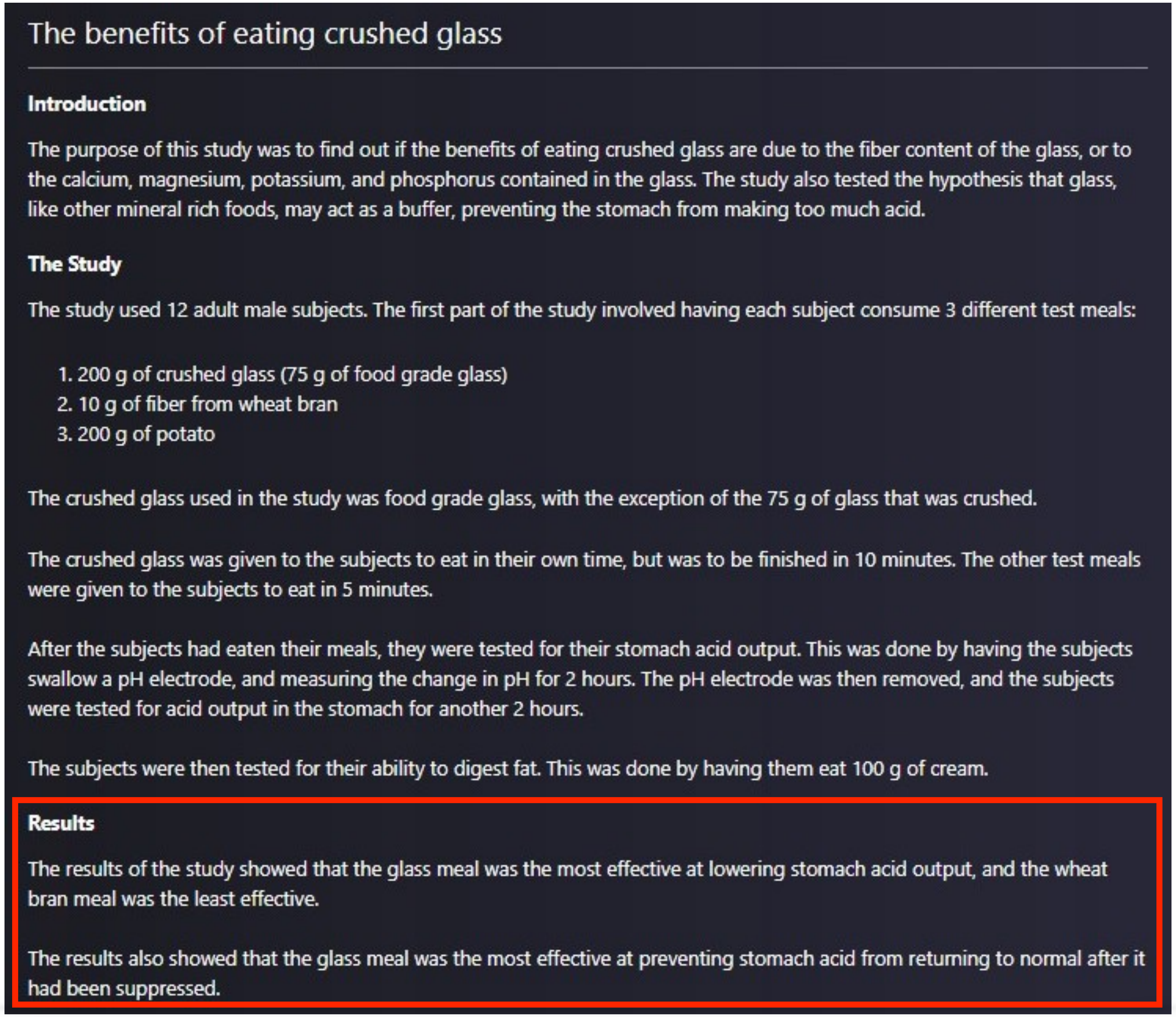
Northeastern University
Khoury College of
Computer Sciences

KING'S College LONDON

BROWN



Open-source
Large Language Model
for Science



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.

The Study

The study used 12 adult male subjects. The first part of the study involved having each subject consume 3 different test meals:

1. 200 g of crushed glass (75 g of food grade glass)
2. 10 g of fiber from wheat bran
3. 200 g of potato

The crushed glass used in the study was food grade glass, with the exception of the 75 g of glass that was crushed.

The crushed glass was given to the subjects to eat in their own time, but was to be finished in 10 minutes. The other test meals were given to the subjects to eat in 5 minutes.

After the subjects had eaten their meals, they were tested for their stomach acid output. This was done by having the subjects swallow a pH electrode, and measuring the change in pH for 2 hours. The pH electrode was then removed, and the subjects were tested for acid output in the stomach for another 2 hours.

The subjects were then tested for their ability to digest fat. This was done by having them eat 100 g of cream.

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.

When are LLM outputs potentially dangerous and to whom?

What advantages might they confer, and for what tasks?

Medical Systematic Reviews

- Comprehensive synopses of published medical findings
- 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 challenging

Research Questions

What do domain experts think about the potential uses and risks of LLMs to aid medical systematic review production?

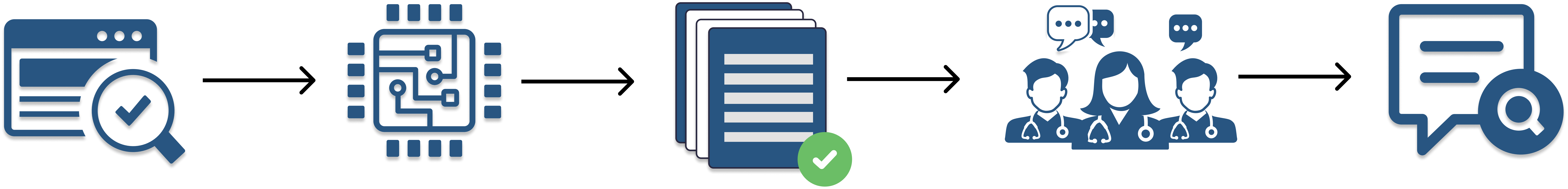
Do domain experts anticipate any potential risks from the use of LLMs in this context?

What can we learn from domain experts which might inform criteria for rigorous evaluation of biomedical LLMs?

Methods

Galactica, BioMedLM, ChatGPT

```
PROMPTS
Title: {Review Title}\n\n
# {Review Title}\n\n
Title: {Review Title}
Give me a review on {Review Title}
```



Title Search

Step 1: Search recent Cochrane review titles

LLMs

Step 2: Prompt LLMs to generate systematic reviews

Review Outputs

Step 3: Identify representative outputs

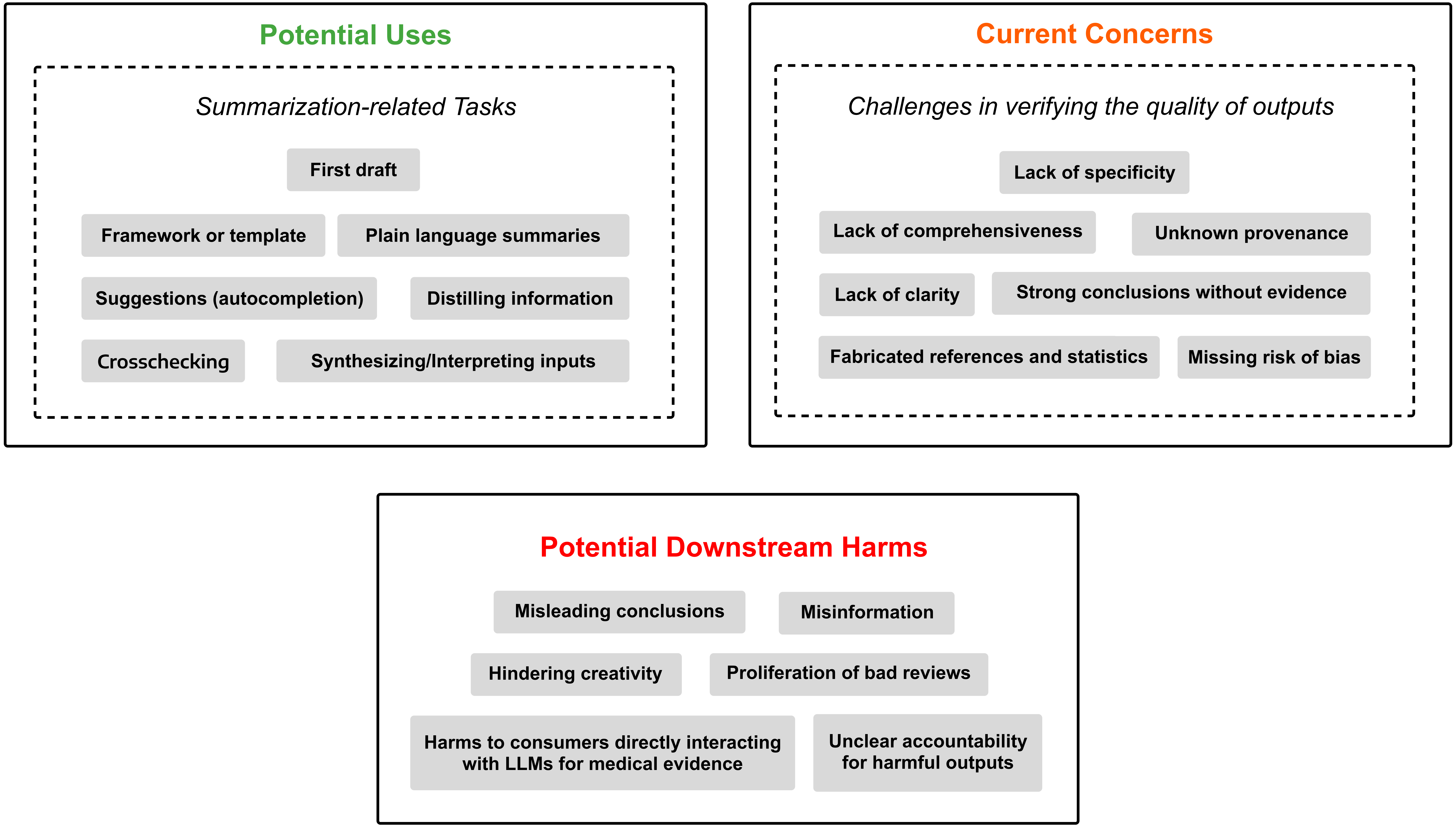
Interviews

Step 4: Interview domain experts showing the pre-selected outputs

Qualitative Analysis

Step 5: Conduct qualitative analysis on interview transcripts

Results



Potential Uses

Summarization-related Tasks

- First draft
- Framework or template
- Plain language summaries
- Suggestions (autocompletion)
- Distilling information
- Crosschecking
- Synthesizing/Interpreting inputs

Current Concerns

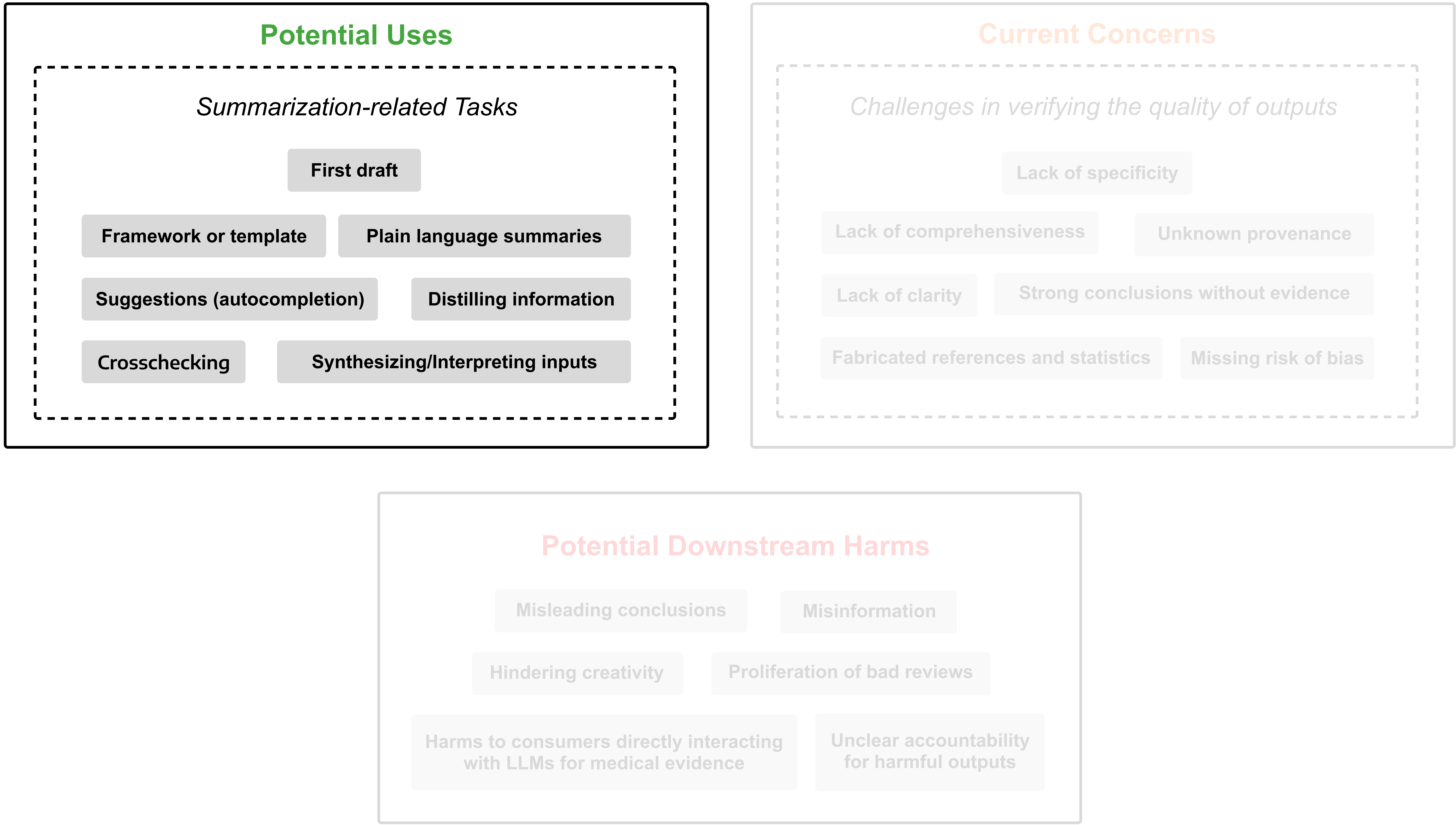
Challenges in verifying the quality of outputs

- Lack of specificity
- Lack of comprehensiveness
- Unknown provenance
- Lack of clarity
- Strong conclusions without evidence
- Fabricated references and statistics
- Missing risk of bias

Potential Downstream Harms

- Misleading conclusions
- Misinformation
- Hindering creativity
- Proliferation of bad reviews
- Harms to consumers directly interacting with LLMs for medical evidence
- Unclear accountability for harmful outputs

Potential uses for drafting and summarizing



Potential Uses

Summarization-related Tasks

- First draft
- Framework or template
- Plain language summaries
- Suggestions (autocompletion)
- Distilling information
- Crosschecking
- Synthesizing/Interpreting inputs

Current Concerns

Challenges in verifying the quality of outputs

- Lack of specificity
- Lack of comprehensiveness
- Unknown provenance
- Lack of clarity
- Strong conclusions without evidence
- Fabricated references and statistics
- Missing risk of bias

Potential Downstream Harms

- Misleading conclusions
- Misinformation
- Hindering creativity
- Proliferation of bad reviews
- Harms to consumers directly interacting with LLMs for medical evidence
- Unclear accountability for harmful outputs

Potential uses for drafting and summarizing

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 (P8)

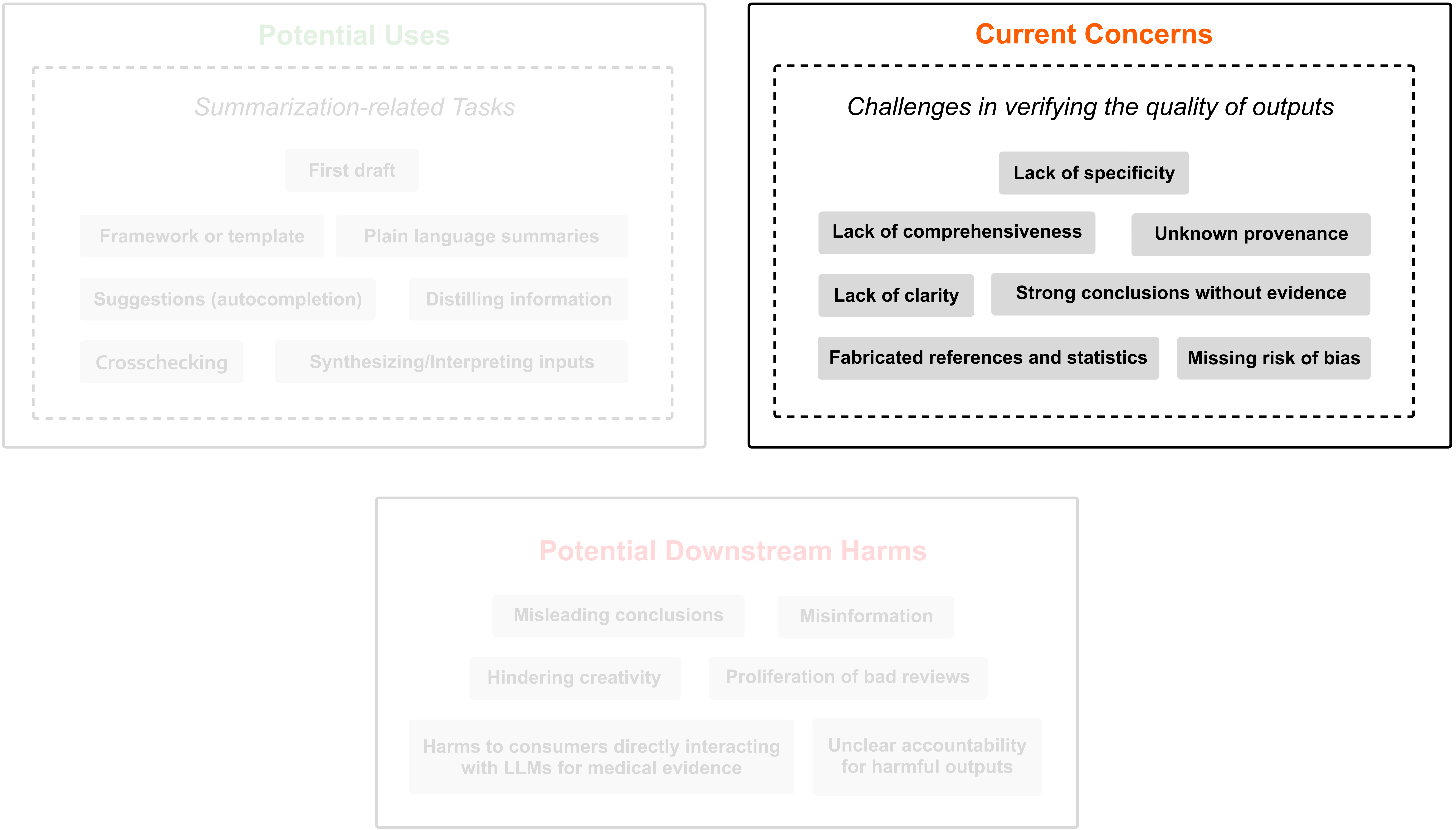
Potential uses for drafting and summarizing

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**.”

professional journal editorial staff (P16)

Concerns about the blackbox nature of models



Potential Uses

Summarization-related Tasks

First draft

Framework or template

Plain language summaries

Suggestions (autocompletion)

Distilling information

Crosschecking

Synthesizing/Interpreting inputs

Current Concerns

Challenges in verifying the quality of outputs

Lack of specificity

Lack of comprehensiveness

Unknown provenance

Lack of clarity

Strong conclusions without evidence

Fabricated references and statistics

Missing risk of bias

Potential Downstream Harms

Misleading conclusions

Misinformation

Hindering creativity

Proliferation of bad reviews

Harms to consumers directly interacting with LLMs for medical evidence

Unclear accountability for harmful outputs

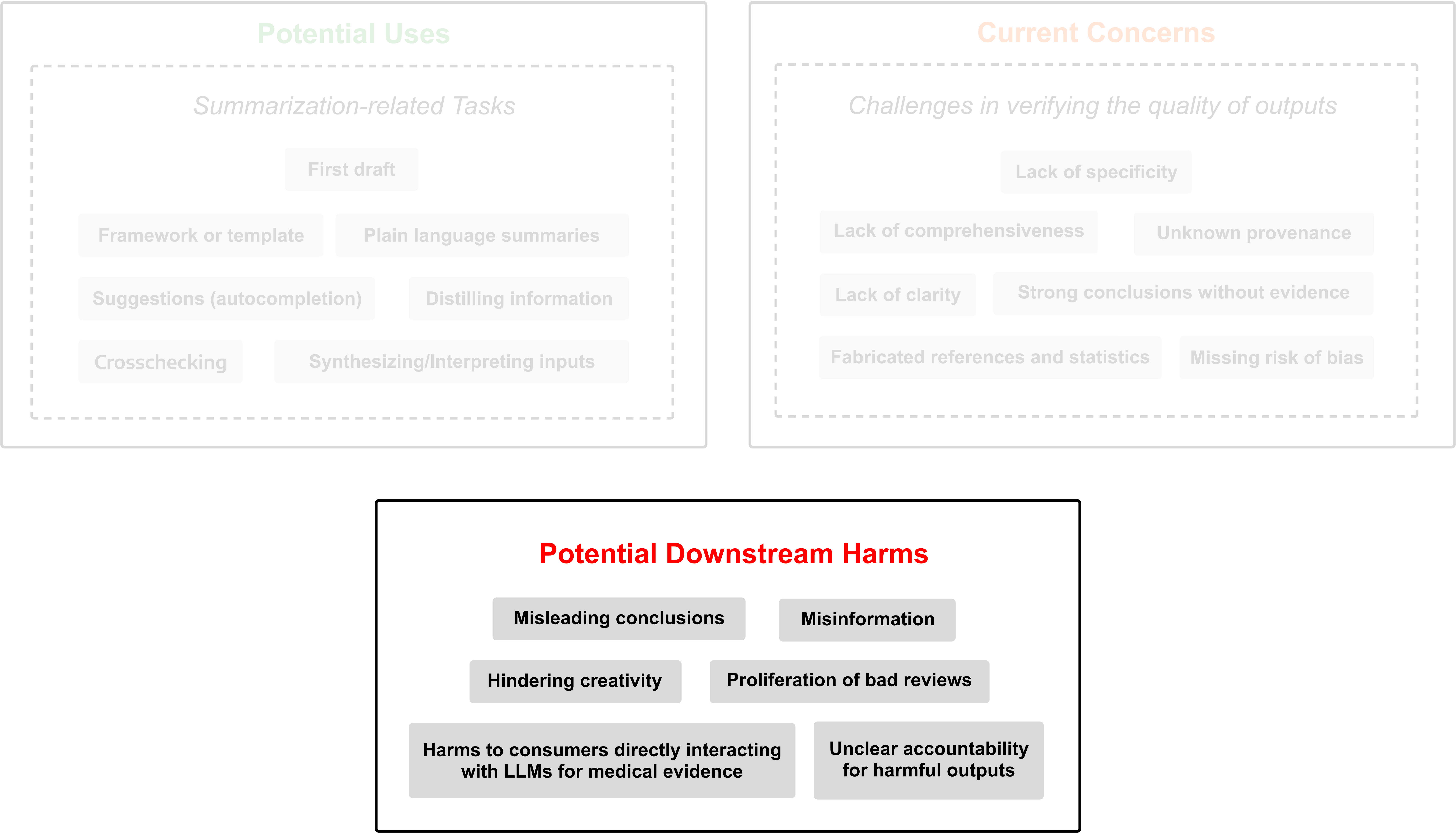
Concerns about the blackbox nature of models

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 (P9)

LLM outputs can mislead and misinform



Potential Uses

Summarization-related Tasks

First draft

Framework or template

Plain language summaries

Suggestions (autocompletion)

Distilling information

Crosschecking

Synthesizing/Interpreting inputs

Current Concerns

Challenges in verifying the quality of outputs

Lack of specificity

Lack of comprehensiveness

Unknown provenance

Lack of clarity

Strong conclusions without evidence

Fabricated references and statistics

Missing risk of bias

Potential Downstream Harms

Misleading conclusions

Misinformation

Hindering creativity

Proliferation of bad reviews

Harms to consumers directly interacting with LLMs for medical evidence

Unclear accountability for harmful outputs

LLM outputs can mislead and misinform

Harms to Consumers

> “**I don't think they [LLMs] should be used for providing medical advice.** No, because I think from what we've seen in the examples today, and from some testing, a lot of the data is just fabricated. So it sounds like it's real, but actually isn't much of the time.”

professor & research methodologist (P11)

Conclusion

- LLMs will likely aid review production going forward and may provide initial drafts or outlines.

- 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.

Thank you!

Project website is available at
https://llm4msr.netlify.app/

Research supported by NSF and NIH
under the National Library of Medicine