Integrating Large Language Models Into an Existing Review Process: Promises and Pitfalls

Edwards M[1], Ferrante di Ruffano L[1]

[1] **York Health Economics Consortium, University of York, York, YO10 5NQ, UK**

INTRODUCTION

The recent development and rise in accessibility of large language models (LLMs) has generated excitement around their possibilities for reducing the resource burden of conducting reviews (Figure 1). Following testing, we assessed the cost, accuracy, and accessibility of LLMs to reviewers, and consider what types of reviews LLMs are currently best suited to assist with.

Figure 1: Stages of a review

METHODS

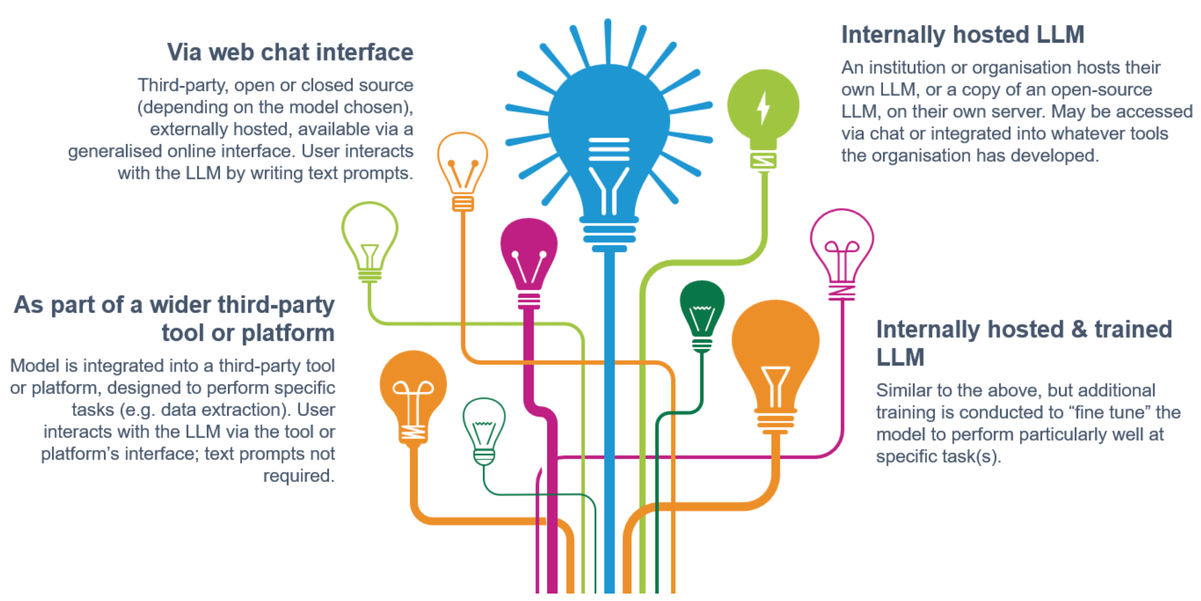
We conducted testing of a LLM, Claude 3 Opus. Access via a chat interface was tested as this form of access, whether via a direct subscription to a model or via an institutional subscription to services such as Amazon Bedrock[1], is open to most reviewers, regardless of setting or scale (Figure 2). We used the tool to conduct:

- High level data extraction for a targeted review
- Highly granulated extraction for a systematic review
- Risk of bias assessment of RCTs

Tasks were conducted using a "zero-shot" approach, i.e. no specific prior examples or training were provided to the model before conducting the requested task.

We assessed the results for accuracy and completeness, and considered the time taken to integrate the LLM into our existing review process. We also considered other routes for accessing LLMs (Figure 2), and the benefits and drawbacks of each.

Figure 2: Options for accessing LLMs



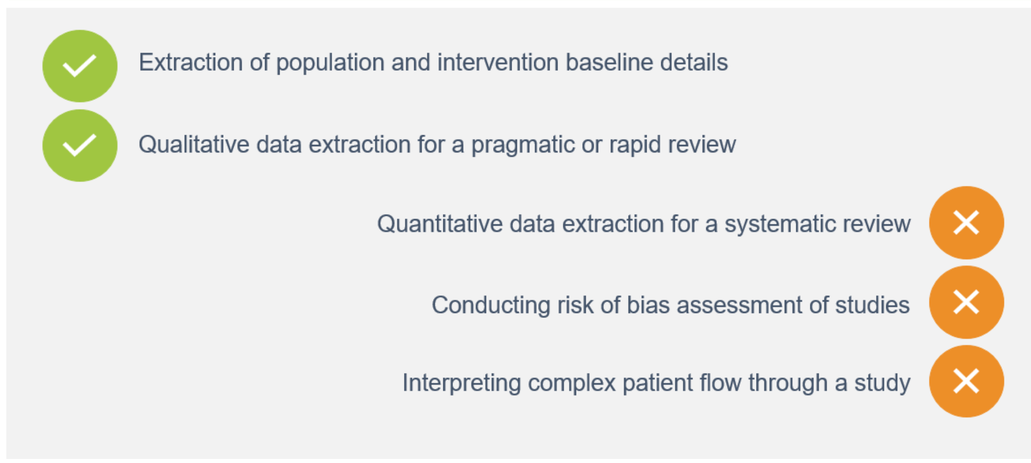
RESULTS

The LLM via a chat interface was highly accessible, inexpensive, and saved significant time in conducting high level qualitative data extraction for a pragmatic review. Outputs were sufficiently standardised and easy to manipulate and integrate into our existing work process (Figure 3).

Extracting accurate granular outcome data for a systematic review proved more difficult, with the model failing to interpret complexities of patient flow and struggling to respond accurately to lengthy, detailed prompts. The time taken for subsequent checking, correcting, and formatting outweighed any time saved (Figure 3).

The model identified some relevant content for conducting risk of bias assessment with the Cochrane Risk of Bias 1 tool[2], although lacked context, and human judgement was still needed to ensure consistency and correctness in final decision making.

Figure 3: LLM chat interface for reviewing tasks



The "zero-shot" use of LLMs via a simple web chat interface can offer significant time savings for targeted reviews, although copyright issues exist in uploading published papers for extraction and / or synthesis, unless the model is accessed on a local server.

Optimal performance for data extraction and risk of bias assessment for systematic reviews is unlikely to be achieved without fine tuning a version of the chosen model with archive data. This process is currently costly, commercial confidentiality must be considered, and the skill set required is outside the scope of many review teams.

Purpose-built tools (either internal to an organisation or created and accessed via a third-party) integrating a LLM for specific reviewing tasks are available (e.g. PITTS.ai[3]) and are usually designed to offer a user-friendly interface. Developers of such tools should ensure that their tools can be integrated into clients' existing processes with the use of standardized import and export formats such as CSV or RIS.

The use of a LLM as part of any kind of review should be fully documented in the methods sections of the review writeup, including prompts used to generate the output. Reviewers should be mindful that questions remain around the stability of LLMs over time[4], and that any outputs of a LLM should always be externally validated and checked for trustworthiness, reliability, and fairness.[5]

CONSIDERATIONS WHEN ACCESSING LLMs FOR REVIEWING

- Type of review and the intended purpose
- Complexity and level of subject expertise required to conduct the necessary synthesis: is additional training of the model required to achieve acceptable performance?
- Skill level of users
- Available budget and time for training
- Location of the model, and any associated copyright issues

REFERENCES

1. https://aws.amazon.com/bedrock/ 2. https://www.bmj.com/content/343/bmj.d5928
3. https://pitts.ai/ 4. https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1710
5. https://www.ema.europa.eu/en/documents/other/factsheet-four-principles-safe-responsible-use-large-language-models_en.pdf

CONTACT US

mary.edwards@york.ac.uk
+44 1904 323437
York Health Economics Consortium
www.yhec.co.uk