Large Language Models and "Zero-Shot Learning" for Data Extraction in a Systematic Review: A Case Study

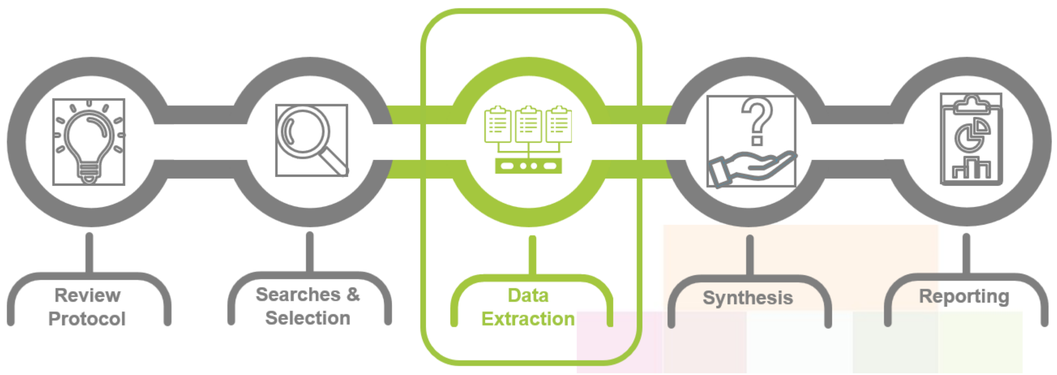
Edwards M[1], Ferrante di Ruffano L[1]

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

INTRODUCTION

A typical systematic review includes extraction (Figure 1) of highly granulated data in a standardized format, a resource intensive part of the review process. We investigated whether the chat interface to a large language model (LLM) could provide time savings in extracting such data while retaining the accuracy necessary for a systematic review.

Figure 1: Data extraction



We set out to test access via a chat interface 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 systematic reviewers, regardless of setting or scale (see Figure 2). However, such access is dependent on the interface provided by the owners or developers of the LLM, and the scope for customisation or integration with reviewers' internal systems tends to be limited.

"Zero-shot" refers to the use of the model to complete a task for which no specific prior examples or training have been offered by the user.

We used Claude Opus 3 to conduct testing. Claude is a family of LLMs developed by Anthropic and based on a "constitutional AI"[2] approach. This approach is designed to enable the training of AI assistants that are both helpful and harmless.

The "constitutional AI" approach is consistent with the aim of systematic reviewing, to produce reliable findings using explicit, systematic methods, selected with a view aimed at minimizing bias[3]. At the time of testing the models developed by Anthropic were generally held to be some of the most reliable easily-accessible models for summarizing quantitative data, although we note that the LLM landscape is constantly changing and individuals and organisations should determine for themselves which model(s) best suit their particular needs.

Figure 2: Routes for accessing a LLM[4]

Via a web chat interface	Third-party, open or closed source (depending on the model chosen), externally hosted, available via a general online interface.
As part of a wider packaged tool (external)	Third party, open or closed source, externally hosted, integrated into a wider system or tool that supports specific tasks for a targeted user base.
Internally hosted	Internal hosting by the reviewers' institution allows for greater security. May be accessed via chat or as part of a wider, internally developed, tool.
Internally hosted & trained	Significant investment and skill required to internally host and train a bespoke model, including integrating reviewers' existing data.

METHODS

A data extraction sheet from a completed review of biologic treatments was selected. A set of prompts was designed to obtain details of the methods, interventions, and populations assessed by three of the included studies. Each paper was uploaded individually, and the results were copied into the original data sheet and compared with those produced and checked by two independent human reviewers. Testing of outcome extraction was also conducted using the same methods.

RESULTS

Despite being offered no explicit training examples, the model successfully extracted details of the intervention and population assessed in each arm (both of which tend to be clearly and consistently reported as a table or in a discrete section of the narrative writeup). However, it struggled to interpret complex patient flow through the studies.

Primary outcomes in the intent to treat population were successfully extracted (these data were clearly described and visible in the papers' abstracts) but extraction of secondary outcomes, subgroups, and outcomes at different timepoints (which were more inconsistently reported in tables, plots, images, or scattered throughout the narrative) proved much less reliable. We found that for extracting outcome data, the subsequent checking, correcting, and formatting of the output outweighed any time saved.

Successful prompts were specific and granular, in line with the level of detail required in the resulting extracted data. Some trial and error was required to construct suitable prompts. Thorough checking of the resulting data was vital.

Figure 3: LLM chat interface for data extraction

- ✓ Successfully extracted details of intervention, including dose, scheduling and duration of treatment in each arm
- ✓ Successfully extracted details of population, including age, gender, duration of disease, and genetic variants, assessed in each arm
- ✓ Primary outcomes in the ITT population were extracted completely and correctly
- ✗ Model struggled to interpret complex patient flow through the studies
- ✗ Extraction of secondary outcomes, subgroups, and outcomes at different timepoints were unreliable and not always correct
- ✗ Not as useful as a human reviewer at spotting key unanticipated information to contextualise the data

CONCLUSIONS

The "zero-shot" use of LLMs via a chat interface is easily accessible, inexpensive, and requires no specialist user skills. Figure 3 summarises the performance of LLMs accessed in this way for extracting different data types and elements.

While such use of LLMs may provide some time savings in extracting basic study data on population baseline characteristics and interventions assessed, such access methods do not lend themselves to the detailed prompts required for successful extraction of the complex patient flow and outcome data required for a systematic review. In addition, the use of externally hosted models raises issues with copyright and confidentiality.

At present, we suggest that more specialist tools, enabling task-specific training of a model, are required to support systematic reviews. The "zero-shot" use of LLMs via a chat interface may be more suited to support rapid or pragmatic reviews of open-source data.

REFERENCES

1. https://aws.amazon.com/bedrock/ 2. https://www.constitutional.ai/
3. https://www.cochranelibrary.com/cdsr/about-cdsr#:~:text=A%20Cochrane%20Review%20is%20a,answer%20a%20specific%20research%20question 4. https://www.ema.europa.eu/en/documents/other/guiding-principles-use-large-language-models-regulatory-science-medicines-regulatory-activities_en.pdf

CONTACT US

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