Large Language Models and “Zero-Shot Learning” for Data Extraction in a Targeted Review: A Case Study

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

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

Abstract #145195

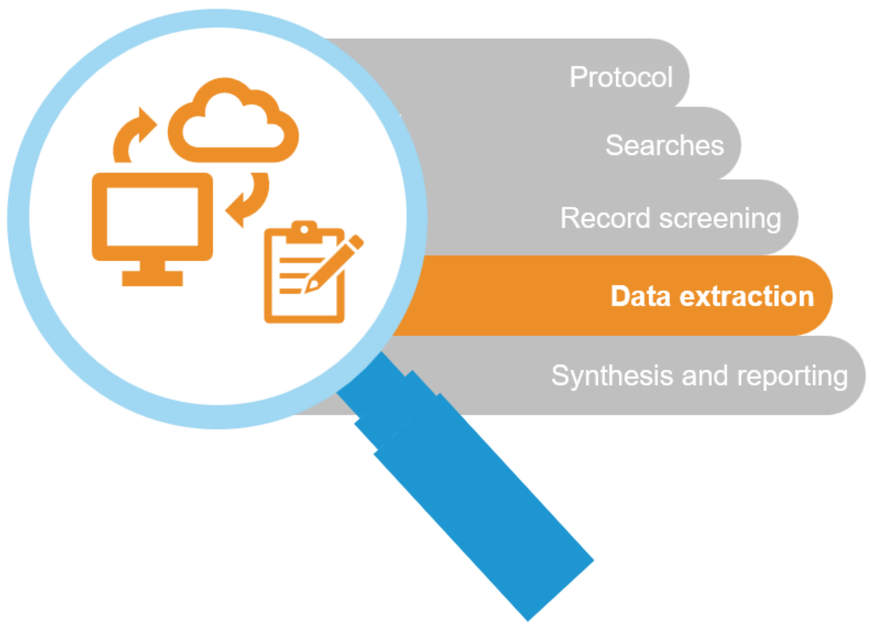
INTRODUCTION

A “targeted”, or “pragmatic” review is one that adapts the conventional systematic review process to take into consideration limited time and/or resources available. This is usually achieved by applying additional limits to the search or eligibility criteria, but additional time and resource savings may be required at the record screening and data extraction phases.

Extraction and presentation of data (Figure 1), even at a high level, is a time-consuming process, and large language models (LLMs) may offer a pragmatic solution for saving time while retaining the accuracy and consistency required. LLMs accessed via a chat interface require minimal user training and perform tasks without any setup overheads beyond an initial phase of prompt engineering.

We assessed the chat interface to Claude 3 Opus[1] for accuracy, consistency, presentation of data, and time savings in the context of high-level extraction for a targeted review.

Figure 1: Stages of a targeted review



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[2], is open to most systematic reviewers, regardless of setting or scale. However, such access is dependent on the interface provided by the owners or developers of the LLM. 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.

METHODS

A targeted review was conducted to investigate disparities in patient characteristics in the diagnosis and treatment of patients with immunodeficiency. Full methods and results of this review will be published at a later date.

We used a chat interface to Claude 3 Opus to extract data from 25 papers, with a human reviewer checking all data points. Study and population details were extracted, plus brief details of any study results or discussion regarding disparities in diagnosis or treatment.

Papers were uploaded in pairs to minimize prompts, with the model explicitly tasked with labelling each set of data points with the name of the relevant PDF from which it was extracted.

Data were extracted using two sets of prompts, each of which was developed and piloted on three papers before reaching a final, standardized form which was then used to extract data from all the remaining papers.

A summary of the first set of prompts, used to extract population baseline details, is shown in Figure 2. As specified in the prompts, data were provided by the interface in a structured format suitable for copying directly into a simple Microsoft Excel sheet for storage, checking, and later synthesis by a human reviewer.

Figure 2: Example of prompt structure

I'm a medical researcher and I need to extract data from some papers.

Provide the results in a structured format with answers separated by carriage returns and labelled with the file name of the PDF from which the answers are extracted. The data points I need for each PDF paper are:

1. **How many patients are reported, and which immunodeficiency / immunodeficiencies do they have?**
2. **Does the paper report the *[baseline characteristic 1]* and *[baseline characteristic 2]* of the patients?**
3. **If the answer to question 2 is "yes", what is the reported *[characteristic 1]* and *[characteristic 2]* of the patients?**

Please extract these data from the first two papers, *[RefID-Author-YEAR]* and *[RefID-Author-YEAR]*

RESULTS

Of the 25 papers from which data were extracted: eleven papers required no edits to the data; five papers required minimal edits; nine papers contained minor errors or omissions in the data.

One paper was extracted correctly but the answers reported by the model also contained additional data drawn from the second paper of the pair. Another pair of papers was extracted by the model and mislabelled, with data for each paper labelled with the file name of the other paper. Following this error, PDFs were uploaded one at a time.

The output was suitably formatted to allow easy transference into an Excel spreadsheet. The main issue identified was the failure of the model to interpret subtle but important differences between baseline characteristics; for example, having extracted data on patient race, the model failed to extract data on patient ethnicity, despite the prompts having requested both pieces of data separately. However, this problem would likely be resolved by providing minimal additional training for the model.

CONCLUSIONS

Even allowing time for human checking and minor correction of the extracted data, use of the LLM via a chat interface enabled extraction and checking of 25 papers in a single day.

Although this kind of access is less customisable than a bespoke tool, it is simple to use, relatively inexpensive, and can still offer significant resource savings in the context of suitable pragmatic reviews and scoping reviews.

In order to obtain the best efficiencies from LLMs, it is likely that review teams will need bespoke trained models, and to re-think their processes (for example moving away from spreadsheet-based data extraction and storage). However, in the interim, chat interfaces offer an easy access point to the capabilities of LLMs and provide outputs of a quality suitable for high level targeted reviews.

REFERENCES

1. https://www.anthropic.com/claude 2. https://aws.amazon.com/bedrock/

CONTACT US

mary.edwards@york.ac.uk

+44 1904 323437

York Health Economics Consortium

www.yhec.co.uk