Assessing the accuracy of AI indexing of abstracts in a literature review

William Letton and Eleanor King, 19/09/23

Introduction

Literature reviews are often time-consuming and expensive, problems that are worsened by the accelerating rate of research publication. Tasks such as abstract screening and data extraction require comprehension of natural language and often specific domain knowledge, making the process resistant to automation.

Recently, large language AI models capable of reasonable levels of natural language comprehension and generation have become widely available. It is unclear what role this kind of automation will eventually play in the literature review process, particularly as the results are inconsistent and the information processing opaque. Here, we present the results of a single experiment comparing human and AI performance on a simple data extraction task.

Methods

Project choice

A recently-completed literature review project on treatments for HIV infection was used. As part of this project, an Evidence Mapper tool had been used to manually index 206 study abstracts in 14 fields. These fields were reviewed for their appropriateness for automated indexing. Fields were removed if they were not based on the abstract text (i.e. 'Location' and 'Year'), or had unclear descriptions relating to a network comparison analysis that were performed as part of the project. This left 8 fields.

100 abstracts from this project were selected with a random number generator for the experiment.

Language model

The large language model used in this experiment was OpenAI's GPT (generative pre-trained transformer) model 3.5, which has been designed to parse and generate natural language. This model was accessed through Application Programming Interface (API) calls to the service. These calls involve submitting a query that includes both the abstract text and details of what information to extract from it. In June 2023, OpenAI introduced a Function Calling option to it's Chat Completion tool. This option is designed to allow the user to specify that the AI should return a response in a format that is appropriate for passing to a programmatic function. This is ideal for data extraction tasks, where the output needs to be in a specified format and without the verbosity that usually accompanies language model output.

Prompt engineering

The input to the Function Calling Chat Completion requires a field name, description, and data type.

Getting the best results from generative language models requires appropriate prompts to be submitted. At present this process is more of an art than a science, and often involves a trial-and-error approach to interacting with the black-box language model that can be both lengthy and frustrating. For the purposes of this experiment, the prompts were created as follows:

1) The Evidence Map generated using human indexing included a Home page, on which each field was described in natural language. The field names and descriptions found here were used as the base for the AI prompts.
2) A field of 'Number of studies' was added. It was known from previous testing that this reduced a known error in the AI responses, where the number of studies in a literature review would be returned instead of the number of subjects in the study. This field was not assessed for accuracy as it wasn't present in the human-indexed Evidence Map.
3) The field names and descriptions were then altered if it was judged that they would be unclear to a naïve person reading them, for example if the description was made up solely of examples, or referenced some other part of the project documentation.
4) A data type was specified for each field.
5) A brief test was performed to assess whether the prompts were producing roughly the right kind of response, regardless of accuracy. This resulted in one further change: adding the option to return 'unknown' as a study type.

The resulting field names, descriptions, and data types are shown in Table 1 below.

Table 1 – instruction data submitted to the language model.

name	description	data_type_primary	data_type_secondary
Disease	The type of HIV if stated e.g., HIV-1.	array	string
HAART regimen with doses	The complete HAART drug regimen with doses (total daily dose)	array	string
Outcomes	The specific outcomes reported with their timeline where possible e.g., viral load at week 96	array	string
Subpopulations	The particular characteristics of the population being studied e.g., co-infection with tuberculosis or hepatitis, pregnant or breastfeeding women, adult, children or infants	array	string
Study type	The study type. Either Randomised Controlled Trial (RCT), non-randomised comparative study, non-comparative study or unknown	string	NA
Study duration	How long the study went on for	string	NA
Study population size	The number of patients included in the study	integer	NA
Number of studies	The total number of studies, articles, references, and publications included in the review	integer	NA
Risk factors	Reported risk factors in the study population for being infected with HIV e.g., fetal exposure	array	string

A function was written to convert the table of field information into the correct JSON string format for submission to the OpenAI API. The query also included the following role prompts:

- "You are a helpful assistant that extracts summaries of research paper abstracts for a database."
- “Extract a summary from the following research paper abstract: ”

This was followed by the abstract text. If the study citation had no abstract field then the title was used instead.

Response temperature

The Chat Completion method submitted to the OpenAI API includes a ‘temperature’ parameter that sets the level of randomness in the response. For creative tasks this can be increased to generate more ‘imaginative’ output. However, for this experiment the value was set to 0.

On submission, 3 of the 100 queries resulted in an error. This was found to be due to the language model returning the data in a format that could not be parsed as a JSON string. To overcome this issue the temperature parameter was gradually increased until the returned data was in the correct format.

Scoring

Following submission, each of the 100 abstracts had a results for each of the 8 fields for both human and AI indexing. A human scorer was used who was not involved in the original literature review. The human and AI indexing were presented to the scorer in a randomised order to reduce scoring bias. The scorer then gave each indexed value a score from the following options:

- Completely accurate
- Largely accurate
- Unclear/park
- Largely inaccurate
- Completely inaccurate

The scorer was also instructed as follows:

- “Due to Excel formatting, blank entries are marked as a "0". Treat these fields as unknown/unclear.”
- “Entries that are unknown/unclear are correct if the correct answer is not given or is not clear in the abstract.”
- “Some records only have a Title text and not an Abstract text. In these cases please score the indexing as if the Title text was the Abstract text.”
- “If you are uncertain what score to give, mark the score as 'unclear/park'.”
- “For study duration, reporting the recruitment period is incorrect.”

The scorer was instructed to mark down responses that contained information that wasn’t in the source text.

Results

The results of the scoring are shown in Figure 1.

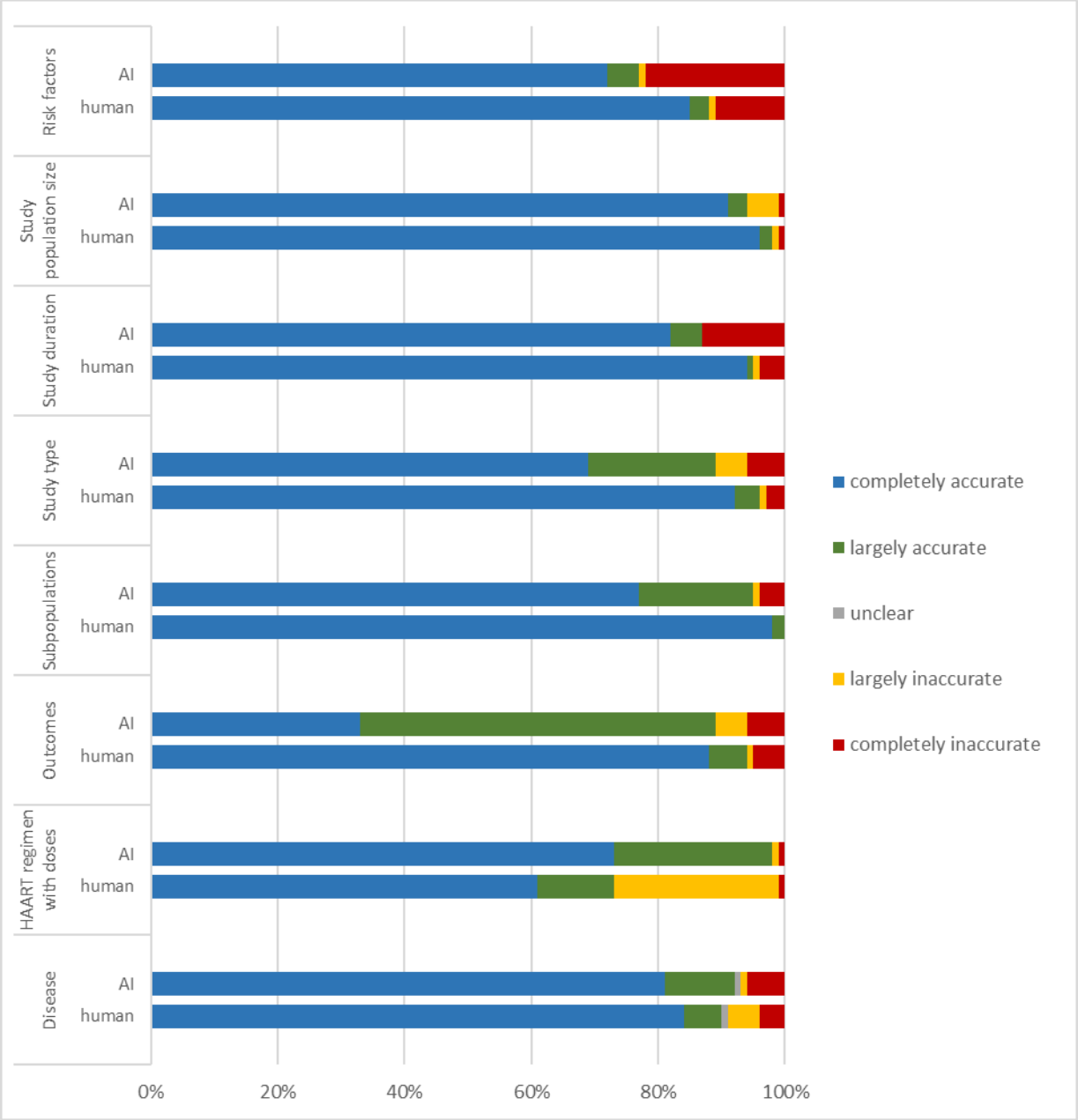

Figure 1 - Accuracy of human and AI indexing by field

It can be seen that both the human and AI indexing contained errors. Only in the case of the 'Subpopulations' field did the human indexing not contain any inaccuracies.

For six of the eight fields the AI indexing was less accurate than the human indexing. However, for 'HAART regimen with doses' field the scorer judged the AI output to be more accurate on average than the human output. For the 'Disease' field the AI scored lower for 'completely accurate' responses, but higher overall for responses judged as 'completely accurate' or 'largely accurate'.

Table 2 shows the proportion of responses on which the human and AI indexing was scored as having different levels of accuracy, and what proportion of those times the AI scored higher than the human.

Table 2 - Summary of differences in accuracy scoring between human and AI indexing by field

Field name	Different scores (%)	AI with higher score when scores differ (%)
Disease	18%	44%
HAART regimen with doses	46%	72%
Outcomes	69%	10%
Subpopulations	23%	4%
Study type	39%	21%
Study duration	17%	18%
Study population size	12%	33%
Risk factors	25%	24%
overall	31%	28%

The proportions vary with field, but on average the human and AI were given different accuracy scores 31% of the time, and of those the AI scored more highly 28% of the time.

Discussion

This experiment suggests that AI can display impressive data extraction potential from abstract text, though it is still less accurate than an expert human at this task. However, as language model performance improves this gap should continue to close. Future research should compare OpenAI's GPT model 3.5 with the updated 4.0. Biomedical-specific models are also in development and available for use that may show better performance on tasks that require domain knowledge.

Future experiments should also aim to compare human and AI performance on a more level playing field. In this experiment the human indexer had more information about the topic and purpose of the review than just what was in the prompts given to the AI.

Both the human and AI made errors in the indexing. The acceptable error-rate will depend on the purpose of the indexing. Where full accuracy is less important, such as during initial scoping, there may be advantages to an AI approach, which is faster and cheaper. In this experiment the cost was approximately £0.11 for the 100 abstracts. However, AI indexing may introduce unexpected biases, which should be investigated. In 28% of cases where the scores differed between the human and AI indexing the AI indexing scored higher. This suggests there may be a role for AI in checking the work done by humans.

A common problem with using language models to extract information from text is their tendency to confidently 'hallucinate' plausible-sounding results. In this experiment, only one case of this hallucination was noted by the scorer, and unblinding revealed this to have been in the human indexing set.

One reason that the human indexer had lower scores on the "HAART regimen with doses" field is that they replaced occurrences of "Tenofovir" with "Tenofovir disoproxil fumarate". This is technically correct, but the scorer had been instructed to penalise the presence of information that was not taken directly from the abstract. The ideal balance between direct fidelity to the source document and introduction of domain knowledge will depend on the project goals.

In this study, the prompt engineering reflected a heuristic method – i.e. would a reasonable human understand what was being requested? Further refinement of prompts may improve performance, and the effect of providing examples should be investigated. The scorer noted that for the "Outcomes" field the AI simply copied any outcomes verbatim from the abstract text, while the human indexer had often performed some interpretation or grouping in line with the interests of the

review. For the field ‘Study duration’, the AI frequently returned the duration of the recruitment period rather than the study period when the latter was not given. This was scored as being inaccurate, though it is consistent with the prompt given, “how long the study went on for”.