August 2024

White Paper

Bridge
Evidence that matters

Artificial Intelligence in Systematic Literature Reviews

Part 2 | AI-aided Full-text Screening

Abstract: In the first research paper of this series on AI in systematic literature reviews (SLRs), we shared our methodology for testing the performance of AI models in title/abstract (ti/ab) screening, with which we achieved high sensitivity (82% to 96% across five different categories of SLR projects). On advancing our program to evaluate the performance of AI models on screening of a large number of full-texts (~2,000) from the same five projects, we obtained a sensitivity of ≥99%. Here, we share our methodology and results from this test of AI-enabled full-text screening (FTS).

Introduction

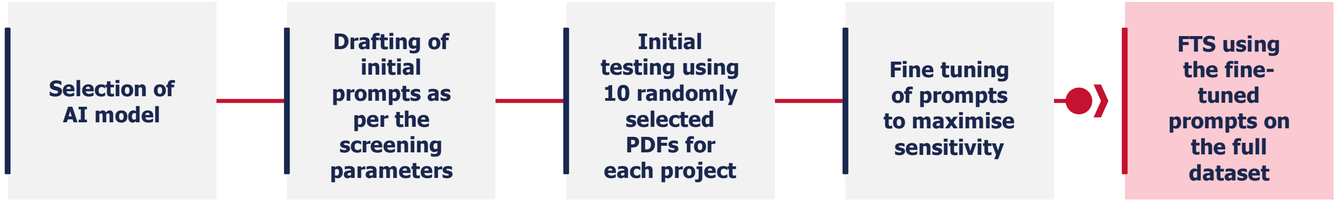
In the first research paper in this series[1], we reported the results of our tests on AI in ti/ab screening. In this update, we describe our tests on AI in the next step in SLRs, FTS. FTS includes a thorough, detailed evaluation of the entire content ("full text") of each publication. Our objective was to use a selected AI model to screen full-text publications for standard parameters (population, intervention, comparator, outcomes, and study design [PICOS]) and also have the AI model *provide the reason* for excluding any full-text publication with reference to the pre-defined inclusion/exclusion criteria.

Methodology

We used our internal 'gold-standard' FTS databases from five SLR projects (already doubled-QC'd by experienced colleagues) as our reference datasets. The five SLRs included were diverse in terms of complexity, study designs, and topic area. Three were SLRs focused on clinical trials, and two were burden of illness SLRs focused on observational studies.

Our approach to evaluating AI performance is shown in Figure 1.

Figure 1: Approach to evaluate AI performance for full-text screening


Selection of AI model
Drafting of initial prompts as per the screening parameters
Initial testing using 10 randomly selected PDFs for each project
Fine tuning of prompts to maximise sensitivity
FTS using the fine-tuned prompts on the full dataset

AI=artificial intelligence; FTS=full-text screening; PDF=portable document format; ti/ab=title/abstract

Of the large language models (LLMs) available at the time of this testing, we selected Generative Pretrained Transformer-4 (GPT-4) model since it had performed very well in our previous test of ti/ab screening.[1] This is also consistent with recent studies that reported the performance of GPT-4 to be superior to older AI models and all other LLMs (including previous versions of GPT) in various natural language processing tasks in English and other languages.[2,3] We accessed the GPT-4 model via the OpenAI API application programming interface (API), using Python programming language to issue prompts. We provided text in the form of PDFs of full publications.

The initial prompts were drafted to obtain either a direct classification decision (i.e., Yes/No) or actual information regarding the parameters that were part of the screening flowchart, and to obtain a rationale for the classification/extraction of information for these parameters. As done in traditional FTS, we also developed an algorithm to provide the reason for exclusion in a hierarchical manner (according to the screening flowchart).

For each project, we initially uploaded text from 10 PDFs. Based on the initial results, we fine-tuned the prompts over repeated tests to obtain maximum sensitivity. Fine-tuning for sensitivity was a key focus for FTS, because of the greater risk posed by false negatives at this step. This approach is also consistent with findings from a recent survey conducted among the authors of SLRs, where the expectation of accuracy from AI was higher for FTS than for ti/ab screening.[4] The fine-tuned prompts were then used on the entire set of FTS publications, and the results were compared against the in-house screening databases.

Figure 2 provides an example of a full-text publication from an SLR focused on the indication of 'borderline personality disorder' as screened by the AI model. The figure includes the rationale provided by the AI for each decision in the screening algorithm.

Figure 2: Example of a full-text publication screened by the AI model

Context: A full-text publication titled "*Health related quality of life for young people receiving dialectical behaviour therapy (DBT): a routine outcome-monitoring pilot*" for the burden of illness SLR on borderline personality disorder (BoPD) was screened using the AI model. The parameters of interest for inclusion were: observational study; conducted among human BoPD patients; and providing data on clinical, treatment, humanistic, or economic burden parameters.

Screening parameters for FTS	Examples of prompts (paraphrased*) asked to AI	AI model output	AI model rationale	Decision for *each parameter*	I/E decision for the *full-text*
Human study	Is the study conducted in humans?	Yes	The study involved adolescent DBT teams collecting outcome data from young people receiving DBT, indicating human participants were involved in the study.	**Include**	**Include**
Study design	What is the study design?	Prospective cohort study	The study involved following a cohort of adolescents over time, assessing them at multiple points (admission and discharge), indicative of a prospective cohort study design.	**Include**	
Disease of interest	Did the study include patients with BoPD?	Yes	The study targeted adolescents with BPD type presentations, and required participants to meet ≥5 BPD criteria, one of which must be recent occurrence of self-harm behaviour.	**Include**	
Outcomes	Did the study provide data on **clinical burden** (e.g., prevalence)?	No	The study does not provide data necessary to calculate the prevalence of BoPD in the underlying population, nor does it aim to measure prevalence.	**Include**	
	Did the study provide data on **humanistic burden** (e.g., HRQoL)?	Yes	The study specifically reported on HRQoL for young people receiving DBT. It used the EQ-5D to assess HRQoL at admission and discharge from the DBT programs.		

AI=artificial intelligence; BPD/BoPD=borderline personality disorder; EQ-5D=EuroQol- 5 Dimension; FTS=full-text screening; HRQoL=health-related quality of life; I/E=inclusion/exclusion; SLR=systematic literature review.

*Only illustrative prompts are shown; the actual prompts were more detailed and had been fine-tuned after repeated testing.
Note: For brevity, details on all outcomes (e.g., incidence, mortality, activities of daily living, costs, etc.) assessed in the project are not shown in Figure 2.

Results

Based on screening of 2,066 full texts across the five projects, the sensitivity and the 'practical sensitivity' (i.e., the proportion of publications that were eventually included/prioritised for final reporting in the project and were correctly identified as such during AI FTS) was very high i.e., ≥99%; while the specificity was relatively low (ranging from 6% to 22%; Table 1). Overall, the accuracy for FTS ranged from 75% to 93% (i.e., proportion of correct matches between the AI decisions and the human decisions in the reference datasets; Table 1).

Table 1: Results of full-text screening for the five SLR projects*

Project topic	Total full texts	Accuracy	Full-text Sensitivity	'Practical' Sensitivity	Specificity	PPV	NPV	TP	FP	TN	FN
Anemia-CKD	450	75%	**99%**	**100%**	6%	75%	78%	330	111	7	2
mCRPC	981	86%	**100%**	**100%**	16%	86%	89%	819	134	25	3
Atopic dermatitis	250	93%	**100%**	**100%**	6%	93%	50%	231	17	1	1
Soft tissue Sarcoma	229	81%	**100%**	**100%**	22%	80%	100%	174	43	12	0
Borderline personality disorder	156	81%	**99%**	**100%**	15%	81%	83%	121	29	5	1

CKD=chronic kidney disease; FN=false negative; FP=false positive; mCRPC=metastatic castration-resistant prostate cancer; NPV=negative predictive value; PPV=positive predictive value; SLR=systematic literature review; TN=true negative; TP=true positive.

*Of the five SLR projects included for this testing, three were clinical trials focused (anaemia-CKD, mCRPC, and soft tissue sarcoma) and two were burden of illness SLRs focused on observational studies (atopic dermatitis and borderline personality disorder).

Note: Please refer to glossary for definitions of the measures mentioned in Table 1.

Reflections on findings and next steps

The sensitivity obtained in our testing is impressive, and provides assurance that no relevant publications are being inadvertently excluded. This supports the utility of integrating AI in the FTS workflow to improve efficiency without compromising quality. The low specificity in our testing implies that human review will still be required, given the reasonably high proportion of false positives; however, the time impact of false positives is relatively low for FTS as compared to ti/ab screening. This is because volumes are typically much higher at ti/ab screening, therefore specificity needs to be high to ensure substantial workload reduction; in contrast, the lower volumes at FTS make specificity less of an issue. For example, in the atopic dermatitis project presented in Table 1, a total of 7,356 citations went through ti/ab screening (see the first paper in this series[1]), but only 250 through FTS. Since accuracy is affected by errors in both sensitivity and specificity, our overall accuracy was lower than the sensitivity due to very few false negatives (i.e., very high sensitivity) but a large number of false positives (i.e., low specificity).

It should be kept in mind that these are initial results, and as we continue to improve our methodology, we expect the specificity to improve whilst retaining the extremely high sensitivity. Finally, in our AI screening process, we plan to combine FTS with initial data extraction and categorisation in a single step, which could potentially translate into substantially improved efficiency in the delivery of SLRs.

We are continuing to explore and test the utility of AI across multiple facets of health economics and outcomes research (HEOR). We will shortly present our findings on initial data extraction and categorisation in an upcoming research paper in this series, and also publish our complete findings on AI-assisted SLRs (ti/ab screening, FTS, and initial data extraction and categorisation) in a peer-reviewed journal. Later this year, we will prepare a paper focused on our 'real-world' experience, incorporating key learnings from our delivery of multiple AI-enabled SLRs to clients in the past year.

Glossary

Accuracy: In the context of full-text screening, this indicates the fraction of publications correctly identified for inclusion and exclusion by AI during full-text screening amongst total full texts.

AI: Artificial Intelligence is a branch of computer science that aims to create systems capable of performing tasks that normally require human intelligence. These tasks include visual perception, speech recognition, decision-making, and translation between languages.

GPT: Generative Pre-trained Transformer refers to a series of language processing AI models developed by OpenAI. These models utilize a transformer architecture for deep learning and are pre-trained on a vast corpus of text data. The "generative" aspect refers to the model's ability to generate coherent and contextually relevant text based on input prompts.

In supervised learning, the algorithm is trained on a pre-defined set of training examples, which then facilitate its ability to reach an accurate conclusion when given new data.

In unsupervised learning, the algorithm is given data without predefined labels and is allowed to find structure in its input on its own.

LLM: Large Language Models are a type of artificial intelligence models that process, understand, generate, and sometimes translate human language. These models are "large" both in terms of the size of their neural network architecture (having a large number of parameters) and the vast amount of data they are trained on. LLMs are often based on transformer architectures and are trained on diverse datasets from the internet or other large text corpora.

NPV: Negative Predictive Value is the proportion of negative test results that are true negatives. NPV = True Negatives / (True Negatives + False Negatives)

PPV: Positive Predictive Value is the proportion of positive test results that are true positives. PPV = True Positives / (True Positives + False Positives)

'Practical' sensitivity: In the context of full-text screening, we have defined practical sensitivity to refer to those true positives that were eventually included for final reporting in the review, i.e., the proportion of final actual positives that were correctly identified as such during AI full-text screening.

Sensitivity: Sensitivity measures the proportion of actual positives that are correctly identified as such.

Sensitivity = True Positives/ (True Positives + False Negatives)

SLR: A systematic literature review is a methodical and comprehensive approach to identifying, evaluating, and synthesizing all relevant research on a specific topic or research question.

Specificity: Specificity measures the proportion of actual negatives that are correctly identified as such.
Specificity = True Negatives / (True Negatives + False Positives)

Authors: Saifuddin Kharawala, Sam Isaacs, Pankdeep Chhabra, Divyanshu Jindal, and Paul Gandhi

References

1. Kharawala S, Issacs S, Jindal D, Gandhi P. Artificial Intelligence in Systematic Literature Reviews Part 1 | AI-aided Title/Abstract Screening. Available at: https://www.bridgemedical.org/site/assets/files/2156/ai_in_literature_reviews_white_paper_31_jan_24-1.pdf.
2. Syriani E, David I, Kumar G. Assessing the ability of ChatGPT to screen articles for systematic reviews. arXiv. 2023. doi:10.48550/arXiv.2307.06464 [posted online July 2023]
3. OpenAI. GPT-4 technical report. arXiv. 2023 http://arxiv.org/abs/2303.08774 [posted online March 2023]
4. Hanegraaf P, Wondimu A, Mosselman JJ, et al. Inter-reviewer reliability of human literature reviewing and implications for the introduction of machine-assisted systematic reviews: a mixed-methods review. BMJ Open. 2024;14(3): e076912. PMID: 38508610; PMCID: PMC10952858.