MSR39

Large Language Models and contextual prioritization of titles and abstracts for systematic review screening

Riaz Qureshi[1,2], Eitan Agai[2]

1 - University of Colorado Anschutz Medical Campus, Denver, CO, USA; 2 - PICO Portal, St. Petersburg, FL, USA

Introduction

Major advancements have been made in the utilization of machine learning (ML) approaches to reorder records for screening in systematic reviews (SRs).

As reviewers screen, ML prioritizes remaining records based on the content of included and excluded titles/abstracts.

This requires a training period for the AI which is also influenced by human user error.

The evidence is still unclear as to the potential utility of Large Language Models (LLMs) for contextual prioritization without a human-decision training period.

Objective: To assess the sensitivity of two approaches to record prioritization using LLMs.

Methods

We used two SRs, completed entirely within the PICO Portal platform, with double human screening at title/abstract and full-text levels.

SR#1 had strict eligibility criteria, 48 final includes, and 12103 initial records

SR#2 had broad eligibility criteria, 84 includes, and 9054 initial records

Approach A – ChatGPT-4o was prompted to identify the most likely includes for a question, given the full eligibility criteria and all titles/abstracts.

Approach B – Text embeddings to compare and score contextual similarity of titles/ abstracts to the inclusion/exclusion criteria separately, combined and normalized scores, then ranked the records by relevance to the question with an inclusion cutoff of > 0.66 normalized score.

PROMPTS

Results

Sensitivities were calculated as the proportion of the human-screened final includes that were predicted as includes by the LLMs.

Table. Results of LLM record prioritization

Approach A (ChatGPT-4o)		Approach B (Embeddings)	
SR#1	**SR#2**	**SR#1**	**SR#2**
Includes = 82	**Includes = 613**	**Includes = 1641**	**Includes = 489**
Incl.% = 0.7%	**Incl.% = 6.8%**	**Incl.% = 13.6%**	**Incl.% = 5.4%**
Sens. = 0.23	**Sens. = 0.32**	**Sens. = 0.75**	**Sens. = 0.60**

Notes: *"Inc.%" = Percent of total records included by prompts; "Sens."= Sensitivity*

After screening between 1% - 14% of all records, systematic reviewers may already identify between 23% - 75% of final includes for their project. Combined with ML-assistance, it should be feasible to switch to single-human screening after ~30% of all records have been screened, saving >50%

Example prompt structure for ChatGPT (Approach A):

"We are conducting a systematic review and pairwise meta-analysis to evaluate the effectiveness of _____ on treating _____.

For this review, we have the following eligibility criteria:

Population: We will include studies of ... We will exclude studies of ...

Intervention: We will include studies that examine ...

Comparator: Comparators will include ...

Study design: We will include ... We will exclude ...

The file that I will upload is a .csv containing _____ records that were retrieved from multiple databases and de-duplicated.

The first column is a record ID, the second column is the title, and the third column is the abstract (the first row is the names of these columns). I want you to look through each record's title and abstract and make a judgement for whether it is likely to be relevant or not to our review question.

Apply the criteria that I specified above and return a new .csv that contains the same three columns, but only the records that are relevant for our systematic review. If a record is missing the title or abstract, ignore the missing information and base your decision on the present information."

Conclusions

- LLMs show great potential utility for "pre-screening" titles/abstracts to increase SR screening efficiency.
- The initial set of records flagged as includes by the LLM often contains more positives than an equal-sized random sample, making it a stronger training set for downstream ML models.
- Implementing a 2-stage workflow consisting of LLM prioritization of likely includes, followed by ML-based filtering significantly boosts sensitivity.
- A combination of LLM pre-screening and ML training could greatly reduce the overall effort and resources required for screening.

DISCLOSURES: Dr. Riaz Qureshi reports consulting with PICO Portal; Mr. Eitan Agai is the founder of PICO Portal **AI Assisted Content Disclosure:** AI was not used in the creation of the abstract or presentation.

REFERENCES: [1] Cao C et al. *Annals of Internal Medicine*. 2025; 178:389-401 | [2] Delgado-Chaves FM et al. *Proceedings of the National Academy of Science*. 2025; 122(2):e2411962122 | [3] Dennstadt F et al. *Systematic Reviews*. 2024; 13:158 | [4] Huotala A et al. *28th International Conference on Evaluation and Assessment in Software Engineering*. 2024. | [5] Li M et al. *Systematic Reviews*. 2024; 13:219 | [6] Luo R et al. *medRxiv*. 2024; DOI: 10.1101/2024.06.03.24308405 | [7] Galli C et al. *Preprints.org*. 2025; DOI: 10.20944/preprints202503.0981.v1