ACCELERATING EVIDENCE GENERATION: LEVERAGING LLMS FOR FULL-TEXT STUDY SELECTION

MSR80

Christopher Olsen, BHSc[1]; Jayson Habib, MPH[1]; Elizabeth Salvo-Halloran, MSc[1]; Sumeet Singh, BScPhm, MSc[1]; Nicole Ferko, MSc[1]

[1]EVERSANA, Victoria, BC, Canada

BACKGROUND AND OBJECTIVES

- The use of AI in systematic literature reviews (SLRs) has shown promise in reducing human labour and delivering results more rapidly, with tools dedicated to various review stages as well as full workflows becoming increasingly available.
- Ensuring rigour during study selection, particularly at the full-text stage, requires contextual understanding and identifying nuanced interpretations,[1] which adds substantial challenge in the implementation of AI tools such as large language models (LLMs).
- Emerging guidance such as PRISMA extensions for AI-assisted reviews emphasize a need for explicit documentation of AI use, validation approaches, and human oversight.[2]

OBJECTIVE: This study sought to validate a proprietary LLM tool in the conduct of study selection at the full-text stage by comparing to a dual-reviewer process from a published SLR in Crohn's disease (CD).

Table 1: Classification metrics

Metric	Definition
Agreement rate	Proportion of eligibility decisions matching between LLM and human review across all records
Precision	Proportion of true includes out of the total number of true includes and false includes
Recall	Proportion of true includes out of total number of true includes and false excludes
Specificity	Proportion of true excludes out of the total number of true excludes and false includes
Negative Predictive Value (NPV)	Proportion of true excludes out of the total number of true excludes and false excludes

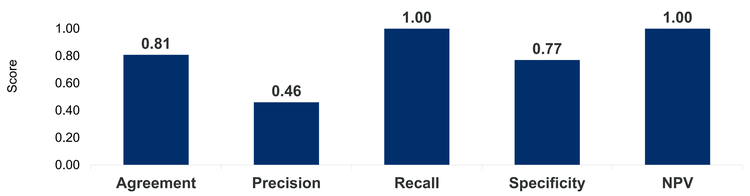
METHODS

- The screening tool was initially developed in an R-based application and leveraged LLMs through application programming interface (API) calls to perform **zero-shot full-text screening.**
- Multiple enhancements were made:
  - The tool was subsequently re-engineered in a **Python-based** environment, incorporating enhanced document processing workflows and more efficient API integration.
  - The updated workflow incorporated optical character recognition (OCR) functionality, which improved record readability and reduced exclusions due to previously unreadable records.
  - The LLM underpinning the original implementation (GPT-4o-2024-11-20) was upgraded to a newer model (**GPT-4.1-2025-04-14**).
- A pilot dataset from a published SLR in CD[3] was used for **iterative prompt engineering** to develop the screening-specific instructions for the AI tool to perform full-text screening.
- Responses for each criterion, justification, and eligibility decision were **evaluated for agreement and accuracy** against human screening decisions (from the published review) using the metrics defined in **Table 1**.

RESULTS

- The screening dataset consisted of 402 records; 10 records were removed based on prior human-assigned exclusion reasons identified during pre-Level 2 screening, as the application does not perform automated duplicate or date-based exclusions.
- The updated tool successfully processed all records without technical failures, including previously unreadable records in the original tool (n=8); for which all screening decisions were accurate (100%).
- The updated assessment of the pilot database resulted in an 80.8% agreement rate, perfect recall and NPV, and specificity of 0.77, confirming that all records predicted for exclusion were correctly excluded (**Figure 2**).
- The model correctly excluded 259/336 records (77.1%; NPV=1.0) and predicted inclusion for 143 records (TP=66; FP=77) which underscore the tool's ability to substantially reduce manual screening burden while maintaining conservative exclusion behavior (**Figure 2**).
- Prioritizing recall over precision enhanced model performance and accuracy by ensuring relevant papers were included for data extraction (n=66).
- Estimated human time savings were in excess of 40%, given that agreement rate was similar to that of the original dual-human process.

Figure 2: Full-text Screening Metrics

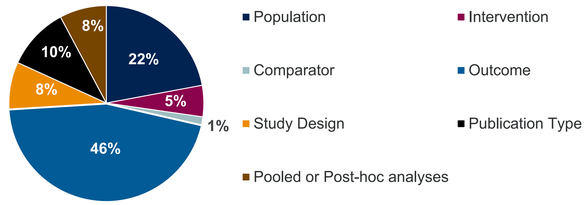


- With the additional capabilities of including previously unreadable records and improved parsing process, the total tool runtime increased compared to the original version, with the added benefit of more context and records for AI screening decisions; LLM API call runtime was similar between tool versions.

Screening Insights

- Pooled data and post-hoc study designs represented a recurring source of disagreement, highlighting a challenge for both AI and human reviewers and over-inclusion of records observed.
- Review of false include decisions by AI indicated areas where contextual understanding is key for determining eligibility of records (**Figure 3**).

Figure 3: Errors in Eligibility by Human Exclusion Reason (FP; n=77)



DISCUSSION

- The LLM tool demonstrated **good performance** with **perfect recall** and 80.8% agreement rate, though it was **over-inclusive** versus human reviewers.
- The application's **high-recall design** minimizes the need for additional re-screening, enabling reviewers to progress more efficiently to downstream SLR tasks (i.e., data extraction).
- **Zero-shot prompting** allows generative LLMs to be applied directly to screening tasks without task-specific training[4], facilitating rapid incorporation of human expertise, and responsiveness to protocol or PICOS changes.
- **Prompt refinements** in the updated application that explicitly addressed the most **nuanced criteria** (i.e., pooled and post-hoc study designs) contributed to improved performance relative to earlier iterations.
- These findings support the role of **AI as a decision-support tool** that augments, rather than replaces, human judgment, particularly for nuanced exclusion criteria requiring contextual interpretation.

CONCLUSION

This study supports the use of AI for improving full-text screening efficiency. Calibrating tools for high sensitivity, even at the expense of specificity, may provide an optimal balance of accuracy and efficiency. Future iterations of the tool based on LLMs with improved reasoning capabilities may further enhance performance.

ABBREVIATIONS

AI = Artificial Intelligence; API = application programming interface; FP = false positive; GPT = Generative Pre-trained Transformer; HEOR = Health Economics and Outcomes Research; LLM = large-language model; NPV = negative predictive value; PICOS = Population, Intervention, Comparator, Outcomes, and Study Design; SLR = systematic literature review; TIAB = title and abstract; TP = true positive.

REFERENCES

1. Moens M, Nagels G, Wake N, Goudman L. Artificial intelligence as team member versus manual screening to conduct systematic reviews in medical sciences. iScience. 2025;28(10):113559. Published 2025 Sep 12. doi:10.1016/j.isci.2025.113559. **2.** Holst D, Moenck K, Koch J, Schmedemann O, Schüppstuhl T. Transparent Reporting of AI in Systematic Literature Reviews: Development of the PRISMA-trAIce Checklist. JMIR AI. 2025;4:e80247. Published 2025 Dec 10. doi:10.2196/80247. **3.** Disher T, Naessens D, Sanon M, et al. One-Year Efficacy of Guselkumab Versus Advanced Therapies for the Treatment of Moderately to Severely Active Crohn's Disease: A Network Meta-Analysis. Adv Ther. 2025;42(6):2708-2727. doi:10.1007/s12325-025-03183-x. **4.** Wang, Shuai, Harrisen Scells, Shengyao Zhuang, Martin Potthast, Bevan Koopman, and Guido Zuccon. "Zero-shot generative large language models for systematic review screening automation." In European Conference on Information Retrieval, pp. 403-420. Springer, Cham, 2024.

Presented at the 2026 ISPOR Conference

Philadelphia, PA, USA
May 17 to 20, 2026

EVERSANA®