MSR217

Evaluating a Large Language Model Approach for Systematic Literature Review Screening with Domain Expert Input

Wei-Hua Huang, Varadraj Poojary, Kimberly Hofer, Mir Sohail Fazeli
Evidinno Outcomes Research Inc., Vancouver, British Columbia, Canada

Background

- Systematic reviews are crucial for synthesizing evidence for health technology assessments, and guiding clinical practice and policy[1]
- However, systematic reviews are time-consuming; this poses challenges for researchers to maintain up-to-date evidence in fast-moving fields[2]
- The growing volume of published studies has heightened demand for more efficient review methodologies
  - Automation tools (e.g., natural language processing) show promise in expediting study screening, data extraction, and quality assessment[3]
- Recent advancements in large language models (LLMs), such as OpenAI's GPT-4, offer potential to automate labor-intensive stages of reviews, improving both efficiency and comprehensiveness[4]

Objective

- To assess the performance of an LLM in conducting full-text screening for systematic reviews with domain expert input
- To determine the feasibility and reliability of LLMs in reducing manual workload and expediting the systematic review process while maintaining accuracy and quality in evidence synthesis

Methods

SYSTEM DEVELOPMENT

- Custom system using GPT-4o was developed to automate full-text screening process in systematic reviews aiming to accurately include or exclude studies based on Population, Intervention, Comparison, and Outcomes (PICO) criteria with minimal human involvement after initial setup phase
- **LLM Sherpa** facilitated parsing/interpretation of large text volumes,[5] breaking input into meaningful components for effective analysis of complex documents (e.g., scientific studies) with nuanced information dispersed across multiple sections
- Two-stage prompt approach with GPT-4o was used to enhance screening accuracy and transparency (**Figure 1**):
  1. **Stage 1: Understanding and Contextualization**
     - GPT-4o was provided research objectives and PICO criteria to build context for decision-making, allowing it to interpret relevant patterns
  2. **Stage 2: Decision-Making and Rationale Generation**
     - The model screened studies using context from Stage 1, providing a rationale for each inclusion/exclusion decision to ensure transparency

Figure 1: Workflow for Automated Screening

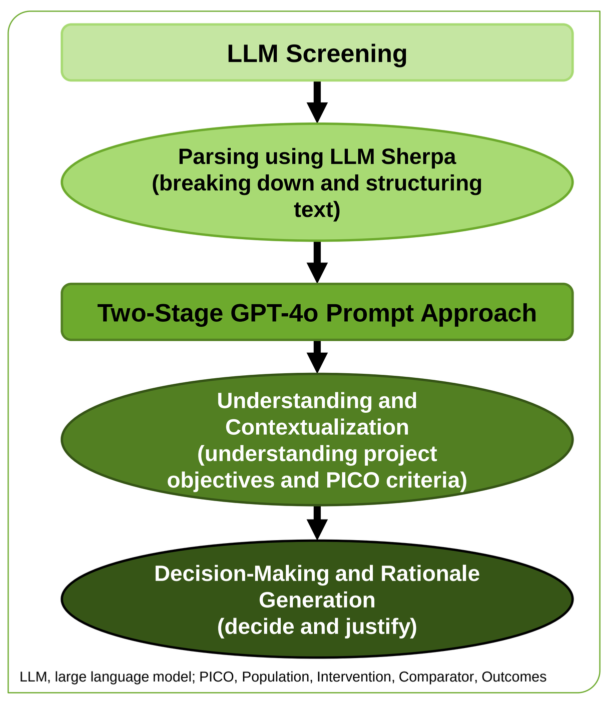
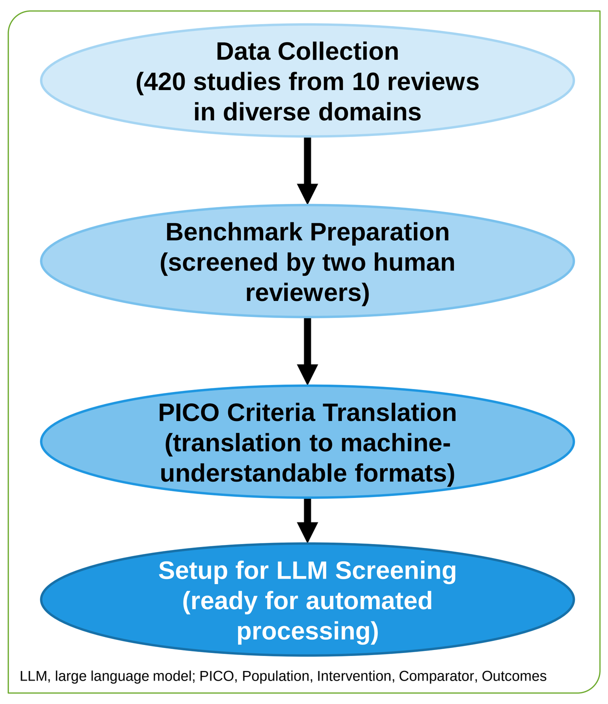


Figure 2: Pre-Screening Preparation



DATA SOURCES AND BENCHMARK FOR EVALUATION

- Final dataset of 420 studies from 10 systematic reviews across cardiology, dermatology, and oncology was used to evaluate performance of LLM
- Each study had been screened by two independent human reviewers, creating a reconciled "benchmark" dataset for reliable comparison of the LLM's decisions against human expert assessments (**Figure 2**)
- **Translation of PICO into machine-understandable formats**
  - PICO criteria were translated into structured, machine-readable formats by a domain expert to facilitate accurate interpretation by LLM; this format included explicit definitions for population, intervention, comparison, and outcome criteria of interest
- **Screening process with LLM**
  - GPT-4o model autonomously screened each study's full-text PDF for relevance based on PICO criteria
  - To ensure consistency, studies were removed from the screening process if they:
    - Were tagged as "duplicate publications" by human reviewers
    - Were excluded by human reviewers due to non-PICO-related reasons (e.g., full-text PDF could not be retrieved)

PERFORMANCE ASSESSMENT

- Performance of GPT-4o was assessed using key metrics[6]:
  1. **Sensitivity (Recall):** Model's ability to correctly identify relevant studies (true positives)
  2. **Positive Predictive Value (PPV):** Proportion of identified relevant studies that are indeed relevant (true positives among all positives)
  3. **Negative Predictive Value (NPV):** Proportion of identified irrelevant studies that are indeed irrelevant (true negatives among all negatives)
- System's output was compared with human screening results to assess agreement, inclusion/exclusion counts, and overall accuracy, as well as consistency, reliability, and generalizability
- Domain experts qualitatively assessed quality of LLM's rationales for inclusion/exclusion decisions to ensure alignment with PICO criteria and logical reasoning standards (**Figure 3**)
- The overall automation workflow is depicted in **Figure 4**

Figure 3: Evaluation and Comparison Process

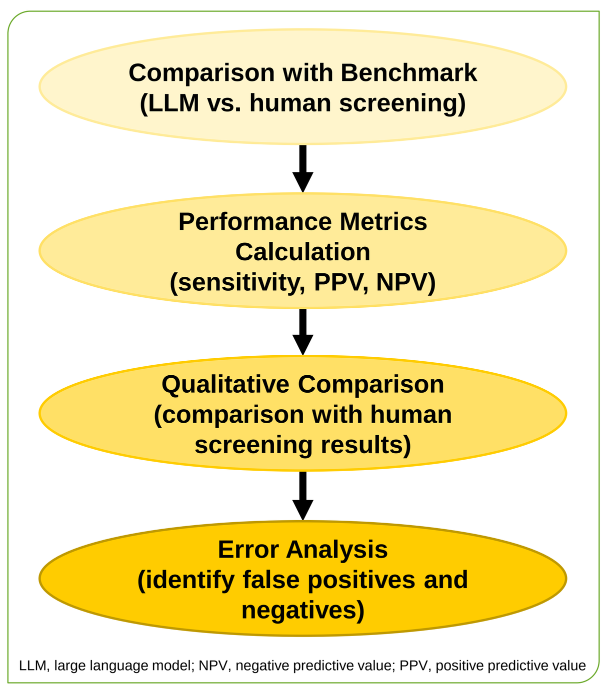
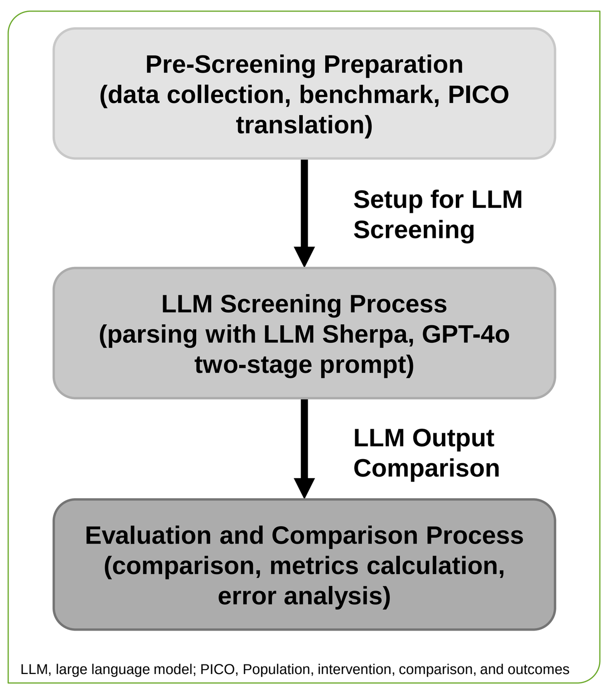


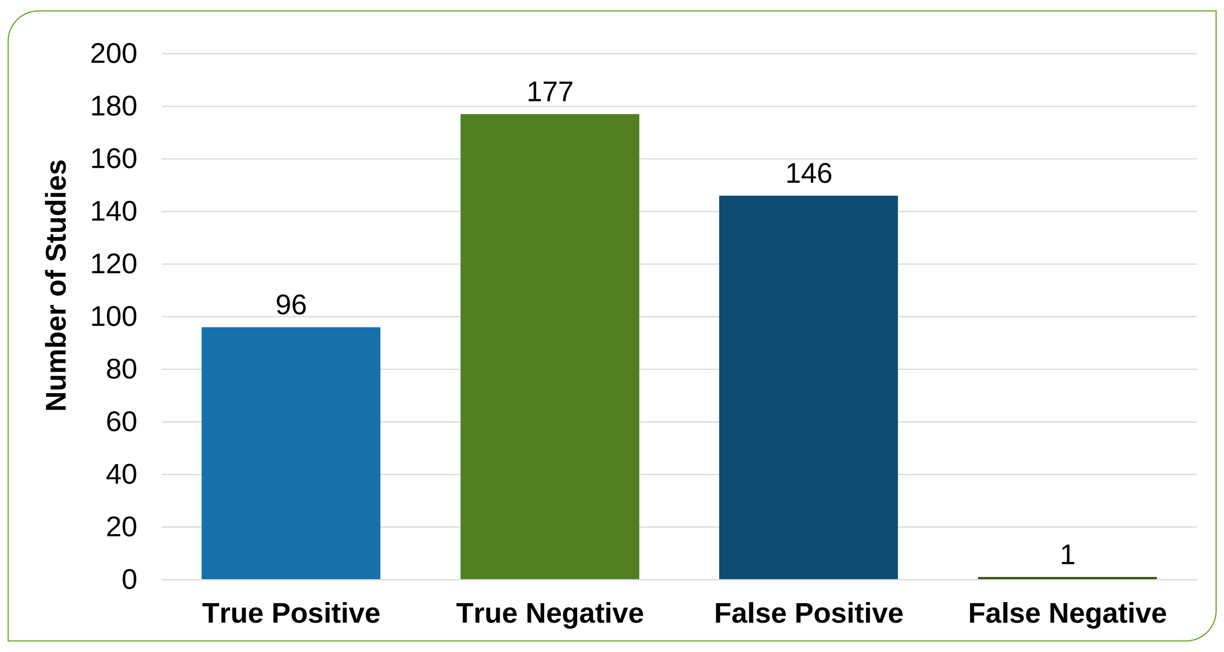
Figure 4: Overall Automation Workflow



Results

- Model showed **99.0% sensitivity** (96/97), effectively identifying studies that should be included in review (**Figure 5**)
  - High sensitivity ensures relevant studies are not overlooked; model is reliable for initial screening and contributes to a comprehensive review
- Model showed **39.6% PPV** (96/242) with 146 false positives identified, reflecting a tendency toward **over-inclusion** (**Figure 5**)
  - High false positive rate requires additional filtering of flagged studies, indicating the need for human review to ensure precision
- Model showed **99.4% NPV** (177/178), indicating **high accuracy** in identifying studies that should be excluded from review (**Figure 5**)
  - High NPV is crucial for minimizing unnecessary inclusions and reducing manual workload for human reviewers, especially in large-scale reviews
- LLM-based approach substantially reduced time spent screening, requiring only 20 minutes per 50 studies compared to 8 hours for a human reviewer (**Figure 6**)
- Thorough error analysis revealed **only one false negative**, underscoring the model's effectiveness in capturing relevant studies (**Figure 7**)

Figure 7: Error Analysis of False Positives and False Negatives



- **Justifications** for inclusion/exclusion were reviewed by domain experts and generally aligned with standard review practices
  - Model's ability to produce understandable and reasonable explanations enhances its utility in the screening process
  - Improves transparency and aiding human reviewers in validating or questioning model's decisions
- Hybrid model **combining LLM screening with human oversight** could provide optimal balance of efficiency and accuracy
  - The high rate of false positives emphasizes the importance of hybrid approach, ensuring balance of precision and recall
  - i.e., LLM would handle initial screening to exclude irrelevant studies; human reviewers would focus on refining final selection, significantly reducing manual workload in large-scale systematic reviews while upholding high standards for evidence synthesis

GPT-4o-based model showed high sensitivity (99.0%) and accuracy (NPV 99.4%) as a full-text study selection tool for literature reviews; time spent reviewing was reduced by 96%

Figure 5: Performance Metrics of LLM Screening

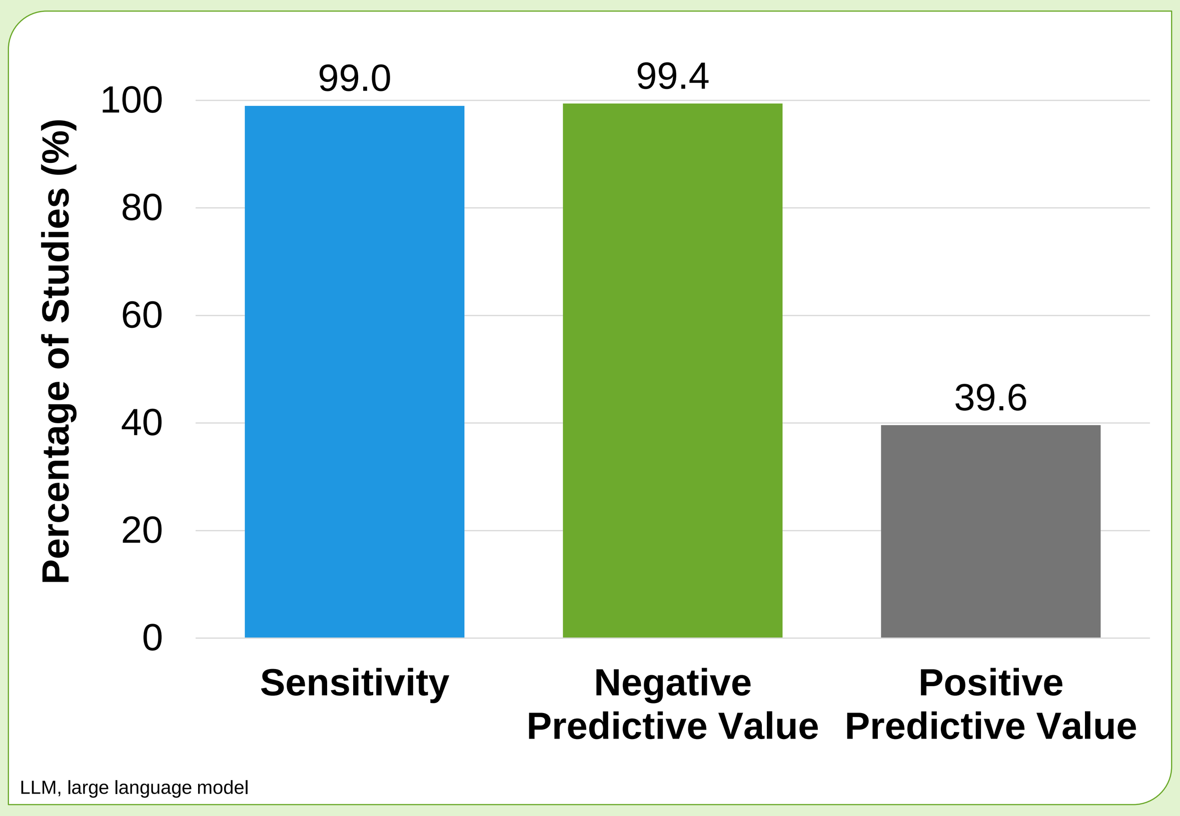
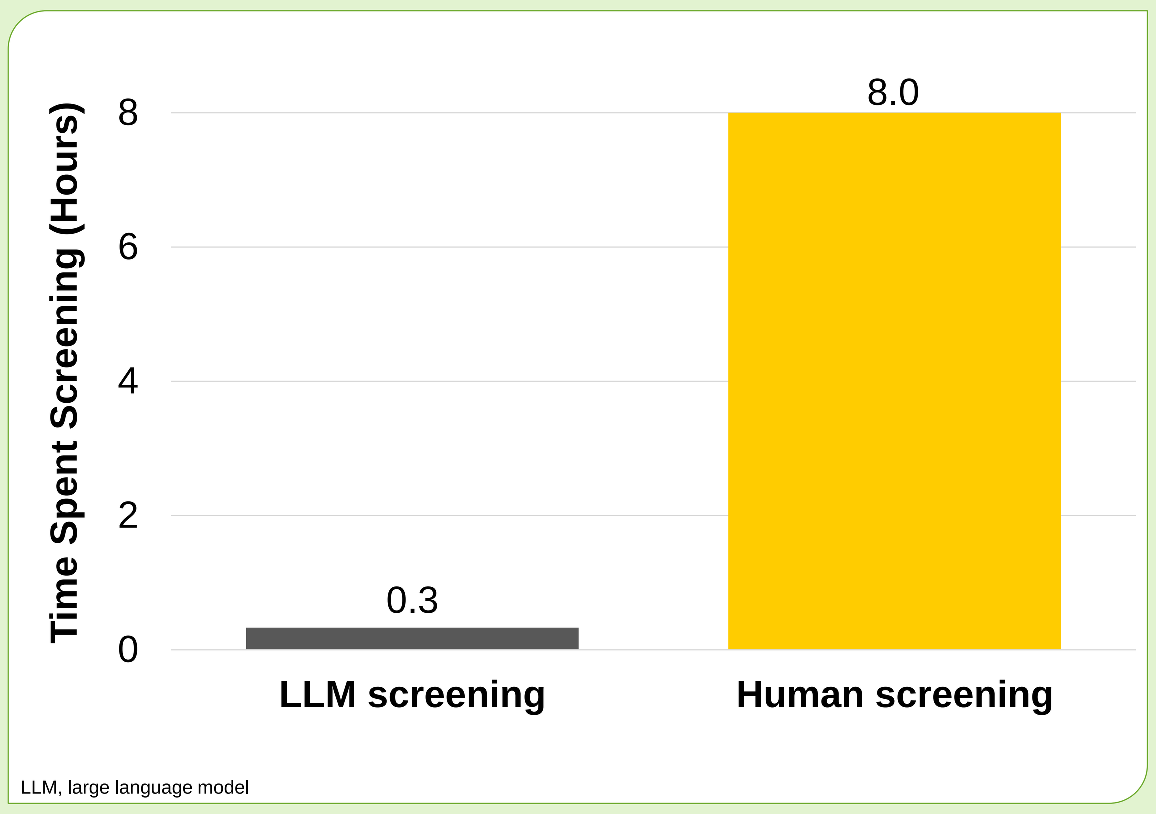


Figure 6: Time for LLM Screening vs. Human for 50 Studies



Conclusions

- Integration of GPT-4o for automating full-text screening in systematic reviews shows significant promise in alleviating the manual workload associated with large-scale reviews
- Model demonstrated high accuracy in exclusion decisions and robust sensitivity in identifying relevant studies, positioning it as a valuable tool in the initial screening stages
  - However, challenges such as over-inclusion and false positives highlight necessity for human oversight to ensure optimal screening
- A hybrid approach that combines LLM-driven automation with expert human review could optimize both efficiency and accuracy in the systematic review process

References

1. Kim JSM, et al. *Syst Rev.* 2022;11:206.
2. Borah R, et al. *BMJ Open.* 2017;7:e012545.
3. Ofori-Boateng R, et al. *Artif Intell Rev.* 2024;57:200.
4. OpenAI GPT-4. 2023. https://openai.com/index/gpt-4-research/
5. LLM Sherpa. 2024. https://github.com/nlmatics/llmsherpa
6. Shahriar S, et al. *Appl Sci.* 2024;14(17):7782.

Acknowledgments

Authors report employment with Evidinno Outcomes Research Inc. (Vancouver, BC, Canada)
Authors would like to thank Ellen Kasireddy of Evidinno Outcomes Research Inc. for her assistance in poster development

EVIDINNO
EVIDENCE NAVIGATION & SYNTHESIS

Email: mfazeli@evidinno.com