Assessing the Performance of Artificial Intelligence for Title-Abstract and Full-Text Screening in Systematic Literature Reviews

Kamboj G[1], Sharma S[1], Malik A[2], **Rathi H**[1,2]

[1]Skyward Analytics, Gurugram, Haryana, India; [2]EasySLR, Gurugram, Haryana, India

SA5

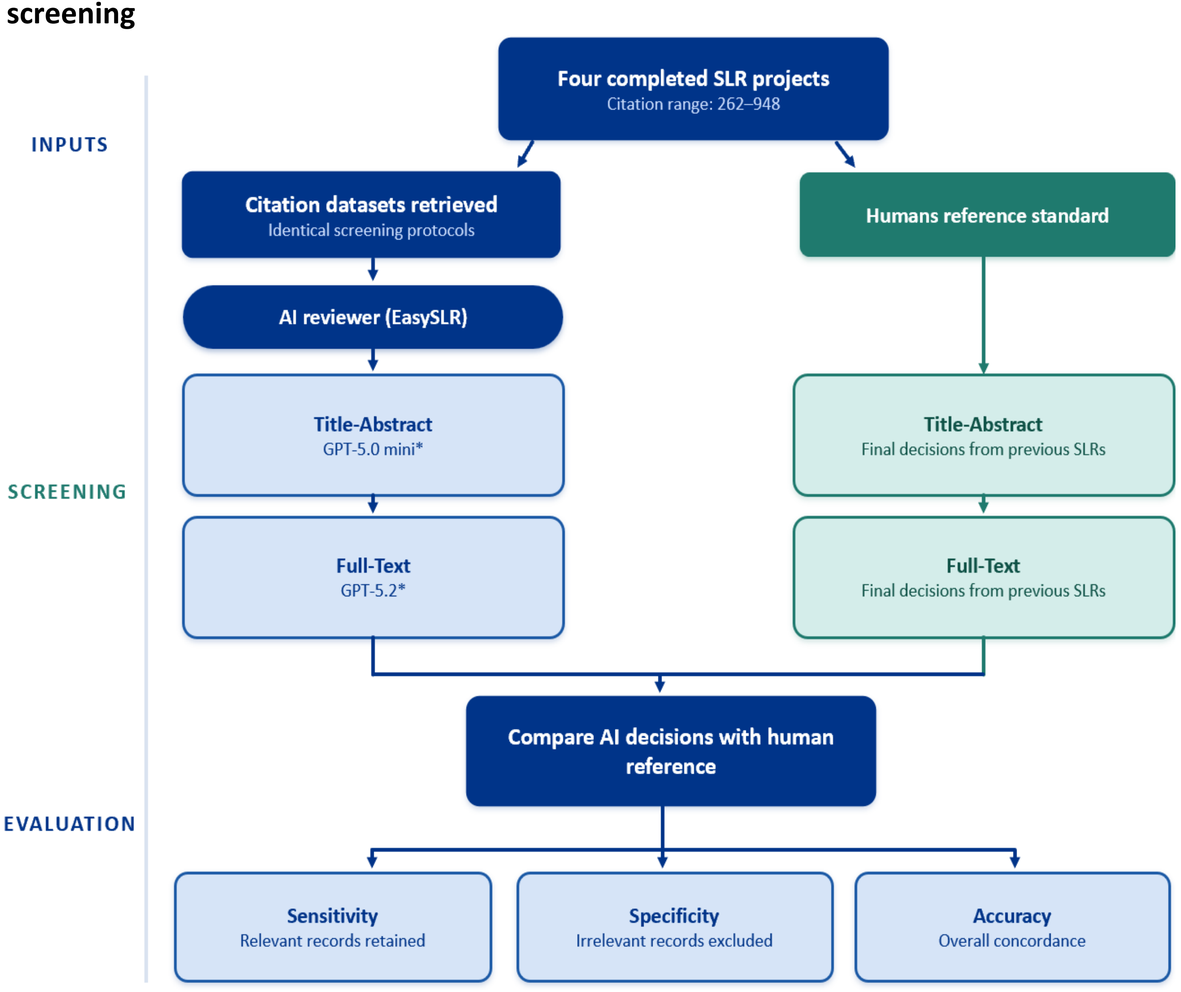
INTRODUCTION

- Systematic literature reviews (SLRs) are a cornerstone of evidence synthesis, providing a structured approach to synthesising the available evidence on a given topic or research question[1]
- Artificial intelligence (AI)-assisted tools are increasingly used in SLRs to enhance efficiency; however, evidence on their performance as independent reviewers at the title-abstract and full-text screening stages remains limited, highlighting the need to evaluate their potential to support human-led evidence synthesis[2,3]
- This study evaluates the performance of an independent AI reviewer for title-abstract and full-text screening using the EasySLR platform against human reviewer decisions

METHODS

- A retrospective validation assessment was conducted across four previously completed SLRs to assess the performance of an independent AI reviewer at the title-abstract and full-text screening stages
- The methodological workflow used to evaluate the performance of an independent AI reviewer is presented in Figure 1
- Human screening decisions from the previous completed SLRs served as the reference standard for evaluating AI performance
- Screening was conducted independently by the AI reviewer within the EasySLR™ platform
- The AI reviewer was provided with the same screening protocols used by the human reviewers
- Title-abstract screening was performed using the OpenAI GPT-5.0 mini model, while full-text screening was performed using the OpenAI GPT-5.2 model
- AI screening decisions were recorded separately for each screening stages

Figure 1. Methodological workflow for evaluating AI reviewer performance in SLR screening



Abbreviations: AI, artificial intelligence; SLRs, systematic literature reviews; Tiab, title-abstract.
*Model selection was stage-dependent: GPT-5.0 mini at title-abstract for volume efficiency, GPT-5.2 at full text where lower record counts allowed a higher-cost model.

- **Performance Evaluation:** The performance of the AI reviewer was assessed separately for each SLR and screening stage using the following measures:
- **Sensitivity:** proportion of studies included by human reviewers that were correctly identified as included by the AI reviewer
- **Specificity:** proportion of studies excluded by human reviewers that were correctly identified as excluded by the AI reviewer
- **Accuracy:** overall proportion of screening decisions that were concordant between the AI and human reviewers

Funding: No external funding was received for this study.

RESULTS

- Across the four SLRs, the number of citations screened ranged between 262–948. At the title-abstract stage, the AI demonstrated high sensitivity (80–100%), with moderate-to-high specificity (60.9–88.5%) and accuracy (63.0–90.1%)
- While the AI reliably captured relevant citations at the title-abstract stage, it generated a variable number of false positives (Figure 2)
- Full-text screening yielded notably stronger and more consistent alignment with human reviewer judgements compared to the title-abstract stage
- At the full-text stage, sensitivity ranged from 83.3% to 100%, specificity was 100%, and accuracy ranged from 91.7% to 100% across the four SLRs
- Overall, high concordance between AI and human screening decisions was observed at the full-text screening stage (Figure 3)

Figure 2. Title-abstract screening performance

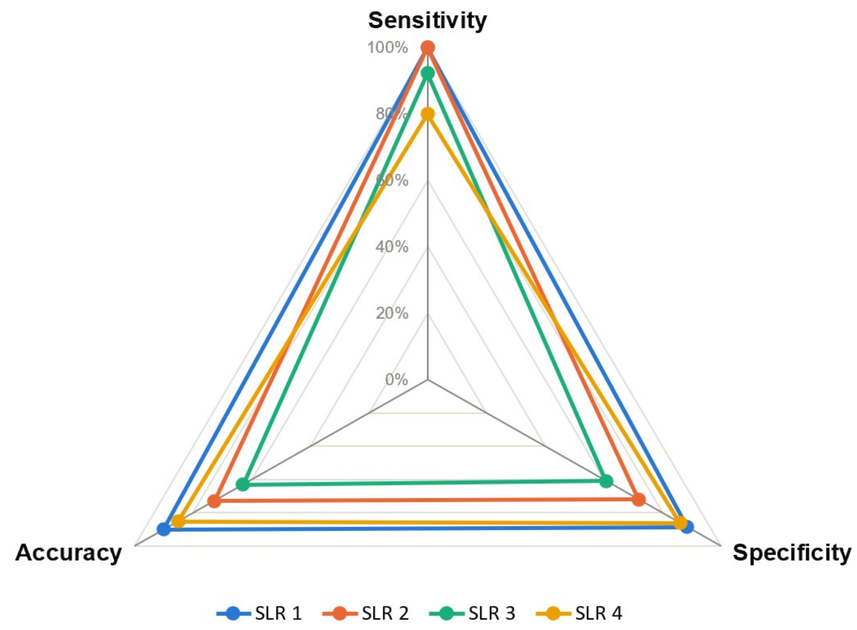
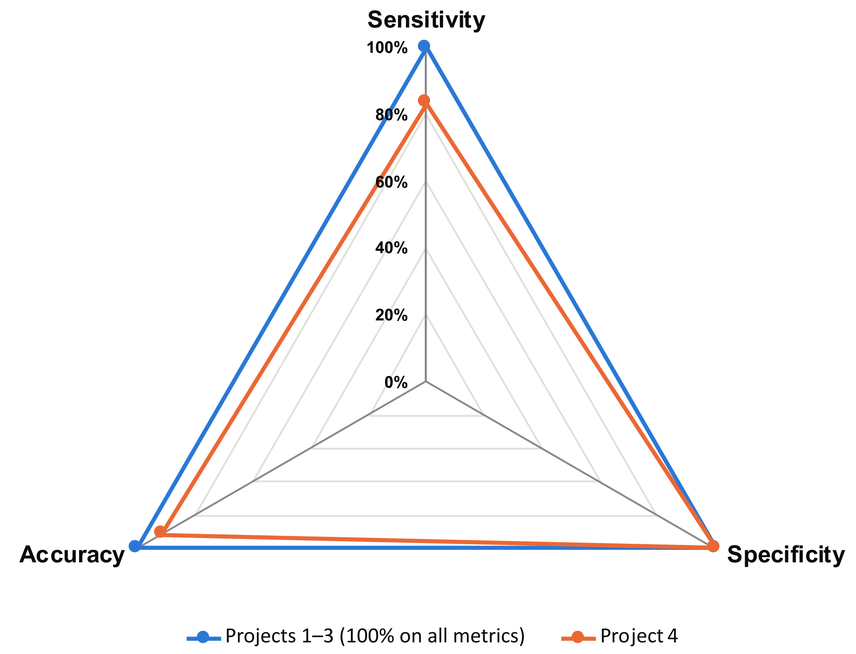


Figure 3. Full-text screening performance



Strengths and Limitations

- Validation across four SLRs using consistent screening protocols enabled a robust comparison of AI and human reviewer decisions
- Evaluation across both title-abstract and full-text screening stages provided a comprehensive assessment of AI screening performance
- The relatively low citation volumes across the four SLRs may limit generalisability, as performance may vary with larger evidence bases
- Findings are specific to the GPT-5.0 mini and GPT-5.2 models

CONCLUSION

The findings indicate that AI may support SLR workflows by improving screening efficiency while maintaining strong agreement with human reviewer decisions, particularly during full-text evaluation. Although AI demonstrated high sensitivity, human oversight remains essential to ensure methodological rigor. These findings should be interpreted cautiously, as performance may vary across research questions. Future research should evaluate AI performance on larger datasets and refine screening protocols to improve AI comprehension.

Poster presented at ISPOR Asia Pacific Summit 2026, Bangkok, Thailand (06–08 September 2026)

REFERENCES:
1. Fabiano N, et al. JCPP Adv. 2024 Apr 23;4(2):e12234. 2. Zhan J, et al. Syst Rev. 2025 Dec 18;15(1):24. 3. Hamel C, et al. BMC Med Res Methodol. 2021 Dec 20;21:285.

easySLR Fly Through Reviews

SKYWARD ANALYTICS