Accelerating Literature Reviews with Large Language Models (LLMs): An Evaluation of Performance and Efficiency

MSR190

Raju Gautam, PhD[1], Saeed Anwar, MPharm[2], Ratna Pandey, MSc[2], Tushar Srivastava, MSc[1]
[1]ConnectHEOR, London, UK; [2]ConnectHEOR, Delhi, India | Email: raju.gautam@connectheor.com

connectHEOR

KEY MESSAGE

AI-human hybrid workflows in EasySLR™ achieved 84%–100% accuracy in title/abstract screening and 90%–100% sensitivity in full-text screening, with efficiency gains of 40%–60% over human-only workflows. AI-only approaches offer up to 300%–500% speed improvement in data extraction, but accuracy remains limited (9%–60%), underscoring the need for hybrid models in high-stakes evidence synthesis. AI in literature reviews is moving toward a co-pilot model, where human expertise and machine efficiency are tightly integrated. The future lies not in full automation, but in **intelligent augmentation**, enabling faster, scalable, and high-quality evidence generation.

BACKGROUND

- Systematic literature reviews (SLRs) are essential for evidence-based healthcare decision-making but are time- and resource-intensive, often requiring extensive manual effort across screening and data extraction stages. This can delay the timely completion of evidence packages, particularly for large evidence bases.
- Advances in artificial intelligence (AI), especially large language models (LLMs), offer opportunities to streamline these processes by automating tasks such as study identification and data extraction[1]. While early applications show improved efficiency, concerns remain regarding accuracy, reliability, and performance in complex tasks[2]. Furthermore, limited information exists across review types. Thus, rigorous evaluation of AI-driven tools is needed to assess their effectiveness and appropriate role in evidence synthesis.

OBJECTIVES

1. Evaluate performance and accuracy of AI based tool for review across key stages of SLR: title/abstract screening, full-text screening, and data extraction.
2. Quantify efficiency gains of AI-only vs. hybrid AI-human workflows compared to traditional human-only reviews.
3. Assess domain-specific variability between clinical and economic literature reviews.

METHODOLOGY

Study Design

A retrospective evaluation framework was implemented to assess both the performance and operational efficiency of a web-based AI-enabled SLR platform (EasySLR™). The analysis included previously completed clinical and economic SLRs to capture variation in review types.

TITLE Abstract Screening (T/A) → Full Text Screening (FT) → Data Extraction (DE)

Performance Metrics

- Accuracy
- Sensitivity (true positive inclusion rate)
- Specificity (true negative exclusion rate)
- Processing speed improvement vs. human-only baseline

Efficiency Assessment

Efficiency was quantified by comparing:

- Time per article (AI vs human)
- Total time saved (minutes and hours) across stages

Dataset Sizes used for evaluation: 2333 articles for (T/A screening) | 104 studies (full-text) | 97 studies (data extraction)

RESULTS

✓ A total of 2,333 articles of clinical and economic review were screened by both human reviewers and AI as shown in Table 1 and 2, which showed generally higher performance in economic studies during screening stages, while data extraction accuracy was substantially lower in clinical studies compared to moderate performance in economic studies.

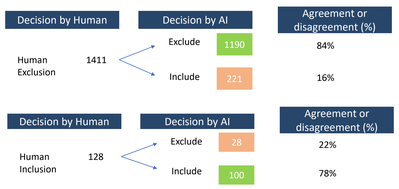
Table 1: Performance metrics (accuracy, specificity, and sensitivity) for the **clinical review (1539 articles for screening)**

Review Step	Efficiency Improvement	Accuracy	Sensitivity	Specificity
Title/abstract screening	150%	84%	78%	84%
Full text screening	11%	70%	90%	70%
Data-extraction (only study details)	500%	9%	NA	NA

Table 2: : Performance metrics (accuracy, specificity, and sensitivity) for the **economic review (794 articles for screening)**

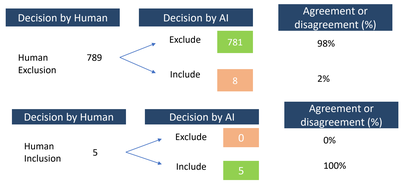
Review Step	Efficiency Improvement	Accuracy	Sensitivity	Specificity
Title/abstract screening	100%	99%	95%	98%
Full text screening	12%	85%	98%	88%
Data-extraction (only study details)	300%	60%	NA	NA

Fig 1: Human vs AI decision concordance in clinical screening



NOTE: All the studies included in final stage (2nd Pass) were included by AI and none were missed.

Fig 2: Human vs AI decision concordance in Economic screening



NOTE: All the studies included in final stage (2nd Pass) were included by AI and none were missed.

RESULTS (CONTINUED)

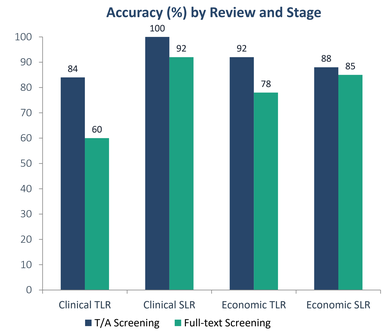


Fig 3: Accuracy (%) across T/A and FT screening.

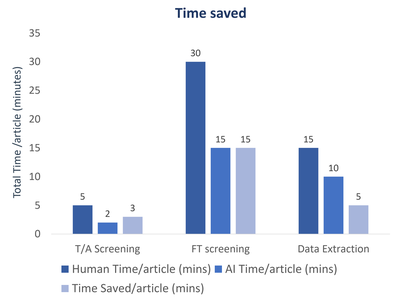
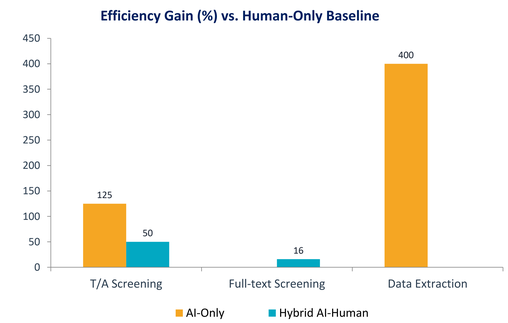


Fig 4: Time per article and time saved (mins) using AI vs human across SLR stages.



	Efficiency	Sensitivity
T/A	100-150%	78-95%
FT	11-12%	90-98%

Fig 5: Efficiency comparison between AI-only and hybrid AI–human approaches across SLR stages.

✓ Figures 1 and 22 illustrate the agreement between human and AI decisions for clinical and economic studies, respectively, demonstrating high concordance across both review types, with particularly strong alignment in economic studies and no missed inclusions at the final stage.

✓ Figure 3 shows that accuracy was consistently high across review stages, with the highest performance observed in clinical SLR (100% at title/abstract and 92% at full-text screening), while accuracy decreased at the full-text stage across both reviews, particularly for clinical TLR (60%) and economic TLR (78%), with economic SLR maintaining relatively stable performance (88% and 85%).

✓ Figure 4 demonstrates that AI substantially reduced time per article across all SLR stages, with the greatest time savings observed during full-text screening (15 minutes), followed by data extraction (5 minutes) and title/abstract screening (3 minutes).

✓ Figure 5 indicates that the hybrid AI–human approach substantially reduced workload compared to AI-only, particularly in title/abstract screening (50 vs. 125) and data extraction, while full-text screening required minimal hybrid input (16), highlighting efficiency gains through selective human involvement.

✓ At the screening stages, the hybrid AI–human approach maintains high accuracy, sensitivity, and specificity (e.g., T/A: 84–100% accuracy; Full-text: up to 92% accuracy) while achieving moderate speed gains (12–60%), suggesting an optimal balance between efficiency and reliability, as shown in Table 3.

✓ AI-only enables higher speed (100–500%) with minimal-to-compromised accuracy, while hybrid maintains accuracy with moderate speed gains (12–60%), as shown in Table 3.

RESULTS (CONTINUED)

Summary of AI-Human Hybrid vs. AI-Only Performance

Stage	Workflow	Accuracy	Sensitivity	Specificity
T/A Screening	Hybrid	84%–100%	70%–97%	84%–100%
Full-text Screen	Hybrid	60%–92%	90%–100%	70%–88%
Data Extraction	AI-Only	9%–60%	N/A	N/A

Table 3: Performance matrix across review stages by workflow (hybrid vs AI-only).

Efficiency Gains vs. Human-Only Workflows

Stage	Approach	Speed Improvement	Accuracy Trade-off
T/A Screening	AI-Only	100%–150%	Minimal
T/A Screening	Hybrid	40%–60%	Maintained
Full-text Screen	Hybrid	12%–20%	Maintained
Data Extraction	AI-Only	300%–500%	Compromised

Table 4: Speed–accuracy trade-offs across review stages for AI-only and hybrid approaches.

Limitations

Key Limitations

- Reduced reliability in clinical reviews: Performance appears weaker compared to economic reviews, indicating domain sensitivity.
- Limited generalizability: Evaluation is based on only four reviews, which may not represent all therapeutic areas or review complexities.
- Retrospective design: Results are dependent on previously completed human reviews, which may introduce bias in benchmarking.
- Potential oversight risks: AI-only approaches may miss relevant studies (sensitivity gaps), impacting review completeness.

While AI-only methods deliver substantial efficiency gains, particularly in title/abstract screening and data extraction, their lower accuracy limits use in complex tasks such as full-text screening and clinical reviews.

CONCLUSIONS

AI tools, especially LLMs, can significantly speed up literature reviews, particularly for repetitive tasks like screening and data extraction. However, their accuracy is not consistent, especially for more complex tasks and clinical studies, so they cannot fully replace human reviewers yet. A combined AI–human approach works best, improving efficiency while maintaining quality. Overall, using AI for simpler tasks and humans for more complex decisions is the most effective strategy, with future improvements expected to make AI even more useful.

Hybrid AI-human workflows provide the best balance of accuracy and efficiency, achieving 84%–100% accuracy in T/A screening with 40%–60% time savings.

AI integration can substantially accelerate literature review workflows, particularly in high-volume stages

Full-text screening remains dependent on human expertise, with only modest gains from AI support

A task-specific implementation strategy, using AI for speed and humans for complex judgment, shall offers the most effective model.

AI performance is domain-dependent, with higher accuracy in structured economic reviews and lower reliability in complex clinical contexts, highlighting the need for domain-specific model adaptation

Continued development, domain adaptation, and prospective validation of AI-SLR tools will be critical to improving applicability in high-stakes evidence synthesis

References:
1. Sanghera, Rohan, et al. "High-performance automated abstract screening with large language model ensembles." Journal of the American Medical Informatics Association 32.5 (2025): 893-904.
2. Zhan, J. et al. Accelerating the pace and accuracy of systematic reviews using AI: a validation study. Syst Rev 15, 24 (2026).

Poster presented at ISPOR in Pennsylvania, Philadelphia, USA, May 17–20, 2026

Financial Disclosure: Authors are employees of ConnectHEOR Limited. No external funding received to conduct this research and no conflict of interest to declare.