A Multi-Model Large Language Framework for Automating Systematic Literature Reviews

MSR78

Ritesh Dubey, Ankita Sood, Vedant Soni, Gagandeep Kaur, Rajdeep Kaur, Barinder Singh
Pharmacoevidence, Mohali, India

Science-led HEOR & Tech-driven AI

INTRODUCTION

- A Systematic Literature Review (SLR) is a structured and predefined process used to systematically identify, screen, and synthesize available evidence to support healthcare research and evidence-based decision-making[1]
- The title and abstract screening phase is the most resource-intensive component of an SLR, typically requiring multiple independent reviewers and substantial manual effort to ensure comprehensive identification of relevant studies[2]
- Recent advances in large language models (LLMs) have enabled partial automation of screening tasks, demonstrating the potential to substantially reduce reviewer workload and accelerate systematic review timelines while maintaining high sensitivity[3]
- In Health Technology Assessment (HTA), SLRs are central to evidence generation for regulatory submissions, where screening efficiency and methodological rigor directly influence downstream evidence appraisal and decision-making[4]
- The first artificial intelligence (AI)-assisted HTA submission accepted by National Institute for Health and Care Excellence (NICE) showed that using AI as a second reviewer for title and abstract screening reduced SLR time and cost by about 50%, supporting feasibility in real-world HTA workflows[5]

OBJECTIVE

- To assess whether additional efficiencies can be achieved through fully automated SLR screening using multiple LLMs and confidence-guided decision outputs, compared to the semi-automated benchmark

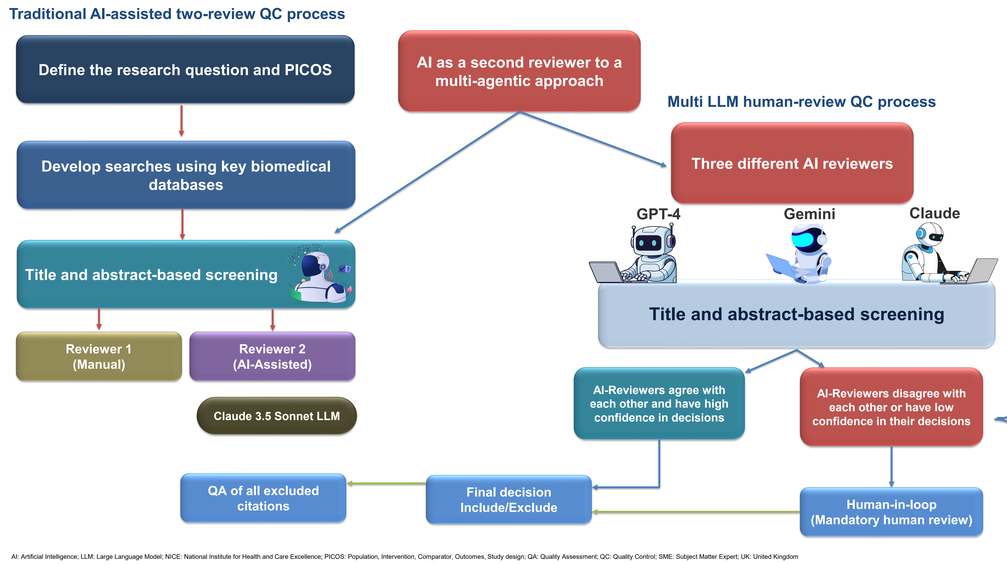
METHODS

- Records from EMBASE®, MEDLINE®, and Cochrane were identified and uploaded into a custom Python-based automated screening interface
- Predefined inclusion and exclusion criteria were specified within the screening protocol prior to model evaluation (refer to Figure 2 for the tool interface)
- Title and abstract screening was performed simultaneously across three LLMs (Claude Sonnet 3.7, Gemini Flash 2.5, and GPT-4o-mini) using predefined inclusion and exclusion criteria (Figure 1)
- A subject matter expert (SME) optimized and fine-tuned the final prompt and conducted quality control (QC) on a sample of AI-processed records to validate screening accuracy
- Each citation was evaluated independently by all models, and screening decisions were finalized only when full agreement was achieved
- Citations with conflicting model decisions or with low confidence were considered as a conflict

RESULTS

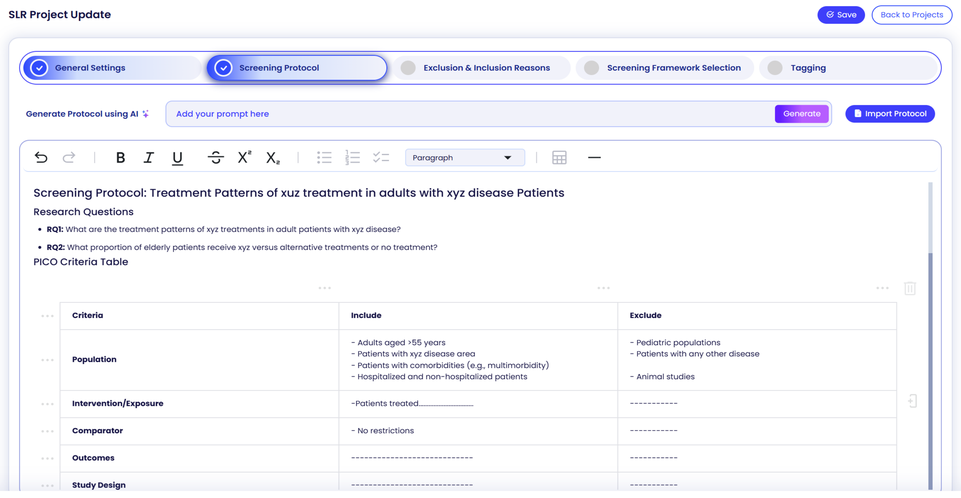
- A three-model consensus mechanism was employed, requiring agreement among LLMs before a final automated decision was recorded, reducing the risk of single-model bias
- Each LLM independently assigned an include/exclude decision alongside a confidence score, enabling a transparent and auditable screening trail
- A total of 1,840 citations underwent automated title and abstract screening using the proposed multi-LLM framework (Figure 3)
- The end-to-end automated screening workflow - from record intake to consensus-based include/exclude decisions - is summarized in Figure 1
- Confidence score-based agreement across three LLMs enabled automated exclusion of the majority of citations without human intervention
- A confidence cut-off of 70% was applied; any decision scoring below this threshold was re-screened by human SMEs to mitigate potential errors
- Only 7% of citations were flagged for manual review due to low model confidence or inter-model decision conflicts

Figure 1: AI as a second reviewer (Methodology aligned to AI-SLR approach accepted by NICE UK) to a multi-agentic approach



AI: Artificial Intelligence; LLM: Large Language Model; NICE: National Institute for Health and Care Excellence; PICOS: Population, Intervention, Comparator, Outcomes, Study design; QA: Quality Assessment; QC: Quality Control; SME: Subject Matter Expert; UK: United Kingdom

Figure 2: First-stage automated title and abstract screening workspace



AI: Artificial Intelligence; PICO: Population, Intervention, Comparator, Outcome; RQ: Research Question; SLR: Systematic Literature Review

RESULTS

- SMEs confirmed that no relevant citations were incorrectly excluded by any of the three LLMs, highlighting strong recall performance
- The observed over-inclusion bias (~1-2%) was directionally conservative, meaning that the framework erred on the side of including uncertain citations rather than risking false exclusions, which is an appropriate behavior for systematic review screening
- Compared with a semi-automated baseline (AI as a second reviewer, ~50% time saving), the proposed approach delivered an additional 40% efficiency gain
- The proposed framework reduced the volume of citations requiring human attention from 1,840 to approximately 129, representing a greater than 90% reduction in manual screening burden
- Time-to-decision per citation was substantially reduced compared to both fully manual and semi-automated approaches, enabling the screening phase to be completed in a fraction of the conventionally expected timeline
- Pairwise agreement rates between LLMs were consistently high across all three model combinations, indicating that consensus was not driven by any single dominant model but reflected genuine tri-model alignment

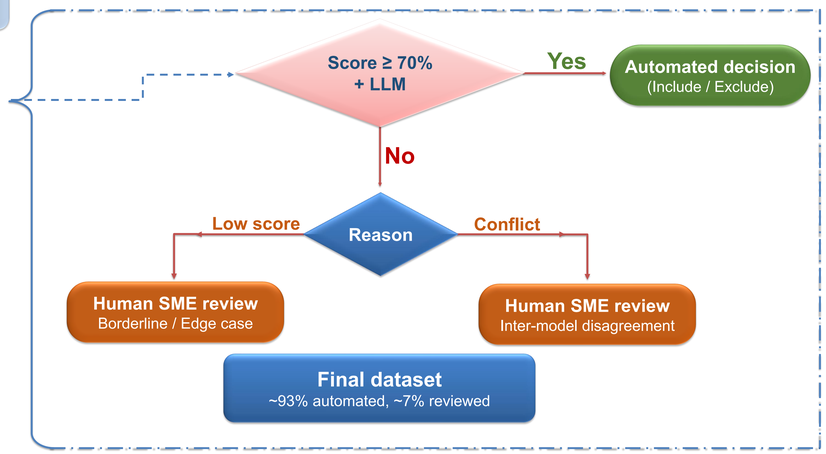
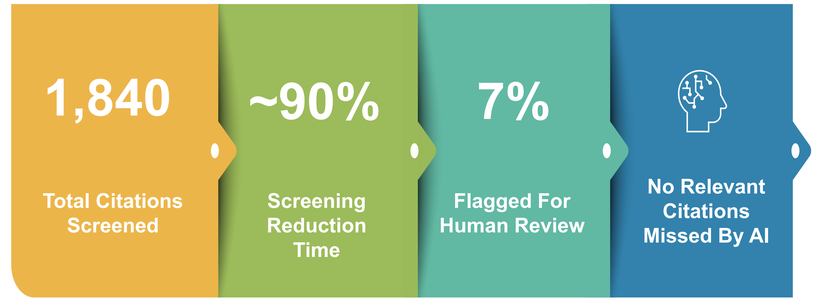


Figure 3: Key screening outcomes



AI: Artificial Intelligence

CONCLUSIONS

- The multi-LLM, confidence-guided framework demonstrated the feasibility of fully automated SLR screening, substantially reducing manual workload while maintaining high reliability
- The approach reduced manual screening burden by >90%, with only ~7% of citations requiring human review due to low confidence or inter-model disagreement
- High inter-model agreement and absence of false exclusions confirmed strong recall, with a conservative over-inclusion bias supporting safe application in systematic review workflows
- Compared with semi-automated screening, the framework delivered additional efficiency gains, supporting scalable implementation, with continued human oversight to ensure methodological rigor and regulatory alignment

References

1. Mahuli et al. Br Dent J. 2023;235(2):90-92; **2.** Issaiy et al. BMC Med Res Methodol. 2024;24(1):78; **3.** Sanghera R, et al. J Am Med Inform Assoc. 2025;32(5):893–904; **4.** Singh B et al. Value in Health, Volume 28, Issue S2, 2025; **5.** Makhija et al. Value Health. 2025;28(12):S496.

Correspondence: Barinder Singh; barinder.singh@pharmacoevidence.com

Disclosure: AS, RD, VS, GK, RK and BS, the authors declare that they have no conflict of interest

Acknowledgements: The authors wish to thank AK (Adarsh Kumar), RA (Rubal Arora), & RS (Rythem Sharma) for their valuable support in drafting this poster