A Comparative Analysis of Large Language Models in Title and Abstract Screening Guided by Human Oversight

MSR106

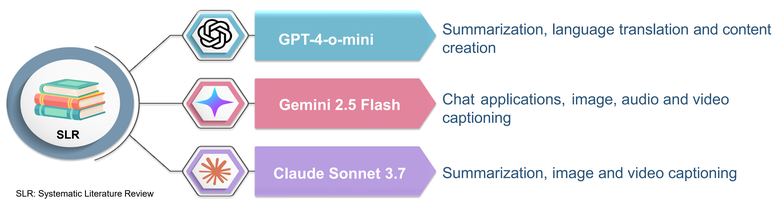
Ankita Sood, Ritesh Dubey, Sunil Kumar, Marjana Bharali, Gagandeep Kaur, Rajdeep Kaur, Barinder Singh
Pharmacoevidence, Mohali, India

Evidence Pharmaco®
Science-led HEOR & Tech-driven AI

INTRODUCTION

- Systematic literature reviews (SLRs) are considered the gold standard for synthesizing clinical evidence, providing the highest level of evidence to inform clinical guidelines and decision-making in evidence-based healthcare[1]
- Title and abstract screening is the most resource-intensive phase of an SLR, typically requiring substantial manual effort to ensure high sensitivity and comprehensive study identification[2]
- The volume of published clinical literature has grown at an average annual rate exceeding 10%, further intensifying the screening workload and making traditional manual approaches increasingly unsustainable[3]
- Recent advances in large language models (LLMs) have demonstrated the potential to automate title and abstract screening, mitigating screening burden while accelerating the systematic review process[4,5]
- Prior evaluations of LLM-assisted screening highlight the importance of human oversight to ensure accuracy, reproducibility, and alignment with systematic review standards[6,7]
- In a prior study, we evaluated first-generation LLMs- Claude Sonnet 3.5, Gemini Flash 1.5, and GPT-4, for automated title and abstract screening, showing that Artificial Intelligence (AI)-assisted screening with expert oversight achieves over 90% accuracy while significantly reducing manual effort[6-8]

Figure 1: Automated screening workflow



SLR: Systematic Literature Review

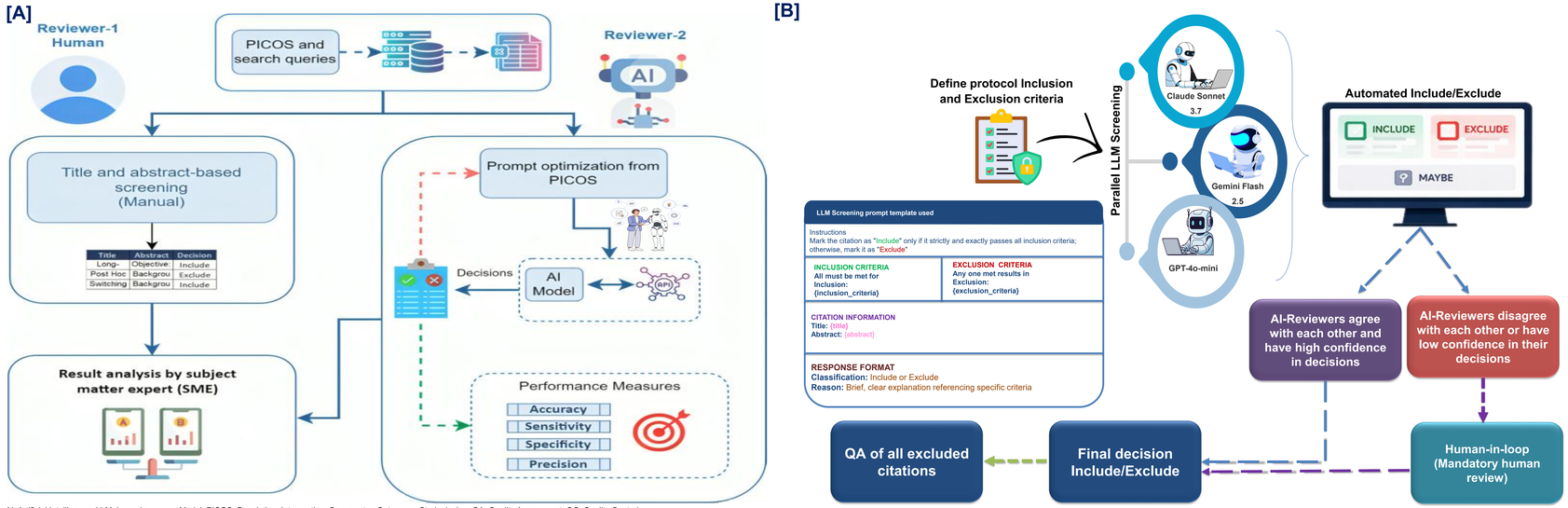
OBJECTIVE

- To evaluate the performance of next-generation LLMs - Claude Sonnet 3.7, Gemini Flash 2.5, and GPT-4o-mini - in automating title and abstract screening for SLRs
- To assess alignment with the National Institute for Health and Care Excellence (NICE) and the Canadian Drug Agency–Common Drug Review (CDA-AMC) guidelines

METHODS

- EMBASE®, MEDLINE®, and Cochrane databases were searched to identify relevant randomized controlled trials (RCTs) in a psychiatric disorder
- A structured search strategy was applied using predefined inclusion and exclusion criteria aligned with NICE and CDA-AMC guidelines[9,10]
- Title and abstract screening was performed to identify potentially relevant studies
- A Python-based interface was developed to facilitate automated interaction between input datasets and three LLMs - Claude Sonnet 3.7, Gemini Flash 2.5, and GPT-4o-mini - for title and abstract screening
- The interface iterated through each record, passing the title, abstract, and eligibility criteria to each LLM for independent evaluation
- A standardized uniform prompt framework ensured consistent screening decisions across all models
- A parser function was used to extract model outputs as “Include” or “Exclude”
- Final screening decisions were accepted when all three models reached consensus; records with discordant outputs were escalated for manual review
- The SME conducted quality control (QC) on a sample of AI-processed records to validate model outputs and assess overall performance
- A subject matter expert (SME) with over a decade of domain knowledge optimized and fine-tuned the final prompt
- The automated screening workflow produced consistent title and abstract screening outcomes across all three LLMs, supported by structured expert oversight and a consensus-based decision framework (Figure 2)

Figure 2: Systematic workflow diagram for title and abstract screening [A] Traditional AI-assisted two-review QC process; [B] Tested Multi-LLM two review QC process



AI: Artificial Intelligence; LLM: Large Language Model; PICOS: Population, Intervention, Comparator, Outcomes, Study design; QA: Quality Assessment; QC: Quality Control.

RESULTS

- Records with discordant screening decisions were automatically flagged for focused manual review, ensuring transparent and robust screening outcomes without compromising review efficiency
- All three next-generation LLMs demonstrated high performance in title and abstract screening, with accuracy and sensitivity exceeding 90% - meeting acceptable thresholds for automated screening applications (Figure 3)
- Among the evaluated models, Claude Sonnet 3.7 achieved the highest accuracy at 97.34%, followed by Gemini Flash 2.5 at 95.05% and GPT-4o-mini at 93.48%. A similar ranking was observed for sensitivity, with Claude Sonnet 3.7 attaining 98.79%, Gemini Flash 2.5 at 94.86%, and GPT-4o-mini at 93.66%

IMPLICATIONS

Accurate & Scalable

- All three LLMs exceeded 93% accuracy - surpassing accepted thresholds for automated screening
- Multi-LLM automation enables rapid, repeatable, and HTA-ready evidence synthesis

Sensitive & Improved

- Claude Sonnet 3.7 achieved 98.79% sensitivity - minimizing the risk of missing relevant studies
- Next-generation models outperformed their predecessors, confirming measurable gains across model generations

Adaptive & Efficient

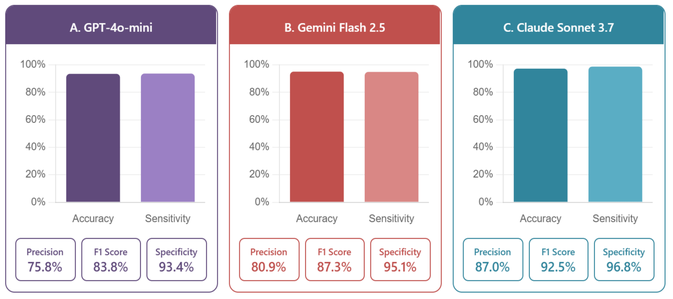
- Consensus-based framework manages screening uncertainty by escalating discordant records for expert review
- AI-driven automation significantly reduces manual screening effort while maintaining review quality

Rigorous & Robust

- Expert oversight and quality control ensured methodological alignment with NICE and CDA-AMC standards
- Consistent performance across three different LLM architectures confirms generalizability of the approach

- Compared to our prior evaluation of first-generation LLMs (ISPOR Europe 2024), next-generation models demonstrated a notable improvement in overall screening performance. Claude Sonnet 3.7 (97.34%) outperformed its predecessor, Claude Sonnet 3.5 (94.69%) in accuracy, and showed a marked improvement in sensitivity (98.79% vs 88.59%). Similarly, Gemini Flash 2.5 (95.05%) exceeded Gemini Flash 1.5 (96.02%) in sensitivity while maintaining comparable accuracy. GPT-4o-mini (93.48%) performed comparably to GPT-4 (95.00%)
- These findings suggest that advances in LLM architecture translate into meaningful gains in screening accuracy and sensitivity, reinforcing the value of periodic re-evaluation of AI tools as model generations evolve

Figure 3: Comparison of all LLM model for SLR screening phase



CONCLUSIONS

- The findings highlight the growing potential of next-generation LLMs to meaningfully enhance title and abstract screening in systematic reviews, with improvements in efficiency, consistency, and scalability, thereby supporting more streamlined and resource-efficient evidence synthesis processes
- Building on consistent performance improvements observed across two generations of LLMs, future efforts should focus on prospective validation across diverse therapeutic areas and review types, integration of adaptive consensus thresholds, and exploration of fully autonomous screening pipelines-while maintaining mandatory human expert oversight to ensure regulatory alignment and methodological rigor as AI capabilities continue to evolve

References

1. Grant MJ & Booth A. Health Inf Libr J. 2009;26(2):91–108; **2.** Chai KEK et al. Syst Rev. 2021;10(1):93; **3.** Zhao X et al. Eur J Med Res. 2022;27(1):95; **4.** Li Y et al. arXiv. 2024. arXiv:2512.11261; **5.** Giannos P. Information. 2025;16(5):378
6. Sood A et al. Value in Health, Volume 27, Issue S2, 2024; **7.** Oami T et al. JAMA Netw Open. 2024;7(7):e2420496; **8.** Oami T et al. Res Synth Methods. 2025. doi:10.1017/rsm.2025.10014; **9.** NICE. Developing NICE Guidelines: The Manual - Ch.6. 2022;
10. CDA-AMC. Finding the Evidence: Literature Searching Tools. 2024

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

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

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

ISPOR, May 17-20, 2026, Philadelphia, PA, United States