MSR170

Adapting Agentic Large Language Models (aLLM) Trained in Solid Tumors for Systematic Literature Review (SLR) in Hematological Malignancies: Validation in Multiple Myeloma (MM) and Chronic Lymphocytic Leukemia (CLL)

Rhiannon Campden[1], Rozee Liu[1], Eddie Xiaole Liu[2], Anna Forsythe[1]

[1]Oncoscope-AI, Miami, FL, USA; [2]Independent, Toronto, ON, Canada

BACKGROUND

- We have previously published data on agentic large language model (aLLM) systems demonstrating strong performance in automating systematic literature reviews (SLRs) in solid tumors[1,2]
- However, extending these systems to hematological malignancies presents distinct challenges, including differences in therapeutic classes, endpoints, disease definitions, and study designs

OBJECTIVES

- → This study evaluates the adaptation and validation of a Real-time AI-assisted Living SLR (REAL-SLR) system originally developed for solid tumors
- → Here, we assess the ability of the aLLM to accurately annotate inclusion/exclusion decisions for individual Population, Intervention/Comparator, Outcomes, and Study design (PICOS) criteria for studies in Multiple Myeloma (MM) and Chronic Lymphocytic Leukemia (CLL)

METHODS

- Our aLLM system comprises multiple autonomous LLMs operating without direct supervision, including GPT-5, GPT-4.1, Gemini 2.5 Pro, and Claude Sonnet 4.5, designed to emulate trained human reviewers
- Models were adapted using hematology-specific treatment guidelines and annotation manuals aligned with PRISMA and Cochrane standards and structured around the Population, Intervention/Comparator, Outcomes, and Study Design (PICOS) framework
- Inclusion and exclusion decisions were recorded independently for each PICOS element and benchmarked against expert human screening
- An iterative refinement process was applied to the annotation manual until >95% accuracy performance and <1% overall false negative thresholds and were achieved
- False negative rates were defined as the number of studies falsely marked as exclude by the aLLM per individual PICOS criteria. Overall false negative rates were calculated based on the results for all four PICOS criteria

RESULTS

- In MM, the aLLM reviewed 800 abstracts, achieving an initial overall accuracy of 93.7% with an overall false negative rate of 3.5% (**Figure 1**)
- Following targeted refinement of instructions addressing hematology-specific interventions, outcomes, and study designs, final accuracy increased to 97.4% (Population 98.3%, Intervention/Comparator 97.3%, Outcomes 97.5%, Study Design 96.9%), exceeding single human reviewer performance (**Figure 1**)
- The overall false negative rate was <0.7%, below the predefined 1% threshold (**Figure 2**)
- In CLL, the aLLM reviewed 891 abstracts, achieving an initial overall accuracy of 93.29% with an overall false negative rate of 2.0% (**Figure 3, 4**)
- A final overall accuracy of 97.3% and a false negative rate of 0.17% was achieved after initial adaptation (**Figure 3,4**)

Figure 1. Overall accuracy of aLLM screening results for MM before and after review and refinement of the model

(True positive screening results)

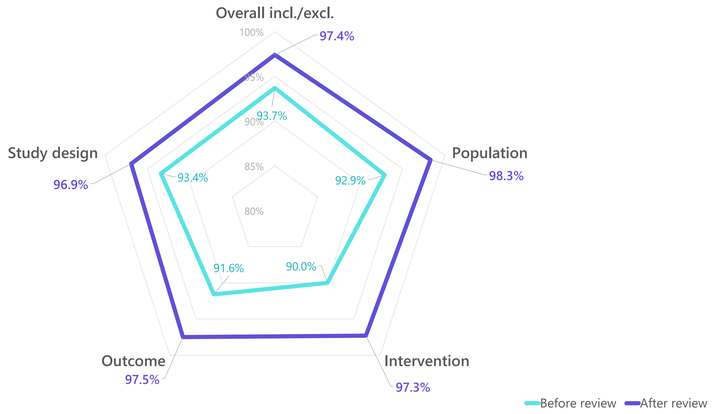


Figure 2. False negative aLLM screening results for MM before and after review

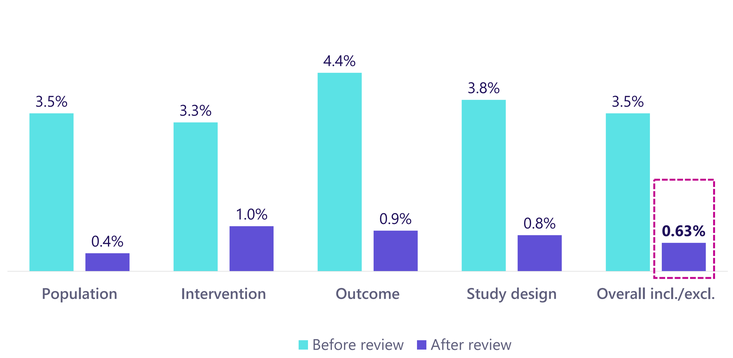


Figure 3. Overall accuracy of aLLM screening results for CLL before and after review and refinement of the model

(True positive screening results)

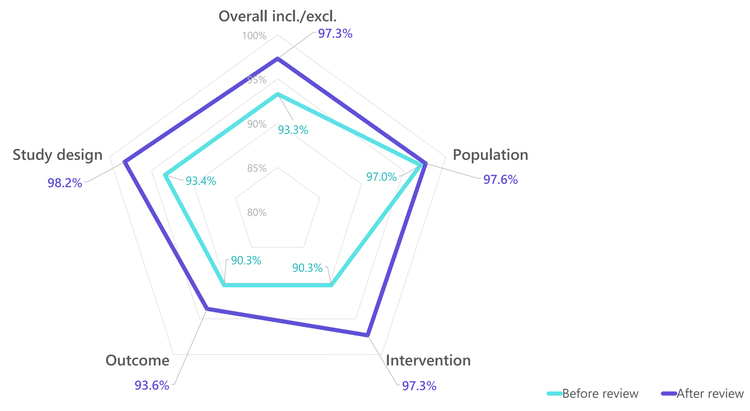
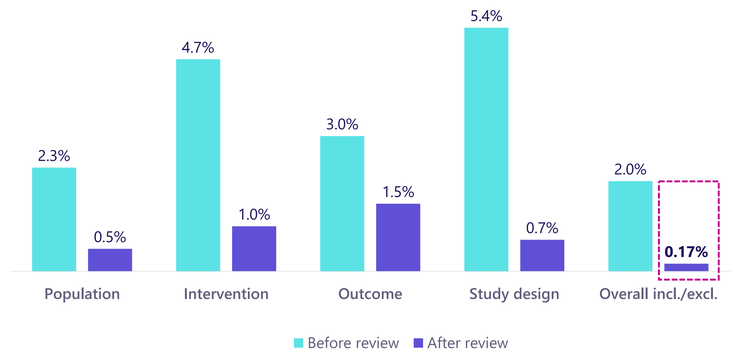


Figure 4. False negative aLLM screening results for CLL before and after review



CONCLUSIONS

- → Agentic LLM systems originally trained in solid tumors can be successfully adapted to hematological malignancies through disease-specific instruction and governance
- → This approach enables accurate, scalable, and real-time SLRs in MM and CLL, supporting living evidence generation for health economics and outcomes research (HEOR), health technology assessment (HTA), and clinical decision making

REFERENCES

1. Sarkisian S, Liu RJ, Liu E, Forsythe A. A living systematic literature review (L-SLR) for non–small-cell lung (NSCLC), prostate (PC), and breast cancer (BC), built with an agentic text annotation system powered by large language models (LLM) to assist treatment decision making [abstract]. *J Clin Oncol.* 2025;43:e13677. doi:10.1200/JCO.2025.43.16_suppl.e1365
2. Liu RJ, Forsythe A, Rege JM, Kaufman PA. Real-time clinical trial data library in non-small cell lung (NSCLC), prostate (PC), and breast cancer (BC) to support informed treatment decisions: now a reality with a fine-tuned large language model (LLM) [abstract]. *J Natl Compr Canc Netw.* 2025;23(3.5):BIO25-024. doi: 10.6004/jnccn.2024.7156

ONCOSCOPE AI
REAL-TIME ONCOLOGY LIBRARY

Presented at ISPOR 2026 | May 17-20, 2026 | Philadelphia, PA, USA | Poster Session 4

See more Oncoscope AI publications