MSR221

Validation of an Agentic Large Language Model (LLM) System in the Extraction Stage of a Real-Time AI-assisted Living Systematic Literature Review (REAL-SLR): a Solution to Instant and Easy access to Clinical Trial Data (CTD)

Rozee Liu[1], Rhiannon Campden[1], Eddie Xiaole Liu[2], Triston Grayston[3], Oscar Correa[3], Anna Forsythe[1]

[1]Oncoscope-AI, Miami, FL, USA; [2]Independent, Toronto, ON, Canada; [3]Eviviz Inc., Vancouver, BC, Canada

BACKGROUND

- Health economics and outcomes research (HEOR) professionals face challenges staying current with clinical trial data
- De novo systematic literature reviews (SLRs) are often time-intensive with tight deadlines and large corpus of publication hits from searches.
- Data extraction of various types are required for SLRs.

OBJECTIVES

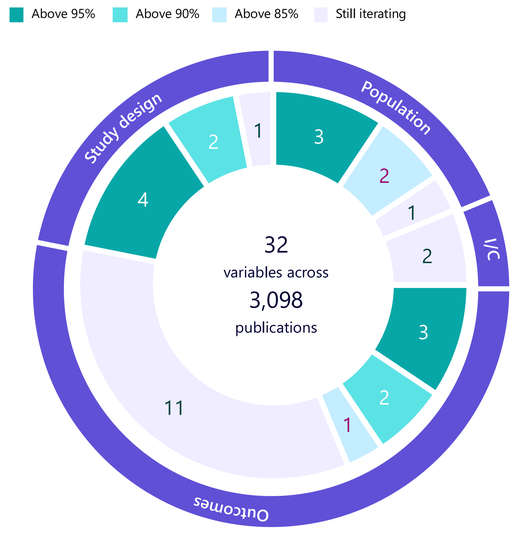
→ To evaluate the feasibility of using an agentic large language model (LLM) for clinical trial data, assessing extraction accuracy and potential time savings with a Real-Time AI-assisted Living Systematic Literature Review (REAL-SLR), following Cochrane's initiative of Adherence to the Responsible AI use in evidence SynthEsis (RAISE)[1]

CONCLUSIONS

→ The agentic LLM system with a RAG architecture demonstrated high accuracy in extracting publications

→ These findings suggest the system can enable real-time clinical data generation, supporting faster evidence development for HEOR decision-making and potentially improving patient access

Figure 1. Distribution of accuracy by PI/COS and accuracy threshold



I, intervention; C, comparator

METHODS

- An agentic large language model (LLM) system was developed to autonomously generate clinical trial data extraction annotations without human input
- The system combined multiple LLMs (OpenAI GPT-5 and GPT-4.1, Gemini 2.5 Pro, and Claude Sonnet 4.5) in a matrix of Cochrane-compliant[1] SLR processes (transparency, reproducibility, and research integrity)
- The agentic system was designed to emulate trained human reviewers by following a standardized annotation manual, decomposing tasks into subtasks, and recording reasoning for traceability
- In addition, a retrieval-augmented generation (RAG) architecture with semantic embeddings, parallel population, intervention/comparator, outcome, study design (PICOS)-aligned extraction chains, and study-type-adaptive prompting was implemented
- Annotations were generated for 32 extraction variables, and accuracy was evaluated against human annotations in publications for prostate (PC) and breast cancer (BC)

Table 1. Annotation variables and accuracies by PI/COS

Population	
Clinical stage	99.00%
Histology	96.00%
Sub-population	95.00%
Biomarker	85.00%
Risk	83.00%
Treatment path	Iterating
Intervention/Comparator	
Interventions	Iterating
Intervention category	Iterating
Study design	
Randomization	100.00%
Phase	100.00%
Name	95.18%
Registration	100.00%
Follow-up	94.44%
N	90.36%
Analysis type	Iterating
Outcomes - Efficacy	
OS measure	95.95%
Landmark survival	95.80%
Median survival	83.60%
Observation time	Iterating
Primary progression measure	Iterating
Median primary progression	Iterating
Landmark primary progression	Iterating
Other progression measures	Iterating
Median other progression	Iterating
Landmark other progression	Iterating
Response measured	Iterating
Response data	Iterating
Outcome – QOL/Subgroup/Safety	
QOL measure(s)	98.80%
Safety data	94.00%
QOL data	90.36%
Subgroups names	Iterating
Subgroup data	Iterating

Abbreviations: N, number; OS, overall survival; QOL, quality of life

Table 2. Overall survival (OS) data extraction formats

Annotation variable	Annotation format			
OS measure	Only abstracts with actual data reported are annotated with "OS" below in this column. Those that only mentioned "OS" as an endpoint (or other forms of mentioning) without data reported should be "NR"			
	RCT	**RCT, Cohorts**	**Non-RCT**	**Non-RCT, Cohorts**
OS median	"XX months vs. XX months; HR 0.XX; CI 0.XX-0.XX; p=0.XXX"	"Cohort 1: XX months vs. XX months; HR 0.XX; CI 0.XX-0.XX; p=0.XXX; Cohort 2: XX months vs. XX months; HR 0.XX; CI 0.XX-0.XX; p=0.XXX"	"XX months"	"XX months; XX months"
OS landmark	"XX% vs. XX%; p=0.XXX"	"Cohort 1: XX % vs. XX%; p=0.XXX; Cohort 2: XX % vs. XX%; p=0.XXX"	"XX%"	"XX months; XX months"

Abbreviations: CI, confidence interval; HR, hazard ratio; p, p-value; RCT, randomized controlled trial

RESULTS

- Our agentic LLM system generated annotations for 32 extraction variables for 3,098 (1,200 PC, 1,898 BC) publications
- Fourteen out of 32 variables achieved above 90% accuracy, 10/14 of which were above 95%. (**Figure 1, Table 1**)
- In addition to accuracy, our agentic system was able to produce annotations exactly as instructed
- As an example, overall survival (OS) extraction includes 3 variables: OS measure, median, landmark. Standardized annotation formats for these variables have been specified (**Table 2**)
- Only those matching the specified formats were considered a correct annotation. The accuracy of these 3 variables were 95.95%, 83.60%, and 95.80%

REFERENCES

1. Thomas J, Hair K, Noel-Storr, A, et al. Responsible use of AI in evidence SynthEsis (RAISE): recommendations for practice (version 3; updated 13 March 2026). In: Open Science Framework (https://osf.io/), Washington DC: Center for Open Science. DOI 10.17605/OSF.IO/FWAUD (accessed April 24 2026 via https://osf.io/cqa82)

ONCOSCOPE AI
REAL-TIME ONCOLOGY LIBRARY

Presented at ISPOR 2026 | May 17-20, 2026 | Philadelphia, PA, USA | Poster Session 5

See more Oncoscope AI publications