Traditional versus generative AI: A rapid systematic review assessing accuracy and efficiency of AI in title/abstract screening

Hardy E[a], Peddle A[a], Peatman J[a], Ross J[a], Lang S[a]
[a]Petauri Evidence, Bicester, UK

ISPOR poster acceptance code: MSR203
Abstract ID: 7890

Objectives

- Assess the accuracy and efficiency of artificial intelligence (AI) tools for title/abstract (t/a) screening for a systematic literature review (SLR) informing health economics and outcomes research (HEOR)
- Identify any factors influencing AI tool performance

Methodology

A detailed protocol was submitted to the Open Science Framework.

Electronic database searches were conducted in Embase® from inception to March 2025 and supplemented with desktop research.

Two independent reviewers screened all articles at title and abstract and at full-paper stages on Laser AI. One reviewer extracted data and a second checked the data.

AI platforms were categorised as generative large language models (LLMs) or commercially available tools (i.e. machine learning and natural language processing models). Prototype or in-house algorithms/models were excluded.

Quantitative data for measures of accuracy and efficiency were extracted and grouped by tool. A thematic analysis of factors influencing accuracy and efficiency was also conducted. No limitations were placed on measures of accuracy or efficiency eligible for extraction.

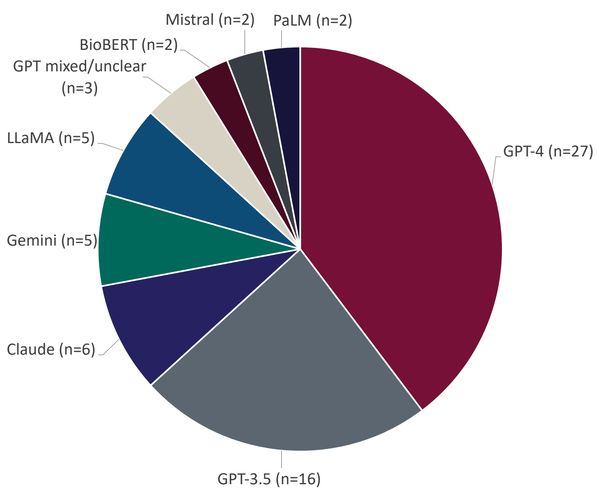
Results

A variety of LLMs and commercially available AI tools are available to support t/a screening.

A total of 88 records that reported accuracy and/or efficiency of AI tools for t/a screening were identified (46 tools, 41 LLMs; 1 both). 25 records compared >1 tool or LLM (one study directly compared LLMs versus tool).

Figure 1: Frequency of reporting of relevant data for LLMs (A) and commercially available tools (B) – >1 only

A



Generative Pre-trained Transformer (GPT)-4 models were the most frequently reported LLM tool, followed by GPT-3.5 (Figure 1A).

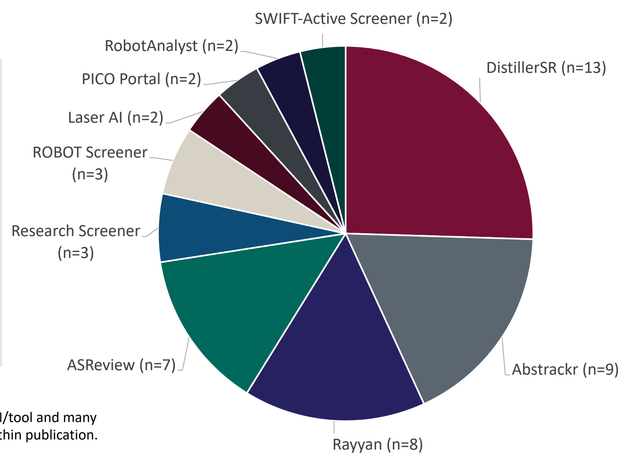
All other LLMs have been grouped by foundational model for the purposes of this summary.

1 study only: AI21 Ultra, Falcon, Flan-T5, GPT-3, Moonshot AI, OpenHermes-neural-chat, and SciFive.

B

DistillerSR, Abstrackr, Rayyan, and ASReview were the most frequently reported commercially available AI tools (Figure 1B).

1 study only: Colandr, EPPI-Reviewer, LiteRev, Nested Knowledge, SWIFT-Review, and SYMPRO.



Some records reported more than one LLM/tool and many reported more than one version of LLM within publication.
n = number of records.

Results continued

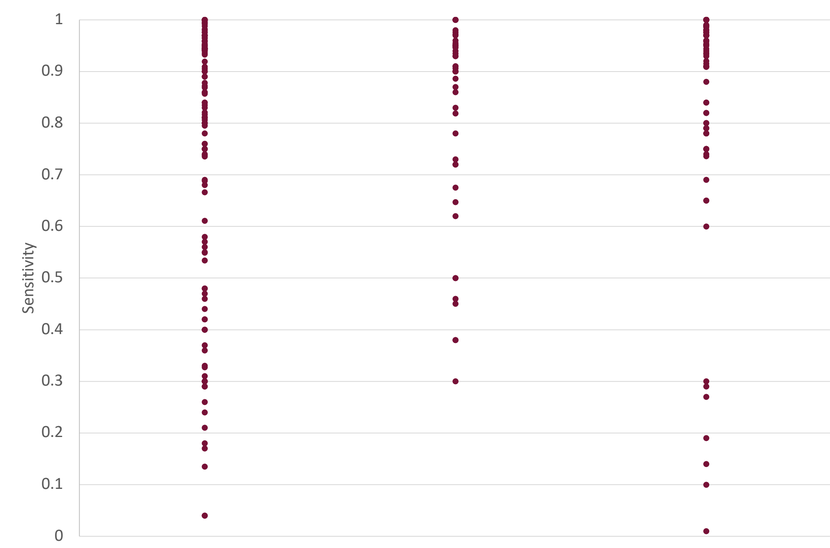
AI use cases and outcome definitions were highly heterogenous across publications. Time and cost savings were evident with use of AI, however the impact on accuracy was less consistent.

Accuracy (n=85: LLMs, n=42; tools, n=44)

To manage heterogeneity between studies, reported metrics were assigned to one of the following headings: accuracy, sensitivity/recall, specificity, precision, F1, and other.

Most frequently reported: sensitivity/ recall (n=57; Figure 2)

Figure 2: Range of sensitivity/recall values for t/a screening reported by AI platform type



Overall data prioritised if available. If not, data were presented as reported within publication (e.g. often several data points reported for one study). If required, data were converted to report as a proportion (0–1).

Efficiency (n=47: LLMs, n=15; tools, n=33)

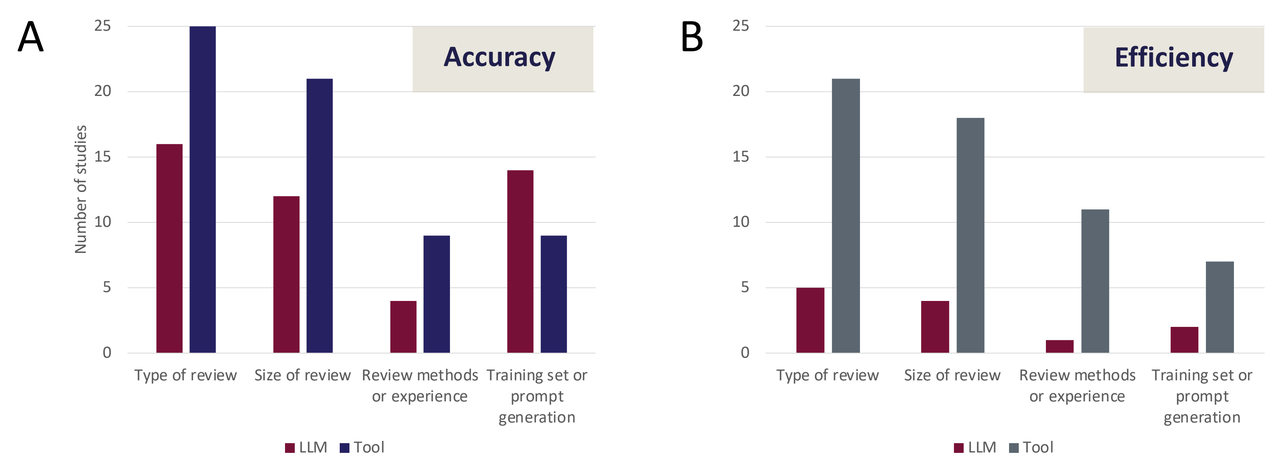
Reported metrics were assigned to one of the following headings: workload savings, time savings, and cost savings.

Most frequently reported: Time savings (n=33)

- Within aligned studies reporting % reduction in time (n=7), savings due to AI ranged from **36–91%**
- Time-based conclusions were inconsistently reported across publications (n=29)

Consistent factors modulating performance metrics for both LLMs and tools included review type/complexity, size, methodology, and training set/prompt generation methods.

Figure 3: Thematic analysis of reported factors influencing accuracy (A) or efficiency (B) of AI platforms for t/a screening



Studies comparing multiple LLMs or tools reported mixed conclusions regarding optimal tool selection.

Conclusion

Integration of AI technology into SLR workflows offers clear efficiency savings versus conventional methodology; however, the conclusions regarding comparative accuracy are less clear.

Researchers considering the use of AI should identify similar use cases and manage expectations based on potential modulating factors.

Clear guidelines on AI methodology across HEOR are essential to validate and quantify both accuracy and efficiency.

Scan for a video walkthrough

Scan for Protocol

Scan for PRISMA and included records

ISPOR Europe 2025
November 9–12
Glasgow, Scotland

Booth no. 608

Abbreviations

AI, artificial intelligence

GPT, Generative Pre-trained Transformer

HEOR, health economics and outcomes research

LLM, large language model

PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SLR, systematic literature review

t/a, title/abstract