SA20

Expediting Evidence Synthesis: A Review of Recent Research Evaluating Artificial Intelligence's Performance in Evidence Synthesis and Summarization of Clinical Literature

Fadi Manuel[1], Rachel Black[1], Tyler Reinsch[2], Jiawei Chen[1], Danny Yeh[1]

[1]AESARA Inc., Chapel Hill, NC, USA
[2]Arysana Inc., Chapel Hill, NC, USA

BACKGROUND

The annual rate of published research has grown steadily over the past decades leading to increasing volumes of available literature.[1]

Synthesis of clinical literature can often be a time-consuming and resource-intensive process due to the number of publications that need to be reviewed, aggregated, analyzed, and interpreted.[2]

Artificial intelligence (AI) can potentially expedite the evidence synthesis process for health economics and outcomes researchers and is gaining traction for this use.

It is unclear, however, which AI tools are being used and whether they effectively improve the efficiency and quality of evidence synthesis.

OBJECTIVE

To review recent literature evaluating the performance of AI and large language models (LLMs) in the evidence synthesis of clinical research

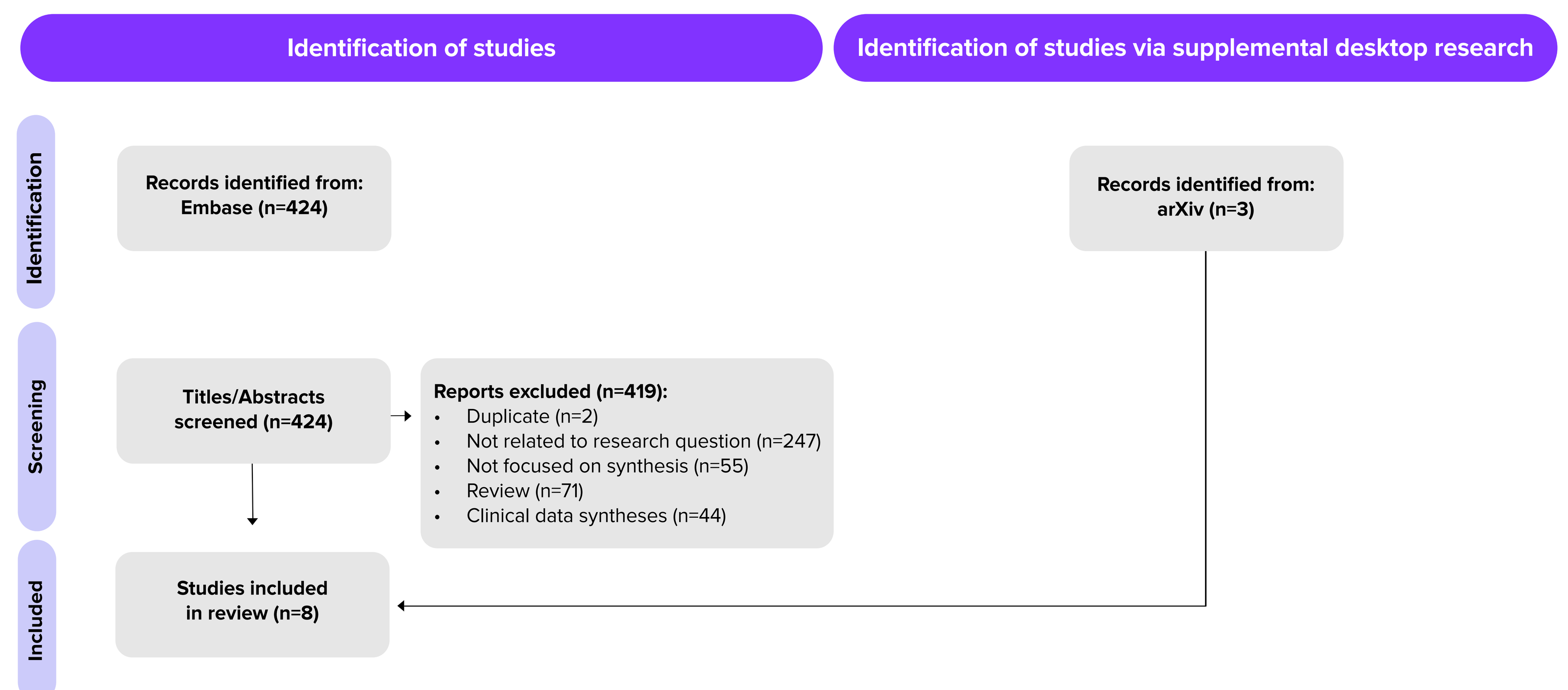
METHODS

A literature review was conducted in EMBASE for articles published since 2022 that describe the performance of LLM tools in clinical literature synthesis

Additional articles were identified through citation searching and supplemental desktop research on arXiv (Figure 1)

Key information was captured including the name of tools used, the type of evidence synthesized, and the methods for evaluating the tool's performance

Figure 1: Search Breakdown



RESULTS

A total of 8 studies were identified; the majority (n=6) focused on evidence summarization, while the remaining 2 evaluated the extraction of pre-specified data (Figure 2)

The clinical literature synthesized included systematic reviews (n=3), randomized controlled trials (n=3), clinical trial report (n=1), and review article (n=1) (Figure 3).

Figure 2: Type of evidence synthesis assessed

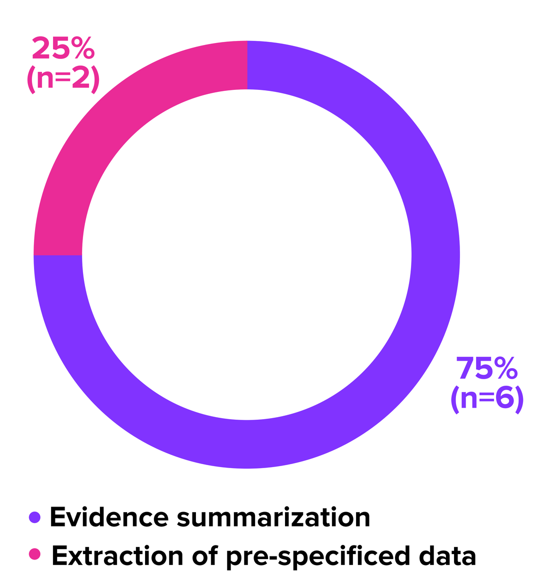
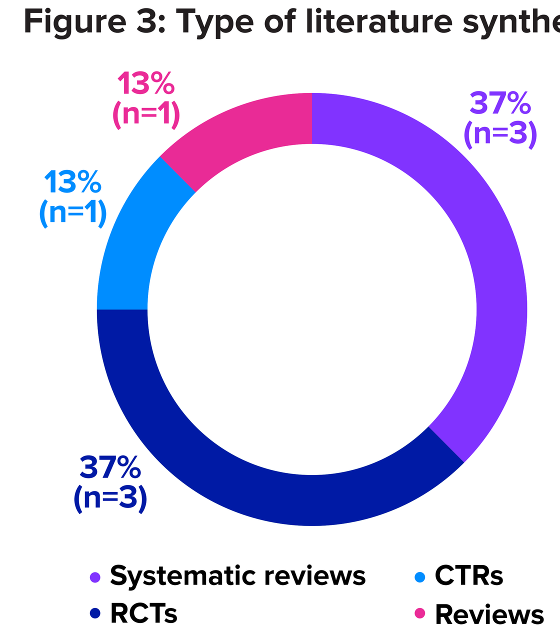
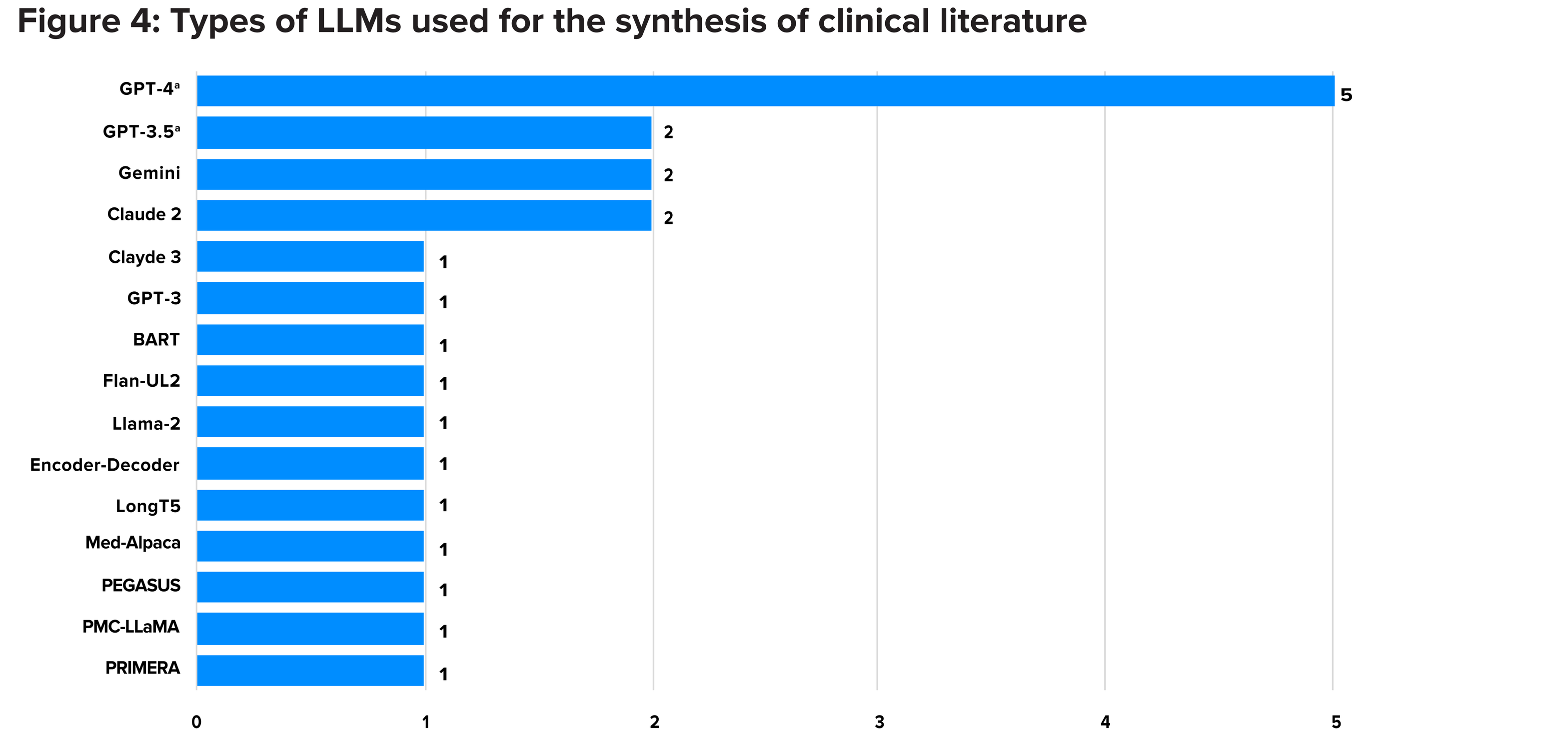


Figure 3: Type of literature synthesized



A total of 17 tools were identified, with some utilizing the same type of LLM (GPT-4), such as Ref AI,[3] ScholarAI,[3] and TrialMind.[4] GPT-4 was the most commonly used LLM (Figure 4)

Figure 4: Types of LLMs used for the synthesis of clinical literature



[a]Includes turbo version

Assessments focused on accuracy comprehensiveness missing data

Human evaluations of accuracy (n=4) resulted in the following scores (Table 1):
- Claude 2: 40%-97%
- GPT-4: 40%-87%
- Gemini: 44%-64%

Table 1: Overview of Assessment Methods

Publication	Tools assessed	Type of evidence synthesis assessed	Type of literature synthesized	Type of assessment	Results of human evaluation
Konet 2024[5]	• Claude 2 • GPT-4	Extraction of pre-specified data	RCT	Accuracy (human)	• Claude 2: 96.3% • GPT-4: 69%
Li 2024[3]	• Ref AI (GPT-4 Turbo • ChatGPT-4 • ScholarAI (GPT-4) • Gemini	Evidence summarization	Review	Accuracy, comprehensiveness, reference integration (human)	• RefAI: 83% • ChatGPT-4: 58.3% • ScholarAI: 62.7% • Gemini: 44%
Li 2024[6]	• GPT3.5-turbo • GPT-4 • Gemini-1.0-pro • Flan-UL2 • Med-Alpaca • PMC-LLaMA	Evidence summarization	RCT	Precision, recall, accuracy (ROUGE, BLEU, METEOR) Completeness, correctness, coherence (human)	**Correctness:** • GPT3.5-turbo: 69% • GPT-4: 72% • Gemini-1.0-pro: 64% • Flan-UL2: 44% • Med-Alpaca: 44% • PMC-LLaMA: 45%
McMinn 2023[7]	• Longformer-Encoder-Decoder • BART • PEGASUS	Evidence summarization	CTR	Accuracy (ROUGE, METEOR)	Not applicable
Shahib 2024[8]	• GPT3	Evidence summarization	RCT	Accuracy, coherence, usefulness (human)	Overall score was not generated for tool
Sun 2024[9]	• Claude 2 • GPT-4	Extraction of pre-specified data	Systematic reviews	Accuracy (human)	• Claude 2: 46-97% • GPT-4: 40-87%
Wang 2024[4]	• TrialMind (GPT4+-Claude3.5) • GPT-4	Evidence summarization	Systematic reviews	Accuracy (human)	**Win rate[a] (across 5 studies)** • TrialMind: 62.5%-100% • GPT-4: 0%-37.5%
Zhang 2024[10]	• PRIMERA • LongT5 • Llama-2 • GPT-3.5	Evidence summarization	Systematic reviews	Precision, recall, accuracy (ROUGE-L, METEOR) Consistency, comprehensiveness, specificity, readability (human)	**Win rate[b]** • LongT5: 60% • Llama-2: 59% • PRIMERA: 55% • GPT-3.5: Not reported

[a]Ratio of summaries evaluated as better for each respective tool
[b]Ratio of machine-generated summaries evaluated as better than the baseline

DISCUSSION

The use of AI will continue to revolutionize how research is conducted, resulting in increased efficiency. Human oversight remains vital for validation and addressing errors.

Potential time savings can lead to cost reduction that can provide additional support for other activities to bolster a biopharmaceutical product's value story.[11]

Due to variations in assessment methodologies across studies, no single LLM tool was identified as the definitive choice for evidence synthesis. However, as AI tools continue to evolve, improvements in accuracy and completeness are anticipated in the future.

CONCLUSION

In this review, GPT-4 was the most commonly tested tool. Future assessments should also quantify the potential time and cost saving through AI.

Numerous methods of assessment were observed, highlighting the need for a standardized assessment checklist to appropriately appraise the performance of LLM tools in evidence synthesis.

With the release of newer models (eg, GPT-4.5, Claude 3.7), continued assessment of AI tools will be essential to determine the feasibility of broader use in research.

REFERENCES

1. Zhao X, Jiang H, Yin J, et al. Changing trends in clinical research literature on PubMed database from 1991 to 2020. Eur J Med Res. 2022;27(1):95. Published 2022 Jun 20. doi:10.1186/s40001-022-00717-9
2. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91-108. doi:10.1111/j.1471-1842.2009.00848.x
3. Konet A, Thomas I, Gartlehner G, et al. Performance of two large language models for data extraction in evidence synthesis. Res Synth Methods. 2024;15(5):818-824. doi:10.1002/jrsm.1732
4. Li Y, Zhao J, Li M, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. J Am Med Inform Assoc. 2024;31(9):2030-2039. doi:10.1093/jamia/ocae129
5. Li J, Deng Y, Sun Q, et al. Benchmarking Large Language Models in Evidence-Based Medicine. IEEE J Biomed Health Inform. Published online October 21, 2024. doi:10.1109/JBHI.2024.3483816
6. McMinn D. Original abstracts from the 2023 European Meeting of ISMPP. Curr Med Res Opin. 2023;39(suppl 1):41. doi: 10.1080/03007995.2023.2184562
7. Shaib C, Li ML, Joseph S, Marshall IJ, Li JJ, Wallace BC. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). arXiv. Preprint posted online May 11, 2023. doi.org/10.48550/arXiv.2305.06299
8. Sun Z, Zhang R, Doi SA, et al. How good are large language models for automated data extraction from randomized trials? medRxiv. 2024:2024.02.20.24303083. doi.org/10.1101/2024.02.20.24303083
9. Wang Z, Cao L, Danek B, Jin Q, Lu Z, Sun J. Accelerating clinical evidence synthesis with large language models. arXiv. Preprint posted online October 28, 2024. doi.org/10.48550/arXiv.2406.17755
10. Zhang G, Jin Q, Zhou Y, et al. Closing the gap between open source and commercial large language models for medical evidence summarization. npj Digital Medicine. 2024/09/09 2024;7(1):239. doi.org/10.1038/s41746-024-01239-w
11. Naylor NR, Hummel N, de Moor C, Kadambi A. Potential Meets Practicality: AI's Current Impact on the Evidence Generation and Synthesis Pipeline in Health Economics. Clin Transl Sci. 2025;18(4):e70206. doi:10.1111/cts.70206

ABBREVIATIONS IN TABLES AND FIGURES

AI, artificial intelligence; CTR, clinical trial report; LLM, large language model; RCT, randomized controlled trial.

ACKNOWLEDGEMENT

Emma Edwards developed the graphics for this poster.

CONTACT INFORMATION

Fadi Manuel
Manager, Value Evidence, AESARA
E-mail: fadi.manuel@aesara.com

Presented at: ISPOR International Conference, May 13-16, 2025, Montreal, Quebec, CA

Download poster here

aesara.com