Performance of large language model clinical data extraction by data domain: A rapid systematic review

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

ISPOR poster acceptance code: SA75
Abstract ID: 8028

Background

Large language models (LLMs) have the potential to increase efficiencies relative to manually conducted systematic reviews; however, caution remains to ensure gold standards are not compromised (1-4).

Objectives

- Collate studies reporting LLM data extraction of clinical publications
- Explore performance of LLM extraction according to data domain
- Identify any factors influencing LLM extraction performance

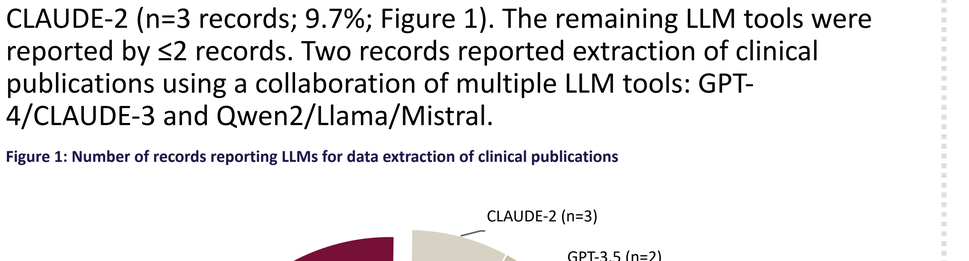
Methods

A rapid systematic review was conducted to identify records reporting LLM data extraction of clinical publications. To be considered for inclusion, clear reporting of the LLM used to perform data extraction was required. A two-step approach was then used for subsequent analyses.

Protocol

Step 1: Map LLMs for all included records

Step 2: Records reporting quantitative LLM extraction performance by data domain were eligible for further analysis. The data domains of interest were based on those commonly extracted from clinical publications:



Results

PRISMA and included records

A total of 31 records that reported data extraction of clinical publications using an identifiable LLM were identified.

Of these included records, 15 reported the performance of LLM data extraction by data domain.

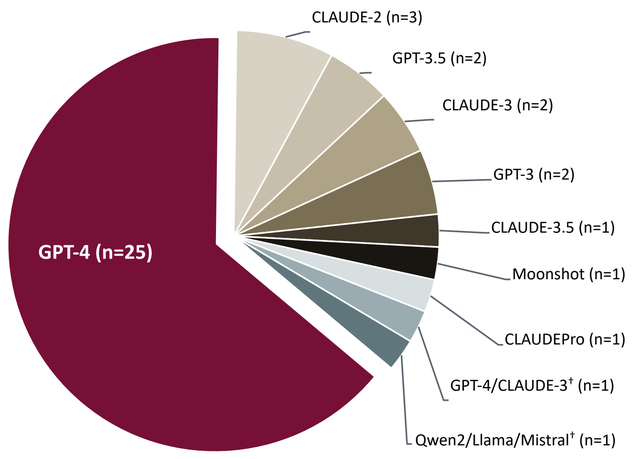
A variety of LLM tools were identified for data extraction

31 records

Across the 31 records, 10 different LLM tools were identified.

Generative Pre-trained Transformer (GPT)-4 models were the most commonly reported LLM tool (n=25 records; 80.6%), followed by CLAUDE-2 (n=3 records; 9.7%; Figure 1). The remaining LLM tools were reported by ≤2 records. Two records reported extraction of clinical publications using a collaboration of multiple LLM tools: GPT-4/CLAUDE-3 and Qwen2/Llama/Mistral.

Figure 1: Number of records reporting LLMs for data extraction of clinical publications

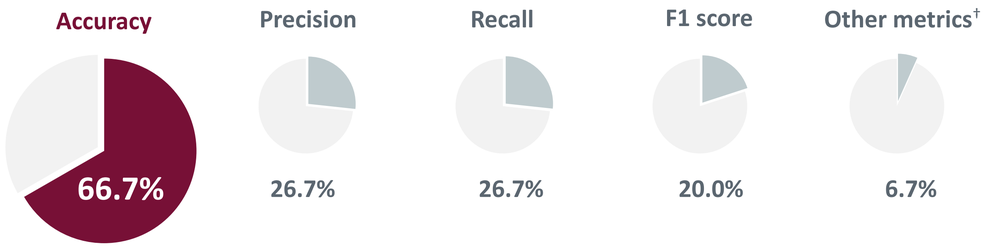


Some records reported more than one LLM tool.
LLM tools were grouped by top-line model to address heterogenous reporting of LLM versions: GPT-4 (GPT-4 Turbo, GPT-4 Omni, GPT-4 Vision, GPT-4); CLAUDE-2 (version not reported); GPT-3.5 (GPT-3.5 Turbo); CLAUDE-3 (CLAUDE-3 Sonnet, CLAUDE-3 Opus); GPT-3 (version not reported); CLAUDE-3.5 (CLAUDE-3.5 Sonnet); Moonshot (Moonshot v1 128k); CLAUDEPro (CLAUDE-3 Opus + CLAUDE-3.5 Sonnet); GPT-4/CLAUDE-3 (GPT-4 Turbo/CLAUDE-3 Opus); Qwen2/Llama/Mistral (versions not reported).
† Multiple LLM approach.
n = number of records.

Heterogenous reporting metrics for LLM performance

15 records

Figure 2: Reporting metrics for LLM extraction performance across records reporting performance by data domain (n=15)



Some records reported more than one metric for measuring LLM performance.
† Gwet's AC2, categorical decision match (e.g. 'good', 'poor', 'no match'), inaccuracy, error rate.
Displays % of records (n=15).

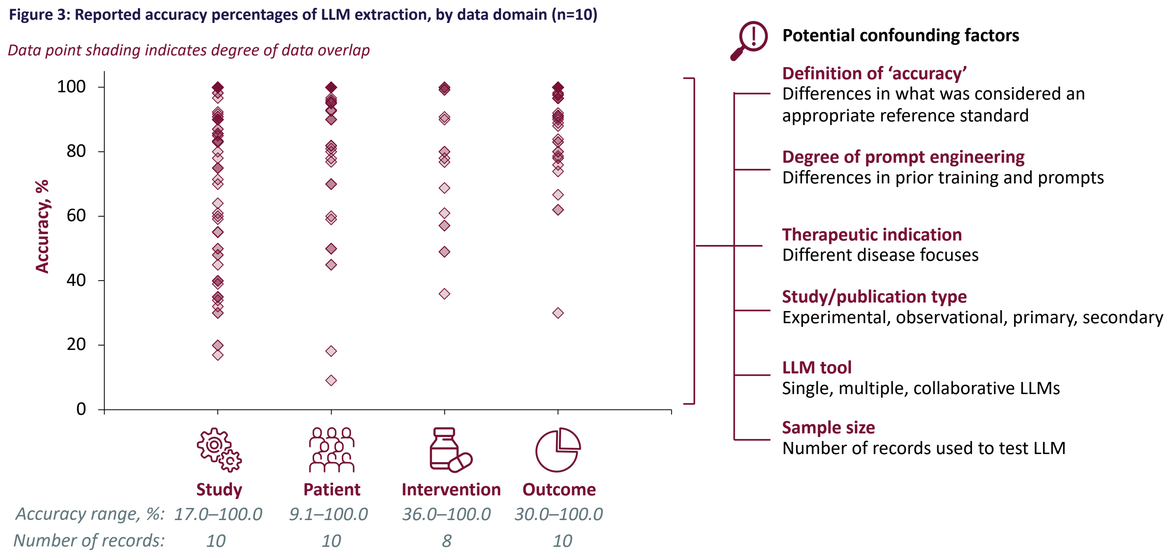
Accuracy was the most common reporting metric (Figure 2), often measured as a "decision match percentage" with a human reviewer.

Reported accuracy was highly variable by data domain

10 records

Reported accuracy varied greatly for each data domain (Figure 3). Records differed in relation to various factors, which can be considered potential confounding factors.

Figure 3: Reported accuracy percentages of LLM extraction, by data domain (n=10)



Records reported accuracy for various data types, which were then grouped according to study, patient, intervention, and outcome.

Data type/format, publication language, and use of single or multiple LLMs can impact LLM extraction performance

15 records

Key factors impacting LLM extraction performance were identified through a thematic analysis (Table 1).

Table 1: Thematic analysis – Reported factors impacting LLM performance, by LLM tool†

LLM	Key factors impacting LLM performance			
	Data type	Data format	Publication language	Single versus multiple LLM
GPT-4	NR	Free text had better accuracy than table or figure text	NR	NR
GPT-3	NR	Free text had better accuracy than table text	NR	NR
CLAUDEPro	String data had better accuracy/recall than numerical/mixed data	NR	NR	NR
CLAUDE-3.5	NR	NR	English language performed better than Chinese language	NR
GPT-4/CLAUDE-3	NR	NR	NR	Multiple LLMs performed better than individual LLMs

† Thematic analysis was conducted across the 15 records included for further analysis to identify key factors reported to impact LLM performance. Not all records provided insights into factors impacting LLM performance. As the table only presents LLM tools with reported impacting factors, not all records (and associated LLM tools) are represented in the table.

Conclusions

- GPT-4 models were the most commonly reported LLM tool for data extraction of clinical publications
- Substantial heterogeneity exists in the reporting of LLM extraction performance
- LLM data extraction has potential for high accuracy across all four data domains; however, no data domain was found to show reliable accuracy rates
- Standardisation in assessing LLM data extraction performance, and full transparency in its reporting, is required to support future research and guide implementation in evidence synthesis processes

Scan for a video walkthrough

ISPOR Europe 2025
9–12 November
Glasgow, Scotland

Booth no. 608

References

1. Lieberum JL, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use–a scoping review. J Clin Epidemiol. 2025;181:111746.
2. Luo X, et al. Potential roles of large language models in the production of systematic reviews and meta-analyses. J Med Internet Res. 2024;26:e56780.
3. Yun HS, Marshall IJ, Trikalinos TA, Wallace BC. Appraising the potential uses and harms of LLMs for medical systematic reviews. arXiv preprint arXiv:2305.11828. 2023.
4. Forero DA, Abreu SE, Tovar BE, Oermann MH. Large Language Models and the Analyses of Adherence to Reporting Guidelines in Systematic Reviews and Overviews of Reviews (PRISMA 2020 and PRIOR). J Med Syst. 2025;49(1):80.

Abbreviations

GPT, Generative Pre-trained Transformer

LLM, large language model

NR, not reported

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