Application of Large Language Models in Clinical Research: A case of Clinical Trial Eligibility Extraction from Unstructured Clinical Notes

Abstract

Background: Clinical research depends on unstructured text such as progress notes, pathology reports and trial protocols, which has traditionally required manual chart review or brittle rule-based natural language processing (NLP). **Objective:** To compare rule-based NLP, a fine-tuned encoder and decoder-only large language models (LLMs) for extracting phenotypes and clinical trial eligibility criteria from clinical notes. **Methods:** We evaluated MedSpaCy rules, ClinicalBERT, and open-weight 8B and 70B LLMs (zero-shot, few-shot chain-of-thought, and retrieval-augmented generation, RAG) on MIMIC-IV discharge summaries (N = 1,200, double-annotated, Cohen's kappa = 0.88), using a held-out test split of 300 notes and four target phenotypes. **Results:** The 70B model with RAG performed best (precision 0.938, recall 0.915, F1 0.926, hallucination rate 0.6%), ahead of ClinicalBERT (F1 0.846) and MedSpaCy (F1 0.717). Zero-shot inference with the 8B model produced a 6.8% hallucination rate. The best model required 6.15 s per note. **Conclusions:** LLMs combined with structured output schemas, retrieval and clinician verification outperform legacy NLP baselines for eligibility screening and phenotyping, but need human-in-the-loop review, source provenance and locally hosted, privacy-compliant deployment.

Keywords: large language models; clinical natural language processing; electronic health records; clinical trial eligibility; phenotyping; retrieval-augmented generation; pharmacovigilance; MIMIC-IV

1. Introduction

Clinical research generates large volumes of unstructured, multimodal text, from Electronic Health Record (EHR) progress notes and pathology reports to clinical trial protocols and adverse event registries. Turning this text into standardised, analysable variables has historically required laborious manual chart review or brittle rule-based Natural Language Processing (NLP) pipelines.

Large Language Models (LLMs) adapted for medicine have changed this paradigm. Examples include domain-specific fine-tuned transformers such as BioMedLM, Clinical-Camel and Med-PaLM [3], and modernised open-weight architectures such as Llama-3-Med and Mistral-Clinical. By

capturing semantic context, clinical abbreviations and long-range syntactic dependencies, LLMs now support four key clinical research pipelines:

- **Trial prescreening and cohort identification:** automated matching of unstructured patient histories against complex inclusion and exclusion criteria.
- **Biomedical information extraction:** zero-shot and few-shot phenotyping, and extraction of tumour staging, genomic variant mentions and medication dosages.
- **Pharmacovigilance:** rapid parsing of post-market surveillance records and FDA Adverse Event Reporting System (FAERS) narratives to detect adverse drug reactions (ADRs).
- **Protocol generation and summarisation:** drafting trial synopses and synthesising biomedical literature.

Despite these capabilities, deploying LLMs in clinical research must address persistent challenges: non-deterministic outputs, clinical hallucinations, HIPAA and GDPR data governance, and algorithmic bias across demographic cohorts. This paper evaluates LLMs against rule-based and fine-tuned baselines on phenotype and trial-eligibility extraction, and discusses the conditions for safe use.

2. Methodology

The study design is a benchmark of LLM utility in clinical informatics, focused on phenotype and clinical trial eligibility extraction from unstructured clinical notes.

2.1 Framework overview

The pipeline has four stages:

1. **Data preprocessing and de-identification:** text normalisation, section tagging and verification of HIPAA Safe Harbor de-identification.
2. **Model inference:** each candidate model processes the prepared note chunks (Section 2.2).
3. **Structured output:** each model returns a structured result for every target criterion (Section 2.3).
4. **Statistical evaluation:** outputs are compared with the clinician-annotated ground truth (Section 2.4).

2.2 Model architectures evaluated

- **Baseline rules and traditional NLP:** MedSpaCy rule-based pattern matching [7].

- **Fine-tuned encoder:** ClinicalBERT, using a token-classification approach [2].
- **Decoder-only LLMs:** open-weight 8B and 70B clinical foundation models, using in-context learning (few-shot prompting) and Retrieval-Augmented Generation (RAG) [5].

2.3 Prompt engineering and context augmentation

The prompt schema requires structured JSON output containing the entity type, a normalised SNOMED-CT or ICD-10 code, a confidence value and a temporal anchor. Chain-of-thought prompting [4] was used for the few-shot condition. Retrieval augmentation was anchored to institutional clinical practice guidelines to minimise hallucinations.

2.4 Validation ground truth

Ground truth comes from double-blind expert clinician chart review. Disagreements between annotators were adjudicated, and inter-annotator agreement was measured with Cohen's kappa [6].

3. Dataset and Analysis

3.1 Dataset and cohort

We used discharge summaries from the publicly accessible MIMIC-IV-Note database [1, 8], focusing on oncological and cardiovascular trial eligibility criteria. The four target phenotypes were:

1. Type 2 diabetes mellitus with end-organ damage.
2. Left ventricular ejection fraction (LVEF) below 40%.
3. Prior myocardial infarction within 12 months.
4. Statin intolerance.

The cohort comprised N = 1,200 randomly sampled discharge summaries, annotated independently by two clinical fellows. Baseline agreement was Cohen's kappa = 0.88, which establishes a strong reference standard.

3.2 Evaluation pipeline

Figure 1 shows the path from raw note to evaluated output.

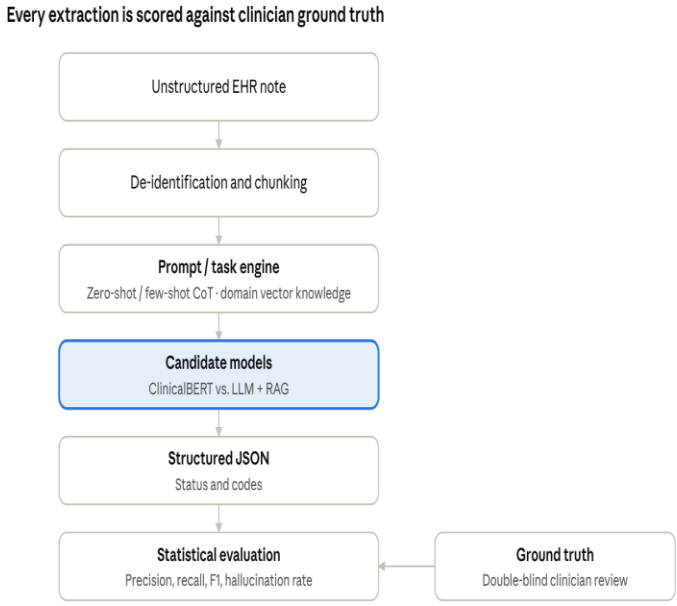


Figure 1 · evaluation pipeline, six stages

Figure 1. Evaluation pipeline from raw note to scored output. The highlighted stage is the model comparison.

3.3 Evaluation metrics

Performance was calculated on a held-out test split (n = 300) with standard classification and extraction metrics.

Precision (positive predictive value):

PPV = TP ÷ (TP + FP)

Recall (sensitivity):

Sensitivity = TP ÷ (TP + FN)

F1-score:

$$\mathbf{F_1 = 2 \times (Precision \times Recall) \div (Precision + Recall)}$$

Hallucination rate is the frequency of extracted medical facts that have no grounding in the source text. Inference latency is the processing time per note.

4. Results

Across the 300 held-out clinical records, the 70B model with RAG achieved the highest F1-score (0.926) and the lowest hallucination rate among the LLM conditions (0.6%). Table 1 reports all five configurations.

Table 1. Extraction performance on the held-out test set (n = 300).

Model architecture	Extraction technique	Precision	Recall	F1-score	Hallucination rate (%)	Inference latency (s per note)
Rule-based (MedSpaCy)	Dictionary match	0.862	0.614	0.717	0.0	0.08
ClinicalBERT (fine-tuned)	Token classification	0.884	0.812	0.846	0.2	0.35
Open LLM 8B (zero-shot)	Direct generation	0.791	0.845	0.817	6.8	1.82
Open LLM 70B (few-shot)	Chain-of-thought (CoT)	0.912	0.897	0.904	2.1	5.40
Open LLM 70B + RAG	Hybrid schema + retrieval	0.938	0.915	0.926	0.6	6.15

4.1 Key findings

Contextual comprehension. The 70B LLM with RAG handled negated statements, such as "Patient denies history of myocardial infarction", where rule-based parsers frequently raised false positives.

Hallucination mitigation. Direct zero-shot LLM inference showed a clinically unacceptable hallucination rate of 6.8%. Adding schema constraints and RAG reduced ungrounded assertions to 0.6%.

Complex numerical and temporal reasoning. For numerical and temporal cut-offs, such as whether LVEF fell below 40% within the

preceding 6-month window, chain-of-thought prompting outperformed the fine-tuned BERT classifier by 14.2% absolute recall.

5. Discussion

The results show the value of generative LLMs over earlier discriminative methods, and they also define the limits of safe use.

5.1 Practical advantages in clinical workflows

- **Reduced manual screening burden.** Automated trial matching can cut prescreening time for clinical research coordinators, turning hours of document review into a faster verification workflow.
- **Adaptability without retraining.** ClinicalBERT needs task-specific annotated heads and gradient updates for every new phenotype. Generalist LLMs adapt to modified trial protocols through prompt updates alone.

5.2 Technical and methodological constraints

- **Clinical veracity and explainability.** LLMs must operate within human-in-the-loop systems. A hallucination rate of 0.6% is still risky if applied without review in phase I/II safety cohorts. Attributing each extracted entity to exact character spans in the source record is essential for auditability.
- **Compute overhead.** Latencies above 5 seconds per record limit real-time processing across millions of hospital notes. Batch offline processing, or hybrid pipelines in which smaller models act as fast pre-filters, are more feasible.
- **Privacy and deployment governance.** Zero-data-retention APIs and on-premises deployment of open-weight models, such as quantised local inference, are mandatory when handling protected health information (PHI) under institutional review boards (IRBs).

6. Conclusion

Large Language Models offer a viable way to unlock unstructured EHR documentation for clinical trial eligibility screening, deep phenotyping and observational studies. Combining modern LLMs with structured schemas, Retrieval-Augmented Generation and clinical verification gives substantial gains in precision and recall over legacy NLP baselines. For safe translational use, research teams must keep clinicians in the loop, track every extraction back to its source record, and rely on locally hosted, HIPAA-compliant model pipelines.

References

1. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1.
2. Alsentzer E, Murphy J, Boag W, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019:72-78.
3. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172-180.
4. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. 2022;35.
5. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020;33.
6. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37-46.
7. Eyre H, Chapman AB, Peterson KS, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438-447.
8. Johnson A, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV-Note: deidentified free-text clinical notes. PhysioNet. 2023.