A Scalable Method for Validated Data Extraction from Electronic Health Records with Large Language Models

Timothy J. Stuhlmiller, AJ Rabe, Jeff Rapp, Penelope Manasco, Alaa Awawda, Hiba Kouser, Hugh Salamon, Donald Chuyka, William Mahoney, Kenny K. Wong, Glenn A. Kramer, Mark A. Shapiro

xCures, Inc. Oakland, CA, USA

Abstract

Purpose: Extracting and structuring relevant clinical information from electronic health records (EHRs) remains a challenge due to the heterogeneity of systems, documents, and documentation practices. Large Language Models (LLMs) provide an approach to processing semi-structured and unstructured EHR data, enabling classification, extraction, and standardization.

Methods: Medical documents are processed through a structured data pipeline to generate normalized FHIR data. Unstructured data undergoes preprocessing, including optical character recognition, document parsing, text chunking and embedding. Embedding enables search and classification which facilitate document retrieval for extraction. LLMs perform named entity recognition and relation extraction, with outputs mapped to FHIR R4 and OMOP and harmonized with pre-structured data for interoperability. Model performance is evaluated through human validation and automated consistency checks. Iterative refinement, error analysis, and standardized schema selection optimize use for analytics and downstream workflows.

Results: An LLM-based schema for medication extraction, validated on 34 patients, achieved ~95% accuracy and F1-score across 7 data fields for 5,789 extracted medications. Deployment across 11,115 patients extracted 2.6 million medication records, increasing total medications by 27% and distinct drug ingredients by 31% over structured data in the EHR. LLM extraction increased oncology medications by 60%, distinct oncology therapies by 64%, and the number of patients with structured oncology medication data by 33%. The LLM enhanced data completeness, improving availability of indication for prescription (61% vs. 31%) and discontinuation reason (17% vs. 0%), outperforming pre-structured data in key clinical variables.

Conclusion: An LLM-powered extraction process that employs embeddings, machine learning classification, schema-based extraction, and mapping of extracted information to healthcare data standards, achieves a significant gain in clinically relevant information over pre-structured data available in the EHR.

Introduction

Healthcare data interoperability has increased the sharing of clinical information, but due to the diversity of systems and lack of standardized documentation processes, finding key information remains challenging at the point-of-care. AI solutions may help clinicians quickly find accurate information, but only if they are able to quickly and accurately parse large amounts of records from disparate sources. To this end, we have developed a scalable pipeline to assist clinicians with information retrieval tasks.

Patient medical information is scattered across numerous systems from different practices and institutions. This makes information retrieval challenging even when one can access records across different EHR systems. Furthermore, freetext medical notes are designed to facilitate clinical communication, while the structured data in an EHR is designed for billing purposes. Thus, accessing information in notes is essential for continuity of clinical care.

However, clinical notes are semantically and linguistically complex. They include orders which may or may not have been filled, specialized forms of negation, and extensive use of context-dependent abbreviations, which have all proven to be challenging for natural language processing. To overcome this, a wide variety of methods have been tested in academic settings, with rule-based and statistical methods, such as conditional random fields, as well as hybrid and machine-learning methods all able to achieve good performance.[1,2] However, rules-based systems are difficult to develop and can lack real-world robustness. Machine-learning methods have suffered from limited public datasets upon which to train and the cost and expertise required to create labeled training data. Notably, those systems fall well short of the 95% accuracy and F1-scores proposed for clinical use following the original I2B2 challenge.[2] More recently, specialized LLMs have shown strong abilities on these and more complex extraction tasks.[3] Here, we present a highly scalable solution using general purpose LLMs for accurate clinical text classification and structured data extraction that can be implemented and adapted quickly, achieves targeted performance, enables efficient and effective data retrieval, and integrates with standard healthcare data models.

Methods

Data Collection and Preprocessing

Patient documents can be received in various forms such as image files (.pdf, .tif, .jpg, among others), Fast Healthcare Interoperability Resources (FHIR), or Consolidated Clinical Document Architecture (C-CDA) files. C-CDA documents are xml format but contain instructions for structured (tabular) data, as well as semi-structured data, such as free-text notes or note sections, and unstructured data, such as embedded image

files. We observed that embedded images account for up to 32% of C-CDA data received from EHRs. Structured C-CDA data is converted to FHIR R4,[4] preserving tabular structure and coding before running additional coding checks and concept normalization, which does not require LLM use. Semi-structured and unstructured data undergo preprocessing before being routed to LLM-based extractors.

Preprocessing includes document parsing, optical character recognition (OCR) for image files, and chunking to break text into relatively homogenous sections. Standardized metadata extraction is performed for C-CDA files, ensuring interoperability with existing clinical data structures. Metadata may be used to inform and improve downstream tasks such as classification, disambiguation, or mapping.

Embedding Model and Text Chunking

To facilitate semantic search, document retrieval, and text classification, documents are divided into sections based on sections within the C-CDA format, headings within those sections, and for long sections within headings, are further chunked based on length and optionally content types. Chunking strategies include splitting by paragraph, sentence, or predefined section headings, while preserving table structures. For PDF or other image files, chunking is initially performed by page, with splitting strategies described above implemented to facilitate retrieval and classification while maintaining context.

A text embedding model (currently OpenAI text-embedding-3-small) generates a vector representation for each chunk, capturing semantic content.[5] In a sample of 41,900 patient EHR records, there were an average of 231 documents per patient. Our chunking strategy resulted in an average of 34.6 chunks per document, or 7,986 chunks stored as embedding vectors per patient. Thus, for this sample, about 335 million vectors were produced for classification and retrieval tasks.

Classification and Semantic Search

Across the tens of millions of documents received from various sources, document labels are highly diverse and reflect different EHR developer or implementer choices. For example, “Plan of Care”, “Plan of Treatment”, and “Selected Encounter” all appear to be different labels given to the same type of information in different systems. Separately, most documents contain a “Results” section which contains important findings from that encounter, but that subheading is not useful for identifying the clinical context of the underlying data. Unsupervised learning (using k-means clustering and hdbscan) with silhouette analyses suggested an optimal number of document classes at 26, 75, or 152 clusters. Based on inspection of documents, we chose the smaller number of classes and implemented classification of chunks or sections into 26 predefined information types, such as Diagnosis, History of Present Illness, Plan of Care, Social History, Genomics, and Medications, among others. An additional

document axis of therapeutic topic best explained the fit for larger numbers of document types (i.e., Clinic Note may exist as many defined subtypes like Radiology Clinic Note, Oncology Clinic Note, Surgery Clinic Note, etc.).

These labels were used to develop a document classification algorithm with supervised machine-learning (random forest) applied to the embedding vectors following a standard train-test-validate random-split approach using Python's skikit-learn package. Human review of the structured source labels confirmed the accuracy of the training, testing, and validation datasets. Using the embedding vectors from standardized section labels in C-CDA to train the model allows chunks from unstructured data, such as image files, to be mapped to the same rubric as the structured EHR data. Before deployment, micro F1 scores were calculated on the validation set of 21,837 records (Table 1).

Document and Schema Selection Logic for Extraction

Extracting clinical concepts from unstructured or semi-structured text involves two main steps: identifying relevant documents and applying task-specific extraction prompts.

Extraction prompts include instructions for structured outputs provided to LLMs to ensure alignment with a predefined format. Due to limited attention, LLMs perform best when tasked with extracting a focused set of related concepts. For instance, extracting medication details—such as name, dose, route, regimen, indication, and reason for discontinuation—into one schema enhances accuracy.

Semi-structured and unstructured file sections are selected for extraction based on semantic search using schema-specific terms. Extraction schemas are then applied based on the retrieved document classifications. For C-CDA documents, this involves using section headers and the random forest embeddings classifier to apply the appropriate extraction schema or schemas, since the information may be homogenous (Administered Medications, Family Medical History) or heterogenous, requiring multiple passes with different extraction schemas (Assessment and Plan, Clinic Note). For unstructured OCR documents, schema selection relies exclusively on the embeddings classifier (Table 1) since these documents lack standardized metadata. This approach ensures that extraction logic is tailored to document content, optimizing for accuracy and consistency while minimizing hallucinations.

LLM Training and Iteration

It has been observed that LLMs are effective zero-shot or one-shot clinical information extractors.[6] In our experience, highly accurate extraction with LLMs is possible with zero-shot prompting, so long as task-specific prompts and schemas are well designed, but greater accuracy is achieved with some in-context learning. Accordingly, we accumulate data from manually annotated clinical documents and synthetic examples generated to improve model generalization within prompts. Aggregated collection of erroneous extractions is very important to identify and improve upon weaknesses in the

overall system, which can be due to retrieval (conflict between document type and extract task) or extraction (extracted information conflicts with human expectation). A/B testing is conducted to evaluate different prompts, different LLMs, and hyperparameters to maximize accuracy and reproducibility.

Consistency checks are then implemented to verify that extractions remain stable across different model runs. Discrepancies trigger further investigation and model refinement, with version control allowing for model roll-back.

Error Analysis and Model Refinement

Recurring extraction errors are identified through a combination of automated error detection and manual review. Common failure modes include incorrect entity linking, misclassification of clinical terms, and missed extractions. Error analysis informs prompt modifications, feature engineering adjustments, and dataset augmentation to improve model robustness.

Validation

The extraction model's performance was assessed independently by clinically-trained human reviewers who undertake a manual review of the medical documents utilized for extraction and the corresponding extraction outputs. A second reviewer performs a QC on the first reviewer's work. Discrepancies are refereed by a third reviewer who engaged additional clinical experts for evaluating complex situations. Reviewers classified each field of extracted data, and we derived True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN) from which the Accuracy [(TP + TN) / (TP + TN + FP + FN)], Precision [TP / (TP + FP)], Recall [TP / (TP + FN)], and F1 Score [2 x TP / (2 x TP + FP + FN)] were calculated on a per-data field basis.

We have observed that LLMs can perform Named-entity Recognition and Relation Extraction tasks with exceptional accuracy, but that the models can sometimes give rise to "hallucinations."[7] In extraction tasks, these are commonly in the form of *correct inferences that are not explicitly stated in the source text*, such as deducing a medication's route of administration from its approved formulation (e.g., inferring "PO" in "APAP 325mg PO" when only "APAP 325mg" is mentioned). This has been described as a lack of faithfulness rather than a lack of factualness.[8] When measuring LLM performance herein, we are focused on faithfulness, thus only explicitly stated, verifiable extractions are deemed valid. In the above example, PO would be considered a False Positive, while APAP and 325mg would be considered True Positives.

Extraction models are reviewed on an ongoing basis with version-controlled updates and random sampling to control for potential model drift.

Mapping to FHIR

C-CDA are initially passed through a pipeline to generate pre-structured medication

data and medical document metadata, mapped to FHIR R4. LLM extraction outputs are also mapped to FHIR R4 resources for interoperability whenever possible. If a schema is not mapped to FHIR, its data is stored in extension fields according to other published extension profiles such as mCODE, offering a standardized naming convention for extension fields.[9] If no suitable extension profiles are available, a custom data model is constructed.

OMOP Normalization

After extraction, further processing is required to integrate these concepts with the broader medical record. FHIR resources mapped from C-CDA contain a code and an Observational Medical Outcomes Partnership (OMOP) Standardized Vocabulary System (e.g., LOINC or SNOMED), which are utilized to identify the equivalent OMOP Concept ID.[10] LLM-based extractions and a subset of structured data from CCDA result in FHIR resources that do not have OMOP mappings. The raw text extracted from these fields are first matched to concepts using the Intelligent Medical Objects (IMO) Precision Normalize database, and if unsuccessful, are manually mapped to the closest OMOP Concept ID.[11] This human coding step ensures that missing details are accurately supplied through maintained, version-controlled resources rather than the static internal knowledge of LLMs.

When normalizing medication data to OMOP, the majority of records are available in the RxNorm vocabulary. In the results below, references to “medication record” indicate a single instance of a medication from any document and the “drug ingredient” references the associated drug ingredient in the RxNorm vocabulary for the mapped medication concept.

Clinical Concept Mapping

Normalized data is grouped into high-level clinical concepts utilizing key nodes in the OMOP ontology. All concepts below these nodes are included in the concept, with some manual exclusions, and occasionally with the inclusion of concepts that are not represented by OMOP. Examples of clinical concepts for cancer medications include “immunotherapy,” “cytotoxic chemotherapy,” and “steroid.” Such groupings facilitate advanced analytics and support downstream workflows.

Results

Medical records in C-CDA and image formats were utilized for LLM extractor development and deployment according to the methods herein (Figure 1). A schema was designed to extract medications from clinical documents. Schema elements included the medication name, start and end dates, indication, discontinuation reason, dose, and route of administration. A cohort of 34 patients were selected for initial validation. An embeddings-based search of terms such as “chemotherapy,”

“immunotherapy” and “medications” yielded 260 medical documents on which to deploy the extraction process. Encounter location type was available for 86% of medical documents, most commonly oncology-related visits (35%). Provider specialty was available for 75% of medical documents.

A total of 5,789 medications were extracted, yielding 3,002 distinct medication records after removing duplicates (same name, start date, and end date). Extracted data elements were manually validated against source records to calculate accuracy, precision, recall, and F1 score, achieving ~95% for all measures across all data fields (Table 2).

The LLM medications extractor was then deployed to the medical records of 11,115 patients with confirmed cancer diagnoses and compared with medication data available pre-structured from C-CDA (Table 3). A significant amount of medication data was available for these patients in structured format after conversion from C-CDA to FHIR, yielding over 9.5 million individual medication records and 1,728 distinct drug ingredients from 476,525 medical documents. Separately, LLM-based extraction yielded ~2.6 million medication records and 2,001 distinct drug ingredients from 169,071 medical documents after normalization to OMOP vocabularies. Document selection for the LLM-based extraction identified 84,637 additional medical documents for data extraction beyond those containing pre-structured medications (18% more), and generated an increase in 27% more total medications and 31% more distinct drug ingredients than what was available in the C-CDA medication tables. Importantly, a review of oncology-specific medications showed that only 5,716 patients (51%) had a structured oncology medication in their medical record. The LLM medications extractor identified 80,325 medical documents for extraction (of which 71% did not contain pre-structured medications) across 7,246 patients (65%) to extract 235,425 medication records and 304 distinct drug ingredients. This yielded a 33% increase in the number of patients with at least one oncology medication, a 60% increase in total oncology medication records, and a 64% increase in distinct oncology drug ingredients. The LLM extracted 122 unique oncology therapies/medications not found in any structured C-CDA table. The LLM also extracted enhanced medication variables, outperforming the pre-structured C-CDA oncology medication data for the indication for prescription (61% vs. 31%) and reason for medication discontinuation (17% vs. 0%).

Discussion

Our findings demonstrate the effectiveness of LLM-based extraction for retrieving and structuring medication data from diverse medical record formats, significantly augmenting structured EHR data with important clinical information. The LLM-based approach increased both the total number of medication records and the distinct

number of drug ingredients extracted, particularly enhancing oncology-specific medication retrieval. This highlights the potential for LLMs to bridge gaps in structured clinical documentation by extracting clinically relevant information from unstructured or semi-structured sources.

Our approach achieved high F1 scores across multiple extracted data elements (average ~95%). However, some limitations warrant consideration. The accuracy of LLM-based extraction is contingent on high-quality input documents, and variations in OCR performance for scanned documents, such as faxes embedded within the medical record, may introduce errors or omissions. Additionally, while LLMs exhibited strong performance in the NER and RE tasks (identifying and correctly associating a dose, route, etc. with a specific medication), strict validation steps were necessary to mitigate hallucinations and ensure that only explicitly stated extractions were retained. While developed with a focus on F1 scores, we note that different applications may require optimization for one metric over another.

Careful prompt engineering with version control remains a critical step to ensure the fidelity of LLM output. Schemas should be optimized for specific tasks and document types for reliable clinical data extraction that integrates with existing coding systems. This highlights the importance of maintaining interoperability through mappings to standard terminologies like FHIR and OMOP. Our approach ensured that extracted data aligned with established clinical dictionaries, reducing the risk of data fragmentation or misclassification.

Future work will explore fine-tuning LLMs for domain-specific extractions, refinement of document chunking strategies and text classifications, extension of schema elements, and adopting additional validation metrics. For example, our current validation is heavily biased against the LLM, with two human abstractors and a referee evaluating the LLM results. Double data entry with a referee was a common approach for validation of data entry in clinical trials, with a mean reported error rate of 0.14%.[12] While this method establishes a gold standard for measuring LLM performance, in future work it may be useful to establish inter-rater reliability between human data abstractors in order to position LLM extraction relative to expert human performance. Despite these additional considerations, our results provide compelling evidence that LLM-based extraction can meaningfully enhance clinical data accessibility and completeness. The methodology presented here enhances interoperability, supports advanced analytics, and provides an efficient and scalable approach to managing complex healthcare data.

Acknowledgments

We would like to acknowledge Sabrina Irizarry, Hannah Hayes, and Joshua Marker.

Conflicts of Interest

All authors are employees or paid consultants of xCures, Inc.

References

1. Fraile Navarro D, Ijaz K, Rezazadegan D, Rahimi-Ardabili H, Dras M, Coiera E, Berkovsky S. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. Int J Med Inform. 2023 Sep;177:105122. doi: 10.1016/j.ijmedinf.2023.105122. Epub 2023 Jun 5. PMID: 37295138.
2. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):524-7. doi: 10.1136/jamia.2010.003939. PMID: 20819856; PMCID: PMC2995676.
3. Chen A, Yu Z, Yang X, Guo Y, Bian J, Wu Y. Contextualized medication information extraction using Transformer-based deep learning architectures. J Biomed Inform. 2023 Jun;142:104370. doi: 10.1016/j.jbi.2023.104370. Epub 2023 Apr 24. PMID: 37100106; PMCID: PMC10980542.
4. HL7 FHIR Release 4 Resource List. Accessed 24 Feb 2025. http://hl7.org/fhir/R4/resourcelist.html
5. OpenAI. New embedding models and API updates. Accessed 21 Feb 2025. https://openai.com/index/new-embedding-models-and-api-updates/
6. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large language models are few-shot clinical information extractors. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 2022. doi: 10.18653/v1/2022.emnlp-main.130.
7. Lee K, Firat O, Ashish Agarwal A, Fannjiang C, Sussillo D. Hallucinations in Neural Machine Translation. ICLR 2019. Submitted 27 Sept 2018. https://openreview.net/pdf?id=SkxJ-309FQ
8. Ye H, Liu T, Zhang A, Hua W, Ji W. Cognitive Mirage: A Review of Hallucinations in Large Language Models. Submitted 13 Sep 2023. arXiv:2309.06794. doi: https://doi.org/10.48550/arXiv.2309.06794
9. HL7 International / Clinical Interoperability Council. Minimal Common Oncology Data Elements (mCODE) Implementation Guide. Accessed 24 Feb 2025. https://build.fhir.org/ig/HL7/fhir-mCODE-ig/
10. Observational Health Data Sciences and Informatics (OHDSI). Standardized Vocabularies. Accessed 24 Feb 2025. https://github.com/OHDSI/Vocabulary-v5.0/wiki/Standardized-Vocabularies

11. IMO Health. HEalthcare Data Standardization | IMO Precision Normalize. Accessed 24 Feb 2025. https://www.imohealth.com/imo-precision-normalize/.
12. Garza MY, Williams T, Ounpraseuth S, Hu Z, Lee J, Snowden J, Walden AC, Simon AE, Devlin LA, Young LW, Zozus MN. Error rates of data processing methods in clinical research: A systematic review and meta-analysis of manuscripts identified through PubMed. Int J Med Inform. 2025 Mar;195:105749. doi: 10.1016/j.ijmedinf.2024.105749. Epub 2024 Dec 4. PMID: 39647291.

Figures

Figure 1. Extraction pipeline for development, validation, and deployment of LLMs. Patient documents are received in multiple formats, including C-CDA, FHIR, and image-based files, requiring preprocessing steps such as document parsing, OCR, and structured metadata extraction for interoperability. Text is segmented into chunks using section-based and semantic chunking strategies, followed by embedding-based classification. Targeted semantic search and document classification algorithms are utilized for retrieval of appropriate documents for LLM model training and deployment. Performance metric thresholds are utilized to graduate LLMs from development to human validation to large-scale deployment. LLM-extracted data is mapped to FHIR R4 resources, normalized to OMOP concepts, and integrated with tabular C-CDA data ensuring standardized representation and enhancing data completeness.

Table 1. Performance of document classifier machine learning algorithm

Classification	F1 score
allergies_adverse_reactions_alerts	0.98
assessment_and_plan	0.98
demographics	0.92
diagnosis	0.89
diagnostic_test_or_lab_results	0.80
disease_biomarkers	0.98
family_medical_history	0.98
functional_status	0.98
genomics	0.98
history_of_present_illness	0.97
insurance	0.99
interventional_radiology	0.99
medical_equipment_and_implants	0.99
medications	0.84
oncology	0.88
pathology	0.96
pharmacology_progress	1.00
physical_exam	0.96
plan_of_care	0.97
procedures	0.94
radiation	0.99
radiology	0.91
reason_for_visit	0.98
social_history	0.95

surgery	0.96
vital_signs	0.96

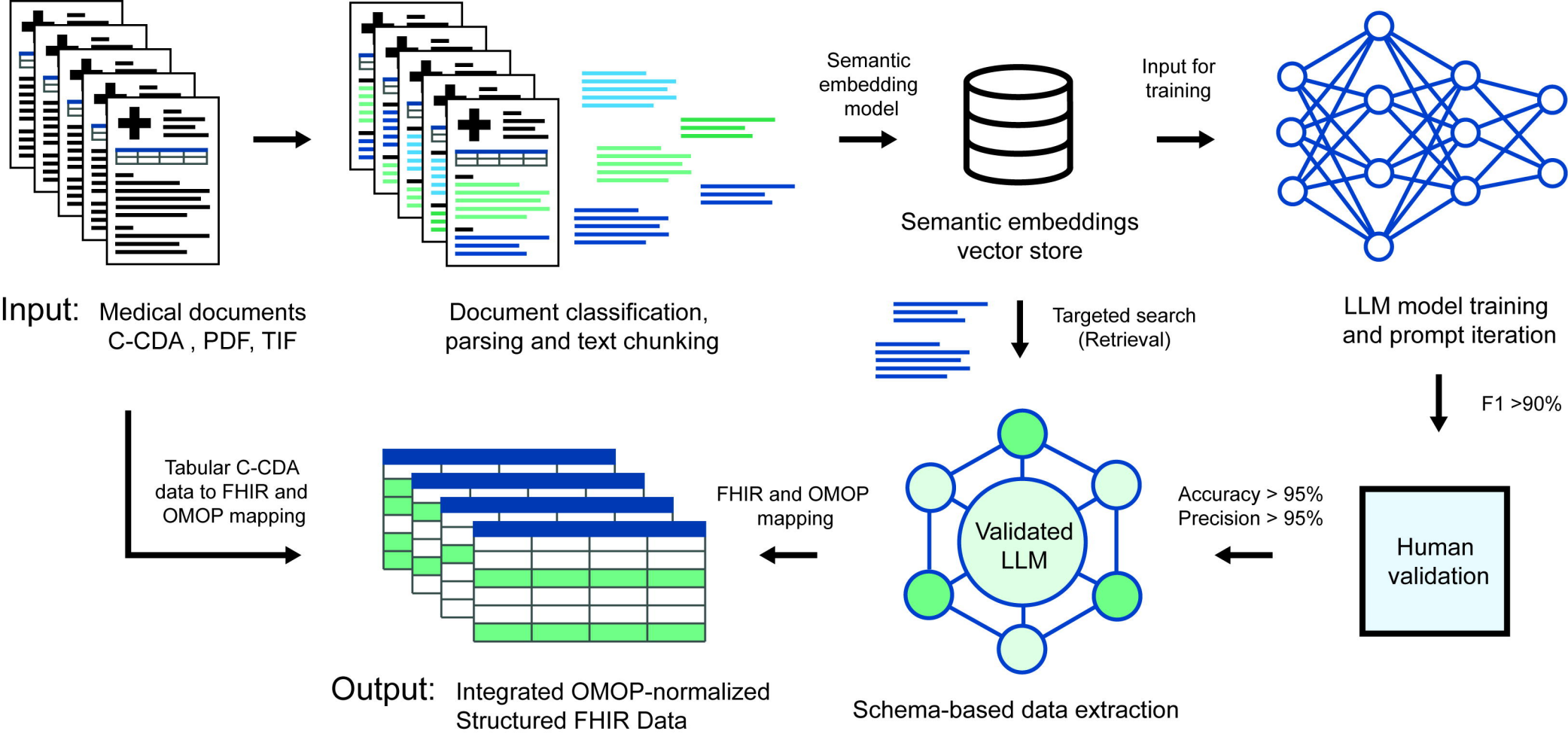
Table 2. Performance metrics of medications extractor

Data field	Accuracy	Precision	Recall	F1
medication_name	1.00	1.00	1.00	1.00
start_date	0.92	0.97	0.91	0.94
end_date	0.94	0.95	0.84	0.89
indication	0.91	0.92	0.91	0.92
discontinuation_reason	0.99	0.91	0.91	0.91
dose_information	0.96	0.98	0.97	0.98
route_text	0.96	0.99	0.97	0.98

Table 3. Information gain from deployed LLM medications extractor

Subset	Data Field	C-CDA pre-structured	LLM	Combined dataset	Gain from LLM (%)
Any medication	**Patients**	11,115	11,115	11,115	
Any medication	**Documents utilized**	476,525	169,071	561,162	18%
Any medication	**Total medication records**	9,643,633	2,610,152	12,253,785	27%
Any medication	**Distinct medication records**	221,712	31,105	241,645	9%
Any medication	**Distinct drug ingredients**	1,728	2,001	2,267	31%
Any medication	**Median distinct drug ingredients per subject**	58	44	67	16%
Oncology-specific*	**Patients**	5,716	7,246	7,589	33%
Oncology-specific*	**Documents utilized**	117,636	80,325	174,459	48%
Oncology-specific*	**Total medication records**	391,666	235,425	627,091	60%
Oncology-specific*	**Distinct medication records**	18,303	2,412	20,199	10%
Oncology-specific*	**Distinct drug ingredients**	190	304	312	64%
Oncology-specific*	**Median distinct drug ingredients per subject**	3	4	4	33%

*Data subsetted to just medications with a clinical concept of: cytotoxic chemotherapy, targeted therapy, immunotherapy, hormone therapy, cell therapy, cancer vaccine.


Input: Medical documents
C-CDA , PDF, TIF
Document classification,
parsing and text chunking
Semantic
embedding
model
Semantic embeddings
vector store
Input for
training
LLM model training
and prompt iteration
F1 >90%
Human
validation
Accuracy > 95%
Precision > 95%
Validated
LLM
Schema-based data extraction
Targeted search
(Retrieval)
FHIR and OMOP
mapping
Tabular C-CDA
data to FHIR and
OMOP mapping
Output: Integrated OMOP-normalized
Structured FHIR Data