HARMON-E : Hierarchical Agentic Reasoning for Multi-modal Oncology Notes to Extract Structured Data

Shashi Kant Gupta[1], Arijeet Pramanik[1], Jerrin John Thomas[1], Regina Schwind[1], Lauren Wiener[3], Avi Raju[3], Jeremy Kornbluth[3], Yanshan Wang[2†], Zhaohui Su[3†], Hrituraj Singh[1*†]

[1] Triomics, New York, USA.
[2] University of Pittsburgh, Pittsburgh, USA.
[3] Ontada, Boston, USA.

*Corresponding author(s). E-mail(s): hrituraj@triomics.com;
Contributing authors: shashi.gupta@triomics.com;
arijeet.pramanik@triomics.com; jerrin.thomas@triomics.com;
regina@triomics.com; lauren.wiener@mckesson.com;
avi.raju@mckesson.com; jeremy.kornbluth@mckesson.com;
yanshan.wang@pitt.edu; zhaohui.Su@mckesson.com;
†These authors contributed equally to this work.

Abstract

Unstructured notes within the electronic health record (EHR) contain rich clinical information vital for cancer treatment decision making and research, yet reliably extracting structured oncology data remains challenging due to extensive variability, specialized terminology, and inconsistent document formats. Manual abstraction, although accurate, is prohibitively costly and unscalable. Existing automated approaches typically address narrow scenarios—either using synthetic datasets, restricting focus to document-level extraction, or isolating specific clinical variables (e.g., staging, biomarkers, histology)—and do not adequately handle patient-level synthesis across the large number of clinical documents containing contradictory information. In this study, we propose an *agentic* framework that systematically decomposes complex oncology data extraction into modular, adaptive tasks. Specifically, we use large language models (LLMs) as reasoning agents, equipped with context-sensitive retrieval and iterative synthesis capabilities, to exhaustively and comprehensively extract structured clinical variables from

real-world oncology notes. Evaluated on a large-scale dataset of over **400,000** unstructured clinical notes and scanned PDF reports spanning **2,250** cancer patients, our method achieves **an average F1-score of 0.93**, with 100 out of 103 oncology-specific clinical variables exceeding 0.85, and critical variables (e.g., biomarkers and medications) surpassing 0.95. Moreover, integration of the agentic system into a data curation workflow resulted in 0.94 direct manual approval rate, significantly reducing annotation costs. **To our knowledge, this constitutes the first exhaustive, end-to-end application of LLM-based agents for structured oncology data extraction at scale.**

Keywords: electronic health records, large language models, oncology data extraction, artificial intelligence, clinical natural language processing, agentic language modeling, cancer, structured data, automated abstraction, clinical informatics, biomarkers, healthcare data curation, medical text mining, computational oncology, EHR notes, unstructured clinical data, patient records, medical document processing, clinical decision support, information extraction

1 Introduction

Traditional efforts toward automating the extraction of structured data from clinical text have relied heavily on rule-based systems or shallow machine learning models (e.g., Conditional Random Fields, Support Vector Machines), each requiring extensive domain-specific feature engineering [1–4]. In the last decade, however, transformer-based models [5], such as Bidirectional Encoder Representations from Transformers (BERT)[6] and its domain-specific variants like ClinicalBERT [7–9], have significantly improved upon classical approaches across a range of clinical natural language processing (NLP) tasks. More recently, the approach has shifted from fine-tuning domain specific models [10, 11] to using the generalized abilities of frontier large language models (LLMs) like GPT-4 and GPT-5[12, 13] to extract key concepts from EHR records [14].

Nevertheless, prior works often assume relatively uniform data (such as a single cancer diagnosis or one type of note), single-document inputs, or a very limited set of concepts. Published studies have demonstrated entity extraction for discrete variables such as tumor stage from pathology reports [15, 16], receptor or genomic biomarkers from pathology or genomic reports [17, 18], and initial regimens or lines of therapy for select cancers [19, 20]. However, these approaches extract variables independently without synthesizing findings across multiple documents to resolve contradictions or complete partial information. This study is distinct in that it targets patient-level synthesis across heterogeneous, longitudinal records, where oncology-specific concepts must be inferred by collating and contextualizing information scattered across multiple notes and data types, rather than relying on single-step extraction from a uniform document.

To address these challenges, we introduce `HARMON-E` : **H**ierarchical **A**gentic **R**easoning for **M**ulti-source **O**ncology **N**otes to **E**xtract structured data

Our framework systematically decomposes oncology data extraction into modular, iterative steps that leverage large language models (LLMs) as "reasoning agents" that interleave retrieval with multi-step inference to reconcile conflicting evidence, normalize temporal references, and derive implicit variables not explicitly stated in the text - such as inferring treatment discontinuation dates from adverse event narratives (Figure 1). Concretely, `HARMON-E` combines context-aware indexing, adaptive retrieval methods (including vector-based search and rule-based actions), and LLM-driven data synthesis into a unified workflow. By adopting a hierarchical, agentic approach, we mitigate common pitfalls in extracting complex, interdependent oncology data and achieve end-to-end alignment with standardized pre-defined set of clinical variables comprising 103 attributes across 16 entity types-including Biomarker, Medication, Diagnosis, Staging, Surgery and Radiation entities as outlined in Table 1 and Figure 3.

We validate `HARMON-E` on a large-scale real-world dataset of over 400,000 unstructured clinical notes and scanned PDFs belonging to 2,250 cancer patients. Our evaluation covers 103 oncology-specific variables—ranging from histopathological findings and biomarker statuses to treatment patterns and disease progression. `HARMON-E` consistently delivers high accuracy, with 100 out of 103 variables exceeding an F1-score of 80%, and crucial fields such as biomarker assessments and medication data surpassing 95% accuracy. Beyond these metrics, when integrated into a data curation platform, our system demonstrates a direct manual approval rate of 94.1%, substantially reducing human annotation burden without sacrificing data quality.

Our contributions can be summarized as follows -

1. **High-accuracy LLM-based agentic workflows.** We provide the first evidence that large language models, organized into an agentic workflow, can exceed 95% F1-score on key oncology concepts at scale, validated against a corpus of 940,923 data points derived from 400,000 clinical notes covering 2,250 patients.
2. **Real-world integration and user acceptance.** We also demonstrate that when integrated into an interactive data curation platform, 94% of the extracted data points get directly accepted by professional oncology abstractors without modification, substantially reducing manual review time.
3. **Novel evaluation framework.** We propose an evaluation methodology which moves beyond traditional named-entity recognition metrics by aligning performance assessment with the real-world requirements of oncology data curation and quality monitoring.

2 Related Work

Early Clinical NLP and Rule-Based Methods: Early clinical NLP relied primarily on rule-based systems, utilizing domain-specific lexicons and hand-crafted patterns [3, 21, 22]. These systems, though precise in certain scenarios, required substantial manual effort and were brittle when encountering variability [1, 23]. Notable examples include MedXN for medication extraction [24], comprehensive clinical NLP toolkits like CLAMP [25], fracture identification from radiology reports [26], and sudden cardiac death risk factor extraction [27]. Statistical machine learning models, notably

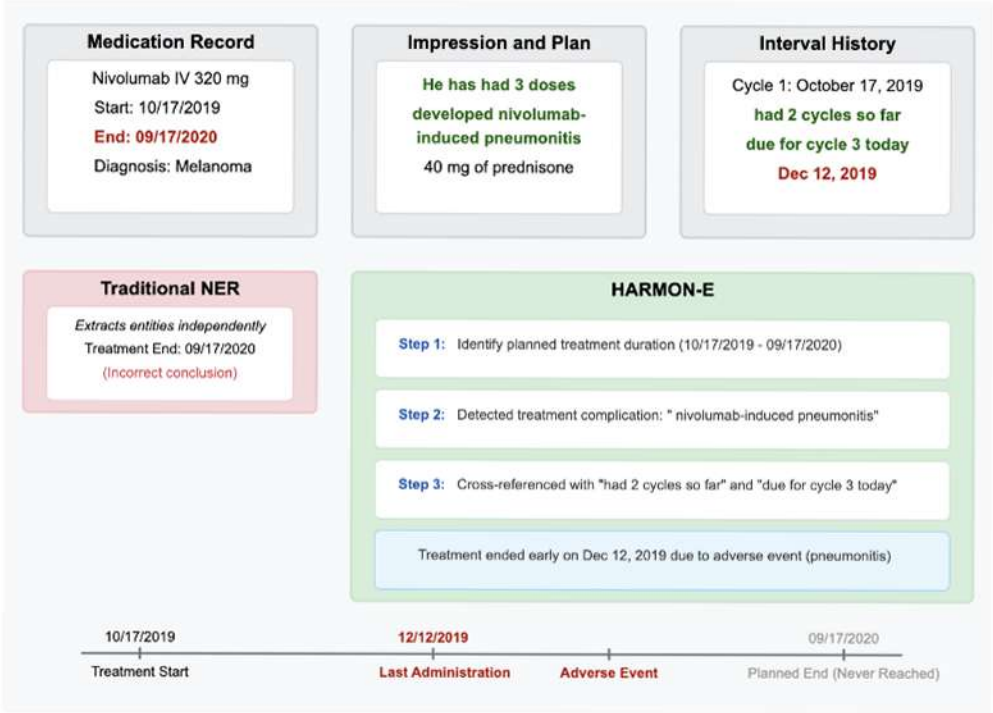

Fig. 1: **Comparative analysis of HARMON-E versus traditional Named Entity Recognition (NER) approaches for medical document processing**. A traditional NER system (lower left, red box) would incorrectly extract the planned end date (09/17/2020) without contextual understanding. In contrast, our agentic system (lower right, green box) performs multi-step reasoning: first identifying the planned treatment duration, then detecting the adverse event (pneumonitis), cross-referencing information about completed treatment cycles, and finally concluding that treatment actually ended on December 12, 2019 due to the adverse event—a conclusion impossible with traditional single-pass methods. (This is not real patient data and is for illustrative purposes only.)

Conditional Random Fields (CRFs) and Support Vector Machines (SVMs), subsequently improved performance on clinical text tasks like de-identification and entity extraction [1]. Early language modeling approaches showed promise for identifying relevant information in clinical notes [28], though still required task-specific adaptation. Nevertheless, adapting these models to heterogeneous oncology datasets posed significant challenges due to diverse terminologies and complex documentation [29]. The evolution of clinical information extraction methods has been comprehensively reviewed by Wang et al. [30], documenting the systematic transition from rule-based to neural approaches.

Deep Learning and Transformer Models in Clinical NLP: Transformer architectures, particularly BERT [5, 6], significantly advanced clinical NLP by providing powerful language representations. Domain-specific models such as BioBERT [8], ClinicalBERT [7], and SciBERT [11] further enhanced clinical NLP tasks, achieving state-of-the-art performance in clinical entity recognition and relation extraction [31, 32]. Comprehensive evaluations of these models on biomedical text-mining tasks demonstrated their superiority over traditional approaches [33]. However, oncology data, characterized by extensive and contextually complex documentation, poses unique challenges inadequately addressed by traditional single-pass transformers. Approaches employing hierarchical transformers and multi-document summarization have attempted to mitigate these limitations [9, 34–36] but still fall short of fully replicating human abstraction.

Large Language Models in Healthcare: Recent advances in large language models (LLMs), including GPT-4, and GPT-5[12, 13], have demonstrated impressive capabilities in summarization, medical question answering, and clinical decision support [37, 38]. This has led to several research works exploring the capabilities of these models in extracting structured data from notes [14, 39, 40]. Recent multi-institutional efforts have shown promise in extracting social determinants of health from clinical notes using LLMs, achieving F1 scores over 0.9 [41]. Work on extracting functional status information, including mobility assessments, from clinical notes has demonstrated the potential of LLMs for capturing complex clinical concepts [42, 43]. Despite this progress, single-pass LLM approaches struggle with multi-document EHR data, potentially losing critical context or misinterpreting conflicting information. Moreover, most research work so far has been limited to a limited set of oncology concepts [9, 36, 39, 44] instead of scaling it to a comprehensive RWD dataset dictionary.

Agentic and Iterative NLP Frameworks: Recent NLP frameworks have increasingly adopted iterative reasoning approaches. The ReAct framework introduced by Yao et al. [45] enables LLMs to interleave reasoning and retrieval actions, improving accuracy on complex reasoning tasks. Techniques such as chain-of-thought prompting [46] and self-consistency decoding [47] further enhance multi-step reasoning capabilities. Additionally, integrating external tools or retrieval systems, as explored in Toolformer [48] and web-augmented retrieval methods [49], has shown promise. Yet, systematically integrating these approaches for clinical NLP, particularly oncology-specific data extraction, remains an open challenge due to accuracy and validation constraints.

Autonomous & Agentic LLM Systems in Oncology: Oncology specific AI research has also begun moving from single-prompt LLM use toward *agentic* pipelines that interleave reasoning and tool calls. Ferber *et al.* developed an autonomous GPT-4–driven clinical-decision agent that orchestrates vision models, knowledge bases, and web search to solve multimodal oncology cases, raising accuracy from 30% (plain GPT-4) to over 87% on complex vignettes [50]. Sandhu *et al.* proposed an open-source modular framework that couples rule-based components with an agentic layer to generate comprehensive breast-cancer notes and benchmark treatment recommendations against NCCN guidelines [51]. Outside healthcare, Zhang and Elhamod

proposed *Data-to-Dashboard*, a multi-agent architecture that detects domain context, extracts concepts, performs iterative self-reflection, and automatically builds analytic dashboards [52]. While these works demonstrate the promise of hierarchical agents, they tackle relatively *narrow* tasks (tens of test cases, disease-specific notes, or generic analytics) rather than end-to-end patient-level abstraction across thousands of heterogeneous documents.

Domain-Specific LLMs for Oncology. Parallel efforts tailor foundation models and pipelines to oncology data. *CancerBERT* pretrains BERT on cancer corpora with a cancer-specific vocabulary and reports significant gains for extracting breast-cancer phenotypes (e.g., receptor status, site/laterality) over general and clinical BERT baselines from EHR notes and pathology reports[53]. *OncoBERT* explores transfer learning on oncology notes for outcome prediction and phenotype structuring, reporting improvements on site and clinical T-staging tasks and highlighting interpretability considerations for radiation oncology workflows [9, 10]. Beyond encoders, registry-oriented systems deliver patient-level abstraction: the original *DeepPhe* extracts phenotypes across entire EMRs, while *DeepPhe-CR* exposes API services integrated into cancer-registrar tools for computer-assisted abstraction, with usability studies and deployment guidance for registry workflows [4, 54]. Patient-level staging at treatment initiation has also been demonstrated in Veterans Affairs data using rule-based NLP with validated roll-up for multiple myeloma [55]. In parallel, registry–EHR fusion cohorts constructed via NLP underscore the need for multi-source abstraction [56]. A complementary line targets narrow modalities or procedures—e.g., *Woollie*, a radiology-focused LLM trained on 39k impressions that attains strong progression-prediction AUROC of 0.97 for progression prediction and outperforms general LLMs on medical benchmarks, and hybrid agentic pipelines coupling rules with GPT-4-Turbo for detailed spine-surgery variable extraction [57, 58]. Orthogonal oncology IE efforts focus on single-entity extraction from pathology and clinical notes, including TNM staging and receptor status [15, 16, 18], systemic treatment identification and line-of-therapy inference [19, 20], and broader RWD curation reviews and tooling [17, 59].

Most oncology LLM studies target narrow concept sets rather than comprehensive data dictionaries [9, 36, 39, 44]. These systems typically optimize for task- or modality-specific accuracy using single-pass encoders or rule-based components; they provide limited mechanisms for cross-entity dependency resolution, large-scale deduplication across documents, and curator-ready validation—gaps our agentic workflow aims to close.

To overcome this, `HARMON-E` aim to mirror the iterative reasoning process of human abstractors. By combining state-of-the-art linguistic models and structured retrieval actions, our proposed approach promises robust, scalable extraction from heterogeneous oncology records. To our knowledge, this is the first agentic LLM framework that delivers both *breadth* (comprehensive data dictionary) and *depth* (patient-level consistency) at scale, bridging the gap between prototype agents or niche LLMs and production-grade oncology data curation.

Table 1: Clinical Entities, Definitions, and Corresponding Attributes

Entity Name	Definition of Entity	Attributes
Biomarker	Consolidated entity for all biomarker testing, including genetic biomarkers, microsatellite instability, copy number alterations, rearrangements, tumor marker tests, and tumor mutation burden.	biomarker_test_date, result_date, biomarker_tested, gene_studied, gene1, copy_number_type, method, code, aminoacid_change, aminoacid_changetype, stain_percent, value, value_quantity, value_unit, interpretation, molecular_abnormal
Biopsy	Describes the biopsy specimen (e.g., liquid or tissue) and dates.	type, result_date, ordered_date, collect_date, insufficient_tissue
Clinical Trial	Derived from clinical trial attributes. Indicates patient enrollment in a trial.	ct_flag, ct_start_date, ct_end_date
Comorbidities	Captures health conditions (Charlson Comorbidity Index) co-existing with primary cancer.	comorb_date, comorb_condition, comorb_condition_present
Distant Metastasis	Indicates metastatic spread beyond the primary site after initial diagnosis.	status_date, status, associated_diagnosis, body_site
Diagnosis	Details about the primary diagnosis, including histology and site.	diag_date, condition, body_site, histology
Family History	Details the family cancer history for a patient.	relationship, condition, onset
Imaging	Describes imaging services a patient has received.	start_date, modality, body_site
Medication	Captures systemic treatments (drugs, therapies), start/end dates, routes, etc.	status, start_date, end_date, treatment_intent, medication, route, termination_reason, baseline_dosage, baseline_dosage_units, baseline_dosage_quantity, baseline_dosage_freq, baseline_dosage_duration, baseline_cycle_length, dose_change_reason, changed_dosage, changed_dosage_units, changed_dosage_quantity, changed_dosage_freq, changed_dosage_duration, changed_cycle_length, dose_formula, dose_formula_unit, treatment_sequence
Nicotine Use Status	Describes patient's smoking or nicotine usage.	code, type, use, use_unit, use_frequency, start_date, end_date
Patient Status	Covers demographics, vital status, and disposition.	vital_status, last_contact_date, hospice_date, relapse_flag
Radiation	Captures radiation therapy details.	modality, start_date, end_date, total_dose_delivered_value, total_dose_delivered_unit, fractions_delivered, body_site
Recurrence Status	Tracks local or metastatic recurrence of disease.	status_date, status, associated_diagnosis, body_site
Staging	Describes TNM staging at diagnosis.	stage_date, stage_type, tumor_category, nodes_category, metastases_category, stage_value
Surgery	Dates and types of surgical procedures received.	surgery_date, surgery_type, outcome
Surgery Observations	Specific to surgical procedures and outcomes.	observe_date, code, value, value_units

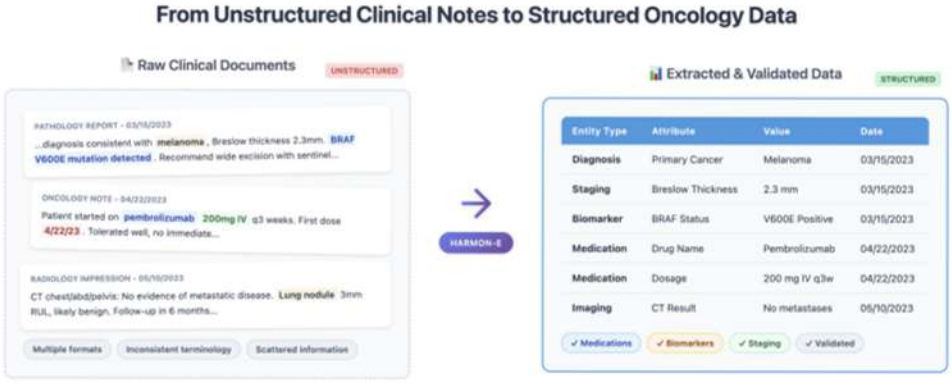

Fig. 2: **The HARMON-E transformation pipeline.** Unstructured clinical documents from multiple sources (left) containing medications, biomarkers, staging information, and other oncology data are processed through the HARMON-E agentic extraction pipeline (center) to produce standardized, structured database entries (right). Each piece of clinically relevant information is extracted, validated, and organized into predefined entity-attribute pairs suitable for clinical research and decision support. The system processes heterogeneous inputs including progress notes, pathology reports, radiology impressions, and scanned PDFs, transforming approximately 180 documents per patient into comprehensive structured records.

3 Methods

3.1 Problem Formulation

We consider a set of patients

$$\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$$

and a predefined collection of *clinical entities* relevant to oncology (e.g., `Biomarker`, `Cancer Related Medication`, `TNM Stage Group`). Each entity E_j consists of a collection of *attributes*

$$\mathcal{A}(E_j) \;=\; \{\, a_{j,1},\, a_{j,2},\, \ldots,\, a_{j,k_j} \},\qquad k_j := |\mathcal{A}(E_j)|.$$

where each attribute can be a categorical label (with a fixed value set), a date, a numeric value, or unstructured text. Here, k_j denotes the number of attributes defined for entity E_j and can differ by entity (see Table 1). Given a corpus of unstructured or semi-structured oncology documents

$$\mathcal{D}_i = \{\, d_1, d_2, \ldots, d_m \} \quad \text{for each patient } p_i,$$

Fig. 3: The tree structure illustrates the decomposition of a patient record into six primary entity categories (Medication, Biomarker, Diagnosis, Staging, Surgery, and Radiation), each containing multiple typed attributes. Representative examples from real oncology data are provided in italics below each attribute. This structured schema ensures consistency across heterogeneous clinical documentation and enables validation of extracted data against clinical standards. The complete model encompasses 16 entity types with 103 distinct attributes (subset shown for clarity).

our goal is to automatically *extract* all valid instances of these entities. Equivalently, we want a structured output $\mathcal{S}_i$ for every patient p_i containing all detected entities $\{E_1, \ldots, E_J\}$ and their attribute values. Thus,

$$\mathcal{S}_i = \bigcup_{j=1}^{J} \hat{E}_j,$$

where J is the total number of clinical entity types considered, and each $\hat{E}_j$ is a set of extracted instances of type E_j. Table 1 lists 16 oncology entity types considered in this study.

We evaluate the quality of extraction by comparing the system outputs against a gold standard dataset. For each entity type, we assess:

- **Entity-Level Recall:** Do we capture *all* correct entity instances mentioned in patients' notes?
- **Entity-Level Precision:** Do we avoid producing extra or incorrect entity instances?
- **Attribute-Level Accuracy:** Conditioned on having a correct entity instance, are all attributes (e.g., dates, biomarker results) accurate?

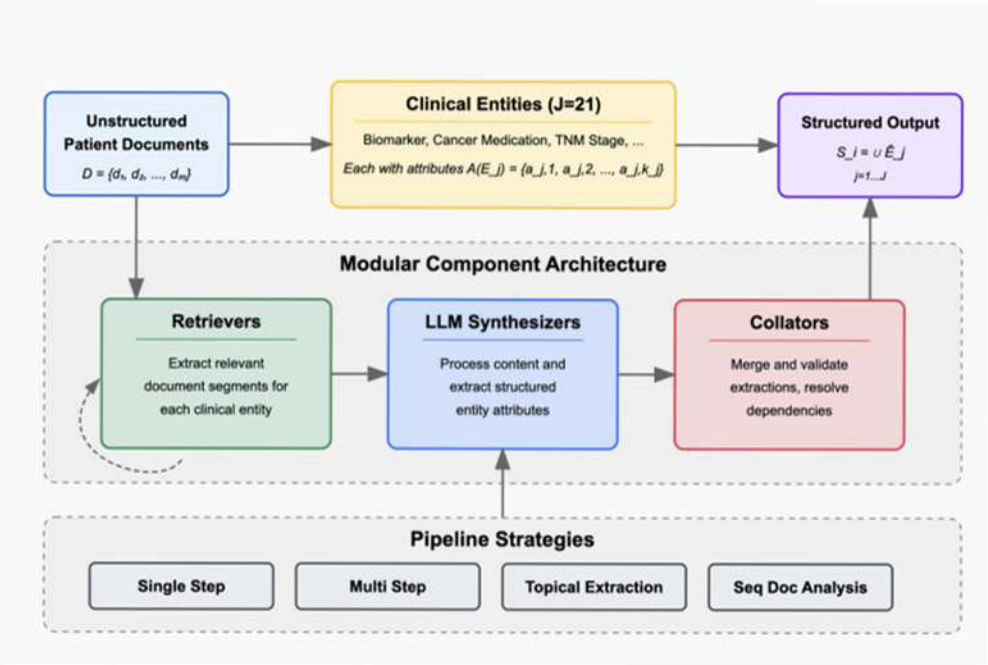

Fig. 4: System workflow of HARMON-E. The system accepts raw, unstructured patient documents (HTML notes and/or scanned PDFs normalized as in Sec. 3.2) and processes them through three main components: *Retrievers*, *LLM Synthesizers*, and *Collators*. Retrievers extract relevant segments for each targeted oncology entity; LLM Synthesizers transform these segments into structured attribute-value pairs; and Collators merge, validate, and resolve any dependencies among the extracted entities. The framework supports multiple pipeline strategies—from single-step to multi-step or topical extraction—accommodating a diverse range of real-world workflows. The final output is a patient-level structured data record spanning multiple oncology concepts (e.g., biomarkers, medications, TNM staging).

We further perform alignment between predicted entity instances and ground-truth references (e.g., via root-based or weighted matching) to compute these metrics systematically as discussed in the Section 4.

3.2 Document ingestion and normalization

Scanned PDFs are transcribed to page-level Markdown using a vision–language model fine-tuned on clinical documents. We then invoke two LLM calls for (i) page-to-document segmentation and doc typing, and (ii) metadata extraction (encounter date, report title, identifiers). The resulting normalized text feeds the retrieval–synthesis–collation pipeline.

3.3 Architecture

We design a modular, domain-focused architecture for extracting oncology-specific data variables at scale. The architecture decomposes the problem into three *components* and integrates them into one of several *pipeline* strategies, each specializing in different document-processing scenarios.

3.3.1 Key Features

- **Modular Component Architecture:** We expose interchangeable building blocks (*retrievers*, *LLM synthesizers*, and *collators*) to accommodate different data formats and use cases.
- **Multiple Pipeline Strategies:** We implement multiple component combinations (*Single-Step*, *Multi-Step*, *Topical Extraction*, *Sequential Document Analysis*), each with distinct retrieval-extraction flows.
- **Flexible Configuration System:** Users can seamlessly modify pipeline parameters (e.g., queries, prompts, pattern matchers) without altering core logic.
- **Strong Typing & Validation:** Entity attributes are rigorously typed (categorical, date, numeric, etc.) to ensure consistent outputs and validate data integrity.
- **Dependency Management for Complex Extractions:** Certain extractions rely on previously resolved attributes (e.g., a medication collator may need a confirmed diagnosis date). Our architecture manages these dependencies explicitly.

3.3.2 Core Data Models

At the heart of the system are *entities* and *attributes*. Each entity E_j includes:

$$\mathcal{A}(E_j) \;=\; \{\,(n_{j,i},\, t_{j,i},\, V_{j,i})\,\}_{i=1}^{k_j},$$

where $n_{j,i}$ is an attribute name (e.g., `biomarker_tested`), $t_{j,i}$ is its type (e.g., `Date`, `Integer`, or `Categorical`), and $V_{j,i}$ is an optional finite set of valid values for categorical attributes (e.g., {`Positive, Negative`}).

We also incorporate *strong typing* by rejecting any extractions that violate the declared schema (e.g., a numeric field with an invalid string). In practice, this ensures consistent representation across different pipeline stages.

3.3.3 Main Components

1. Retrievers

A retriever receives text input and yields a list of *candidate chunks* (snippets) relevant to the extraction goal. Clinical notes are chunked deterministically into sentence-bounded windows, each capped at M characters, with a one-sentence overlap between consecutive windows. We consider two broad categories:

- **Vector Retriever**: Embedding-based similarity search to identify chunks that semantically match a query (e.g., a disease name or biomarker).
- **Regex Retriever**: Regex-based pattern matching to capture well-defined textual cues (e.g., specific drug names, standard biomarkers).

How queries are defined. Retrieval queries are *entity-conditioned*. For a target entity E_j and attribute subset $\mathcal{S}$, the user configures $\{q_1, \ldots, q_m\}$, where m is the number of queries generated for that extraction call.
Example queries. *Diagnosis (surgery):* "Has the patient undergone any resection?", "Is there a mention of mastectomy?", "What is the surgery date?"
Biomarker (BRAF): "When was BRAF last tested?", "What was the BRAF test result?", "How did the lab interpret the BRAF result?"

Algorithmically, a *Vector Retriever* can be summarized as:

Algorithm 1 Vector Retriever (Abstract)

Require: A set of chunk texts, an embedding function $\phi(\cdot)$, a user query q, and retrieval size k.
Compute the embedding $\phi(q)$ for the query.
For each chunk text d, compute $\phi(d)$.
Rank chunks by cosine similarity $\langle \phi(q), \phi(d) \rangle$.
Return top k most similar chunks.

For each query q_r, `top_k` (k_r) is the number of highest-scoring chunks that is retained after ranking; larger k_r increases recall at the cost of more noise and latency.

Examples: surgery-identification queries use k_r=16 to catch sparse mentions; surgery-detail queries (e.g., date, margins) use k_r=8; biomarker result-line queries use small k_r (e.g., 4) since evidence is localized.

2. LLM Synthesizer

Given a chunk or chunks of text, the *LLM Synthesizer* (e.g., an LLM such as GPT-4) is prompted to extract a structured representation. The prompt typically includes instructions about the schema to be returned. For instance, we might request that the LLM produce a JSON-like object with `name`, `dosage`, and `start_date` fields for a `CancerMedication` entity. Concretely:

1. **Concatenate:** The chunk text plus any instructions (e.g., templates, format specification).
2. **Infer:** The LLM extracts attribute values from the chunk's content.
3. **Emit:** A structured representation that respects the required fields and data types.

3. Collators

A collator accepts multiple partial extractions and merges or filters them into a final, validated structure. Here, a collator is a deterministic post-processing module (not an LLM unless explicitly stated) that canonicalizes values, enforces type/value-set constraints, deduplicates, and resolves conflicts via simple precedence rules. Typical collator functions include:

- *Deduplication:* If multiple extractions have the same `name` and `date`, unify them.
- *Validation:* Check whether attributes match type constraints or known value sets.

- *Conflict Resolution:* If two extractions have contradictory attributes, apply domain logic to keep the most recent or plausible instance.

Collation can be formalized as:

$$\text{Collate}\big(\{R_1, \ldots, R_m\}\big) = \widehat{R},$$

where R_i are sets of attribute-value pairs from the LLM Synthesizer, and $\widehat{R}$ is a single consolidated set of entity instances. If the collator depends on other entity data (e.g., a known diagnosis date), it can be passed in as an auxiliary input, ensuring consistent domain constraints across different entity types. Collators can be chained together to produce complex logical strategies.

3.3.4 Pipeline Types

We integrate these components into four canonical pipelines for this specific project, each reflecting a distinct approach to retrieving, synthesizing, and merging entity information. The modular components, however, allows us to create even more complex pipelines which are out of the scope of this work:

(a) Single-Step Pipeline

A straightforward, single-pass strategy. The pipeline first retrieves all relevant chunks using either vector or regex methods. The LLM Synthesizer then processes each chunk in one shot to produce the entity attributes, and a collator merges the results. This is well suited for entities that appear in well-defined contexts or are consistently mentioned in the text such as biomarkers.

(b) Multi-Step Pipeline

A multi-stage approach for more complex or heterogeneous entities. First, we *identify* which variants or subtypes are mentioned (e.g., enumerating possible medications). Next, for each identified subtype, we *extract* more detailed attributes with a targeted LLM prompt. The collator then merges these partial extractions. This approach reduces confusion for the LLM when multiple entity types are possible.

(c) Topical Extraction

We split a patient's documents into coherent *topics* (e.g., *radiology reports*, *molecular testing*, *social history*) and systematically process each topic with a specialized prompt. This is particularly useful when distinct sections of text require different domain knowledge or extraction strategies, but must ultimately be combined into a single patient-level record.

(d) Sequential Documents Analysis

We treat each document as an individual unit, performing retrieval or chunking within that document, then calling the LLM. The collator merges the partial outputs across all documents, preserving the document lineage. This is valuable when the timing and

source of information is critical, such as for pathology or surgical reports that must be chronologically tracked.

Pipeline selection policy

We choose the pipeline by entity complexity and context: Single-Step for localized cues with few attributes (e.g., `Biomarker`); Multi-Step for entities with variants and many attributes (e.g., `Medication`); Topical when sections require specialized prompts (e.g., radiology vs. molecular testing); Sequential Documents when provenance and chronology are critical (e.g., `Staging`, surgery timelines).

3.3.5 Dependency Management

When the extraction for an entity (e.g., `Cancer Medication`) requires context from an upstream entity (e.g., `Primary Cancer Condition`), the pipeline enforces a dependency order among collators. For instance, if the medication collator depends on a confirmed diagnosis date, the pipeline first finalizes the `Primary Cancer Condition` collations before generating medication instances. This ensures consistent references (e.g., no medication entry without a corresponding diagnosis period).

In summary, our methods combine modular retrieval (via patterns or embeddings), LLMs (for robust text-to-structure synthesis), and domain-driven collation (for validation, deduplication, and dependency resolution). This general design accommodates a wide array of oncology-specific extractions, from straightforward biomarkers to multi-document treatments and staging workflows. Implementation specifics, including representative prompt templates and retrieval parameters, are detailed in Supplementary Sections **S-2** and **S-3**.

4 Evaluation

In this section, we outline our approach to assessing the performance of `HARMON-E` in the extraction of oncology variables from clinical notes. Our evaluation targets three core objectives:

1. **Entity-Level Recall:** Confirming that the pipeline correctly identifies all entities recorded in the ground truth.
2. **Precision:** Ensuring the pipeline does not introduce false positives, i.e., entities not present in the ground truth.
3. **Attribute-Level Accuracy:** Verifying that all attribute values match the ground truth for each correctly extracted entity.

Together, these objectives confirm that the pipeline not only captures all relevant information but does so accurately at both the entity and attribute levels.

4.1 Entity Alignment Methods

After running the pipeline, we must determine whether the extracted entities correspond to those in the ground truth data. This alignment step is crucial for measuring entity-level recall, precision, and attribute-level accuracy. Specifically, we employ two methods: Root-based alignment and Weighted alignment.

Fig. 5: **Overview of Two Entity Alignment Methods. (A) Root-Based Alignment:** An alignment is established only if both entities share the same *root attribute* (e.g., "`medication`" with value "Trastuzumab"), making other attributes (such as start dates) irrelevant for basic alignment. **(B) Weighted Alignment:** Each attribute (e.g., `surgery_type`, `surgery_date`, `body_site`) contributes a partial score based on a predefined weight. If the sum of matching attributes meets or exceeds a threshold (e.g., 0.9), the ground-truth and predicted entities are aligned.

4.1.1 Root-Based Entity Alignment

For certain entity types, a specific attribute—the *root attribute*—uniquely identifies that entity. We thus enforce the condition that if these root attributes do not match, the entities cannot be aligned.

$$\text{align}_{\text{root}}(e_1, e_2) = \begin{cases} 1 & \text{if root}(e_1) = \text{root}(e_2), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Root-based alignment applies to entities whose identity is fixed by a single field (Table 2; e.g., `biomarker_tested` for Biomarker Summary, `medication` for Cancer Related Medication).

4.1.2 Weighted Entity Alignment

For entities where the root attribute may be ambiguous or prone to lexical variation, we adopt a weighted alignment strategy. Each attribute is assigned a weight reflecting its importance in identifying that entity. The alignment score between a ground truth entity e_1 and an extracted entity e_2 is computed by:

Table 2: Alignment schemes by entity (anchors and implications).

Entity	Scheme	Anchor / decisive fields	Implication
Biomarker Summary	Root	`biomarker_tested`	Different biomarkers never align even if dates match (e.g., `BRAF` vs `NRAS`).
Cancer Related Medication	Root	`medication`	Drug names must match; other fields are irrelevant for the root check (e.g., `Trastuzumab` vs `Paclitaxel` do not align).
Cancer Related Surgery	Weighted	`surgery_date`, `body_site` > `surgery_type`	Decisive fields dominate; lexical variants of type may still align (e.g., "wide local excision" vs "re-excision").
Staging	Weighted	`stage_date`, `stage_value`, `stage_type`	Prioritize date and value; resolves minor notation differences (e.g., `pT2N0M0` vs `pT2 N0 M0` align; `cT2N0M0` vs `pT2N0M0` do not).

$$\text{align}_{\text{weighted}}(e_1, e_2) = \sum_{i=1}^{k} w_i \cdot \text{match}(a_{1i}, a_{2i}), \tag{2}$$

where w_i is the importance weight of attribute i, and $\text{match}(a_{1i}, a_{2i})$ indicates a match of attribute values. A threshold τ then determines alignment:

$$\text{aligned}(e_1, e_2) = \begin{cases} 1 & \text{if } \text{align}_{\text{weighted}}(e_1, e_2) \geq \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For entities with lexical variation, we use weighted alignment: attributes receive weights under the guidance of oncology experts, with higher weights on clinically decisive fields (Table 2; e.g., `surgery_date`, `body_site`) and lower weights on descriptive fields (e.g., `surgery_type`). The pairwise score is Eq. (2); alignment holds when it exceeds a threshold τ.

4.1.3 Entity Alignment Process

The entity alignment process follows specific constraints regardless of whether root-based or weighted alignment is employed. The fundamental requirement is the uniqueness constraint, which ensures that each ground truth entity aligns with at most one extracted entity, and vice versa. This one-to-one mapping is essential to prevent artificial inflation of recall metrics.

Once entities are aligned, the process identifies a driver attribute that serves as the anchor for computing entity-level recall and precision metrics. The selection of this driver attribute depends on the alignment method used. In root-based alignment, the root attribute itself—such as `biomarker_tested` or `medication`—naturally serves as the driver. For weighted alignment, the system selects the attribute that received the highest weight assignment as the driver attribute. This driver attribute then becomes the primary reference point for evaluating how well the extraction system captures the essential characteristics of each entity.

4.2 Manual Evaluation

Beyond automated metrics, we developed a manual evaluation protocol to assess real-world pipeline performance through clinical expert review. This approach acknowledges that divergences between model outputs and ground truth may stem from either genuine model errors or inconsistencies in the original manual abstractions.

Given the resource-intensive nature of comprehensive manual review, we implemented a strategic sampling method that prioritizes cases of disagreement. We quantified divergence through a disagreement scoring mechanism, formally defined as DS(p) in Supplementary Methods S4, that counts mismatched attributes between pipeline outputs and ground truth for each patient. Patients were ranked by their disagreement scores, with the top 50 highest-scoring cases per entity type selected for expert review. This targeted approach concentrates human expertise where it provides maximum insight into system performance limitations.

Clinical abstraction specialists with oncology domain expertise reviewed selected patient records following a standardized protocol. The review process minimized bias by presenting complete patient records without indicators of data source (pipeline vs. ground truth). Reviewers classified each extracted item as correct, incorrect, or missing—categories detailed in Supplementary Methods S4 along with the complete evaluation workflow.

Two complementary performance metrics were derived from this classification: an Acceptance Score and a Missing Rate. These metrics, whose mathematical formulations are provided in Supplementary Methods S4, characterize both extraction accuracy and completeness—critical dimensions for clinical deployment.

4.3 Dataset

Dataset Description: Our study draws upon a real-world dataset of 2,250 melanoma patients, each contributing an average of approximately 180 unstructured clinical documents. These documents originated from hundreds of distinct healthcare institutions, reflecting considerable variability in document formatting, language use, and clinical notation styles. In total, the corpus is composed of roughly 50% HTML-based notes exported directly from the EHR systems, with the remaining 50% comprising scanned PDF files. Even though the system was tested on Melanoma patients, the same system can be used to get these results by only modifying the valuesets while using same prompts.

To streamline data preparation, all HTML-based documents were treated as single-page entities, irrespective of their actual text length or content density. In contrast, scanned PDF documents were processed at the physical page level, preserving the pagination structure that closely mirrors real-world clinical workflows. In Figure 6, we plot the frequency distribution of per-patient page counts (combining HTML "pages" and PDF pages) for the testing subset. Notably, the distribution exhibits a wide range, with some patients having as few as tens of pages while others have more than a thousand. This spread underscores the variable nature of oncology documentation: some patients receive only a limited number of reports (e.g., routine follow-up or localized

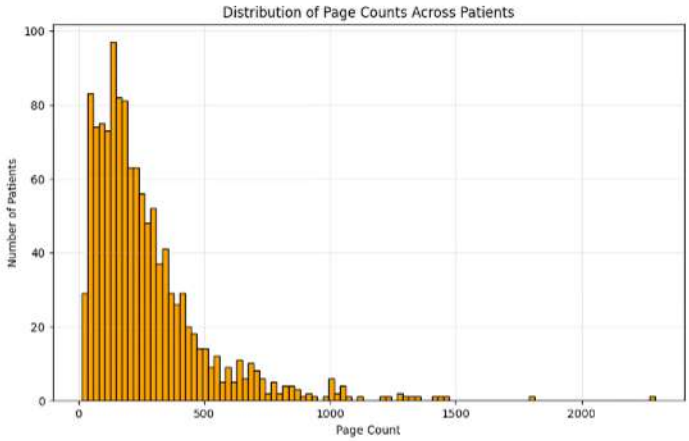

Fig. 6: **Distribution of the Number of Pages per Patient.** Each bar represents the number of patients grouped by their total page count, where HTML-based documents are treated as single pages while PDF pages are counted individually.

treatment), whereas others undergo extensive diagnostic workups, multi-line therapies, and second opinions, producing voluminous records. Such variability in document length and format is precisely what makes the automation of data abstraction both challenging and clinically valuable.

Ground Truth: The dataset was curated by a team of qualified oncology data abstractors (ODS-Cs), employing standardized abstraction guidelines refined over a two-year span. The abstractors recorded over one hundred oncology-specific data elements at the patient level, including (but not limited to) primary cancer diagnoses, treatment regimens (e.g., chemotherapies, targeted therapies, immunotherapies), diagnostic test results (e.g., biomarker findings), disease progression events, and comorbidities. This labor-intensive manual curation process allowed for the creation of a rich, high-fidelity "ground truth" dataset, forming the reference standard against which we benchmark our extraction pipeline.

4.4 Cohort Selection

We analyzed a retrospective melanoma cohort assembled from participating practices under active data-use agreements. The analytic set comprised **2,250** adult patients with histologically confirmed melanoma and available longitudinal unstructured documentation (clinical notes and scanned reports).

Development–test split.

The final eligible population ($n = 2\,250$ unique patients) was randomly partitioned into **50 %** development ($n = 1\,125$) and **50 %** hold-out test ($n = 1\,125$). Randomisation was stratified by (i) health-system cluster, (ii) year of index diagnosis, and (iii) AJCC stage at diagnosis to ensure balanced distribution of site-specific coding styles and disease severity.

The first half was used iteratively for:

- *Prompt Development and Refinement:* Crafting and optimizing large language model (LLM) prompts to ensure comprehensive coverage of target oncology entities and minimize ambiguity in free-text interpretation.
- *Pipeline Configuration and Tuning:* Adjusting individual components—such as retrievers, entity collators, and conflict-resolution logic—to accommodate the diverse formats and terminologies present in both HTML notes and PDF scans.
- *Model Selection:* Comparing the performance of candidate large language models, embedded retrieval methods, and specialized domain rules to choose the pipeline configuration that maximized extraction accuracy while preserving computational efficiency.

Once these elements were established, we used the remaining 1,125 patients as our held-out test cohort for final evaluation. This two-phase approach (development and test) was designed to prevent data leakage and ensure that performance metrics accurately reflect the system's ability to generalize to new patients and institutions.

Table S1 compares key baseline characteristics of the hold-out cohort against the SEER melanoma registry.

5 Results

We summarize in Table 3 the automated evaluation results for each *Entity* averaged over all its attributes. For each entity, we report **Precision**, **Recall**, and **F1**-score, computed according to the alignment methods described in Section 4.1. Furthermore, we present attribute level results in Fig 7. Overall, the pipeline demonstrates robust performance across a wide range of oncology-related attributes, frequently achieving F1-scores above 90%. Below, we highlight notable trends and address a few areas with relatively lower scores.

Table 3: Averaged Results For Different Entities

Entity Name	Precision	Recall	F1
Biomarker	0.9890	0.9722	0.9806
Biopsy	0.8953	0.8631	0.8789
Clinical Trial	1.0000	0.9615	0.9800
Comorbidities	0.8000	1.0000	0.8889
Distant Metastasis	1.0000	0.8182	0.9000
Diagnosis	0.9846	0.9600	0.9722
Family History	1.0000	0.8738	0.9323
Imaging	0.6940	0.9337	0.7962
Medication	0.9722	0.9525	0.9620
Nicotine Use Status	1.0000	1.0000	1.0000
Patient Status	1.0000	1.0000	1.0000
Radiation	0.9481	0.9778	0.9627
Recurrence Status	0.9651	0.7445	0.8405
Staging	0.8771	0.9625	0.9178
Surgery	0.9937	0.9633	0.9782
Surgery Observations	0.9939	0.9899	0.9919

Table 4: Patient-level performance on the hold-out cohort (n = 1 125). Values are macro-averages across the 16 entities; ± denotes 95% confidence intervals were computed using the Bag-of-Little-Bootstraps (BLB): 10 subsets of size m=128 and 100 bootstrap replicates per subset

Configuration	Precision(%)	Recall(%)	F1 (%)
GPT-4o Single-Step	78.2±0.6	70.7±0.7	74.3 ±0.6
HARMON-E (w/o collator)	82.1±0.5	90.7±0.6	86.2±0.5
HARMON-E (full)	**94.1±0.4**	**92.4 ±0.5**	**93.2±0.4**

5.1 Baselines and Ablation Analysis

Robust benchmarking of patient-level extraction is challenging because most prior work in clinical NLP reports *note-* or *sentence-level* metrics. Our gold standard dataset enables the **patient-level** validation, which is closer to real-world use cases. We constructed three reference configurations that satisfy the same output schema and unit-of-analysis:

1. **GPT-4o Single-Step.** All notes for a patient are concatenated (truncated to 32 k tokens) and passed to GPT-4o (June-2024 snapshot) with a single prompt requesting the full JSON schema. No retrieval, no self-reflection, no collator.
2. **No-Collator Ablation.** Identical to HARMON-E except consolidation, validation, and dependency-resolution rules are disabled; LLM generations are used "as-is."
3. **HARMON-E.** Full HARMON-E pipeline with distinct configuration for each entity

All baselines were run on the same 1 125-patient hold-out cohort. Metrics are macro-averaged across the 16 entities, 95% confidence intervals were computed using the Bag-of-Little-Bootstraps (BLB) due to compute constraints: 10 subsets of size m=128 ($\approx N^{0.7}$) and 100 bootstrap replicates per subset with multinomial weights to size N; percentile intervals aggregated across subsets.

Table 4 presents precision, recall and F1. HARMON-E exceeds the GPT-4o single-step by +20 percentage points (pp) in F1 ($p < 10^{-3}$), driven mainly by improvements in date-sensitive attributes (Medication +28.6 pp). Disabling the collator degrades macro-F1 by 6.3 pp, confirming the value of our dependency-aware post-processing.

The GPT-4o baseline required a ∼5× larger average prompt (29.4 k vs 6.1 k tokens per patient) and 11.2× higher inference-time latency (median = 94 s vs 8.4 s). Token savings are principally from targeted retrieval.

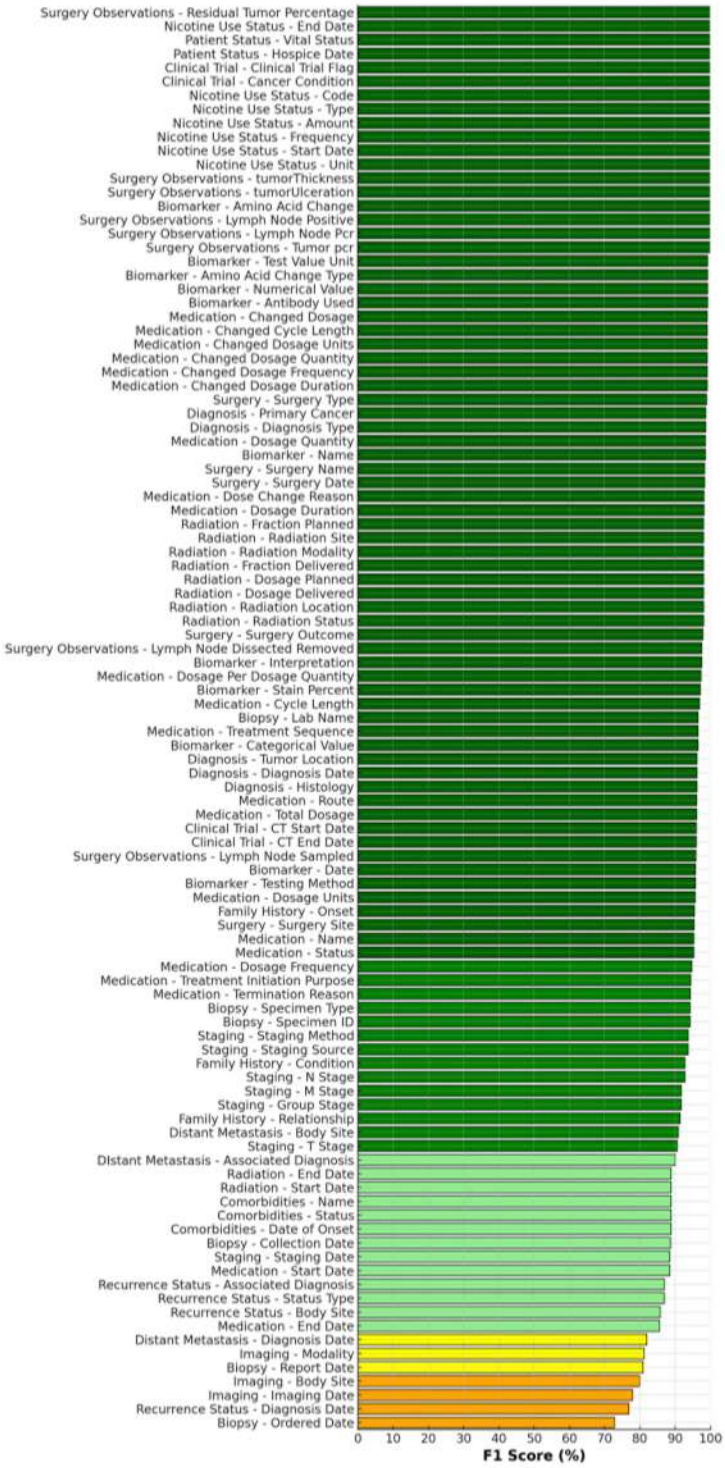

Fig. 7: F1-Scores of the evaluated attributes

5.2 Entity-Level Performance

Diagnosis-Related Entities

For `Diagnosis` (`diagnosisType`, `primaryCancer`, `diagnosisDate`, `tumorLocation`, `histology`), the pipeline achieves F1-scores of 95–98%. Diagnosis dates are occasionally challenging, but the model still maintains over 95% F1, a testament to the pipeline's date normalization and validation steps.

The `Biomarker` entity, including attributes such as `name`, `testingMethod`, `date`, `interpretation`, `categoricalValue`, and `numericalValue`, consistently shows F1-scores in the mid-to-upper 90%s. Amino acid change details (`aminoAcidChange`, `aminoAcidChangeType`) also approach or exceed 98%. This underscores the pipeline's robustness in capturing structured molecular test findings from sometimes lengthy pathology or genetic reports.

TNM staging attributes (`tStage`, `nStage`, `mStage`, `groupStage`) have F1-scores generally between 90% and 95%. `stagingDate`, however, is relatively lower (88.42% F1). We observe that temporal references to staging can appear in summary paragraphs or at multiple time points, leading to partial confusion.

`Recurrence Status` attributes (`statusType`, `associatedDiagnosis`, `diagnosisDate`, `bodySite`) exhibit moderate-to-high performance, with `statusType` reaching 86.87% F1 and `diagnosisDate` at 76.77%. `Distant Metastasis` attributes (e.g., `associatedDiagnosis`, `diagnosisDate`, `bodySite`) consistently outperform recurrence, hovering near 90%. The slight drop in `diagnosisDate` for recurrence stems from complexities in distinguishing the actual date of recurrence which can be inferred from imaging, biopsy or even clinician judgement. If the margin of error is increased to +-14 days, the system achieves $\approx$ 89% F-1 score.

Treatment-Related Entities

The `Cancer Related Medication` entity exhibits consistently strong results, with many attributes (`name`, `treatmentSequence`, `route`, `dosageQuantity`, etc.) achieving F1-scores of 95% or higher. Several fields (`changedDosageQuantity`, `changedDosage`, `changedDosageUnits`, and related attributes) reach or exceed 99% F1, indicating that the pipeline accurately captures even nuanced treatment modifications.

By contrast, date-related attributes (e.g., `startDate`, `endDate`) exhibit slightly lower performance. For instance, `endDate` achieves an F1-score of 85.47%. While still high, This aligns with known difficulties in extracting date information from unstructured text, potentially due to ambiguous or partial documentation. If the margin of error is increased to +-7 days, the system achieves $\approx$ 92% highlighting the ability of the system to land at a nearby date, even if not completely correct. This is due to the difficulties in analyzing the date of the last dosage of the medication which is very often derived and not explicitly stated.

The pipeline excels in `Surgery`, where `surgeryType`, `surgeryName`, `surgeryDate`, and `surgeryOutcome` all exceed 95% F1, indicating robust detection of procedural information. `surgerySite` stands at 95.34% F1, which still signifies strong performance.

Under `Surgery Observations`, most attributes (e.g., `tumorThickness`, `tumorPCR`, `lymphNodePCR`) are extracted with near-perfect precision and recall. Even for more

complex fields like `lymphNodeSampled` ($\sim$ 96% F1), the pipeline demonstrates minimal error rates, suggesting that specialized prompts for pathology details are effective.

`Radiation` attributes (`radiationModality`, `radiationSite`, `dosageDelivered`, etc.) frequently approach 98–99% in F1, demonstrating high precision and recall. The date-related fields (`startDate`, `endDate`) again pose slight challenges (both at 88.89% F1), but remain in a range considered acceptable for large-scale automated extractions.

Other Entities

The attributes `name`, `status`, `dateOfOnset` under the `Comorbidities` entity exhibit solid performance near 88.89% F1. Although lower than some other entities, the pipeline still captures the majority of comorbidity data correctly despite variability in how clinical notes reference chronic conditions.

`Family History` attributes—`relationship`, `condition`, and `onset`—achieve F1-scores in the low-to-mid 90% range. The pipeline occasionally struggles with ambiguous family relationships or missing explicit onset dates, but it still yields a high level of accuracy overall.

5.3 Summary of Automated Evaluation

In conclusion, the 73 out of 103 *Entity–Attribute* pairs surpass 90% F1-score, highlighting the effectiveness of the agentic multi-step approach outlined in Section 3.3. Entities with more complex or date-dependent attributes (e.g., certain `Biopsy` fields, some `Imaging` details) see moderate performance dips, underscoring the inherent challenges of inconsistent or ambiguous date references in clinical text. Nonetheless, these results illustrate that the proposed `HARMON-E` pipeline provides robust and comprehensive extraction of oncology-specific data, setting the stage for efficient downstream curation and analysis.

6 Discussion

Our proposed `HARMON-E` pipeline systematically addresses the challenge of extracting fine-grained oncology data from heterogeneous EHR notes. The strong performance reported in Table 3, with a approximately 75% *Entity–Attribute* pairs exceeding 90% F1-score, underscores the effectiveness of a multi-step, agentic approach to clinical information extraction. Here, we examine the factors behind these results, the limitations inherent to current methods, and potential avenues for future refinement.

6.1 Robustness of the Agentic Approach

The crux of our pipeline's success lies in the hierarchical and *modular* decomposition of complex tasks. Rather than forcing a single model to infer all oncology attributes from large, fragmented documents, we divide the workload into discrete steps such as *retrieval*, *LLM-based synthesis*, and *collation*. This strategy closely mimics human workflows where different abstraction tasks (e.g., medications vs. biomarkers) demand different retrieval contexts and domain-specific prompts.

The high scores across most medication attributes, notably around dosage adjustments and medication status, point to the efficacy of using agentic iterative prompts. In addition, the pipeline's ability to detect multiple lines of therapy and handle partial or conflicting data suggests that multi-step retrieval—where each pass narrows the focus—reduces confusion that might otherwise arise in single-shot extractions.

6.2 Why not ClinicalBERT/SciBERT?

Those models (and their fine-tuned variants) emit token-level BIO tags or short relation triples per document; transforming such outputs into patient-level, longitudinal abstractions requires bespoke heuristics (e.g., cross-document clustering, conflict resolution, temporal alignment) that vary across data partners and are *not* publicly standardised. Any comparison would therefore conflate intrinsic model performance with pipeline engineering choices and is unlikely to offer actionable insights. Despite that, we ran additional experiments with self-designed postprocessing script and have summarized the results in **Supplementary S7**.

6.3 Manual Adjudication Results

To complement the automated evaluations, we integrated the pipeline outputs into a data curation platform (as shown in Fig. 8) and conducted a manual review, or *adjudication*, of 13,609 individual data points spanning `Radiation`, `Surgery`, `Medication`, `Staging`, and `Diagnosis` entities. As shown in Figure 9 (a representative dashboard excerpt), reviewers with oncology data abstraction expertise assessed each extracted entry in a patient's record. Specifically, they had three options for each item:

- **Approve (Correct)**: The pipeline output required no edits.
- **Edit (Incorrect)**: The pipeline output existed but needed to be modified (e.g., a wrong date or attribute value).
- **Add (Missing)**: The pipeline missed an item or field that reviewers deemed relevant.

The adjudication process demonstrated that **94.1%** of extracted items were *directly approved* without changes. This high approval rate highlights the real-world robustness of the pipeline. In contrast, **4.2%** of data points were categorized as *missing entries*, signifying that the pipeline had overlooked relevant details. An additional **1.7%** required *edits* to correct partially incorrect information.

Figure 9 also breaks down the adjudication outcomes by entity type. `Medication`, `Radiation`, and `Diagnosis` exhibit particularly high direct approvals (routinely over 90%). Where missing data did occur, it was often tied to ambiguous dates or undocumented changes. Edits tended to involve small discrepancies such as inaccurate route or dosage in medication entries or minor staging detail mismatches.

6.4 Challenges with Date Fields and Context Ambiguity

Despite the overall strong performance, several date-related attributes (`startDate`, `endDate`, `stagingDate`, etc.) manifest lower recall or precision relative to simpler

Fig. 8: Manual adjudication interface for expert validation of HARMONE extractions. The data curation platform displays extracted clinical entities in a three-panel layout: (left) hierarchical list of all extracted entities organized by type (Diagnosis, Staging, Distant Metastasis, etc.) with entity counts and completion indicators; (center) detailed view of the selected entity showing all extracted attributes, with the example showing a Primary Melanoma diagnosis dated February 11, 2019 (Deidentified/Shifted date), at anatomical location "Upper Extremity/Shoulder"; (right) source clinical note with automatic highlighting of the relevant text passage from which the information was extracted, enabling traceable validation. The highlighted yellow section shows the exact source text supporting the extraction, while the orange boxes indicate the specific data points under review. Clinical experts can directly approve extractions using the "Mark as complete" button, edit incorrect values inline, or add missing information not captured by the automated pipeline. This interface facilitated the review of 13,609 data points across 50 high-disagreement patients, achieving a 94.1% direct approval rate as described in Section 6.3.

categorical fields. This is partly because real-world oncology notes often contain multiple, sometimes conflicting date references: for instance, the medication's original *prescription date* may differ from the *administration start date*, and not all providers consistently document time points in an unambiguous format.

Similarly, entities like `Imaging` can appear sporadically across radiology reports, discharge summaries, and referral letters, each containing references to different imaging sessions. The pipeline may over-detect or conflate sessions if a single retrieval chunk contains synonyms or abbreviations pointing to multiple studies. Further refinement of chunking strategies and more sophisticated context tracking could address these limitations, for example by leveraging specialized date resolution modules or advanced temporal reasoning prompts within the LLM.

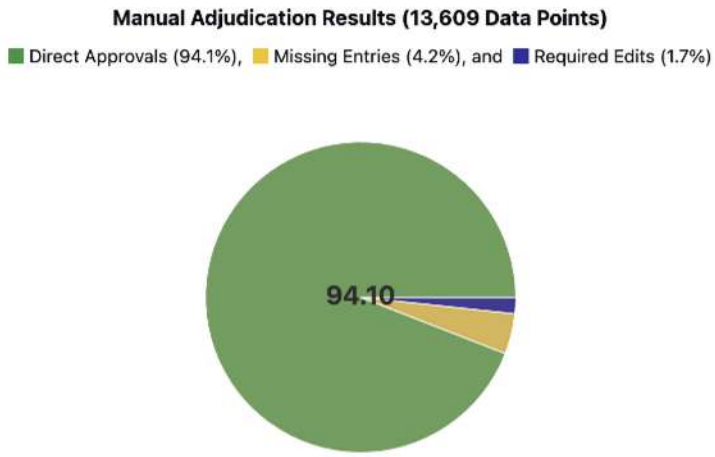

Fig. 9: HARMON-E demonstrates high direct-approval rates (94.1%), with low missing (4.2%) and required-edit (1.7%) rates further corrobarting the automated metrics.

6.5 Implications for Clinical Research and Real-World Evidence

The breadth and depth of high-accuracy extraction demonstrated by HARMON-E show promise for scaling up real-world data (RWD) curation in oncology. Automating the abstraction of TNM staging, biomarker statuses, and multi-line therapies has the potential to expedite clinical trial matching, pharmacovigilance, and precision medicine initiatives. Indeed, if integrated into routine EHR workflows, such an agentic pipeline could accelerate retrospective analyses of large cancer cohorts while maintaining data fidelity akin to manual chart reviews.

Moreover, by capturing a broad range of attributes (e.g., `familyHistory`, `comorbidities`, `nicotineUseStatus`), the system can support comprehensive epidemiological studies that hinge on robust phenotypic representations. The precision seen in medication changes (e.g., `dosageQuantity`, `doseChangeReason`) also paves the way for granular analysis of treatment patterns over time.

7 Conclusion

We presented HARMON-E, a modular, agentic framework for extracting structured oncology data from heterogeneous, often voluminous EHR notes. By combining domain-aware retrieval, large language models for context-sensitive synthesis, and robust collation, HARMON-E demonstrates state-of-the-art performance on a comprehensive range of oncology related clinical attributes. Automated evaluation highlights F1-scores exceeding 90% for most attributes, and our multi-step strategy effectively addresses the challenges of conflicting or fragmented documentation.

In automating the generation of curated, patient-level oncology data, HARMON-E offers an impactful tool for both clinical research and operational workflows. Our

ongoing efforts focus on extending the pipeline's applicability to other solid tumors, refining date extraction modules, and incorporating active learning mechanisms for continuous improvement. Ultimately, by reducing the resource-intensive burden of manual abstraction, this framework stands to accelerate real-world evidence generation and support more personalized cancer care at scale.

Declarations

Data Availability

This study is retrospective, and no new data was generated. Due to access restrictions and the risk of re-identification, study data will not be shared externally.

Code Availability

Pipeline configuration files, pseudo-code for helper scripts and representative prompt templates are released in the Supplementary. The authors agree to provide code snippets at a reasonable request by non-competing entity.

Acknowledgements

The authors gratefully acknowledge the oncology data abstraction and informatics teams at Ontada for their assistance in data access and clinical validation. The authors also appreciate the engineering and data operations teams at Triomics for their work in implementing the HARMON-E infrastructure, including large-scale data ingestion, retrieval optimization, and validation pipelines.

Author Contributions

Hrituraj Singh and Yanshan Wang contributed to the conceptualization of the study. Methodology was developed by Hrituraj Singh, and Shashi kant Gupta. Software and implementation were carried out by Shashikant Gupta, Jerrin John Thomas, and Arijeet Pramanik.. All authors reviewed and approved the final manuscript.

Funding

This research was conducted with institutional support from Triomics and Ontada. No external funding was specifically dedicated to this study.

Competing Interests

Hrituraj Singh, Shashikant Gupta, Arijeet Pramanik, Regina Schwind and Jerrin John Thomas are employees of Triomics, Inc.

Lauren Wiener, Avi Raju, Jeremy Kornbluth, and Zhaohui Su are employees of Ontada LLC.

Yanshan Wang declares no competing interests.

References

[1] Uzuner, O., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association **18**, 552–556 (2011)

[2] Eftimov, T., Koroušić Seljak, B., Korošec, P.: A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. PloS one **12**(6), 0179488 (2017)

[3] Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. Journal of the American Medical Informatics Association **17**(5), 507–513 (2010)

[4] Savova, G.K., Tseytlin, E., Finan, S., Castine, M., Miller, T., Medvedeva, O., Harris, D., Hochheiser, H., Lin, C., Chavan, G., Jacobson, R.S.: DeepPhe: A natural language processing system for extracting cancer phenotypes from clinical records. Cancer Research **77**(21), 115–118 (2017) https://doi.org/10.1158/0008-5472.CAN-17-0615

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

[6] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[7] Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. (2019)

[8] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[9] Preston, S., Wei, M., Rao, R., Tinn, R., Usuyama, N., Lucas, M., Gu, Y., Weerasinghe, R., Lee, S., Piening, B., Tittel, P., Valluri, N., Naumann, T., Bifulco, C., Poon, H.: Toward structuring real-world data: Deep learning for extracting oncology information from clinical text with patient-level supervision. Patterns **4**(4), 100726 (2023) https://doi.org/10.1016/j.patter.2023.100726

[10] Lin, H., Ginart, J.B., Chen, W., Interian, Y., Gong, H., Liu, B., Upadhaya, T., Lupo, J., Hong, J., Braunstein, S., et al.: Oncobert: building an interpretable transfer learning bidirectional encoder representations from transformers framework for longitudinal survival prediction of cancer patients (2023)

[11] Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. In: EMNLP (2019)

[12] OpenAI: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[13] OpenAI: Introducing GPT – 5. Accessed: 2025-10-18 (2025). https://openai.com/index/introducing-gpt-5/

[14] Bhattarai, K., Oh, I.Y., Sierra, J.M., Tang, J., Payne, P.R.O., Abrams, Z., Lai, A.M.: Leveraging gpt-4 for identifying cancer phenotypes in electronic health records: A performance comparison between gpt-4, gpt-3.5-turbo, flan-t5, llama-3-8b, and spacy's rule-based and machine learning-based methods. JAMIA Open **7**(3), 060 (2024) https://doi.org/10.1093/jamiaopen/ooae060

[15] Abedian, S., Sholle, E.T., Adekkanattu, P.M., Cusick, M.M., Weiner, S.E., Shoag, J.E., Hu, J.C., Campion, T.R.J.: Automated extraction of tumor staging and diagnosis information from surgical pathology reports. JCO Clinical Cancer Informatics **5**, 1054–1061 (2021) https://doi.org/10.1200/CCI.21.00065

[16] Kefeli, J., Berkowitz, J., Acitores Cortina, J.M., Tsang, K.K., Tatonetti, N.P.: Generalizable and automated classification of TNM stage from pathology reports with external validation. Nature Communications **15**(1), 8916 (2024) https://doi.org/10.1038/s41467-024-53190-9

[17] Gauthier, M.-P., Law, J.H., Le, L.W., Li, J.J.N., Zahir, S., Nirmalakumar, S., Sung, M., Pettengell, C., Aviv, S., Chu, R., Sacher, A., Liu, G., Bradbury, P., Shepherd, F.A., Leighl, N.B.: Automating access to real-world evidence. JTO Clinical and Research Reports **3**(6), 100340 (2022) https://doi.org/10.1016/j.jtocrr.2022.100340

[18] Pironet, A., Poirel, H.A., Tambuyzer, T., De Schutter, H., Walle, L., Mattheijssens, J., Henau, K., Van Eycken, L., Van Damme, N.: Machine learning-based extraction of breast cancer receptor status from bilingual free-text pathology reports. Frontiers in Digital Health **3**, 692077 (2021) https://doi.org/10.3389/fdgth.2021.692077

[19] Zeng, J., Banerjee, I., Henry, A.S., Wood, D.J., Shachter, R.D., Gensheimer, M.F., Rubin, D.L.: Natural language processing to identify cancer treatments with electronic medical records. JCO Clinical Cancer Informatics **5**, 379–393 (2021) https://doi.org/10.1200/CCI.20.00173

[20] Meng, W., Mosesso, K.M., Lane, K.A., Roberts, A.R., Griffith, A., Ou, W., Dexter, P.R.: An automated line-of-therapy algorithm for adults with metastatic non–small cell lung cancer: Validation study using blinded manual chart review. JMIR Medical Informatics **9**(10), 29017 (2021) https://doi.org/10.2196/29017

[21] Friedman, C., Shagina, L., Lussier, Y.: Automated encoding of clinical documents based on natural language processing. Journal of the American Medical Informatics Association **11**, 392–402 (2004)

[22] Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G.: A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of Biomedical Informatics **34**, 301–310 (2001)

[23] Uzuner, O., Bodnari, A., Shen, S., Forbush, T., Pestian, J., South, B.R.: Evaluating the state of the art in coreference resolution for electronic medical records. Journal of the American Medical Informatics Association **19**(5), 786–791 (2012)

[24] Sohn, S., Clark, C., Halgrim, S.R., Murphy, S.P., Chute, C.G., Liu, H.: Medxn: an open source medication extraction and normalization tool for clinical text. Journal of the American Medical Informatics Association **21**(5), 858–865 (2014)

[25] Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., Xu, H.: Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association **25**(3), 331–336 (2018)

[26] Tibbo, M.E., Wyles, C.C., Fu, S., Sohn, S., Lewallen, D.G., Berry, D.J., Maradit Kremers, H.: Natural language processing of radiology reports for identification of skeletal site-specific fractures. BMC medical informatics and decision making **19**(1), 73 (2019)

[27] Moon, S., Liu, S., Scott, C.G., Samudrala, S., Abidian, M.M., Geske, J.B., Noseworthy, P.A., Shellum, J.L., Chaudhry, R., Ommen, S.R., *et al.*: Automated extraction of sudden cardiac death risk factors in hypertrophic cardiomyopathy patients by natural language processing. International journal of medical informatics **128**, 32–38 (2019)

[28] Zhang, R., Pakhomov, S., Melton, G.B.: Automated identification of relevant new information in clinical narrative using language modeling. Journal of biomedical informatics **49**, 255–261 (2014)

[29] Murff, H.J., FitzHenry, F., Matheny, M.E.: Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA **306**(8), 848–855 (2011)

[30] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical information extraction applications: a literature review. Journal of biomedical informatics **77**, 34–49 (2018)

[31] Peng, N., Zhang, Y., Jin, W.: Transfer learning-based approach for clinical named entity recognition with limited training data. Journal of Biomedical Informatics

95, 103218 (2019)

[32] Si, Y., Sun, Y., Li, H., Zhang, Z.: Deep learning for clinical information extraction: A survey. Journal of Biomedical Informatics **94**, 103196 (2019)

[33] Peng, Y., Chen, Q., Lu, Z.: An empirical study of multi-task learning on bert for biomedical text mining. arXiv preprint arXiv:1908.11692 (2019)

[34] Ding, Y., Shen, D., Huang, Z., Chen, W.: Cogran: A hierarchical transformer model for multi-document summarization. arXiv preprint arXiv:2004.07840 (2020)

[35] Fabbri, A., Fan, P., Gupta, R., Sedoc, J., Khalid, B.: Multi-document summarization. arXiv preprint arXiv:1908.08376 (2019)

[36] Adamson, B., Waskom, M., Blarre, A., Kelly, J., Krismer, K., Nemeth, S., Gippetti, J., Ritten, J., Harrison, K., Ho, G., Linzmayer, R., Bansal, T., Wilkinson, S.C., Amster, G., Estola, E., Benedum, C.M., Fidyk, E., Estévez, M., Shapiro, W., Cohen, A.B.: Approach to machine learning for extraction of real-world data variables from electronic health records. Frontiers in Pharmacology **14**, 1180962 (2023) https://doi.org/10.3389/fphar.2023.1180962

[37] Kung, A.Y., Li, J.X., Liu, M.Y., Yuan, M., Luo, Y., Zhang, Y., Zhang, Y., Lu, Y., Li, J., Sun, Y., et al.: Performance of a large language model at medical question answering. arXiv preprint arXiv:2303.13257 (2023)

[38] Agrawal, A., Chen, Y., Zhang, Y., Li, Y., Zhang, Y., Lu, Y., Li, J., Sun, Y., Li, J.X., Kung, A.Y.: Large language models for clinical decision support: A systematic review. arXiv preprint arXiv:2303.13258 (2023)

[39] Wong, C., Preston, S., Liu, Q., Gero, Z., Bagga, J., Zhang, S., Jain, S., Zhao, T., Gu, Y., Xu, Y., et al.: Universal abstraction: Harnessing frontier models to structure real-world data at scale. arXiv preprint arXiv:2502.00943 (2025)

[40] Porter, R., Diehl, A., Pastel, B., Hinnefeld, J.H., Nerenberg, L., Maung, P., Kerbrat, S., Hanson, G., Astorino, T., Tarsa, S.J.: Llmd: A large language model for interpreting longitudinal medical records. arXiv preprint arXiv:2410.12860 (2024)

[41] Keloth, V.K., Selek, S., Chen, Q., Gilman, C., Fu, S., Dang, Y., Chen, X., Hu, X., Zhou, Y., He, H., Fan, J.W., Wang, K., Brandt, C., Tao, C., Liu, H., Xu, H.: Social determinants of health extraction from clinical notes across institutions using large language models. npj Digital Medicine **8**, 287 (2025)

[42] Fu, S., Jia, H., Vassilaki, M., Keloth, V.K., Dang, Y., Zhou, Y., Garg, M., Petersen, R.C., St Sauver, J., Moon, S., Wang, L., Wen, A., Li, F., Xu, H., Tao, C., Fan, J., Liu, H., Sohn, S.: Fedfsa: Hybrid and federated framework for functional status ascertainment across institutions. Journal of Biomedical Informatics

152, 104623 (2024)

[43] Kaster, L., Hillis, E., Oh, I.Y., Aravamuthan, B.R., Lanzotti, V.C., Vickstrom, C.R., Brain Gene Registry Consortium, Gurnett, C.A., Payne, P.R.O., Gupta, A.: Automated extraction of functional biomarkers of verbal and ambulatory ability from multi-institutional clinical notes using large language models. Journal of Neurodevelopmental Disorders **17**, 24 (2025)

[44] Stuhlmiller, T.J., Rabe, A., Rapp, J., Manasco, P., Awawda, A., Kouser, H., Salamon, H., Chuyka, D., Mahoney, W., Wong, K.K., et al.: A scalable method for validated data extraction from electronic health records with large language models. medRxiv, 2025–02 (2025)

[45] Yao, S., Zhao, J., Yu, D.: React: Synergizing reasoning and acting in language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[46] Wei, J., Xiong, D., Ma, Z., Huang, D., Mihaylov, T., Gong, C., Zhang, S., Yang, W., Wang, X., Liu, Y., et al.: Chain-of-thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 (2022)

[47] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: ACL 2023 (2022)

[48] Schick, T., Hoffmann, Y., Borgeaud, S., Hennigan, T., Wojak, G., Rhodes, A.R., Zemel, R., Sutskever, I.: Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.03328 (2023)

[49] Nakano, R., Choi, Y., Lee, K., Choi, J., Lee, K.: Webgpt: Browser as an oracle. arXiv preprint arXiv:2112.09622 (2021)

[50] Ferber, D., Nahhas, O.S.M.E., Wölflein, G., Wiest, I.C., *et al.*: Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nature Cancer **6**, (2025) https://doi.org/10.1038/s43018-025-00991-6

[51] Sandhu, A., Kim, E.J., colleagues: Open-source modular ai coupled with agentic ai for comprehensive breast cancer note generation and guideline-directed treatment comparison. In: Proceedings of the 2025 American Society of Clinical Oncology Annual Meeting. Journal of Clinical Oncology, vol. 43, p. 13685 (2025). https://doi.org/10.1200/JCO.2025.43.16_suppl.e13685

[52] Zhang, R., Elhamod, M.: Data-to-dashboard: Multi-agent llm framework for insightful visualization in enterprise analytics. arXiv preprint **arXiv:2505.23695** (2025)

[53] Zhou, S., Wang, N., Wang, L., Liu, H., Zhang, R.: Cancerbert: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. Journal of the American Medical Informatics Association **29**(7), 1208–1216 (2022) https://doi.org/10.1093/jamia/ocac040

[54] Hochheiser, H., Finan, S., Yuan, Z., Durbin, E.B., Jeong, J.C., Hands, I., Rust, D., Kavuluru, R., Wu, X.-C., Warner, J.L., Savova, G.K.: Deepphe-cr: Natural language processing software services for cancer registrar case abstraction. JCO Clinical Cancer Informatics (2023) https://doi.org/10.1200/CCI.23.00156

[55] Goryachev, S.D., *et al.*: Natural language processing algorithm to extract multiple myeloma stage from oncology notes in the veterans affairs healthcare system. JCO Clinical Cancer Informatics **8**, 2300197 (2024) https://doi.org/10.1200/CCI.23.00197

[56] Ling, A.Y., Kurian, A.W., Caswell-Jin, J.L., Sledge, G.W., Shah, N.H., Tamang, S.R.: Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open **2**(4), 528–537 (2019) https://doi.org/10.1093/jamiaopen/ooz040

[57] Zhu, M., Lin, H., Jiang, J., *et al.*: Large language model trained on clinical oncology data predicts cancer progression. npj Digital Medicine **8**, 397 (2025) https://doi.org/10.1038/s41746-025-01780-2

[58] Dagli, M.M., Ghenbot, Y., Ahmad, H.S., *et al.*: Development and validation of a novel ai framework using nlp with llm integration for relevant clinical data extraction through automated chart review. Scientific Reports **14**, 26783 (2024) https://doi.org/10.1038/s41598-024-77535-y

[59] Zeng, Z., Deng, Y., Li, X., Naumann, T.J., Luo, Y.: Natural language processing for ehr-based computational phenotyping. IEEE/ACM Transactions on Computational Biology and Bioinformatics **16**(1), 139–153 (2019) https://doi.org/10.1109/TCBB.2018.2810898

Supplementary Information

S-1 Environment and Dependencies

```
Language models :   gpt-4o-2024-05-13, Qwen2-7B, o1-preview-2024-09-12
Embeddings      :   text-embedding-3-large
Vector store    :   FAISS v1.8.1  (Inner Product)
Python          :   3.11.4
openai          :   1.26.0
pydantic        :   2.8.1
tiktoken        :   0.6.0
rich            :   13.7.0
faiss-cpu       :   1.8.1
```

All LLM calls were executed on an Azure OpenAI private endpoint inside a HIPAA-compliant VNet; no PHI left the secure boundary.

S-2 Prompt Templates

Stage	Representative Template Fragment[1]

Biomarker Single-Step	`system: You are an expert molecular pathologist ...`
	`user : <<SNIPPET>>`
	`assistant: Return JSON {biomarker_tested,...,test_date}.`

Medication Multi-Step–1 (enumerate drugs)	`system: You are an oncology pharmacist ...` `List all distinct systemic agents mentioned in the snippet.`
Medication Multi-Step–2 (attributes per drug)	`system: For the drug "{{DRUG}}" extract` `{start_date,end_date,route,dose_change_reason}.`
Collation / Deduplication	`system: Merge entries if medication + start_date` `differ by <=7 days OR share identical event_id.`
Self-reflection	`assistant: If any required attribute is NULL,` `re-read context and attempt one retry.`

To ensure methodological transparency while respecting intellectual property considerations, abbreviated versions of the prompts used in this study are provided in the supplementary materials. These abbreviated prompts capture the essential structure and intent of the interactions with the large language model, sufficient for understanding our methodology. Each entity in our synthesis pipeline underwent multiple processing stages, each requiring distinct prompting strategies. Complete prompt templates with full instructions, examples, and specific parameters can be made available to researchers seeking to replicate our findings upon reasonable request and execution of a non-disclosure agreement. Interested parties may contact the corresponding author for access to these materials.

S-3 End-to-End Pipeline Configuration

```
{
  "pipeline_name": "harmon-e_melanoma_v1",
  "entities": [
    {
      "name": "Biomarker",
      "retriever": {
        "type": "vector",
        "embedding_model": "text-embedding-3-large",
        "k": 12,
        "query_template":
          "Find passages describing laboratory or genomic tests for melanoma."
      },
      "synthesizer": {
        "llm": "gpt-4o-2024-05-13",
        "prompt_file": "prompts/biomarker_single_step.txt",
        "max_tokens": 600
      },
      "collator": {
        "rules": ["deduplicate_by_root: biomarker_tested",
                  "prefer_latest: result_date"]
      }
    },

    {
      "name": "CancerRelatedMedication",
      "retriever": {
        "type": "regex+vector",
        "patterns":
          ["(?i)(nivolumab|pembrolizumab|ipilimumab|vemurafenib)"],
        "k": 20
      },
      "synthesizer": [
        {
          "stage": "enumerate",
          "prompt_file": "prompts/medication_stage1_list.txt"
        },
        {
          "stage": "detail",
          "prompt_file": "prompts/medication_stage2_detail.txt",
          "loop_over": "{{ENUMERATED_DRUGS}}"
        }
      ],
      "collator": {
        "rules": [
          "merge_if_name_and_start<=7d",
          "infer_end_date_from_last_administration",
          "set_status_discontinued_if_end_date<today-28d"
        ]
      }
    }
  ],

  "post_processors": [
    "validate_against_schema",
    "iso8601_date_normalizer",
    "convert_units"
  ],

  "evaluation": {
    "alignment_method": "root_or_weighted",
    "metrics": ["precision", "recall", "f1"],
    "date_tolerance_days": 7
```

[1]Temperature = 0, top_p = 0.1 for all calls.

```
  }
}
```

```
"""
run_harmone.py - Minimal driver to execute the JSON pipeline.
"""
import json, pathlib
from harmone.engine import Pipeline   # lightweight wrapper in SI

cfg_path = pathlib.Path("harmon-e_pipeline.json")
pipe     = Pipeline.from_config(json.loads(cfg_path.read_text()))

for patient_dir in pathlib.Path("/data/melanoma_notes/").iterdir():
    result = pipe.run(
        patient_id = patient_dir.stem,
        note_paths = list(patient_dir.glob("*.txt"))
    )
    pipe.save_json(result, out_dir="outputs/")
```

This sample configuration file illustrates the structure of a typical HARMON-E pipeline, showing how different modules can be connected and parameterized to create a complete workflow. The accompanying driver script provides a straightforward example of how to load and execute such a pipeline configuration programmatically.

S-4 Manual Evaluation Protocol

This section provides the detailed methodology for the manual evaluation protocol summarized in Section 4.2.

Disagreement-Based Sampling.

For each patient p, we calculated a Disagreement Score $\mathrm{DS}(p)$ representing the total number of mismatched attributes between pipeline outputs and ground truth:

$$\mathrm{DS}(p) = \sum_{j=1}^{N} \delta_j \tag{1}$$

where $\delta_j = 1$ if attribute j is mismatched and 0 otherwise across N total attributes. Patients were ranked by $\mathrm{DS}(p)$ in descending order, with the top 50 highest-scoring patients selected per entity type for expert review.

Review Categories.

Clinical experts classified each extracted item into one of three categories:

- **Correct**: Pipeline output accurately reflects clinical documentation and requires no modification
- **Incorrect**: Pipeline output contains errors requiring editing or deletion
- **Missing**: Clinically relevant information present in the patient record but absent from pipeline output

Table 1: Step-by-step implementation of the manual evaluation protocol

Step	Action	Implementation Details
1	Calculate Disagreement Score	• Compare each attribute across all entity instances • Count mismatches between pipeline and ground truth • Compute $\mathrm{DS}(p) = \sum_{j=1}^{N} \delta_j$
2	Select Review Cohort	• Sort patients by $\mathrm{DS}(p)$ in descending order • Select top 50 patients per entity type $E \in \mathcal{E}$ • Ensure minimum 5 instances per entity type per patient
3	Conduct Blinded Review	• Present complete patient EHR via curation platform • Remove all indicators of data source (pipeline vs. ground truth) • Expert classifies each item: Correct / Incorrect / Missing • Expert adds any clinically relevant missing information
4	Compute Metrics	• Tally classifications per patient-entity pair • Calculate Acceptance Score (Eq. 2) • Calculate Missing Rate (Eq. 3)

Performance Metrics.

The **Acceptance Score** quantifies the proportion of pipeline-extracted items requiring no modification:

$$\text{Acceptance Score} = \frac{\sum_{p \in P} \sum_{E \in \mathcal{E}} n_{\text{correct}}(p, E)}{\sum_{p \in P} \sum_{E \in \mathcal{E}} n_{\text{extracted}}(p, E)} \tag{2}$$

The **Missing Rate** measures the completeness of extraction:

$$\text{Missing Rate} = \frac{\sum_{p \in P} \sum_{E \in \mathcal{E}} n_{\text{missing}}(p, E)}{\sum_{p \in P} \sum_{E \in \mathcal{E}} [n_{\text{extracted}}(p, E) + n_{\text{missing}}(p, E)]} \tag{3}$$

Variable definitions:

- $P \subseteq \mathcal{P}$: Set of patients selected for review
- $\mathcal{E}$: Set of entity types evaluated
- $n_{\text{correct}}(p, E)$: Count of correct extractions for patient p, entity type E
- $n_{\text{incorrect}}(p, E)$: Count of incorrect extractions
- $n_{\text{missing}}(p, E)$: Count of missing items identified by reviewers
- $n_{\text{extracted}}(p, E) = n_{\text{correct}}(p, E) + n_{\text{incorrect}}(p, E)$: Total extractions

S-5 Dataset Statistics

Table 2: Comparison of baseline characteristics between the SEER 2025 melanoma cohort (2018–2022 diagnoses) and the values reported for the HARMON-E hold-out set. Differences $|\Delta| > 3$ percentage-points are typeset in **bold**. Some numbers are not exact and have been derived from various sources

	Distribution (%)		
Characteristic	**SEER 2025**	**Dataset**	**Δ (pp)**
Median age, years [a]	66 (IQR 59–74)	67	–
Sex – Male	57.7	61.1	+3.4
AJCC Stage I & II	77.0	48.9	**−28.1**
AJCC Stage III	9.5	42.5	**+33.0**
AJCC Stage IV	4.7	1.8	**-2.9**

Notes.
[a] Age summarised as median (inter-quartile range); no Δ computed.
AJCC = American Joint Committee on Cancer; IQR = inter-quartile range; pp = percentage-points.

We extracted reference statistics from SEER public database along with few research papers that summarize melanoma specific statistics. The table 2 summarizes the statistics.

Table 3: Macro-averaged performance of prior baselines on the hold-out melanoma cohort. Not all pipelines could be modified to approximate all data points in the schema

Pipeline	Prec. (%)	Rec. (%)	F1 (%)	Date Err.* (%)
cTAKES 4.0 + rule post-proc	63.7	54.2	58.6	27.8
SciSpacy + CRF aggregation	67.3	61.8	64.4	30.3
ClinicalBERT (fine-tuned, note-level)	66.5	68.9	67.7	26.2

Notes. Prec. = macro-precision; Rec. = macro-recall. (*) Date Err. = percentage of extracted date attributes deviating by $> \pm 14$ d from the reference.

S-6 Comparison with Prior Baseline Systems

To contextualise `HARMON-E`'s performance, we re-implemented three representative extraction pipelines that are commonly cited in clinical-NLP literature. Not all variables were supported - so we just averaged over whatever could be processed using each method. All baselines were executed on the identical 1 125-patient hold-out cohort described in Section 4.4. Because none of the legacy systems natively consolidate information across hundreds of notes, we added a lightweight post-processor that (i) clusters identical concepts within a 14-day window and (ii) propagates the most frequent attribute value. The configurations and macro-level results are summarised in Table 3.