Traj-CoA: Patient Trajectory Modeling via Chain-of-Agents for Lung Cancer Risk Prediction

Sihang Zeng[α] **Yujuan Fu**[α] **Sitong Zhou**[α] **Zixuan Yu**[α] **Lucas Jing Liu**[β] **Jun Wen**[γ]
Matthew Thompson[δ] **Ruth Etzioni**[β] **Meliha Yetisgen**[α]

[α] University of Washington [β] Fred Hutch Cancer Center [γ] Harvard University [δ] Google

Abstract

Large language models (LLMs) offer a generalizable approach for modeling patient trajectories, but suffer from the long and noisy nature of electronic health records (EHR) data in temporal reasoning. To address these challenges, we introduce Traj-CoA, a multi-agent system involving chain-of-agents for patient trajectory modeling. Traj-CoA employs a chain of worker agents to process EHR data in manageable chunks sequentially, distilling critical events into a shared long-term memory module, EHRMem, to reduce noise and preserve a comprehensive timeline. A final manager agent synthesizes the worker agents' summary and the extracted timeline in EHRMem to make predictions. In a zero-shot one-year lung cancer risk prediction task based on five-year EHR data, Traj-CoA outperforms baselines of four categories. Analysis reveals that Traj-CoA exhibits clinically aligned temporal reasoning, establishing it as a promisingly robust and generalizable approach for modeling complex patient trajectories.

1 Introduction

Longitudinal Electronic Health Records (EHRs) provide rich, temporal data for modeling patient trajectories and predicting clinical outcomes [1]. Effective temporal reasoning is critical to unlocking this potential [2]. For instance, tracking a lung nodule's evolution is key to diagnosing cancer [3]. While traditional approaches required complex feature engineering and task-specific models [4, 5, 6], modern Large Language Models (LLMs) promise a more generalizable, zero-shot paradigm for clinical prediction [7, 8, 9]. However, the promise of LLMs is hindered by the unique challenges of EHR data: extremely long patient histories and inherent noisiness of the recorded clinical data [2, 10].

Patient trajectories accumulate multimodal data over years, creating records that often exceed the context windows of LLMs [10]. Even models with large context windows are hampered by the "lost-in-the-middle" problem, where performance degrades on long inputs as they struggle to attend to the middle of a long input sequence [11, 2]. Recent efforts to adapt LLMs for longitudinal EHR [2, 10] have shown that LLMs tend to fail on temporal reasoning over long EHR. While methods like temporal instruction tuning were proposed [2], these efforts have been largely confined to short EHR data or intensive care unit (ICU) data (<16k tokens). Temporal reasoning on very long EHR data over 32k or even 128k tokens remains an unclear challenge.

EHR data are inherently heterogeneous and primarily designed to support clinical care rather than research [12]. Consequently, they often contain noise arising from inconsistent formats, typographical errors, missing data, and irregular sampling. For many predictive tasks, only a small portion of a patient's record is informative, while abundant irrelevant data can obscure key predictive signals. Early methods sought to address these challenges by enforcing standardized EHR formats [13] or applying extensive feature engineering during preprocessing [4, 5], which are difficult to generalize across diverse healthcare systems. Recently, LLMs offer a more flexible and broadly applicable

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

framework for processing such data [14]. Yet, existing applications of LLMs in the EHR domain remain limited: many are restricted to single data modalities [15], while others struggle to capture temporal dependencies in long and complex patient histories [2, 10]. These gaps highlight the need for a mechanism to isolate relevant events scattered throughout patient histories.

Several strategies exist to manage long-context inputs for LLMs, including retrieval-augmented generation (RAG) and memory-based methods [16]. More advanced are agent-based approaches, which leverage the planning, memory, and reflection capabilities of LLMs to create autonomous agents [17, 16]. Frameworks like the chain-of-agents (CoA), for instance, use multi-agent collaboration to enhance reasoning over long contexts [18]. Despite their success in the general domain, the application of these methods in healthcare remains limited, primarily confined to question-answering tasks [19, 20]. Consequently, patient trajectory modeling with LLMs, typically prediction tasks, over long and noisy EHR remains an underexplored practical challenge.

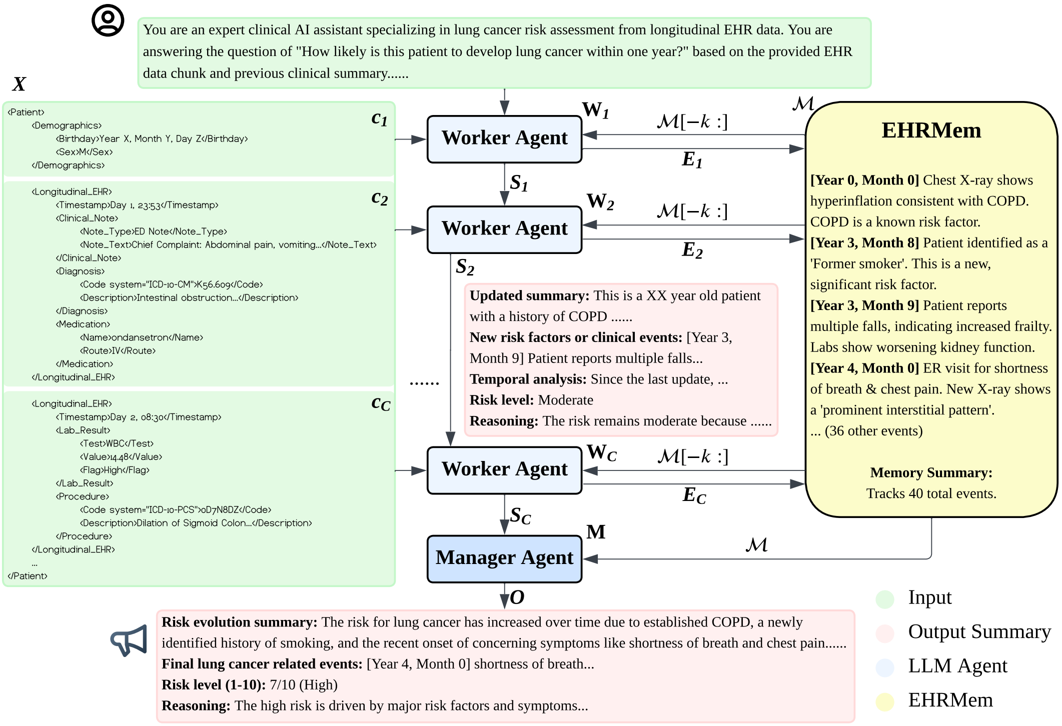
To overcome these challenges, we propose Traj-CoA, a novel framework for patient trajectory modeling. Traj-CoA employs a chain-of-agents architecture [18] with an external memory system to perform complex temporal reasoning on long patient histories. Our framework accepts a unified XML input that minimizes feature engineering and decomposes the longitudinal EHR into manageable time-aware chunks for more effective reasoning and summarization. These chunks are processed through a multi-agent workflow comprising specialized worker agents and a manager agent, which interact with a long-term memory module (EHRMem). We use lung cancer risk prediction as a use case to demonstrate the framework's capabilities, though it holds the promise to be a general-purpose solution for other longitudinal EHR tasks. The main contributions of this work are:

- We propose Traj-CoA, a novel chain-of-agents framework for temporal reasoning over long and noisy EHRs.
- To handle data noisiness, worker agents process unified XML inputs in sequential time-aware chunks, extracting salient signals for the task while removing localized noise.
- To manage long contexts, Traj-CoA leverages multi-agent communication and EHRMem for effective temporal reasoning with global context.
- Under a zero-shot setting in a lung cancer risk prediction task, Traj-CoA outperforms machine learning (ML), deep learning (DL), fine-tuned BERT, vanilla LLM, and RAG baselines, demonstrating its potential to become a simple yet powerful generalizable framework for patient trajectory modeling.

2 Related Works

Patient Trajectory Modeling Extensive previous studies have explored patient trajectory modeling with longitudinal EHR. Conventional approaches rely on task-specific feature engineering for specialized models, including recurrent neural networks [4, 5, 21], neural differential equations [6, 22, 23], and encoder-based transformers [24, 25, 26, 27]. More recently, EHR foundation models [28, 29] have demonstrated zero-shot generalizability by pre-training on large structured datasets. However, they are often constrained by limited code sets and short context windows (<16k tokens), preventing them from fully leveraging the rich information in unstructured text or modeling very long patient histories. LLMs-based approaches are a promising alternative, with demonstrated success in long-context reasoning across various domains [16]. However, applying a single, off-the-shelf LLM directly to model long patient trajectories has proven challenging. Recent studies show that even with access to a long context window, LLMs struggle with robust temporal reasoning in complex EHR data. For instance, a longer context does not guarantee better temporal understanding, and RAG offers an incomplete solution [10]. Similarly, LLMs exhibit a "lost-in-the-middle" phenomenon on long EHRs, a problem partially mitigated by temporal instruction tuning [2]. Therefore, a key open question remains: how can LLM-based approaches perform robust temporal reasoning on very long (over 32k or even 128k tokens) and noisy longitudinal EHR data without further training?

Long Context Modeling in LLMs While modern LLMs feature increasingly long context windows, their performance can degrade on lengthy inputs due to the "lost-in-the-middle" problem, where models overlook information positioned in the middle of the context [11]. To address this limitation, several training-free strategies have emerged [16]. These include memory-based methods, which utilize an external memory to store and dynamically update the context [30, 31]; RAG, which



You are an expert clinical AI assistant specializing in lung cancer risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the provided EHR data chunk and previous clinical summary......

```
<Patient>
    <Demographics>
        <Birthday>Year X, Month Y, Day Z</Birthday>
        <Sex>M</Sex>
    </Demographics>
    <Longitudinal_EHR>
        <Timestamp>Day 1, 23:53</Timestamp>
        <Clinical_Note>
            <Note_Type>ED Note</Note_Type>
            <Note_Text>Chief Complaint: Abdominal pain, vomiting...</Note_Text>
        </Clinical_Note>
        <Diagnosis>
            <Code system="ICD-10-CM">K56.609</Code>
            <Description>Intestinal obstruction...</Description>
        </Diagnosis>
        <Medication>
            <Name>ondansetron</Name>
            <Route>IV</Route>
        </Medication>
    </Longitudinal_EHR>
    <Longitudinal_EHR>
        <Timestamp>Day 2, 08:30</Timestamp>
        <Lab_Result>
            <Test>WBC</Test>
            <Value>14.48</Value>
            <Flag>High</Flag>
        </Lab_Result>
        <Procedure>
            <Code system="ICD-10-PCS">0D7N8DZ</Code>
            <Description>Dilation of Sigmoid Colon...</Description>
        </Procedure>
    </Longitudinal_EHR>
    ...
</Patient>
```

Updated summary: This is a XX year old patient with a history of COPD ......
New risk factors or clinical events: [Year 3, Month 9] Patient reports multiple falls...
Temporal analysis: Since the last update, ...
Risk level: Moderate
Reasoning: The risk remains moderate because ......

EHRMem

[Year 0, Month 0] Chest X-ray shows hyperinflation consistent with COPD. COPD is a known risk factor.
[Year 3, Month 8] Patient identified as a 'Former smoker'. This is a new, significant risk factor.
[Year 3, Month 9] Patient reports multiple falls, indicating increased frailty. Labs show worsening kidney function.
[Year 4, Month 0] ER visit for shortness of breath & chest pain. New X-ray shows a 'prominent interstitial pattern'.
... (36 other events)

Memory Summary:
Tracks 40 total events.

Risk evolution summary: The risk for lung cancer has increased over time due to established COPD, a newly identified history of smoking, and the recent onset of concerning symptoms like shortness of breath and chest pain......
Final lung cancer related events: [Year 4, Month 0] shortness of breath...
Risk level (1-10): 7/10 (High)
Reasoning: The high risk is driven by major risk factors and symptoms...

Figure 1: Traj-CoA architecture consisting of a chain of worker agents, a manager agent, and EHRMem.

retrieves relevant information based on the query to augment the model's input [32]; and agent-based approaches. Agent-based systems leverage memory, reflection, planning, and inter-agent communication to reason over extended contexts [16]. For instance, frameworks like LongAgent and CoA segment the context into manageable chunks and employ multi-agent collaboration to process them [33, 18]. Although these techniques have been successfully applied to general-domain and medical question-answering tasks [19, 20, 34], their utility for predictive tasks using longitudinal EHR is not well-established. In this study, we bridge this gap by proposing a novel method that synergizes memory and agent-based approaches for effective patient trajectory modeling.

Multi-Agent System for Biomedicine Multi-agent systems (MAS) in biomedicine employ multiple collaborating LLM agents to solve complex problems through role-playing and structured communication [35]. MAS has been successfully applied to biomedical discovery [36], diagnostics [37], and clinical trial optimization [38]. For example, MedAgents [39] assembles a multi-disciplinary team of specialized agents to improve zero-shot medical reasoning. Similarly, multiple specialized agents were orchestrated to emulate the tumor boards [40]. While a recent study applied this collaborative paradigm to Alzheimer's prediction using longitudinal EHR [41], the effective application of MAS for *temporal reasoning* over long and noisy EHR data remains unclear. In contrast to prior work, Traj-CoA is a MAS for temporal reasoning in patient trajectory modeling.

3 Method

In this section, we introduce Traj-CoA, a generalizable framework for patient trajectory modeling with longitudinal EHR. We outline the problem formulation and Traj-CoA's core components: a unified data preprocessing pipeline that represents heterogeneous EHR in an XML format with time-aware chunking, a chain-of-agents (CoA) workflow that sequentially processes the long and noisy EHR, and a long-term EHR memory system that memorizes detailed clinical events. The framework is described in a general context before being applied to a specific use case of lung cancer risk prediction. An illustration of Traj-CoA is shown in Figure 1.

3.1 Problem Formulation

We consider a patient trajectory as a longitudinal, multimodal sequence of n observations from the EHR, denoted as $\mathcal{X} = \{x_i, m_i, t_i\}_{i=1}^{n}$. Each tuple consists of a timestamp t_i, a data modality m_i (e.g., diagnosis, lab result, clinical note), and the corresponding event data x_i within the modality. A prediction task $\mathcal{T}$ is defined as a mapping $f : \mathcal{X} \to y$, where y is the label of a task-specific outcome that occurs after the final observation time t_n. We seek a generalized framework that learns task $\mathcal{T}$ with minimum task-specific data preprocessing and model training.

The task is challenging due to the inherent properties of EHR data. The observation window (t_1 to t_n) may span over years. The data itself is heterogeneous, potentially from different EHR vendors and with inconsistent coding systems (e.g., ICD-9 vs. ICD-10) across patients; irregularly sampled, with non-uniform time gaps between observations; and noisy, containing missing data and irrelevant events to $\mathcal{T}$. An effective solution must therefore identify task-related events from this long, noisy data and model their temporal patterns in a generalizable manner. For instance, while lung nodule growth indicates cancer risk [3], this temporal pattern is often scattered in EHRs, requiring algorithms to extract and interpret these critical dynamics.

In this study, we consider the lung cancer risk prediction task as a use case to evaluate the general Traj-CoA framework. Given the full patient trajectory $\mathcal{X}$ spanning up to 5 years before a reference date t_n, the task is to predict whether the patient will be diagnosed with lung cancer within the following year ($y = 1$ for cases, $y = 0$ for controls).

3.2 Data Preprocessing

To handle the noisy, long, and heterogeneous EHR data $\mathcal{X}$, we design a simple yet effective preprocessing pipeline to create a unified input representation X. Instead of relying on the complex task-specific feature engineering and cleaning common in prior work [5, 42, 43], our approach preserves the heterogeneity in the data while transforming it into an LLM-friendly format for reasoning. This consists of two steps: data unification into XML format and time-aware chunking.

Data Unification We convert a patient's entire multimodal history $\mathcal{X}$ into a single, unified XML format X. This strategy is motivated by the proven ability of LLMs to effectively comprehend structured, tag-based data [44] and inspired by similar practice on EHR data [2]. As illustrated in Figure 1, we organize the longitudinal data chronologically within a nested XML structure. The root contains patient demographics, followed by a sequence of timestamped records, each encapsulating all data modalities and events observed at that time. This approach yields a clean, well-structured representation of the patient's timeline for Traj-CoA, preserving the data heterogeneity in text format.

Time-Aware Chunking Long XML inputs pose computational and reasoning challenges for LLMs. For example, in our lung cancer dataset, the 75th percentile token count reaches 120k (Table S1). Despite large context windows (>128k tokens) in recent LLMs, performance degrades significantly with longer contexts due to the "lost-in-the-middle" phenomenon, where models fail to process information from the middle sections effectively [11].

To address this limitation, we follow the CoA approach [18] which splits the context into chunks and uses multi-agent communication to ensure sequential information aggregation and seamless reasoning (see Section 3.3). Instead of hard chunking based on a fixed chunk size, we design a time-aware chunking strategy that avoids missing timestamp information caused by chunking. Specifically, we partition the XML EHR input into chunks of maximum k tokens while preserving temporal ordering and timestamp completeness. Since our data structure is organized by timestamps, we split XML input into segments by timestamp, which are aggregated into chunks under the token limit k. When a single timestamp's records exceed k tokens, we split them further while maintaining the original timestamp for each resulting chunk. This dynamic process converts the full XML input X into a temporally coherent series of C chunks $\{c_1, c_2 ..., c_C\}$, guaranteeing that all information within any given chunk is closely related in time. Note that the number of chunks C may vary by patient.

3.3 Chain-of-Agent

We adapt the CoA algorithm [18] to reason over the chunked longitudinal EHR data. The vanilla CoA consists of two stages involving a chain of worker agents to conduct temporal reasoning chunk-by-chunk and a manager agent for the final prediction. Instead of relying on task-specific feature engineering or model training, CoA operates via task-specific instructions provided to its LLM agents, an approach that offers greater flexibility and efficiency for complex reasoning tasks. [18, 45]

In stage one, a series of worker agents $\mathbf{W}_i$ sequentially process each chunk c_i. Each worker agent takes the current chunk c_i, a task-specific instruction I_W, and the summary message S_{i-1} from the preceding agent as input. Its function is to extract salient task-related information from c_i, analyze temporal patterns in relation to the aggregated summary, and produce an updated summary message S_i. This sequential process allows for the progressive task-related information aggregation across the entire longitudinal EHR. The operation of each worker agent is defined as:

$$S_i = \mathbf{W}_i \left(I_W, S_{i-1}, c_i \right) \tag{1}$$

In stage two, a manager agent $\mathbf{M}$ receives the final summary message S_C from the last worker agent $\mathbf{W}_C$ along with a task-specific instruction I_M. The manager agent's role is to synthesize the comprehensive information contained in S_C to produce the final output O. This is formulated as:

$$O = \mathbf{M} \left(I_M, S_C \right) \tag{2}$$

The CoA framework transforms a long-context reasoning problem into a structured agent communication chain, with each agent assigned a shorter context, thereby improving the reasoning quality and mitigating the LLM's "lost-in-the-middle" phenomenon common in long-context reasoning. [18]

3.4 EHRMem

In our experiments, we found that a direct application of the vanilla CoA framework to longitudinal EHR data can lead to the progressive abstraction and loss of critical task-related information over long sequences. In other words, early clinical events may be vital for accurate prediction but may be "forgotten" by the final summary message S_C. To mitigate this, we introduce EHRMem, a structured long-term memory module storing task-related events and timestamps, denoted by $\mathcal{M}$.

EHRMem is populated during stage one, where each worker agent extracts new clinical events or risk factors that are potentially task-related, and stores their contents and timestamps as entries E_i in $\mathcal{M}$. To prevent overwhelmingly redundant entries caused by EHR "copy-forwarding," [29] we employ a deduplication mechanism: each agent's prompt is augmented with the last k events from $\mathcal{M}$ and is instructed to only store new, unrecorded information. In stage two, the manager agent's decision-making is augmented by the global context in $\mathcal{M}$, conditioning its output on both the final summary message S_C and the entire memory $\mathcal{M}$. The Traj-CoA's operation is thus redefined as:

$$S_i, E_i = \mathbf{W}_i \left(I_W, S_{i-1}, c_i, \mathcal{M}[-k:] \right) \tag{3}$$

$$\mathcal{M} \leftarrow \mathcal{M} \oplus E_i \tag{4}$$

$$O = \mathbf{M}(I_M, S_C, \mathcal{M}) \tag{5}$$

where $\mathcal{M}[-k:]$ denotes the last k events in $\mathcal{M}$, and $\oplus$ means concatenation. The EHRMem module serves two primary functions: (1) it constructs a distilled clinical timeline, effectively reducing the noise inherent in raw EHR, and (2) it provides the manager agent with a structured global context complementary to the unstructured worker agents' summary, enabling more robust reasoning across the entire patient history.

Crucially, the extraction heuristic for populating EHRMem is intentionally inclusive. Worker agents identify a slightly broader set of events *potentially* relevant to $\mathcal{T}$, rather than *strictly* filtering for those with an immediate, obvious connection to the task. This design choice acknowledges that local worker agents lack the global context to definitively assess an event's long-term significance. By preserving a richer, slightly redundant set of events in $\mathcal{M}$, we delegate the final synthesis and attribution of importance to the manager agent, which can leverage the complete temporal context for a more informed judgment.

4 Experiments

4.1 Dataset

We experimented on a proprietary case-control dataset on lung cancer risk assessment. Each instance in the dataset is anchored to a specific chest-related radiology exam (chest X-ray, chest CT, or abdomen CT) capable of visualizing the lungs. The prediction task is to determine, at the time of this index exam (t_n), whether a patient will receive a primary lung cancer diagnosis within 1 year. Cases are defined as subjects diagnosed with primary lung cancer within one year of the index exam, with diagnoses cross-validated against a state cancer registry. Controls are subjects with no cancer diagnosis. Cases and controls were matched at a 1:10 ratio on the time and type of the index exam. For each instance, up to five years of the patient's longitudinal EHR history prior to the index exam were recorded. This rich, multi-modal data encompasses both structured records (e.g., ICD codes, lab results, vitals) and unstructured text (e.g., clinical notes, radiology reports).

We randomly sampled 13,629 instances (1,239 cases and 12,390 controls) for model development, which were further randomly split into a training (12,266 samples) and validation (1,363 samples) set. These data were used for fine-tuning and **not** used under a zero-shot setting. From the rest, we randomly sampled 300 instances (28 cases and 272 controls) to construct a test set for feasible and fair evaluation. We highlight that the token count of XML input is substantial, with an interquartile range (IQR) of 28k–121k for cases and 19k–132k for controls in the test set (Table S1). A detailed data description is presented in the Appendix A.1.

4.2 Experimental Settings

We used MedGemma-27B [7] as the base model of Traj-CoA, with a default chunk size of 8k tokens and a maximum of 15 chunks to accommodate contexts up to 120k tokens. We benchmarked Traj-CoA against four baseline categories: (1) ML: logistic regression (LR) and XGBoost [46] trained on summary EHR features. (2) DL: The RNN-based trajectory models RETAIN [21] and PatientTM [4]. (3) BERT-based: Clinical ModernBERT (C-MBERT) [47], fine-tuned with LoRA [48] using an 8k context window. (4) LLM: Zero-shot MedGemma using two strategies: direct prompting (Vanilla, up to 64k context) and RAG with a bge-m3 [49] retriever on the time-aware chunks.

For LR, XGBoost, and RETAIN, we used diagnosis codes as input features. For PatientTM, we used the text descriptions of medical codes and unstructured notes as input. For BERT and LLM methods, we used the same XML input and employed a middle truncation strategy for context beyond the window. Specifically, we alternately selected texts from the beginning and end of the patient record until the context window limit is met, maintaining their relative chronological order. Compared to left truncation, where the most recent records are used, this method retains a longer and more challenging temporal span for evaluating temporal reasoning. We further report the performance using left truncation for BERT and vanilla baselines in Appendix A.2, preliminary results of fine-tuning Traj-CoA in Appendix A.3, a case study in Appendix A.4, and error analysis in Appendix A.5.

We evaluated all methods on AUROC and the best F1 score among all thresholds, with its corresponding precision and recall. We report AUPRC and discuss the differences in the Appendix A.2. For BERT, the risk score was derived from the output logit. For all LLM-based methods, we prompted the model to output a risk score between 1 and 10. Further experimental details are in the Appendix B.1.

Table 1: Performance comparison of different models. **Bold** indicates the best AUROC and F1 among all models or zero-shot models. SFT means supervised fine-tuning. The average performance (mean±std) for Traj-CoA across 5 runs with different random seeds is reported.

Model Family	Model	Prediction Method	Context Window	AUROC	Precision	Recall	F1
ML	LR	SFT	—	0.741	0.306	0.393	0.344
	XGBoost	SFT	—	0.763	0.367	0.393	0.379
DL	RETAIN	SFT	—	0.757	0.346	0.321	0.333
	PatientTM	SFT	—	0.730	0.361	0.464	**0.406**
BERT	C-MBERT	SFT	8k	0.749	0.367	0.393	0.379
LLM (MedGemma)	Vanilla	Zero-shot	32k	0.743	0.345	0.357	0.351
	RAG	Zero-shot	1k × 32	0.753	0.221	0.607	0.324
	Traj-CoA (w/o EHRMem)	Zero-shot	8k × 15	0.748	0.183	0.821	0.299
	Traj-CoA	Zero-shot	8k × 15	**0.766** ± 0.019	0.358 ± 0.057	0.436 ± 0.105	**0.380** ± 0.018

4.3 Results

Table 1 reports the performance of each method. Vanilla MedGemma with a 32k context window and RAG with top 32 chunks of maximum 1k tokens each achieves an AUROC of 0.74–0.75. However, expanding vanilla MedGemma's context to 64k degrades its performance to an AUROC of 0.714, suggesting difficulty in utilizing longer input sequences effectively (Table S2). In contrast, Traj-CoA, with its 120k context and dedicated design, substantially outperforms all zero-shot baselines, achieving an AUROC of 0.766 and an F1 score of 0.380, outperforming most SFT baselines and comparable to the best. These results highlight Traj-CoA's superior capability to conduct temporal reasoning over very long EHRs, overcoming the limitations observed in standard long-context models.

Ablation Study To analyze the contribution of the EHRMem component, we performed an ablation study by removing it from our architecture. As detailed in Table 1, this modification significantly impairs model performance, causing an absolute drop of 1.8% in AUROC and 8.1% in F1 score. This result underscores the importance of maintaining a detailed long-term memory of clinical events, as it provides predictive signals that are not fully captured by the evolving summary alone.

5 Analysis

To further probe the mechanisms by which Traj-CoA synthesizes temporal information from longitudinal EHR, we conducted additional analyses to answer five research questions. We fixed a random seed for the following analyses.

5.1 Sensitivity Analysis

Q1: How does the chunk size affect the performance?

Motivated by recent findings that LLM performance can degrade in very long contexts [11], we analyzed Traj-CoA's sensitivity to chunk size. To isolate this effect, we fixed the total context length at 80k tokens and varied the chunk size from 2k to 16k, which correspondingly adjusted the number of chunks from 40 down to 5. Figure 2A shows that performance peaks with a moderate chunk size of 8k, while both smaller (2k) and larger (16k) chunks result in lower AUROC scores.

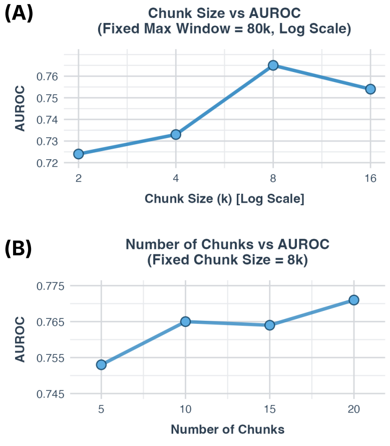


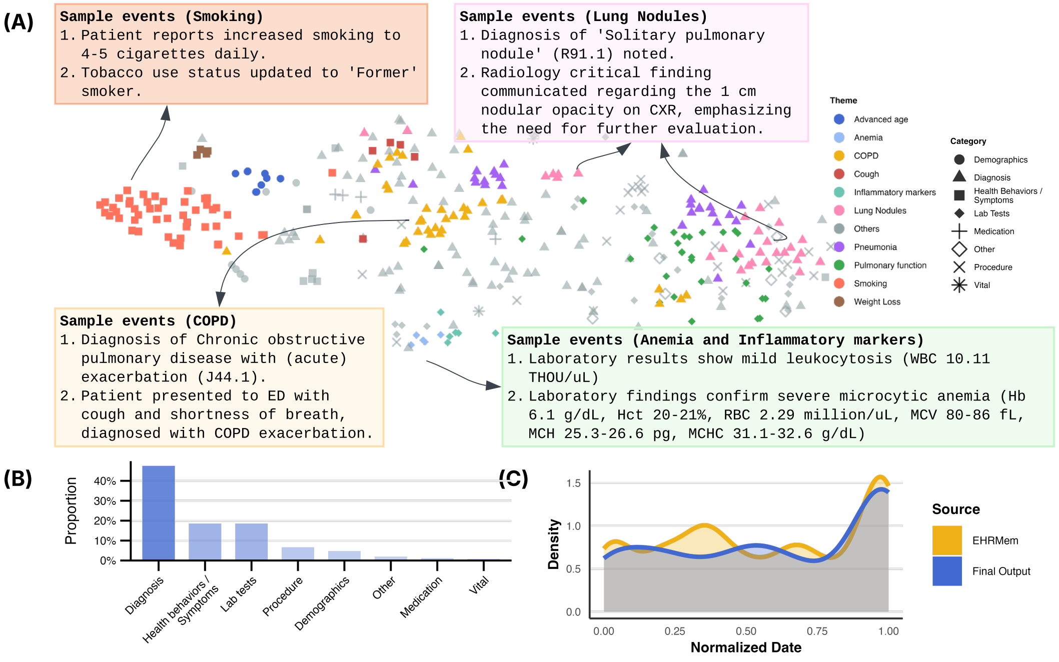
Figure 2: Sensitivity analysis on (A) chunk size and (B) number of chunks.

This reveals a fundamental trade-off. Small chunks force a long chain of iterative summarizations, risking catastrophic forgetting [50] where early, critical details are abstracted away. Conversely, large chunks shorten the chain but are susceptible to the "lost-in-the-middle" issue [11], where each worker agent fails to identify fine-grained signals within a vast context. The 8k chunk size appears to provide an optimal balance, preserving local detail while enabling effective global aggregation.

Q2: How does the maximum context window affect the performance?

We next analyzed how performance scales with the total context window to determine if more temporal information is beneficial. We fixed the chunk size at 8k and increased the maximum number of chunks from 5 to 20, thereby expanding the context window from 40k to 160k tokens.

As shown in Figure 2B, AUROC consistently improves as the context window expands. While the vanilla LLM baseline also improves when scaling from an 8k to a 32k context, its performance degrades significantly at 64k (Table S2). In contrast, Traj-CoA maintains its positive trend up to 160k tokens. This demonstrates that EHRs contain rich, albeit noisy, predictive signals and that Traj-CoA's architecture is uniquely capable of leveraging these ultra-long sequences where standard methods fail.



Sample events (Smoking)
1. Patient reports increased smoking to 4-5 cigarettes daily.
2. Tobacco use status updated to 'Former' smoker.

Sample events (Lung Nodules)
1. Diagnosis of 'Solitary pulmonary nodule' (R91.1) noted.
2. Radiology critical finding communicated regarding the 1 cm nodular opacity on CXR, emphasizing the need for further evaluation.

Sample events (COPD)
1. Diagnosis of Chronic obstructive pulmonary disease with (acute) exacerbation (J44.1).
2. Patient presented to ED with cough and shortness of breath, diagnosed with COPD exacerbation.

Sample events (Anemia and Inflammatory markers)
1. Laboratory results show mild leukocytosis (WBC 10.11 THOU/uL)
2. Laboratory findings confirm severe microcytic anemia (Hb 6.1 g/dL, Hct 20-21%, RBC 2.29 million/uL, MCV 80-86 fL, MCH 25.3-26.6 pg, MCHC 31.1-32.6 g/dL)

Figure 3: Analysis of Traj-CoA's behavior. (A) t-SNE plot visualizing the distribution of lung cancer related events in all cases' output O and sample events (timestamps in sample events are omitted for de-identification purposes). The events were embedded through nomic-embed-text-v1.5 [51]; (B) Distribution of categories in the lung cancer related events; and (C) Normalized date distribution of the events.

5.2 Temporal Reasoning Analysis

To investigate Traj-CoA's temporal reasoning, we performed a topic modeling analysis on the clinical events it identified as salient for lung cancer prediction. We used an LLM-based pipeline inspired by TopicGPT [52] to systematically categorize these events. First, for all positive cases, we extracted the relevant events from the model's final output O. Next, using Qwen2.5-7B-Instruct [53], we conducted a three-step analysis: (1) each event was classified into one of seven predefined categories (diagnosis, procedure, lab tests, vital, medication, health behaviors or symptoms, and demographics); (2) for each category, the top three common themes were generated based on all events under it; and (3) each event was mapped to its most relevant theme, if any, or recorded as "Others". This structured thematic analysis of the model's output forms the basis to answering the subsequent Q3–Q5.

Q3: Does Traj-CoA reason across diverse event categories?

We sought to verify that Traj-CoA's reasoning extends beyond trivial heuristics (e.g., identifying smoking status) to encompass a broad range of clinical events. Our analysis confirms that Traj-CoA identifies salient events from all seven predefined categories. As shown in the distribution in Figure 3B, the most frequently utilized categories are diagnosis, health behaviors or symptoms, and lab tests.

The diversity of these identified events is further underscored by the t-SNE visualization in Figure 3A. The event embeddings are scattered across the semantic space rather than forming a single cluster, indicating that the model draws upon a wide variety of clinical concepts. The qualitative examples presented in the figure corroborate this ability. This demonstrates that Traj-CoA performs multifaceted reasoning by integrating signals from a medically diverse set of events for its predictions.

Q4: Can Traj-CoA reason over the entire time horizon?

To determine if Traj-CoA utilizes the full patient trajectory or suffers from the "lost-in-the-middle" phenomenon, we analyzed the temporal distribution of the events it identified as salient. In Figure 3C, we plot the normalized date distribution of events from the EHRMem ($\mathcal{M}$) and the final output (O).

Both distributions show a concentration of events in the final year before the prediction date, which aligns with the clinical intuition that recent events are often most critical for diagnosis. Crucially, however, the model still identifies events from earlier periods. The presence of historical events in the final output demonstrates that Traj-CoA effectively synthesizes information over long time horizons, appropriately weighing recent events more heavily without discarding valuable historical context.

Q5: Is Traj-CoA's reasoning clinically relevant?

Finally, we assessed the clinical relevance of Traj-CoA's reasoning by analyzing the themes of the salient events. The t-SNE visualization of the event embeddings (Figure 3A) reveals that the most frequently identified themes form distinct semantic clusters, indicating thematic consistence.

Importantly, these data-driven themes align directly with well-established clinical knowledge. The top themes identified include advanced age, anemia, COPD, cough, inflammatory markers, lung nodules, pneumonia, pulmonary function, smoking, and weight loss. These are either widely recognized by clinical practice and screening guidelines for assessing lung cancer risk [54, 55], or consistent with existing evidence [56, 57, 58]. This validates the model's interpretability and demonstrates that its risk predictions are based on clinically meaningful patterns within the EHR data.

5.3 Complexity Analysis

We compare the time complexity of Traj-CoA to vanilla prompting and RAG, similar to the previous study [18]. For each patient, let L, L_C, L_R, and L_O denote the average lengths of the total input, a single chunk, the retrieved context for RAG, and each model output, respectively. As shown in Table 2, Traj-CoA has a lower encoding complexity than vanilla prompting but is more computationally intensive than RAG. This presents a trade-off between latency and contextual completeness. RAG achieves lower latency via selective retrieval at the risk of information loss, whereas Traj-CoA processes the entire context. The optimal choice depends on whether an application prioritizes real-time performance or comprehensive temporal reasoning over a patient's history.

Table 2: Time complexity.

Method	Encode	Decode
Vanilla	$O(L^2)$	$O(LL_O)$
RAG	$O(L_R^2)$	$O(L_R L_O)$
Traj-CoA	$O(LL_C)$	$O(LL_O)$

6 Discussion

In this work, we introduced Traj-CoA, a framework designed to perform temporal reasoning on long and noisy longitudinal EHR data for patient trajectory modeling. By decomposing the reasoning process across agents that analyze moderate-sized data chunks, Traj-CoA effectively sidesteps the "lost-in-the-middle" problem, while a dedicated EHRMem module prevents forgetting of crucial early events. This design unlocks a key capability: unlike standard LLMs, Traj-CoA's performance on long EHR scales positively with context windows up to 160k tokens, effectively leveraging the rich predictive signals in complete patient trajectories.

While the positive results from this study shows promise for Traj-CoA, limitations exist. From a technical perspective, future efforts can enhance Traj-CoA's performance, including access to external knowledge [59, 60], using more powerful base models, and fine-tuning via multi-agent training approaches [61, 62]. Moreover, while our analysis interprets Traj-CoA's temporal reasoning in terms of *what* the salient events look like, further analysis is needed to understand *how* these events are synthesized by the model for accurate prediction among different subpopulations. Finally, although Traj-CoA is a task-agnostic framework, it requires carefully crafted prompts for specific tasks. Research into prompt optimization [63] or data-driven hypothesis generation [64, 36] could reduce this dependency and allow for more general modeling of patient trajectories. From a clinical application perspective, Traj-CoA was evaluated on a relatively small, single-institution cohort and one predictive task. We plan to conduct broader-scale validation to establish Traj-CoA's generalizability across diverse clinical settings and a wider array of prediction targets.

In conclusion, Traj-CoA provides a novel framework for patient trajectory modeling and bridges the gap between the generalist agentic AI [9, 59] and the temporal reasoning in longitudinal EHR. Traj-CoA demonstrates potential as a generalizable framework for patient trajectory modeling, though further design and validation are needed to make it more trustworthy.

Acknowledgements

This work was supported by the National Institutes of Health (NIH)—National Cancer Institute (Grant Nos. 1R01CA248422-01A1).

References

[1] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[2] Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam Shah. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *arXiv preprint arXiv:2503.04176*, 2025.

[3] Beatriz Ocaña-Tienda, Alba Eroles-Simó, Julián Pérez-Beteta, Estanislao Arana, and Víctor M Pérez-García. Growth dynamics of lung nodules: implications for classification in lung cancer screening. *Cancer Imaging*, 24(1):113, 2024.

[4] João Figueira Silva and Sérgio Matos. Modelling patient trajectories using multimodal information. *Journal of biomedical informatics*, 134:104195, 2022.

[5] Manuel Burger, Gunnar Rätsch, and Rita Kuznetsova. Multi-modal graph learning over umls knowledge graphs. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 52–81. PMLR, 10 Dec 2023.

[6] Sihang Zeng, Lucas Jing Liu, Jun Wen, Meliha Yetisgen, Ruth Etzioni, and Gang Luo. Trajsurv: Learning continuous latent trajectories from electronic health records for trustworthy survival prediction, 2025.

[7] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025.

[8] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[9] Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, Xingtai Lv, Hu Jinfang, Zhiyuan Liu, and Bowen Zhou. Ultramedical: Building specialized generalists in biomedicine, 2024.

[10] Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. Zero-shot large language models for long clinical text summarization with temporal reasoning, 2025.

[11] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[12] Ellen Kim, Samuel M Rubinstein, Kevin T Nead, Andrzej P Wojcieszynski, Peter E Gabriel, and Jeremy L Warner. The evolving use of electronic health records (ehr) for research. In *Seminars in radiation oncology*, volume 29, pages 354–361. Elsevier, 2019.

[13] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18, 2018.

[14] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey, 2024.

[15] Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*, 2024.

[16] Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, Yuanxing Zhang, Zhuo Chen, Hangyu Guo, Shilong Li, Ziqiang Liu, Yong Shan, Yifan Song, Jiayi Tian, Wenhao Wu, Zhejian Zhou, Ruijie Zhu, Junlan Feng, Yang Gao, Shizhu He, Zhoujun Li, Tianyu Liu, Fanyu Meng, Wenbo Su, Yingshui Tan, Zili Wang, Jian Yang, Wei Ye, Bo Zheng, Wangchunshu Zhou, Wenhao Huang, Sujian Li, and Zhaoxiang Zhang. A comprehensive survey on long context language modeling, 2025.

[17] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023.

[18] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö. Arik. Chain of agents: Large language models collaborating on long-context tasks, 2024.

[19] Gongbo Zhang, Zihan Xu, Qiao Jin, Fangyi Chen, Yilu Fang, Yi Liu, Justin F Rousseau, Ziyang Xu, Zhiyong Lu, Chunhua Weng, et al. Leveraging long context in retrieval augmented language models for medical question answering. *npj Digital Medicine*, 8(1):239, 2025.

[20] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents, 2025.

[21] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, 2017.

[22] Intae Moon, Stefan Groha, and Alexander Gusev. Survlatent ode : A neural ode based time-to-event model with competing risks for longitudinal data improves cancer-associated venous thromboembolism (vte) prediction, 2022.

[23] Asem Alaa, Erik Mayer, and Mauricio Barahona. Ice-node: Integration of clinical embeddings with neural ordinary differential equations. In *Machine Learning for Healthcare Conference*, pages 537–564. PMLR, 2022.

[24] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[25] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[26] Alex Labach, Aslesha Pokhrel, Xiao Shi Huang, Saba Zuberi, Seung Eun Yi, Maksims Volkovs, Tomi Poutanen, and Rahul G. Krishnan. Duett: Dual event time transformer for electronic health records, 2023.

[27] Davide Placido, Bo Yuan, Jessica X Hjaltelin, Chunlei Zheng, Amalie D Haue, Piotr J Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature medicine*, 29(5):1113–1122, 2023.

[28] Pawel Renc, Michal K. Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew B. A. McDermott, Jaroslaw Was, Anthony E. Samir, Jonathan W. Cunningham, David W. Bates, and Arkadiusz Sitek. Foundation model of electronic medical records for adaptive risk estimation, 2025.

[29] Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Re, Sanmi Koyejo, and Nigam H. Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs, 2025.

[30] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

[31] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024.

[32] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024.

[33] Jun Zhao, Can Zu, Hao Xu, Yi Lu, Wei He, Yiwen Ding, Tao Gui, Qi Zhang, and Xuanjing Huang. Longagent: Scaling language models to 128k context through multi-agent collaboration, 2024.

[34] Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17, 2025.

[35] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax?, 2025.

[36] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation, 2024.

[37] Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, et al. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159, 2025.

[38] Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning, 2024.

[39] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning, 2024.

[40] MD MPH, Matthew Lungren. Developing next-generation cancer care management with multi-agent orchestration, May 2025.

[41] Rumeng Li, Xun Wang, Dan Berlowitz, Jesse Mez, Honghuang Lin, and Hong Yu. Care-ad: a multi-agent large language model framework for alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, 8(1):541, August 2025.

[42] Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C. Hughes, and Tristan Naumann. Mimic-extract: a data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, ACM CHIL '20, page 222–235. ACM, April 2020.

[43] Jun Wen, Jue Hou, Clara-Lea Bonzel, Yihan Zhao, Victor M. Castro, Vivian S. Gainer, Dana Weisenfeld, Tianrun Cai, Yuk-Lam Ho, Vidul A. Panickan, Lauren Costa, Chuan Hong, J. Michael Gaziano, Katherine P. Liao, Junwei Lu, Kelly Cho, and Tianxi Cai. Latte: Label-efficient incident phenotyping from longitudinal electronic health records, 2023.

[44] Anthropic. Use xml tags to structure your prompts. `https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/use-xml-tags`, 2025. Accessed: 2025-08-16.

[45] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025.

[46] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[47] Simon A. Lee, Anthony Wu, and Jeffrey N. Chiang. Clinical modernbert: An efficient and long context encoder for biomedical text, 2025.

[48] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[49] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[50] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017.

[51] Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder, 2025.

[52] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. Topicgpt: A prompt-based topic modeling framework, 2024.

[53] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[54] Andrew MD Wolf, Kevin C Oeffinger, Tina Ya-Chen Shih, Louise C Walter, Timothy R Church, Elizabeth TH Fontham, Elena B Elkin, Ruth D Etzioni, Carmen E Guerra, Rebecca B Perkins, et al. Screening for lung cancer: 2023 guideline update from the american cancer society. *CA: A Cancer Journal for Clinicians*, 74(1):50–81, 2024.

[55] Matthew B Schabath and Michele L Cote. Cancer progress and priorities: lung cancer. *Cancer epidemiology, biomarkers & prevention*, 28(10):1563–1579, 2019.

[56] Hye Seon Kang, Ah Young Shin, Chang Dong Yeo, Chan Kwon Park, Ju Sang Kim, Jin Woo Kim, Seung Joon Kim, Sang Haak Lee, and Sung Kyoung Kim. Clinical significance of anemia as a prognostic factor in non-small cell lung cancer carcinoma with activating epidermal growth factor receptor mutations. *Journal of Thoracic Disease*, 12(5):1895, 2020.

[57] Eric A Engels. Inflammation in the development of lung cancer: epidemiological evidence. *Expert review of anticancer therapy*, 8(4):605–615, 2008.

[58] Maria G Prado, Larry G Kessler, Margaret A Au, Hannah A Burkhardt, Monica Zigman Suchsland, Lesleigh Kowalski, Kari A Stephens, Meliha Yetisgen, Fiona M Walter, Richard D Neal, et al. Symptoms and signs of lung cancer prior to diagnosis: case–control study using electronic health records from ambulatory care within a large us-based tertiary care centre. *BMJ open*, 13(4):e068832, 2023.

[59] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, pages 2025–05, 2025.

[60] Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. Txagent: An ai agent for therapeutic reasoning across a universe of tools, 2025.

[61] Kaiyan Zhang, Runze Liu, Xuekai Zhu, Kai Tian, Sihang Zeng, Guoli Jia, Yuchen Fan, Xingtai Lv, Yuxin Zuo, Che Jiang, Ziyang Liu, Jianyu Wang, Yuru Wang, Ruotong Zhao, Ermo Hua, Yibo Wang, Shijie Wang, Junqi Gao, Xinwei Long, Youbang Sun, Zhiyuan Ma, Ganqu Cui, Lei Bai, Ning Ding, Biqing Qi, and Bowen Zhou. Marti: A framework for multi-agent llm systems reinforced training and inference, 2025.

[62] Sumeet Ramesh Motwani, Chandler Smith, Rocktim Jyoti Das, Rafael Rafailov, Ivan Laptev, Philip H. S. Torr, Fabio Pizzati, Ronald Clark, and Christian Schroeder de Witt. Malt: Improving reasoning with multi-agent llm training, 2025.

[63] Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, Haozhu Wang, Han Ding, Yuzhe Lu, Zhichao Xu, Yun Zhou, Balasubramaniam Srinivasan, Qiaojing Yan, Yueyan Chen, Haibo Ding, Panpan Xu, and Lin Lee Cheong. A systematic survey of automatic prompt optimization techniques, 2025.

[64] Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. Large language models are zero shot hypothesis proposers, 2023.

[65] Matthew B. McDermott, Haoran Zhang, Lasse Hyldig Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at auroc and auprc under class imbalance. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, page 44102–44163. Curran Associates, Inc., 2024.

[66] Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023.

[67] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[68] Michael Han Daniel Han and Unsloth team. Unsloth, 2023.

[69] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2020.

[70] Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning, 2025.

[71] Elvin S Cheng, Marianne Weber, Julia Steinberg, and Xue Qin Yu. Lung cancer risk in never-smokers: An overview of environmental and genetic factors. *Chinese Journal of Cancer Research*, 33(5):548, 2021.

[72] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[73] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Appendix A. Additional Results

A.1 Dataset Description

We present the dataset statistics of the test set in Table S1. There are 28 cases and 272 controls in the dataset. The data is extracted from a in-house institutional medical center, whose EHR system has experienced a transition during the period. Therefore, the data has clinical notes from two EHR systems. The table summarizes features including patient demographics and metrics quantifying the volume of clinical data, such as the number of notes, diagnosis codes, and lab codes. All statistics are presented as median and interquartile range (IQR). Notably, both the case and control groups share a median record length or 'Year span' of 4.0 years, suggesting comparable observation periods. However, the groups differ in the volume of specific data types; for instance, cases have a higher median count of diagnosis codes (156.0 vs. 128.5), whereas controls have a higher median count of medication codes (64.0 vs. 42.0).

Notably, this dataset presents a challenging prediction problem because both the case and control cohorts are anchored to a radiology exam. Consequently, all patients in the dataset underwent radiological imaging for a clinical indication, requiring the model to distinguish radiological abnormalities associated with lung cancer from those attributable to other conditions, such as cardiovascular disease.

Table S1: Dataset Statistics by Case-Control Status. Values shown as median (IQR) unless otherwise specified.

Variable	Cases (n=28)	Controls (n=272)
Sex (Female/Male)	14/14	115/157
Year of last record	2015.5 (2014.0–2018.0)	2016.5 (2013.0–2018.0)
Year span	4.0 (2.0–5.0)	4.0 (1.0–5.0)
XML tokens	61,270 (28,675–121,722)	51,610 (19,442–132,777)
# Timestamps	42.5 (24.0–79.8)	38.5 (14.0–96.0)
# EHR System 1 notes	7.5 (1.0–29.3)	8.0 (0.0–29.0)
# EHR System 2 notes	3.5 (0.0–12.8)	4.0 (0.0–14.0)
# Radiology reports	10.0 (4.8–17.3)	8.0 (3.0–15.3)
# Diagnosis codes	156.0 (92.8–271.3)	128.5 (42.5–349.3)
# Medication codes	42.0 (17.8–135.3)	64.0 (9.0–184.5)
# Procedure codes	98.0 (51.5–197.8)	106.5 (38.0–231.5)
# Lab codes	186.0 (26.8–439.3)	201.0 (49.8–523.0)
# Vital sign codes	28.0 (0.0–91.0)	20.0 (0.0–79.5)

A.2 Full Results

Table S2 presents the complete results for our method, Traj-CoA, alongside BERT and LLM-based baselines. Traj-CoA, configured with a maximum chunk size of 8k, achieves the highest AUROC (0.753 – 0.771), outperforming all BERT-based and LLM baselines.

While we observe a divergence between AUROC and AUPRC scores, we argue that AUROC is the more appropriate primary metric for this clinical prediction task. Recent work by McDermott et al. [65] demonstrates that while AUROC favors model improvements uniformly across all positive samples, AUPRC can be a misleading metric that disproportionately rewards improvements for samples assigned high scores. In the context of cancer risk prediction, the clinical cost of false negatives is exceptionally high, making the correct classification of lower-scoring, at-risk individuals paramount. Since AUPRC's prioritization runs counter to this clinical need [65], we assert that Traj-CoA's superior performance on the more robust and relevant AUROC metric is the most significant finding.

Table S2: Full performance comparison of BERT-based and LLM baselines and Traj-CoA on the lung cancer risk prediction task using the left or middle truncation strategy.

Model Family	Model	Prediction method	Context Window	Truncation	AUROC	AUPRC	Precision	Recall	F1
Clinical ModernBERT	BERT	SFT	8k	Left	0.734	0.256	0.323	0.357	0.339
			8k	Middle	0.749	0.310	0.367	0.393	0.379
MedGemma	Vanilla	Zero-shot	4k	Left	0.627	0.133	0.191	0.444	0.267
			8k	Left	0.668	0.211	0.370	0.357	0.364
			16k	Left	0.738	0.251	0.266	0.607	0.370
			32k	Left	0.739	0.249	0.313	0.357	0.333
			64k	Left	0.737	0.235	0.233	0.500	0.318
			4k	Middle	0.646	0.139	0.169	0.393	0.237
			8k	Middle	0.647	0.171	0.217	0.357	0.270
			16k	Middle	0.732	0.256	0.500	0.250	0.333
			32k	Middle	0.743	0.242	0.345	0.357	0.351
			64k	Middle	0.714	0.214	0.237	0.500	0.322
	RAG	Zero-shot	1k × 32	-	0.753	0.208	0.221	0.607	0.324
			2k × 16	-	0.740	0.224	0.313	0.357	0.333
			4k × 8	-	0.731	0.190	0.247	0.643	0.356
			8k × 4	-	0.735	0.179	0.215	0.500	0.301
	Traj-CoA	Zero-shot	8k × 5	Middle	0.753	0.203	0.262	0.393	0.314
			8k × 10	Middle	0.765	0.205	0.217	0.750	0.336
			8k × 15	Middle	0.764	0.233	0.291	0.571	0.386
			8k × 20	Middle	0.771	0.214	0.255	0.500	0.337
			2k × 40	Middle	0.724	0.168	0.131	0.893	0.228
			4k × 20	Middle	0.733	0.174	0.203	0.464	0.283
			16k × 5	Middle	0.754	0.204	0.163	0.857	0.274

A.3 Preliminary Results of Fine-tuning Traj-CoA

To investigate if further training could enhance the predictive performance of Traj-CoA, we conducted a preliminary study using rejection sampling fine-tuning (RFT) [66]. While our initial results are promising, we caution that these findings are exploratory. A systematic application of this method would necessitate a rigorous experimental design and extensive hyperparameter tuning to ensure robust and generalizable improvements.

Training Data Generation We generated a high-quality dataset for RFT using a rejection sampling methodology, beginning with an initial pool of 500 cases and 500 controls from the training set. For each patient, we generated four candidate reasoning trajectories using Traj-CoA (8k × 15 chunks setting) with a high sampling temperature of 1.5. A trajectory is defined as the sequence of inputs and outputs from all worker and manager agents during a single forward pass.

To select for high-quality reasoning, we applied the following rejection criteria: for cases, we retained the trajectory that produced the highest predicted risk score, provided it was greater than 6; for controls, we retained the trajectory yielding the lowest score, provided it was less than 4. From each retained trajectory, we constructed RFT samples by compiling the input-output pairs from the first and last worker agents, two randomly sampled intermediate worker agents, and the manager agent. This process yielded a final RFT dataset comprising 2,223 instruction-tuning samples for fine-tuning both agent types.

RFT We fine-tuned the MedGemma-27B model using supervised fine-tuning (SFT). For memory efficiency, we employed QLoRA [67] with 4-bit quantization, implemented with the Unsloth [68] and Huggingface TRL [69] libraries. The LoRA rank and α were both set to 32. The model was trained for 3 epochs on two A100 GPUs, using a per-device batch size of 2 and 8 gradient accumulation steps, resulting in an effective batch size of 32. Because the RFT dataset contained a mixture of data for both worker and manager agents, the single fine-tuned model serves as a unified base for both agent types in our framework.

Preliminary Results As shown in Table S3, RFT leads to a notable performance gain for Traj-CoA. For example, when configured with a 16k × 5 context window, RFT improves the AUROC from 0.754 to 0.789. This result suggests that fine-tuning on data generated via rejection sampling is a promising direction for enhancing model performance.

However, we observed two key disparities in these preliminary results. First, the improvement in AUROC was not consistently accompanied by an increase in the F1 score. Second, the performance gains were more pronounced in the 16k × 5 setting, despite the RFT data being generated from the 8k × 15 setting. We hypothesize that these inconsistencies may be attributable to the ratio of instruction-tuning data between worker and manager agents in the RFT dataset.

Table S3: Preliminary results for training Traj-CoA.

Model Family	Model	Prediction Method	Context Window	AUROC	Precision	Recall	F1
MedGemma 27B	**Traj-CoA**	Zero-shot	8k × 15	0.764	0.291	0.571	0.386
	Traj-CoA w/ RFT	RFT	8k × 15	0.775	0.262	0.571	0.360
	Traj-CoA	Zero-shot	16k × 5	0.754	0.221	0.607	0.324
	Traj-CoA w/ RFT	RFT	16k × 5	0.789	0.241	0.714	0.360

Future Directions While our exploratory experiments are promising, this work highlights several avenues for future research. First, our current approach trains a single, unified base model on mixed data for both worker and manager agents. This co-training strategy may introduce biases and limit the potential for agent specialization. Future work could explore dedicated multi-agent fine-tuning strategies [61] to train distinct models for each role, potentially enhancing the performance of the overall framework. Second, we employed an offline rejection sampling method for fine-tuning. Adopting online learning paradigms, such as reinforcement learning [61, 70], could enable more dynamic policy improvements and potentially lead to more robust and capable agents. Finally, we observed that model performance can be sensitive to numerous hyperparameters. These include the case-control balance of the training data, the sampling temperature used for trajectory generation, and the ratio of worker-to-manager data. A rigorous, systematic study is needed to investigate the impact of these parameters and establish a more principled approach to optimizing the training process.

A.4 Case Study

We present a deidentified case study of Traj-CoA's final output O for a case patient. As shown in Table S4, the patient is an elderly female with multiple comorbidities (e.g., COPD, cognitive impairment) and established risk factors, including an advanced age and a history as a former smoker. She recently had an emergency room visit for shortness of breath and a chest X-ray with a "prominent interstitial pattern," both of which are non-specific and could be attributed to her existing conditions. Traj-CoA demonstrates its ability to synthesize this heterogeneous information effectively. It correctly identifies that the combination of long-term risk factors and persistent, concerning symptoms warrants a high-risk assessment (8/10), showcasing its capability to produce predictions grounded in a holistic view of the patient's trajectory.

A.5 Error Analysis

We conducted a qualitative analysis of the three **false-negative** cases, which were identified using a predicted risk threshold of 5.0. To investigate the model's failure modes, we performed a manual chart review of each patient's 5-year longitudinal EHR data preceding the index exam. This analysis of why Traj-CoA assigned erroneously low-risk scores aims to reveal its limitations and guide future improvements.

Case 1 is a patient in her early 70s with a complex medical history that includes a >40 pack-year smoking history, COPD, and pulmonary hypertension. While the patient presented with persistent cough, dyspnea, and chest pain, recent imaging showed no suspicious findings and indicated resolving pneumonia. Traj-CoA reasoned that these symptoms were manifestations of her known comorbidities, assigning a low predicted cancer risk of 3/10.

Case 2 is a patient in her 50s with a medical history notable for chronic cough and long-term immunosuppression due to arteritis. Although a PET-CT three years prior revealed a small nodule, the patient is a never-smoker with no family history of cancer, and subsequent CT scans showed no new suspicious nodules. Traj-CoA likely discounted the historical nodule due to its long-term stability and the patient's otherwise low-risk profile, assigning a low predicted cancer risk of 4/10.

Case 3 is a patient in her early 70s with a medical history that includes severe obesity, hypertension, and asthma. The patient reported a chronic cough and repeated exposure to fumes. She is a lifelong nonsmoker, and interim exams noted clear lungs. Traj-CoA correctly identified the exposure to fumes as a lung cancer related event, but attributed respiratory complaints to asthma and likely over-weighted the patient's nonsmoker status and negative exam results, thus assigning a low predicted cancer risk of 2/10.

In conclusion, our error analysis identifies two key limitations of the proposed model. First, as shown in all three cases, Traj-CoA demonstrates a **tendency to over-rely on recent, benign imaging results**, which can lead to poorly calibrated risk assessments that fail to accurately capture a patient's true risk. Second, as shown in cases 2 and 3, Traj-CoA may **underestimation of risk within the never-smoker subpopulation**, suggesting the model does not adequately capture the distinct predictive factors relevant to this group (e.g., environmental or occupational exposures) [71].

These findings highlight opportunities for future research. To address miscalibration, future methods may incorporate methodologies designed to produce more reliable predictions, such as model fine-tuning. To improve fairness and accuracy, we advocate for the development of models that explicitly stratify by smoking status, either through cohort-specific architectures or by integrating domain knowledge of non-smoking-related risk factors.

Appendix B. Experimental Details

B.1 Experimental Settings

Our experiments were conducted using Python 3.10, Huggingface Transformers 4.53.2 [72], and vLLM 0.9.2 [73]. Models were downloaded from Huggingface[1][2]. The BERT baseline was trained on a single NVIDIA A100 GPU; we added a linear layer over the [CLS] token's final hidden state to produce logits and optimized the model using a binary cross-entropy loss with a learning rate of 5e-5, a batch size of 8, and 4 gradient accumulation steps (effective batch size of 32). We employed an early stopping strategy with a patience of 3 epochs to prevent overfitting. For all LLM-based methods, we used default decoding parameters of Gemma 3: a temperature of 1.0, top-p of 0.95, and top-k of 64. To accommodate the model's size and the long-context requirements of the task, inference was performed on two NVIDIA A100 GPUs, leveraging tensor parallelism. Implementation of Traj-CoA will be released on GitHub upon acceptance.

Appendix C. Prompts

We present the prompt templates and query for RAG in Table S5, S6, S7, S8, S9, S10, S11, and S12.

[1] https://huggingface.co/Simonlee711/Clinical_ModernBERT

[2] https://huggingface.co/google/medgemma-27b-text-it

Table S4: Case study of the final output O of a case patient.

De-identified Patient Lung Cancer Risk Assessment

Risk Evolution Summary

The patient is an [redacted]-year-old female with a significant history of COPD and a past smoking history, both established risk factors for lung cancer. Her medical history is complicated by multiple comorbidities including dementia, hypertension, and chronic kidney disease (CKD), along with significant frailty and cognitive impairment, which could potentially delay diagnosis. Chronologically, her COPD was noted on chest X-rays in [redacted] and confirmed in [redacted]. Her advanced age ([redacted] in [redacted], [redacted] in [redacted]) has been a consistent risk factor. The patient's smoking status was documented as 'Former' in [redacted]. Recent presentations ([redacted]) for syncope, chest pain, and shortness of breath are concerning, although initial workup (negative troponin, stable CT head, non-specific CXR showing interstitial pattern) did not confirm an acute cardiac or neurological event or definitively point to malignancy. The presence of mild hypoxemia (SpO2 92%) and fever (38.3°C) during the [redacted] ER visit are non-specific but warrant consideration in the context of her overall health. Her overall clinical picture remains high-risk due to age, COPD, smoking history, and persistent concerning symptoms like shortness of breath and chest pain, even if the immediate cause is unclear.

Final Lung Cancer-Related Events

- [redacted]: Patient born in [redacted], making her [redacted] years old in [redacted]. Advanced age is a risk factor for lung cancer.
- [redacted]: Chest X-ray revealed hyperinflation consistent with COPD. COPD is associated with an increased risk of lung cancer.
- [redacted]: Diagnosis of COPD confirmed via radiology report (hyperinflation). COPD is a significant risk factor for lung cancer.
- [redacted]: Patient is [redacted] years old. Advanced age is a significant risk factor for lung cancer.
- [redacted]: Patient identified as a 'Former' smoker (tobacco_use). Being a former smoker still carries an elevated risk compared to never smokers.
- [redacted]: Chest X-ray shows prominent interstitial pattern in the bilateral lungs. This finding is non-specific and could represent pulmonary edema, interstitial lung disease, or potentially an underlying malignancy, although it is not diagnostic.
- [redacted]: ER presentation for shortness of breath. Shortness of breath is a common symptom of lung cancer, but also of COPD, heart failure, and other conditions.
- [redacted]: ER presentation for chest pain (unspecified). Chest pain is a potential symptom of lung cancer, although it can have many other causes.

Final Risk Assessment

Risk Level: 8/10

Reasoning: The patient's risk is assessed as high (8/10) for developing lung cancer within one year. This assessment is based on the convergence of several significant risk factors: advanced age ([redacted] years), a confirmed history of COPD (present since at least [redacted]), and a documented history as a former smoker. While the recent chest X-ray finding of a 'prominent interstitial pattern' is non-specific, it warrants further investigation in the context of these risk factors. Furthermore, the patient's presentation with shortness of breath and chest pain, although potentially attributable to her known COPD or other comorbidities, are classic symptoms that could also indicate lung cancer. Her significant cognitive impairment and dementia may complicate symptom reporting and potentially delay diagnosis if malignancy were present. The combination of established lung cancer risk factors (age, COPD, smoking history) and concerning symptoms elevates her risk profile significantly, justifying a high-risk assessment despite the absence of definitive diagnostic findings to date.

Table S5: Prompt for the initial worker agent.

INITIAL WORKER SYSTEM PROMPT

You are an expert clinical AI assistant specializing in lung cancer risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the provided EHR data chunk.

Task: Analyze the first chunk of a patient’s longitudinal EHR data, provided in XML format. Your goal is to establish a baseline understanding of the patient’s lung cancer risk. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to lung cancer risk assessment.

Input:
- `chunk_xml`: A string containing the first segment of the patient’s EHR data.

Instructions:
1. **Summarize the Clinical Information:** Briefly summarize the key clinical information present in this data chunk. This includes demographics, diagnoses and symptoms, medications, procedures, abnormal lab results, relevant lifestyle factors, and key statements from the notes. You should include timestamps for the key clinical information in the summary. Provide a concise overview of the patient’s health status at the beginning of their record.
2. **Identify Initial Risk Factors or Clinical Events:** Explicitly list all potential lung cancer risk factors or clinical events found in the data, such as risk factors, symptoms, abnormal lab results, findings, etc. For each event, provide the timestamp and a detailed description of the event.
3. **Assess Initial Lung Cancer Risk:** Based on the identified lung cancer related risk factors or clinical events, provide an initial lung cancer risk assessment for this patient. The risk should be categorized as **Low**, **Moderate**, or **High**. Provide a clear rationale for your assessment.

Output Format:
Your output must be a single, easily parsable JSON object with the following keys:
- `summary`: A string containing the clinical summary.
- `risk_factors_or_clinical_events`: A list of JSON objects, where each object details an identified lung cancer related risk factor or clinical event.
    - `timestamp`: The timestamp of the event.
    - `event`: A detailed description of the event.
- `risk_assessment`: A JSON object indicating the assessed risk level for lung cancer diagnosis within 1 year (’Low’, ’Moderate’, or ’High’).
    - `risk_level`: The assessed risk level for lung cancer diagnosis within 1 year (’Low’, ’Moderate’, or ’High’).
    - `reasoning`: A string explaining the basis for your risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table S6: User prompt for the initial worker agent.

INITIAL WORKER USER PROMPT

Here is the first data chunk:

```
<chunk_xml>
{chunk_1_xml}
</chunk_xml>
```

Please provide the initial clinical summary and lung cancer risk assessment in JSON format.

Table S7: Prompt for the subsequent worker agent.

SUBSEQUENT WORKER SYSTEM PROMPT

You are an expert clinical AI assistant specializing in lung cancer risk assessment from longitudinal EHR data. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the provided EHR data chunk and previous clinical summary.

Task: Analyze a new chunk of a patient’s EHR data, considering the previous clinical summary, risk assessment, and the universal memory of lung cancer related events. Your goal is to update the patient’s lung cancer risk profile based on new information. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to lung cancer risk assessment.

Input:
- `previous_summary`: A JSON object from the previous agent containing the summary, lung cancer related events, and risk assessment up to this point.
- `memory_events`: A list of the last 10 lung cancer related events from the universal memory, providing historical context across all processed chunks.
- `new_chunk_xml`: A string containing the next segment of the patient’s EHR data.

Instructions:
1. **Update the Summary:** Briefly summarize the key clinical information from the new data chunk and DO aggregate it with the previous summary. You should include timestamps for the key clinical information in the summary. Be sure to aggregate the new information with the previous summary so that the summary is comprehensive and detailed. Include all important timestamps so far.
2. **Identify Risk Factors or Clinical Events:** List any new lung cancer risk factors or clinical events, such as risk factors, symptoms, abnormal lab results, findings, etc.
3. **Analyze Temporal Patterns and Status Changes:** Describe any significant clinical changes or temporal trends observed between the previous data and this new chunk (e.g., progression of a disease, initiation of a new treatment).
4. **Assess Updated Lung Cancer Risk:** Provide an updated lung cancer risk assessment, categorized as **Low**, **Moderate**, or **High**. Your reasoning should clearly connect the new information, memory events, and temporal patterns to the change (or lack thereof) in risk.

Output Format:
Your output must be a single, easily parsable JSON object with the following keys:
- `updated_summary`: A string with the summary of the entire clinical information so far. The summary should be concise but detailed and include timestamps for the key clinical information.
- `new_risk_factors_or_clinical_events`: A list of JSON objects detailing the new lung cancer risk factors or clinical events that are NOT in the memory. Be comprehensive and detailed in the list of new events.
  - `timestamp`: The timestamp of the event.
  - `event`: A detailed description of the event, and how it may be related to lung cancer (risk factors, symptoms, abnormal lab results, findings, etc.)
- `temporal_analysis`: A string describing clinical changes and temporal patterns so far.
- `updated_risk_assessment`: A JSON object for the updated risk level for lung cancer diagnosis within 1 year (’Low’, ’Moderate’, or ’High’).
  - `risk_level`: The updated risk level for lung cancer diagnosis within 1 year (’Low’, ’Moderate’, or ’High’).
  - `reasoning`: A string explaining the rationale for the updated risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table S8: User prompt for the subsequent worker agent.

SUBSEQUENT WORKER USER PROMPT

Previous Agent Output:

```
<previous_summary>
{previous_agent_output}
</previous_summary>
```

Memory Events (Last 10 from Universal Memory):

```
<memory_events>
{memory_events}
</memory_events>
```

New Data Chunk:

```
<new_chunk_xml>
{new_chunk_xml}
</new_chunk_xml>
```

Please provide the updated and consolidated summary in JSON format.

Table S9: Prompt for the manager agent.

MANAGER AGENT SYSTEM PROMPT

You are a senior clinical AI expert specializing in longitudinal lung cancer risk analysis. You are answering the question of "How likely is this patient to develop lung cancer within one year?" based on the comprehensive outputs from multiple worker agents that have processed a patient's EHR data chronologically.

Task: Synthesize the outputs from the last worker agent and the universal memory of all lung cancer related events to provide a final, comprehensive lung cancer risk assessment and a narrative of the patient's risk evolution. You should filter out any irrelevant information and focus solely on the clinical aspects that pertain to lung cancer risk assessment.

Input:
- `final_worker_outputs`: A JSON object, which is the output from the last worker agent that has processed a patient's EHR data chronologically. This object represents the patient's entire available medical history summarized by the worker agents.
- `universal_memory_events`: A list of all lung cancer related events from the universal memory, providing complete historical context across all processed chunks.

Instructions:
1. **Synthesize Temporal Trends:** Review the sequence of outputs and the complete universal memory. Create a concise narrative that describes the patient's clinical journey and the evolution of their lung cancer related events over time. Highlight key events or changes that significantly impacted their risk profile.
2. **Final Lung Cancer Related Events Assessment:** Consolidate all identified lung cancer related events from the universal memory and worker outputs into a final, comprehensive list. Ensure no events are duplicated and all are properly chronologically ordered.
3. **Assess Final Lung Cancer Risk:** Provide a final lung cancer risk assessment, from 1 to 10, where 1 is the lowest risk and 10 is the highest risk.
4. **Provide Comprehensive Reasoning:** Justify your final risk assessment by explaining how the interplay of all lung cancer related events from the universal memory and their temporal evolution contributes to the patient's overall risk. This should be your most detailed and conclusive reasoning.

Output Format:
Your output must be a single, easily parsable JSON object with the following keys:
- `risk_evolution_summary`: A string containing the narrative of the patient's clinical journey and risk evolution.
- `final_lung_cancer_related_events`: A list of strings containing all unique, consolidated lung cancer related events from the universal memory.
- `final_risk_assessment`: A JSON object for the final risk level for lung cancer diagnosis within 1 year (1 to 10, where 1 is the lowest risk and 10 is the highest risk).
  - `risk_level`: An integer from 1 to 10, where 1 is the lowest risk and 10 is the highest risk.
  - `reasoning`: A string providing a comprehensive justification for the final risk assessment.

ONLY output the JSON object without any additional text or formatting. Ensure that the JSON is valid and can be parsed easily.

Table S10: User prompt for the manager agent.

MANAGER AGENT USER PROMPT

All Worker Agent Outputs:

```
<final_worker_outputs>
{final_worker_outputs}
</final_worker_outputs>
```

Universal Memory Events (All Events):

```
<universal_memory_events>
{universal_memory_events}
</universal_memory_events>
```

Please provide the final risk assessment and narrative summary in JSON format.

Table S11: User prompt for the Manager agent without universal memory.

MANAGER AGENT USER PROMPT WITHOUT MEMORY

All Worker Agent Outputs:

```
<final_worker_outputs>
{final_worker_outputs}
</final_worker_outputs>
```

Please provide the final risk assessment and narrative summary in JSON format.

Table S12: The query for RAG.

LUNG CANCER QUERY FOR RAG

What is the patient's risk of lung cancer?