EHR-RAG: BRIDGING LONG-HORIZON STRUCTURED ELECTRONIC HEALTH RECORDS AND LARGE LANGUAGE MODELS VIA ENHANCED RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic Health Records (EHRs) provide rich longitudinal clinical evidence that is central to medical decision-making, motivating the use of retrieval-augmented generation (RAG) to ground large language model (LLM) predictions. However, long-horizon EHRs often exceed LLM context limits, and existing approaches commonly rely on truncation or vanilla retrieval strategies that discard clinically relevant events and temporal dependencies. To address these challenges, we propose *EHR-RAG*, a retrieval-augmented framework designed for accurate interpretation of long-horizon structured EHR data. *EHR-RAG* introduces three components tailored to longitudinal clinical prediction tasks: *Event- and Time-Aware Hybrid EHR Retrieval* to preserve clinical structure and temporal dynamics, *Adaptive Iterative Retrieval* to progressively refine queries in order to expand broad evidence coverage, and *Dual-Path Evidence Retrieval and Reasoning* to jointly retrieve and reason over both factual and counterfactual evidence to mitigate bias toward evidence supporting only one outcome. Experiments across four long-horizon EHR prediction tasks show that *EHR-RAG* consistently outperforms the strongest LLM-based baselines, achieving an average Macro-F1 improvement of 10.76%. Overall, our work highlights the potential of retrieval-augmented LLMs to advance clinical prediction on structured EHR data in practice.

1 INTRODUCTION

Electronic Health Records (EHRs) are digital longitudinal patient records composed of structured clinical events such as diagnoses, medications, laboratory results, and procedures. Unlike free-text clinical notes, structured EHR data are systematically collected, standardized across care settings, and less subjective in description, making them a reliable and widely available foundation for clinical decision making (Rosenbloom et al., 2011; Ebbers et al., 2022). Beyond individual events, structured EHRs encode rich temporal and categorical information, including event types, ordering, recurrence patterns, and long-term clinical trajectories, that underpins clinical decision-making across the full course of clinical care (Menachemi & Collum, 2011; Cowie et al., 2017). For many non-acute or chronic conditions, patients accumulate long-horizon EHRs spanning years of irregular visits and thousands of heterogeneous events. These trajectories are common in real-world healthcare but difficult to interpret, even for clinicians, and recent studies have highlighted both their modeling challenges and clinical importance (Loh et al., 2025; Wornow et al., 2023; Fries et al., 2025). Effective clinical prediction in this setting requires reasoning over long temporal spans while preserving event-type structure and clinical dependencies.

Traditional machine learning (ML) methods have been widely studied for EHR-based prediction tasks, including length-of-stay estimation (Cai et al., 2016; Levin et al., 2021), readmission prediction (Ashfaq et al., 2019; Xiao et al., 2018), and medication recommendation (Bhoi et al., 2021; Shang et al., 2019). More recently, large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) have shown promise in clinical reasoning due to their strong generalization capabilities to new tasks and their capability to perform multi-step reasoning to provide rationale instead of only final answer to make it more trustworthy for clinician's decision making (Bedi et al., 2025;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Jiang et al., 2023; 2024; Kojima et al., 2022). However, LLMs cannot natively process long-horizon structured EHRs: their fixed context windows are quickly exceeded by multi-visit patient histories, and naive event serialization obscures temporal structure and discards clinically relevant context.

Existing approaches address this limitation primarily through heuristic truncation, such as retaining only recent events (Lin et al., 2025) or selecting events associated with high-frequency codes (Liao et al., 2025). While retrieval-augmented generation (RAG) (Gao et al., 2023) offers a mechanism to selectively retrieve relevant events, existing RAG-based methods typically retrieve isolated subsets of the patient history primarily to satisfy context-length constraints (Zhu et al., 2024b; Kruse et al., 2025; Zhu et al., 2024a). As a result, vanilla RAG in clinical settings often suffers from low retrieval quality, incomplete evidence coverage, and reduced robustness when reasoning over long-horizon EHRs (Li et al., 2025b; Perçin et al., 2025). Despite the prevalence and importance of such data, to the best of our knowledge, no existing RAG framework is explicitly designed to handle their longitudinal, multi-event nature.

In this work, we study clinical prediction with long-horizon structured EHR data and propose ***EHR-RAG***, an enhanced retrieval-augmented framework tailored for this setting. *EHR-RAG* is built around three novel components designed to improve retrieval ***quality***, ***completeness***, and ***robustness*** under strict context constraints: (a) ***E**vent- and **T**ime-Aware **H**ybrid **EHR R**etrieval* (***ETHER***), which preserves event-type structure and temporal dynamics during retrieval; (b) ***A**daptive **I**terative **R**etrieval* (***AIR***), which progressively refines retrieval queries to expand evidence coverage while respecting context limits; and (c) ***D**ual-Path **E**vidence **R**etrieval and **R**easoning* (***DER***), which jointly retrieves and reasons over both factual patient history and counterfactual evidence to improve robustness. Together, these designs enable reliable long-horizon clinical reasoning with LLMs. We evaluate *EHR-RAG* against a range of baseline methods on a long-horizon EHR benchmark. Experimental results demonstrate consistent Macro-F1 improvements across all four clinical prediction tasks, with gains of 3.63% on *Long Length of Stay*, 11.28% on *30-day Readmission*, 16.46% on *Acute Myocardial Infarction*, and 7.66% on *Anemia* relative to the strongest LLM-based baselines. Overall, *EHR-RAG* achieves an average Macro-F1 improvement of 10.76%.

In summary, we make the following contributions:

- We identify and analyze the limitations of existing truncation-based and vanilla RAG approaches on long-horizon structured EHR data, highlighting challenges in evidence completeness and temporal reasoning for reliable clinical prediction.
- We propose *EHR-RAG*, a retrieval-augmented framework explicitly designed for long-horizon EHR reasoning, integrating event- and time-aware retrieval, adaptive iterative retrieval, and dual-path evidence retrieval.
- We conduct extensive experiments demonstrating consistent improvements over strong baselines across multiple long-horizon EHR prediction tasks.

2 Related Work

Clinical Predictive Models. Structured EHR data has supported the development of a wide range of machine learning models for different clinical prediction tasks (Cai et al., 2016; Levin et al., 2021; Ashfaq et al., 2019; Bhoi et al., 2021). Models such as RETAIN (Choi et al., 2016), GRAM (Choi et al., 2017), and KerPrint (Yang et al., 2023) are specifically designed to capture complex temporal and hierarchical patterns within structured EHRs and have shown strong performance across multiple predictive settings. In parallel, recent efforts have focused on scaling clinical prediction through large foundation models, including CLMBR-T-Base (Wornow et al., 2023), Virchow (Vorontsov et al., 2024), and MIRA (Li et al., 2025a). However, traditional predictive models—ranging from task-specific architectures to supervised foundation model training—remain inflexible, requiring labeled data for every downstream task and often failing to generalize beyond the distribution on which they were trained. These limitations are especially problematic in the dynamic and heterogeneous healthcare environment.

LLM-based Clinical Prediction. Recently, the paradigm of clinical prediction has begun to shift toward the use of LLMs (Achiam et al., 2023). LLMs offer greater adaptability and generalization than traditional supervised models, with the capacity to interpret diverse medical information and support more versatile clinical decision-making. Several studies apply LLMs directly to clinical

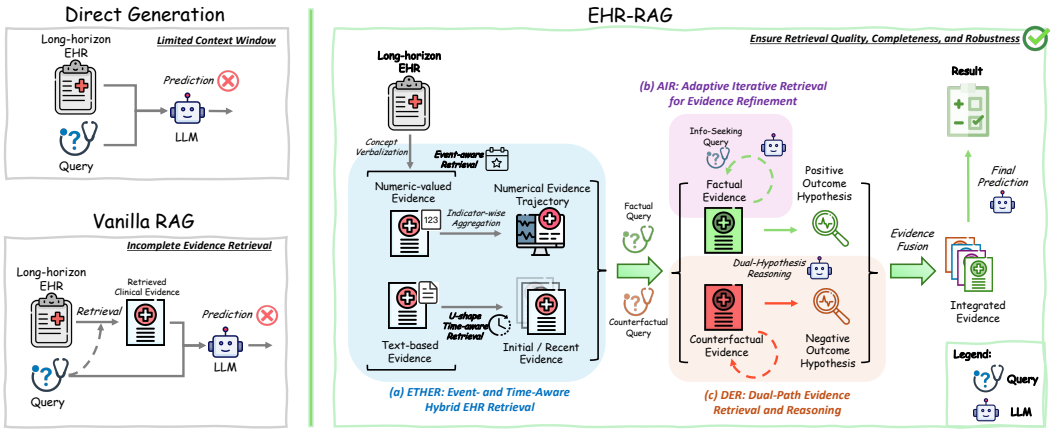

Figure 1: Overview of the ***EHR-RAG*** framework for long-horizon clinical prediction. Compared with *Direct Generation* and *Vanilla RAG*, our framework explicitly addresses context truncation and incomplete retrieval. It integrates ***(a) Event- and Time-aware Hybrid Retrieval (ETHER)***, ***(b) Adaptive Iterative Retrieval (AIR)***, and ***(c) Dual-Path Factual and Counterfactual Reasoning (DER)***, ensuring retrieval quality, robustness, and completeness for reliable clinical prediction.

prediction tasks (Lovon-Melgarejo et al., 2024; Zhu et al., 2024c; Chen et al., 2024; Kruse et al., 2025), while others employ LLMs as encoders for structured EHR data (Hegselmann et al., 2025). Additional efforts explore training or adapting LLMs for specific downstream healthcare tasks (Lin et al., 2025; Jiang et al., 2023; 2024; Yang et al., 2022), and recent work introduces LLM-driven agents to assist with clinical reasoning and EHR interaction (Cui et al., 2025; Shi et al., 2024). Despite these advances, few studies have addressed long-horizon EHR prediction, leaving substantial challenges in enabling LLMs to interpret complex, multi-visit longitudinal patient histories. Our work extends this line of research by developing an LLM-based approach specifically tailored for long-horizon EHR prediction.

Retrieval-Augmented Generation with LLM. RAG has emerged as an effective technique for enhancing LLMs with external knowledge beyond their limited context window (Gao et al., 2023). A variety of retrieval strategies have been proposed to help LLMs accurately access additional relevant information, leading to improved performance across diverse tasks and settings (Asai et al., 2023; Jiang et al., 2025b;a). In the healthcare domain, where clinical information is often overwhelming in volume and complexity, RAG has become increasingly popular for grounding LLM outputs in accurate task-relevant evidence (Neha et al., 2025; Kim et al., 2025). Some studies use RAG to inject medical knowledge and improve healthcare applications (Soman et al., 2024; Niu et al., 2024; Cao et al., 2024a), while others employ RAG to mitigate long-context limitations. For example, REALM (Zhu et al., 2024b) enhances multimodal EHR analysis using LLMs, and EMERGE (Zhu et al., 2024a) integrates RAG into multimodal EHR predictive modeling. Additional work has investigated RAG-based methods for clinical trial and patient matching (Tramontini et al., 2025). However, most of these approaches focus on unstructured EHR data such as clinical notes, rather than structured tabular EHR, which exhibits distinct temporal dynamics and event-type structure. Moreover, vanilla RAG pipelines often suffer from incomplete evidence retrieval in clinical settings. Our work aims to fill this gap by designing a RAG framework specifically tailored for long-horizon EHR data.

3 METHODOLOGY

Figure 1 illustrates *EHR-RAG*, a retrieval-augmented framework designed to enable LLMs to reason over long-horizon structured EHR data under strict context constraints. The framework comprises three core components: *Event- and Time-Aware Hybrid EHR Retrieval (ETHER)*, *Adaptive Iterative Retrieval (AIR)*, and *Dual-Path Evidence Retrieval and Reasoning (DER)*.

3.1 TASK FORMULATION

Given a patient p, we assume access to a longitudinal structured EHR $\mathcal{E}_p = \{e_i\}_{i=1}^{N}$, where each clinical event is represented as $e_i = (c_i, v_i, \tau_i)$, with c_i denoting a clinical concept (e.g., diagnosis, procedure, laboratory test, or medication), v_i the associated value (numeric or textual), and τ_i the timestamp at which the event occurred. Events are ordered chronologically such that $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_N$.

At a prediction time τ^*, the available patient history is

$$\mathcal{E}_p^{\leq \tau^*} = \{e_i \in \mathcal{E}_p \mid \tau_i \leq \tau^*\}. \tag{1}$$

Our goal is to predict a clinical outcome $y \in \mathcal{Y}$ (binary or multiclass) based on this long-horizon EHR context. Examples include prolonged length of stay, laboratory abnormality severity, or risk of acute clinical events.

We assume access to a pretrained LLM $\mathcal{M}$ capable of clinical reasoning but limited by a fixed context window, making directly input of $\mathcal{E}_p^{\leq \tau^*}$ infeasible when N is large. Therefore, we aim to construct a compact, task-relevant evidence context $\mathcal{C} \subset \mathcal{E}_p^{\leq \tau^*}$ such that $\hat{y} = \mathcal{M}(\mathcal{C})$, is consistent with the ground-truth outcome y.

3.2 EVENT- AND TIME-AWARE HYBRID EHR RETRIEVAL

Structured EHR $\mathcal{E}_p^{\leq \tau^*}$ contain heterogeneous event types, including numeric measurements (e.g., lab tests and vital signs) and textual records (e.g., event descriptions and clinical notes). Treating all clinical events uniformly during retrieval is often ineffective. Dense text embeddings are primarily optimized for capturing semantic similarity and are known to be insensitive to precise numerical values and magnitude differences (Wallace et al., 2019; Zhang et al., 2020), which limits their effectiveness for reasoning over numeric clinical measurements such as laboratory results and vital signs. Moreover, raw numerical values alone do not fully convey their clinical significance without appropriate temporal context and longitudinal interpretation. Motivated by these limitations, we propose an event- and time-aware hybrid retrieval strategy that explicitly handles numeric events separately while preserving their temporal structure.

Indicator-wise aggregation of numeric events. All numeric clinical events are first grouped by clinical indicator (i.e., event name), forming a collection of indicator-specific temporal trajectories. Formally, for each indicator k, we define its numeric value trajectory as $\mathcal{V}_k = \{(v_{k,j}, \tau_{k,j}) \mid \tau_{k,j} \leq \tau^*\}$, where $v_{k,j}$ denotes the j-th observed numeric value of indicator k and $\tau_{k,j}$ is its corresponding timestamp. Collectively, the set $\{\mathcal{V}_k\}$ forms an indicator-level numeric representation of the patient, where each $\mathcal{V}_k$ captures the longitudinal evolution of a single clinical measurement. This representation preserves clinically meaningful temporal dynamics while enabling efficient and semantically informed retrieval over numeric EHR data.

Coarse-to-fine indicator selection. Given a task-specific query q_{task}, we first perform coarse-grained retrieval over all indicator-level summaries $\{\mathcal{V}_k\}$ using dense similarity search, where embeddings are computed from indicator names. This step yields a candidate indicator set $\mathcal{K}_{\text{coarse}} = \{k_1, \ldots, k_{N_{\text{coarse}}}\}$, where N_{coarse} denotes the number of indicators retained after coarse retrieval. We then apply LLM-based reranking to select a smaller, task-relevant subset $\mathcal{K}_{\text{fine}} \subset \mathcal{K}_{\text{coarse}}$ such that $|\mathcal{K}_{\text{fine}}| = N_{\text{fine}}$. For each selected indicator $k \in \mathcal{K}_{\text{fine}}$, we retain the N_{recent} most recent measurements occurring prior to the prediction time τ^* to ensure temporal relevance, forming an indicator-specific numeric evidence trajectory $\mathcal{E}_k = \{(v_{k,j}, \tau_{k,j}) \in \mathcal{V}_k \mid \tau_{k,j} \leq \tau^*\}$. Collectively, these indicator-level evidence trajectories constitute the final numeric evidence collection $\mathcal{E}_{\text{num}} = \{\mathcal{E}_k \mid k \in \mathcal{K}_{\text{fine}}\}$.

U-shape time-aware retrieval for textual events. Non-numeric clinical events (e.g., diagnoses, procedures, and clinical notes) are serialized at the event level and segmented into overlapping temporal chunks, with each chunk embedded and indexed in a vector store. Given a task-specific query q_{task} and prediction time τ^*, we first retrieve a candidate pool of K_{cand} textual chunks using dense semantic similarity search, and then jointly consider semantic relevance and temporal proximity to select textual evidence.

Prior work has explored incorporating temporal information to improve retrieval quality, for example by introducing time-aware relevance weighting or decay functions (Cao et al., 2024b; Abdallah

et al., 2025). However, in clinical settings, temporal importance is not strictly monotonic: while recent events are often critical, early events corresponding to disease onset or initial admission can be equally informative. This observation motivates a U-shaped time-aware retrieval strategy that explicitly emphasizes both recent and early clinical evidence.

For each candidate textual chunk c with timestamp τ_c, we compute a semantic similarity score $s_{\text{sem}}(q_{\text{task}}, c)$ using dense embeddings, together with a U-shaped temporal relevance score. Let τ_{first} denote the earliest timestamp in the patient record. The temporal relevance score is defined as

$$s_{\text{time}}(\tau_c) = \max\left(\exp\left(-\frac{\tau^* - \tau_c}{\tau_{\text{recent}}}\right), \exp\left(-\frac{\tau_c - \tau_{\text{first}}}{\tau_{\text{early}}}\right)\right). \tag{2}$$

This U-shaped formulation assigns higher importance to events occurring close to the prediction time τ^* or near the beginning of the patient trajectory, while downweighting mid-history events that are less informative for the current prediction. The final retrieval score for each textual chunk is computed as a convex combination of semantic relevance and temporal importance,

$$s(c) = \alpha\, s_{\text{sem}}(q_{\text{task}}, c) + (1 - \alpha)\, s_{\text{time}}(\tau_c), \tag{3}$$

where $\alpha \in [0, 1]$ controls the trade-off between semantic similarity and temporal relevance. Candidate chunks are ranked by $s(c)$, and the top-K_{final} chunks are selected and temporally ordered to form the final textual evidence set $\mathcal{E}_{\text{text}} = \{c_1, \ldots, c_{K_{\text{final}}}\}$.

By combining indicator-level numeric evidence $\mathcal{E}_{\text{num}}$ with U-shaped time-aware textual evidence $\mathcal{E}_{\text{text}}$, *ETHER* ensures broad semantic coverage while preserving event-type distinctions and clinically meaningful temporal structure under a compact evidence budget.

3.3 Adaptive Iterative Retrieval

Single-pass retrieval is often insufficient for long-horizon EHRs, where relevant clinical evidence may be temporally dispersed or only indirectly related to the initial query. To address this limitation, we progressively expands evidence coverage in a controlled and targeted manner.

Starting from an initial query $q^{(0)} = q_{\text{task}}$, we retrieve an initial evidence set $\mathcal{E}^{(0)}$. At each iteration t, the LLM assesses whether the current evidence set $\mathcal{E}^{(t)}$ is sufficient to answer the clinical prediction task. If so, the retrieval process terminates. Otherwise, the LLM generates a refined query

$$q^{(t+1)} = \mathcal{M}_{\mathcal{R}}\left(q^{(t)}, \mathcal{E}^{(t)}\right), \tag{4}$$

where $\mathcal{M}_{\mathcal{R}}(\cdot)$ denotes an LLM-based query refinement module that identifies a single missing yet clinically salient aspect of the current evidence. Each refined query is explicitly constrained to be concise, focused on one clinical dimension, and non-redundant with previously retrieved information.

Evidence retrieved using the refined query is then merged with the existing context via deduplication and temporal ordering:

$$\mathcal{E}^{(t+1)} = \text{Merge}\left(\mathcal{E}^{(t)}, \text{Retrieve}(q^{(t+1)})\right). \tag{5}$$

This iterative process continues until the evidence is deemed sufficient or a predefined iteration limit is reached. By incrementally expanding the evidence set in a targeted manner, AIR improves retrieval recall while preventing uncontrolled context growth, which is critical for effective long-horizon clinical reasoning.

3.4 Dual-Path Evidence Retrieval and Reasoning

Clinical prediction often involves competing hypotheses (Sox et al., 2024; Elstein et al., 1978), and reasoning along a single evidence pathway can lead to biased or overconfident conclusions (Graber et al., 2005; Croskerry, 2003). To improve robustness, we propose a dual-path evidence retrieval and reasoning strategy.

Specifically, we construct two complementary retrieval queries: a **factual (positive) query** q^+ that seeks evidence supporting the presence of the target outcome, and a **counterfactual (negative)**

query q^- that seeks evidence supporting its absence. Each query is processed independently using the adaptive iterative retrieval mechanism, yielding two textual evidence sets, $\mathcal{E}^+_{\text{text}}$ and $\mathcal{E}^-_{\text{text}}$. For each path, the retrieved textual evidence is combined with the shared numeric evidence $\mathcal{E}_{\text{num}}$, and the LLM is prompted to form an explicit outcome hypothesis:

$$h^+ = \mathcal{M}\big(\mathcal{E}^+_{\text{text}} \cup \mathcal{E}_{\text{num}}\big), \quad h^- = \mathcal{M}\big(\mathcal{E}^-_{\text{text}} \cup \mathcal{E}_{\text{num}}\big), \tag{6}$$

where h^+ and h^- denote the positive and negative outcome hypotheses, respectively. The corresponding evidence sets are then fused into a unified evidence context:

$$\mathcal{E}^{\text{fuse}} = \mathcal{E}^+_{\text{text}} \cup \mathcal{E}^-_{\text{text}} \cup \mathcal{E}_{\text{num}}. \tag{7}$$

Finally, the model is explicitly prompted to compare the *strength*, *directness*, and *clinical relevance* of evidence supporting each outcome hypothesis, and to produce a final prediction conditioned on both hypotheses and the fused evidence:

$$\hat{y} = \mathcal{M}_{\text{dec}}\big(\mathcal{E}^{\text{fuse}}, h^+, h^-\big), \tag{8}$$

where $\mathcal{M}_{\text{dec}}(\cdot)$ denotes the LLM-based decision function that performs comparative hypothesis evaluation and outputs the final clinical prediction.

Overall, this dual-path design enables balanced hypothesis evaluation, mitigates confirmation bias and spurious correlations, and improves the robustness and reliability of long-horizon clinical prediction from structured EHR data.

4 Experiments

4.1 Experimental Setup

To systematically evaluate our proposed method and compare it against baseline approaches under long-horizon clinical prediction settings, we conduct all experiments on the EHRSHOT benchmark (Wornow et al., 2023). Unlike commonly used EHR dataset, such as MIMIC-III/IV (Johnson et al., 2016; 2023), which are largely restricted to ICU or emergency department settings, EHRSHOT captures longitudinal patient records across general hospital care. As a result, patient records in EHRSHOT often span multiple decades and contain thousands of clinical events. On average, EHRSHOT includes 2.3× more clinical events and 95.2× more encounters per patient than MIMIC-IV, making it substantially more challenging and better suited for evaluating long-horizon clinical reasoning.

We select four representative prediction tasks from EHRSHOT: *Long Length of Stay*, *30-day Readmission*, *Acute Myocardial Infarction (Acute MI)*, and *Anemia*. The first two tasks fall under the *Operational Outcomes* category, Anemia belongs to *Anticipating Lab Test Results*, and Acute MI is categorized as *Assignment of New Diagnoses*. The first three tasks are binary classification problems, while Anemia is a four-class classification task. Together, these tasks span diverse clinical objectives and temporal reasoning scenarios, enabling a comprehensive evaluation of long-horizon reasoning on structured EHR data. We report Accuracy, Macro-F1, and per-class F1 scores, as most clinical classification tasks are inherently imbalanced.

We conduct experiments using three LLMs that span both proprietary and open-source settings: *GPT-5*, *Claude-Opus-4.5*, and *LLaMA-3.1-8B*. These models cover a broad range of architectures, training scales, model sizes, and accessibility levels, allowing us to assess the robustness of *EHR-RAG* across different LLM backbones. Additional details of the experimental setup are provided in Appendix D.

4.2 Baselines

We compare *EHR-RAG* with commonly used LLM-based baselines for EHR data, including direct generation and diverse RAG approaches, under matched inference budgets.

- *Direct Generation* (Lin et al., 2025): The LLM directly predicts the clinical outcome from the patient EHR without retrieval. To satisfy context constraints, the EHR is truncated by retaining the most recent events up to the maximum context window, a common heuristic in prior LLM-based EHR studies. This baseline evaluates the performance of naive context truncation without evidence selection.

Table 1: Performance of *EHR-RAG* and other LLM-based baselines on *Long Length of Stay* and *30-day Readmission* using *GPT-5*. We report Accuracy (%), Macro-F1 (%), and per-class F1 scores (%). **Bold** indicates the best performance, and improvements over the strongest baseline are highlighted in green. *EHR-RAG* consistently outperforms other methods on both tasks.

Method	Long Length of Stay				30-day Readmission			
	Accuracy	Macro F1	$F1_{Short}$	$F1_{Long}$	Accuracy	Macro F1	$F1_{No\ Readmit}$	$F1_{Readmit}$
Direct Generation	69.41	66.52	76.36	56.67	46.56	44.37	55.41	33.33
RAG	68.24	64.77	75.82	53.71	45.80	43.40	55.06	31.73
Uniform RAG	65.10	61.50	73.27	49.72	52.67	49.63	62.65	35.42
Rule-based RAG	63.53	56.58	73.95	39.22	57.63	44.08	71.61	16.54
ReAct RAG	65.10	60.85	73.75	47.95	48.09	45.22	57.76	32.67
EHR-RAG (Ours)	**74.12**$_{+4.71}$	**70.15**$_{+3.63}$	**81.03**$_{+4.67}$	**59.26**$_{+2.59}$	**71.76**$_{+14.13}$	**60.91**$_{+11.28}$	**81.50**$_{+9.89}$	**40.32**$_{+4.90}$

Table 2: Performance of *EHR-RAG* and other LLM-based baselines on *Acute Myocardial Infarction* and *Anemia* using *GPT-5*. We report Accuracy (%), Macro-F1 (%), and per-class F1 scores (%). **Bold** indicates the best performance, and improvements over the strongest baseline are highlighted in green. *EHR-RAG* consistently outperforms other methods on both tasks.

Method	Acute MI				Anemia					
	Accuracy	Macro F1	$F1_{No\ MI}$	$F1_{MI}$	Accuracy	Macro F1	$F1_{Low}$	$F1_{Medium}$	$F1_{High}$	$F1_{Abnormal}$
Direct Generation	88.38	57.97	93.72	22.22	44.57	28.38	70.74	17.30	18.48	7.02
RAG	90.04	59.83	94.67	25.00	42.08	20.44	70.43	18.85	12.90	0.00
Uniform RAG	89.21	47.15	94.30	0.00	48.42	32.39	72.65	26.97	22.68	7.27
Rule-based RAG	91.29	52.06	95.42	8.70	44.34	26.21	67.61	17.33	16.30	3.57
ReAct RAG	89.21	56.49	94.22	18.75	46.15	31.64	69.37	27.23	22.68	7.27
EHR-RAG (Ours)	**92.95**$_{+1.66}$	**76.29**$_{+16.46}$	**96.16**$_{+0.74}$	**56.41**$_{+31.41}$	**57.01**$_{+8.59}$	**44.07**$_{+11.68}$	**80.31**$_{+7.66}$	**43.00**$_{+15.77}$	**43.59**$_{+20.91}$	**9.38**$_{+2.11}$

- *RAG* (Gao et al., 2023): A vanilla retrieval approach that retrieves relevant EHR evidence and conditions the LLM on the retrieved context. This baseline reflects the most common single-pass RAG formulation used in recent clinical and non-clinical applications.
- *Uniform RAG* (Liu et al., 2024): A retrieval baseline that uniformly samples EHR evidence without semantic prioritization. This baseline controls for context length and isolates the benefit of relevance-aware retrieval over random selection.
- *Rule-based RAG* (Liao et al., 2025): A heuristic retrieval approach that selects EHR evidence using predefined rules rather than learned relevance signals. In our implementation, events are ranked by occurrence frequency, and the top frequently occurring events are selected. This baseline reflects commonly adopted heuristic filtering strategies for handling long EHR sequences.
- *ReAct RAG* (Yao et al., 2022): A reasoning-and-acting framework that interleaves reasoning steps with retrieval actions. The LLM iteratively generates reasoning traces and retrieval queries to obtain additional EHR evidence, evaluating whether generic iterative retrieval improves long-horizon EHR reasoning without domain-specific design.

All baselines use the same underlying LLM backbone and context budget as *EHR-RAG* to ensure a fair and controlled comparison.

Additionally, in Section E, we compare *EHR-RAG* with classical ML baselines following the EHRSHOT benchmark to assess whether LLM-based reasoning offers benefits beyond classical EHR models:

- *Count-based LR* (Wornow et al., 2023): Logistic regression trained on count-based clinical features extracted from the EHR.
- *CLMBR-based LR* (Wornow et al., 2023): Logistic regression trained on patient representations produced by the pretrained *CLMBR-T-Base* foundation model (Wornow et al., 2023; Steinberg et al., 2021).

4.3 Main Results

Tables 1 and 2 present the main results on the four long-horizon clinical prediction tasks from EHRSHOT, spanning operational outcomes, diagnosis prediction, and laboratory abnormality assessment.

Table 3: Ablation results on the *Long Length of Stay*, *Anemia*, *Acute MI*, and *30-day Readmission* tasks using *GPT-5*. We report Accuracy (%) and Macro-F1 (%). Red text indicates performance drops relative to *EHR-RAG*.

Method	Long Length of Stay		Anemia		Acute MI		30-day Readmission	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
EHR-RAG (Ours)	**74.12**	**70.15**	**57.01**	**44.07**	**92.95**	**76.29**	**71.76**	**60.91**
w/o *ETHER*	72.94 -1.18	68.47 -1.68	53.39 -3.62	40.12 -3.95	87.50 -5.45	74.91 -1.38	67.94 -3.82	55.13 -5.78
w/o *AIR*	72.16 -1.96	68.19 -1.96	54.07 -2.94	41.44 -2.63	86.36 -6.59	74.52 -1.77	67.56 -4.20	56.29 -4.62
w/o *DER*	70.59 -3.53	67.55 -2.60	46.83 -10.18	38.58 -5.49	66.67 -26.28	62.50 -13.79	65.19 -6.57	53.47 -7.44
Vanilla RAG	68.24 -5.88	64.77 -5.38	42.08 -14.93	20.44 -23.63	90.04 -2.91	59.83 -16.46	45.80 -25.96	43.40 -17.51

(i) Operational outcome prediction. As shown in Table 1, *EHR-RAG* consistently outperforms all LLM-based baselines on both *Long Length of Stay* and *30-day Readmission*. For length of stay prediction, *EHR-RAG* achieves the highest accuracy (74.12) and macro-F1 (70.15), with substantial improvements over the strongest baseline across both short-stay and long-stay classes. On the highly imbalanced readmission task, *EHR-RAG* also yields notable improvements in macro-F1 (60.91).

(ii) Clinical diagnosis and lab prediction. Table 2 presents results on *Acute MI* and *Anemia*. For acute MI prediction, *EHR-RAG* substantially improves macro-F1 (76.29) and F1 on the positive MI class (56.41), indicating more effective identification of clinically meaningful risk signals from long-horizon EHRs. On the multi-class anemia task, *EHR-RAG* achieves the best overall accuracy (57.01) and macro-F1 (44.07), with consistent improvements across all severity levels, including the rare abnormal class.

(iii) Baseline analysis. We observe that vanilla RAG can underperform direct generation in some cases, suggesting that retrieving events solely based on semantic similarity, without explicitly modeling recency and temporal context, may omit clinically critical recent evidence and lead to degraded performance. Uniform RAG exhibits unstable behavior, performing well in some cases but poorly in others, suggesting that while information across the patient history contains useful signals, indiscriminate retrieval lacks robustness. Rule-based RAG yields moderate improvements but introduces systematic bias by prioritizing high-frequency events, which can skew predictions toward dominant classes. ReAct RAG occasionally provides marginal gains over RAG by enabling iterative retrieval and broader evidence coverage; however, it does not consistently translate additional retrieval steps into reliable performance improvements.

(iv) Overall comparison. Across all four tasks, *EHR-RAG* consistently outperforms direct generation, vanilla RAG, and other retrieval baselines. These results highlight the importance of jointly modeling event-type structure, temporal dynamics, and evidence diversity through event- and time-aware retrieval, adaptive iterative refinement, and dual-path evidence retrieval and reasoning for accurate and reliable long-horizon clinical prediction from structured EHR data. Overall, these findings demonstrate that effective long-horizon EHR reasoning requires structured, temporally aware, and diversity-preserving retrieval mechanisms rather than generic RAG strategies. Appendix L presents a qualitative case study comparing *EHR-RAG* with direct generation and vanilla RAG.

4.4 Ablation Analysis of Key Components

Table 3 reports the ablation results on *Long Length of Stay* and *Anemia*, analyzing the contribution of each core component of *EHR-RAG*. Red subscripts indicate the absolute performance drop relative to the full framework.

(i) Overall impact of each component. Excluding *Event- and Time-Aware Hybrid EHR Retrieval (ETHER)* leads to clear drops, particularly on *Anemia*, highlighting the importance of preserving event-type distinctions and temporal structure when retrieving long-horizon clinical evidence. Removing *Adaptive Iterative Retrieval (AIR)* results in uniform declines across all metrics, suggesting that iterative query refinement is critical for recovering sparse but task-relevant evidence distributed over time. Finally, ablating *Dual-Path Evidence Retrieval Reasoning (DER)* causes the largest Macro-F1 degradation on *Anemia*, underscoring the value of explicitly modeling factual and counterfactual retrieval and reasoning in complex and multi-class clinical prediction settings. Overall, the full framework achieves the best performance across all metrics, while each ablated variant

Table 4: Performance comparison of *EHR-RAG* and other LLM-based baselines across different LLM backbones (*GPT-5*, *Claude-Opus-4.5*, and *LLaMA-3.1-8B*) on the *Long Length of Stay* task. We report Accuracy (%), Macro-F1 (%), and per-class F1 scores (%). **Bold** indicates the best performance, and improvements over the strongest baseline are highlighted in green.

Method	GPT-5				Claude-Opus-4.5				LLaMA-3.1-8B			
	Accuracy	Macro-F1	$F1_{Short}$	$F1_{Long}$	Accuracy	Macro-F1	$F1_{Short}$	$F1_{Long}$	Accuracy	Macro-F1	$F1_{Short}$	$F1_{Long}$
Direct Generation	69.41	66.52	76.36	56.67	39.22	38.98	35.15	42.80	33.73	33.67	27.47	**39.13**
RAG	68.24	64.77	75.82	53.71	49.02	48.96	50.76	47.15	27.84	25.73	13.21	38.26
Uniform RAG	65.10	61.50	73.27	49.72	47.45	47.39	49.24	45.53	30.59	29.48	20.63	38.33
Rule-based RAG	63.53	56.58	73.95	39.22	45.10	45.03	46.97	43.09	40.78	38.79	49.83	27.75
ReAct RAG	65.10	60.85	73.75	47.95	47.84	47.76	45.71	49.81	28.63	26.92	15.74	38.10
EHR-RAG (Ours)	**74.12**$_{+4.71}$	**70.15**$_{+3.63}$	**81.03**$_{+4.67}$	**59.26**$_{+2.59}$	**56.08**$_{+7.06}$	**55.58**$_{+6.62}$	**60.28**$_{+9.52}$	**50.88**$_{+1.07}$	**46.67**$_{+5.89}$	**44.62**$_{+5.83}$	**55.26**$_{+5.43}$	33.98$_{-5.15}$

exhibits systematic performance drops, demonstrating that all components of *EHR-RAG* contribute meaningfully and synergistically to the overall gains.

(ii) Comparison with vanilla RAG. Vanilla RAG performs substantially worse than all *EHR-RAG* variants, with especially severe degradation on *Anemia* (14.93 accuracy and 23.63 Macro-F1 points). This confirms that naive retrieval alone is insufficient for long-horizon structured EHR reasoning, and motivates the need for the proposed event-aware, iterative, and dual-path design.

4.5 Performance Across Different LLM Backbones

Table 4 compares the performance of different LLM backbones on the *Long Length of Stay* task, including both proprietary and open-source models with varying capacities.

(i) Consistent gains across models. *EHR-RAG* yields consistent performance improvements across *GPT-5*, *Claude-Opus-4.5*, and *LLaMA-3.1-8B*, despite substantial differences in model architecture and training scale. This consistency indicates that the proposed retrieval and reasoning framework generalizes across LLM backbones, rather than relying on the characteristics of any single model.

(ii) Larger gains on smaller models. While *GPT-5* achieves the highest absolute performance, the relative improvements introduced by *EHR-RAG* are more pronounced for *LLaMA-3.1-8B*. Compared to the strongest baseline, *EHR-RAG* improves Macro-F1 by over 6 points on both models, suggesting that its design can partially mitigate the limitations of smaller model capacity.

4.6 Additional Analysis of Efficiency and Comparison with More Baselines

Detailed analyses of computational efficiency are provided in Appendix H. We compare *EHR-RAG* with traditional ML baselines in Appendix E and with a broader set of RAG and memory-centric baselines in Appendix J.

Regarding efficiency, *EHR-RAG* incurs moderately higher computational cost than Direct Generation, which represents the most straightforward way of providing serialized EHR data to LLMs, while achieving substantially stronger predictive performance. In medical prediction tasks, even modest performance improvements can translate into meaningful downstream value; therefore, we view these gains as substantial rather than negligible. Overall, we believe the additional overhead is justified by the improved evidence quality and more reliable long-horizon clinical prediction.

Compared with traditional ML methods, our LLM-based framework is more general and can be adapted to different clinical prediction tasks without task-specific training. Compared with broader RAG and memory-centric baselines, *EHR-RAG* is specifically designed for structured longitudinal EHR data, including heterogeneous clinical events, temporal dependencies, and dispersed evidence. This task-specific design naturally leads to better performance in long-horizon EHR reasoning.

5 Conclusion

In this paper, we introduce *EHR-RAG*, an enhanced retrieval-augmented framework designed to bridge long-horizon EHRs and LLMs. *EHR-RAG* addresses key challenges posed by long-horizon EHR data. Extensive experiments demonstrate that *EHR-RAG* consistently outperforms other LLM-based baselines across diverse long-horizon clinical prediction tasks.

REFERENCES

Abdelrahman Abdallah, Bhawna Piryani, Jonas Wallat, Avishek Anand, and Adam Jatowt. Extending dense passage retrieval with temporal information. *arXiv e-prints*, pp. arXiv–2502, 2025.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.

Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Sławomir Nowaczyk. Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, 97: 103256, 2019.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 2025.

Suman Bhoi, Mong Li Lee, Wynne Hsu, Hao Sen Andrew Fang, and Ngiap Chuan Tan. Personalizing medication recommendation with a graph-based approach. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–23, 2021.

Xiongcai Cai, Oscar Perez-Concha, Enrico Coiera, Fernando Martin-Sanchez, Richard Day, David Roffe, and Blanca Gallego. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, 23 (3):553–561, 2016.

Lang Cao, Jimeng Sun, and Adam Cross. Autord: An automatic and end-to-end system for rare disease knowledge graph construction based on ontologies-enhanced large language models. *arXiv preprint arXiv:2403.00953*, 2024a.

Lang Cao, Zifeng Wang, Cao Xiao, and Jimeng Sun. PILOT: Legal case outcome prediction with case law. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 609–621, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.34. URL https://aclanthology.org/2024.naacl-long.34/.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*, 2024.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 787–795, 2017.

Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, 2017.

Pat Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8):775–780, 2003.

Hejie Cui, Zhuocheng Shen, Jieyu Zhang, Hui Shao, Lianhui Qin, Joyce C Ho, and Carl Yang. Llms-based few-shot disease predictions using ehr: A novel approach combining predictive agent reasoning and critical agent instruction. In *AMIA Annual Symposium Proceedings*, volume 2024, pp. 319, 2025.

Tom Ebbers, Rudolf B Kool, Ludi E Smeele, Richard Dirven, Chrisje A den Besten, Luc HE Karssemakers, Tim Verhoeven, Jasmijn M Herruer, Guido B van den Broek, and Robert P Takes. The impact of structured and standardized documentation on documentation quality; a multicenter, retrospective study. *Journal of Medical Systems*, 46(7):46, 2022.

Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, 1978.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, et al. Lightmem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*, 2025.

Jason Alan Fries, Michael Wornow, Ethan Steinberg, Zepeng Frazier Huo, Hejie Cui, Suhana Bedi, Alyssa Unell, and Nigam Shah. Advancing responsible healthcare ai with longitudinal ehr datasets, 2025. Stanford HAI News.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Mark L Graber, Nancy Franklin, and Ruthanna Gordon. Diagnostic error in internal medicine. *Archives of internal medicine*, 165(13):1493–1499, 2005.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Stefan Hegselmann, Georg von Arnim, Tillmann Rheude, Noel Kronenberg, David Sontag, Gerhard Hindricks, Roland Eils, and Benjamin Wild. Large language models are powerful electronic health record encoders. *arXiv preprint arXiv:2502.17403*, 2025.

Shih-Cheng Huang, Zepeng Huo, Ethan Steinberg, Chia-Chun Chiang, Matthew P Lungren, Curtis P Langlotz, Serena Yeung, Nigam H Shah, and Jason A Fries. Inspect: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*, 2023.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*, 2023.

Pengcheng Jiang, Cao Xiao, Minhao Jiang, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. *arXiv preprint arXiv:2410.04585*, 2024.

Pengcheng Jiang, Lang Cao, Ruike Zhu, Minhao Jiang, Yunyi Zhang, Jimeng Sun, and Jiawei Han. Ras: Retrieval-and-structuring for knowledge-intensive llm generation. *arXiv preprint arXiv:2502.10996*, 2025a.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, R. Tian, S. Kang, Z. Wang, Jimeng Sun, and Jiawei Han. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv: 2503.00223*, 2025b.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Hyunjae Kim, Jiwoong Sohn, Aidan Gilson, Nicholas Cochran-Caggiano, Serina Applebaum, Heeju Jin, Seihee Park, Yujin Park, Jiyeong Park, Seoyoung Choi, et al. Rethinking retrieval-augmented generation for medicine: A large-scale, systematic expert evaluation and practical insights. *arXiv preprint arXiv:2511.06738*, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20715–20735, 2025.

Scott Levin, Sean Barnes, Matthew Toerper, Arnaud Debraine, Anthony DeAngelo, Eric Hamrock, Jeremiah Hinson, Erik Hoyer, Trushar Dungarani, and Eric Howell. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. *BMJ Innovations*, 7(2), 2021.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, et al. Mira: Medical time series foundation model for real-world health data. *arXiv preprint arXiv:2506.07584*, 2025a.

Zhiyuan Li, Haisheng Yu, Guangchuan Guo, Nan Zhou, and Jiajun Zhang. Muisqa: Multi-intent retrieval-augmented generation for scientific question answering. *arXiv preprint arXiv:2511.16283*, 2025b.

Yusheng Liao, Chaoyi Wu, Junwei Liu, Shuyang Jiang, Pengcheng Qiu, Haowen Wang, Yun Yue, Shuai Zhen, Jian Wang, Qianrui Fan, et al. Ehr-r1: A reasoning-enhanced foundational language model for electronic health record analysis. *arXiv preprint arXiv:2510.25628*, 2025.

Jiacheng Lin, Zhenbang Wu, and Jimeng Sun. Training llms for ehr-based reasoning tasks via reinforcement learning. *arXiv preprint arXiv:2505.24105*, 2025.

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. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

De Rong Loh, Elliot D Hill, Nan Liu, Geraldine Dawson, and Matthew M Engelhard. Limitations of binary classification for long-horizon diagnosis prediction and advantages of a discrete-time time-to-event approach: Empirical analysis. *JMIR AI*, 4:e62985, 2025.

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G Moreno, and Lynda Tamine. Revisiting the mimic-iv benchmark: Experiments using language models for electronic health records. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, pp. 189–196, 2024.

Nir Menachemi and Taleah H Collum. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pp. 47–55, 2011.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226, 2025.

Shuai Niu, Jing Ma, Liang Bai, Zhihua Wang, Li Guo, and Xian Yang. Ehr-knowgen: Knowledge-enhanced multimodal learning for disease diagnosis generation. *Information Fusion*, 102:102069, 2024.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: towards llms as operating systems. 2023.

Sezen Perçin, Xin Su, Qutub Sha Syed, Phillip Howard, Aleksei Kuvshinov, Leo Schwinn, and Kay-Ulrich Scholl. Investigating the robustness of retrieval-augmented generation at the query level. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pp. 439–457, 2025.

S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186, 2011.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1126–1133, 2019.

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40 (9):btae560, 2024.

Harold C Sox, Michael C Higgins, Douglas K Owens, and Gillian Sanders Schmidler. *Medical decision making*. John Wiley & Sons, 2024.

Ethan Steinberg, Ken Jung, Jason A Fries, Conor K Corbin, Stephen R Pfohl, and Nigam H Shah. Language models are an effective representation learning technique for electronic health record data. *Journal of biomedical informatics*, 113:103637, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Daniel Leon Tramontini, Shrestha Ghosh, and Carsten Eickhoff. Investigating rag-based approaches in clinical trial and patient matching. *Machine Learning for Health 2025*, 2025.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10): 2924–2935, 2024.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. *arXiv preprint arXiv:1909.07940*, 2019.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Cao Xiao, Tengfei Ma, Adji B Dieng, David M Blei, and Fei Wang. Readmission prediction via deep contextual embedding of clinical concepts. *PloS one*, 13(4):e0195024, 2018.

Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. Kerprint: local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5357–5365, 2023.

Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ digital medicine*, 5(1):194, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales? *arXiv preprint arXiv:2010.05345*, 2020.

Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. Emerge: Enhancing multimodal electronic health records predictive modeling with retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3549–3559, 2024a.

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*, 2024b.

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*, 2024c.

Contents of Appendix

A AI Use Statement 16

B Limitations 16

C Impact Statement 16

D Detailed Experimental Setup 16

E Comparison with Traditional ML Baselines 17

F Statistics of EHR History Lengths in EHRSHOT 18

G Evaluation on an Additional Longitudinal EHR Dataset 19

H Efficiency Analysis 19

I Complementarity Analysis of Retrieval Modules 20

J Comparison with Broader RAG and Memory-Centric Methods 20

K Variance Analysis for Stochastic GPT-based Experiments 21

L Case Study 23

M Metadata for Clinical Prediction Tasks 29

N Prompt Design 30

A AI USE STATEMENT

In this work, we used generative AI tools for interpreting experimental results and assisting with the writing of the manuscript. We have not used generative AI tools for generating datasets, conducting experiments without author verification, or making final research decisions, and other required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for brainstorming, literature search, improving readability, polishing text, organizing paper structure, formatting tables, and formatting references. We have reviewed all AI-assisted work. We manually checked technical claims, citations, experimental results, and manuscript content. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

B LIMITATIONS

Residual hallucination and overconfidence risks. Although retrieval improves grounding, *EHR-RAG* still relies on LLMs for evidence interpretation and final prediction. As a result, the system may still exhibit hallucination or overconfidence, especially when the retrieved evidence is insufficient, noisy, or only partially relevant. Further alignment and reliability-oriented training may therefore be needed to improve calibration and reduce unsupported conclusions.

Future research directions. The current framework is training-free, making it easy to apply across different clinical prediction tasks without task-specific model tuning. A promising direction for future work is to incorporate agentic reinforcement learning to further improve both evidence retrieval and clinical reasoning, allowing the system to learn more adaptive retrieval strategies and more reliable decision-making policies from feedback.

C IMPACT STATEMENT

This work aims to advance large language models for reasoning over long-horizon electronic health records through retrieval-augmented generation. While our framework is developed using publicly accessible, de-identified data, we acknowledge that work involving clinical information raises considerations around privacy, fairness, and the appropriate use of AI-assisted decision support. Our method is designed as a research contribution rather than a deployed clinical system, and it does not make autonomous treatment recommendations. Instead, it seeks to improve technical robustness in temporal retrieval and reasoning, which may support future development of safer and more interpretable clinical AI tools.

D DETAILED EXPERIMENTAL SETUP

Dataset Processing. We use the EHRSHOT benchmark (Wornow et al., 2023), which is available in a MEDS-compatible format[1]. In the raw data, each clinical event is represented as a MEDS code. To make events interpretable to LLMs, we map these codes to natural-language descriptions by constructing a lightweight medical ontology from the Athena vocabulary[2] and a CPT subset via UMLS[3]. We evaluate four tasks, *Long Length of Stay*, *30-day Readmission*, *Acute Myocardial Infarction*, and *Anemia*, whose metadata are summarized in Appendix M. For LLM inputs, we serialize each event using a unified template: `[time] type - description (value: value)`; for example, `[2018-02-05 08:56:00] measurement - Blood Pressure (value: 120 mmHg)`.

LLM Generation Settings. We evaluate three LLMs: *GPT-5* (`gpt-5-2025-08-07`), *Claude-Opus-4.5* (`claude-opus-4-5-20251101`), and *LLaMA-3.1-8B* (`Meta-Llama-3.1-8B-Instruct`). All models are deployed on the Azure AI platform, which complies with the PhysioNet Credentialed Data Use Agreement and responsible use

[1] https://github.com/Medical-Event-Data-Standard
[2] https://athena.ohdsi.org/vocabulary/list
[3] https://uts.nlm.nih.gov/uts/

guidelines for clinical data[4]. For all LLM-based methods, we set the generation temperature to 0 to ensure stable and deterministic outputs, except for *GPT-5*, which requires a temperature of 1. To ensure computational efficiency and fair comparison across models, we fix the maximum context length to 128,000 tokens and keep all other generation hyperparameters at their default values. All prompt templates used in our experiments are provided in Appendix N.

ML Baseline Training Settings. The training setups for both *Count-based LR* and *CLMBR-based LR* strictly follow prior work (Wornow et al., 2023), including model architectures and hyperparameter settings. For each task, we train separate models under varying data regimes with 1, 2, 4, 8, 16, 32, 64, 128, 256 shots per class, as well as using the full training set. The pretrained *CLMBR-T-Base* foundation model (Wornow et al., 2023; Steinberg et al., 2021) is used to generate patient representations for the CLMBR-based LR baseline. All machine learning baselines are trained and evaluated on a single NVIDIA A800 GPU with 80GB of memory.

Baseline Retrieval Settings. For *Direct Generation*, we truncate each patient EHR by retaining the most recent 1,000 events. For *Rule-based RAG*, we first rank all events by the frequency of their associated event codes within the patient history, and then select the top 1,000 most frequently occurring events as retrieved evidence. For all other RAG-based baselines, we adopt dense retrieval using `text-embedding-3-small` from Azure OpenAI as a lightweight text embedding model. Each patient EHR is segmented into chunks of 100 event rows with an overlap of 5 rows between adjacent chunks. For vanilla RAG methods, we retrieve the top 10 chunks per query, while keeping the total number of retrieved event rows comparable across methods. *Uniform RAG* randomly samples chunks and concatenates them until the same event budget is reached, serving as a relevance-agnostic control. For *ReAct RAG*, we retrieve the top 5 chunks per query over 3 iterative retrieval steps, accounting for potential duplicate retrievals across iterations. All retrieval pipelines are implemented using LangChain[5], and cosine similarity between query and document embeddings is used as the retrieval scoring function.

EHR-RAG Settings. For our proposed method, *EHR-RAG*, we use the same dense text embedding model as the RAG baselines to ensure a fair comparison. The basic retrieval configuration follows *ReAct RAG*: at each iteration, we retrieve the top-K_{final}=5 textual chunks per query over three iterative retrieval steps. In the U-shaped time-aware retrieval component, we set the semantic–temporal trade-off weight to $\alpha = 0.75$. The recent temporal window is defined as $\tau_{\text{recent}} = 180$ days, while the early-history window spans up to $\tau_{\text{early}} = 3{,}650$ days (approximately 10 years). For textual retrieval, an initial candidate pool of K_{cand}=100 chunks is retrieved before temporal re-ranking, from which the final evidence set $\mathcal{E}_{\text{text}}$ is selected. For numeric event retrieval, we adopt a two-stage indicator selection strategy with a coarse-grained top-N_{coarse}=30 followed by a fine-grained top-N_{fine}=10. For each selected indicator, we retain the N_{recent}=5 most recent measurements to construct the numeric evidence set $\mathcal{E}_{\text{num}}$. We observe that the overall performance of *EHR-RAG* is not highly sensitive to these hyperparameters; they can be tuned intuitively within a reasonable range without significantly affecting results.

E Comparison with Traditional ML Baselines

Figure 2 compares the Macro-F1 performance of *EHR-RAG* with ML baselines across varying levels of labeled data availability, ranging from few-shot to full-data regimes. The x-axis denotes the number of training samples per class. LLM-based methods are evaluated in a zero-shot setting and are therefore shown as horizontal reference lines.

(i) Performance under low-data regimes. Across all tasks, *EHR-RAG* outperforms ML baselines in most low-shot settings. When only a limited number of labeled samples per class are available (e.g., 1 to 16 shots), both count-based and CLMBR-based LR models experience substantial performance degradation. In contrast, *EHR-RAG* maintains relatively stable performance by leveraging pretrained LLM knowledge together with retrieved EHR evidence, highlighting its strong data efficiency in low-resource settings.

(ii) Comparison with representation-based ML models. As expected, CLMBR-based LR generally outperforms count-based features, particularly as the number of training samples increases,

[4] https://physionet.org/news/post/gpt-responsible-use

[5] https://www.langchain.com

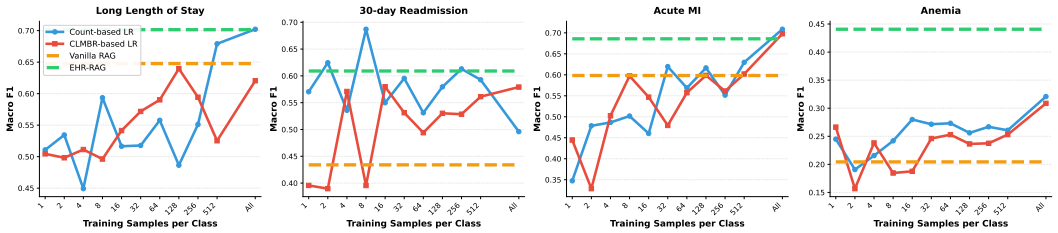

Figure 2: Macro-F1 performance comparison between *EHR-RAG*, vanilla RAG, and traditional machine learning (ML) baselines under varying amounts of labeled training data across four tasks. The x-axis denotes the number of training samples per class, while dashed horizontal lines indicate the zero-shot performance of LLM-based methods. *EHR-RAG* consistently matches or outperforms ML baselines, particularly in low-resource settings.

Table 5: Patient history lengths in EHRSHOT, measured before the prediction time for each clinical event.

Statistic	Number of Events	Serialized Tokens
Mean	6,524.0	180,062
P25	834	23,018
Median	3,082	85,063
P75	7,882	217,543
P95	26,296	725,770
Maximum	78,117	2,156,029

reflecting the advantage of pretrained foundation models in capturing general clinical representations. However, even in medium- and full-data settings, *EHR-RAG* achieves comparable or superior Macro-F1 scores across all tasks. This suggests that retrieval-augmented LLM reasoning can effectively integrate heterogeneous and temporally distributed clinical evidence beyond what is captured by fixed patient representations.

(iii) Robustness on long-horizon and imbalanced tasks. The advantages of *EHR-RAG* are most pronounced on long-horizon and class-imbalanced tasks, such as *Acute MI* and *Anemia*. In these settings, ML baselines struggle to capture sparse yet clinically meaningful signals distributed over extended patient histories. In contrast, *EHR-RAG* leverages strong generalization and zero-shot reasoning capabilities to surface and aggregate relevant evidence, leading to more robust performance.

Overall, these results demonstrate that *EHR-RAG* not only outperforms traditional ML baselines in low-resource settings, but also remains competitive or superior even when large amounts of labeled data are available, underscoring its effectiveness as a general-purpose framework for long-horizon EHR prediction.

F Statistics of EHR History Lengths in EHRSHOT

To characterize the lengths of patient histories in EHRSHOT, we analyzed the records available before the prediction time for each clinical event. Table 5 reports the number of events and the corresponding serialized token counts.

Although some EHR histories fit within modern context windows, a substantial portion remains challenging to process in full: the 75th-percentile history contains approximately 218K serialized tokens, and the 95th percentile reaches approximately 726K tokens. Recent proprietary frontier models, such as GPT and Claude, offer context windows of up to one million tokens and can accommodate many of these histories. However, many locally deployable open-weight models, including variants of Gemma and Qwen, support context windows of approximately 32K–256K tokens. This limitation is particularly relevant in healthcare, where privacy and compliance requirements often favor local deployment.

Table 6: Evaluation on the INSPECT dataset using *GPT-5-nano*. We report Accuracy (%), Macro-F1 (%), and per-class F1 scores (%) on the *Pulmonary Embolism Positive* task using EHR data only.

Method	Accuracy	Macro-F1	F1 (Short)	F1 (Long)
Direct Generation	49.14	49.04	46.85	51.24
Vanilla RAG	33.70	27.45	6.15	48.74
Uniform RAG	40.38	39.83	34.04	45.61
Rule-based RAG	36.00	34.32	23.81	44.83
ReAct RAG	42.86	34.50	11.11	57.89
EHR-RAG (Ours)	**68.97**	**67.58**	**74.29**	**60.87**

G Evaluation on an Additional Longitudinal EHR Dataset

While EHRSHOT is our primary benchmark, its underlying data follows the Medical Event Data Standard (MEDS), which is designed to provide a general and extensible representation of longitudinal EHR data rather than a dataset-specific structure. Therefore, *EHR-RAG* is not tied to a specialized ontology or highly customized format unique to EHRSHOT, and can naturally transfer to other structured EHR datasets.

To further evaluate cross-dataset generalization, we additionally evaluate *EHR-RAG* with *GPT-5-nano* on the INSPECT dataset (Huang et al., 2023), using EHR data only for the *Pulmonary Embolism Positive* prediction task. INSPECT is also a longitudinal clinical dataset, but differs from EHRSHOT in task definition and clinical setting, making it a useful testbed for assessing whether *EHR-RAG* generalizes beyond the original benchmark.

As shown in Table 6, *EHR-RAG* substantially outperforms all LLM-based baselines on INSPECT, achieving 68.97% Accuracy and 67.58% Macro-F1. In contrast, vanilla RAG performs poorly, suggesting that one-shot semantic retrieval is insufficient for this longitudinal clinical prediction task. These results further suggest that *EHR-RAG* does not simply overfit to EHRSHOT's task structure, but can generalize to another longitudinal EHR dataset with a different clinical prediction objective.

H Efficiency Analysis

We analyze the computational efficiency of *EHR-RAG*, as long-horizon EHR reasoning requires processing large numbers of temporally dispersed clinical events. Table 7 reports the average number of input tokens, output tokens, and end-to-end latency per prediction on the *Long Length of Stay* task using *GPT-5-nano*. Compared with single-pass baselines, *EHR-RAG* incurs higher inference cost because it performs multi-stage evidence retrieval and reasoning. Specifically, *EHR-RAG* uses 96.4K input tokens and 585.9 output tokens on average. With parallel execution when applicable, *EHR-RAG* achieves an average latency of 8.23 seconds.

Importantly, this additional cost should be interpreted in the context of the target problem. Long-horizon structured EHR prediction requires recovering clinically relevant signals from thousands of heterogeneous events, where important evidence may be sparse, temporally dispersed, or missed by one-shot retrieval. In this setting, Vanilla RAG is computationally cheaper but often retrieves insufficient evidence, while Direct Generation must process the entire serialized patient history. Compared with Direct Generation, *EHR-RAG* increases total input tokens moderately, from 70.9K to 96.4K, while substantially improving predictive performance. Therefore, although *EHR-RAG* is not designed as the lowest-cost solution, its latency remains practical for asynchronous or less latency-sensitive clinical decision-support scenarios, such as chronic disease risk prediction and longitudinal patient review.

Given the substantial performance gains across tasks, we believe this additional computation represents a worthwhile accuracy–cost trade-off rather than redundant overhead. In medical prediction tasks, even modest improvements can translate into meaningful downstream value; therefore, the gains achieved by *EHR-RAG* are substantial rather than negligible. Overall, the additional overhead is justified by improved evidence coverage, higher evidence quality, and more reliable long-horizon clinical prediction.

Table 7: Efficiency analysis on the *Long Length of Stay* task using *GPT-5-nano*. We report the average number of input tokens, output tokens, and end-to-end latency per prediction. For *EHR-RAG*, parallel execution is used when applicable.

Method	Avg Input Tokens	Avg Output Tokens	Avg Latency
Direct Generation	70,959.6	115.0	2.91s
Vanilla RAG	15,450.5	108.4	2.82s
ReAct RAG	10,431.9	112.8	4.59s
EHR-RAG (Ours, parallel)	**96,426.6**	**585.9**	**8.23s**

Table 8: Event-level pairwise Jaccard IoU between evidence sets retrieved by different modules.

Comparison	Avg. Jaccard	Std.
ETHER $\leftrightarrow$ AIR	0.169	0.058
ETHER $\leftrightarrow$ DER	0.153	0.056
AIR $\leftrightarrow$ DER	0.735	0.149

I Complementarity Analysis of Retrieval Modules

To further examine whether the proposed retrieval modules provide redundant or complementary evidence, we conducted an additional event-level overlap analysis. Specifically, we ran the three modules, ETHER, AIR, and DER, independently, and computed the pairwise Jaccard intersection-over-union (IoU) between their retrieved evidence sets.

As shown in Table 8, ETHER exhibits low overlap with both AIR and DER, with average Jaccard scores below 0.17. This result is expected, as ETHER emphasizes numeric indicators and time-aware initial retrieval, whereas AIR and DER operate over mixed-type evidence chunks. The limited overlap suggests that ETHER retrieves evidence that is largely complementary to the evidence retrieved by the other two modules, rather than redundant.

AIR and DER show a higher degree of overlap, with an average Jaccard score of 0.735. This similarity is also expected because both modules operate over the refined retrieval space. Nevertheless, the two modules are not identical. The remaining differences arise from the additional information introduced by DER's dual-path design, which retrieves and reasons over both factual and counterfactual evidence. In a subset of cases, the Jaccard scores between AIR and DER fall in the range of 0.44–0.54, indicating that DER contributes nontrivial additional evidence beyond AIR alone.

Overall, this analysis provides additional empirical evidence that ETHER, AIR, and DER are not simply redundant stacked modules. Instead, they play complementary roles in evidence retrieval for long-horizon structured EHR prediction, with ETHER contributing distinct time-aware numeric evidence and DER further enriching the refined retrieval space through dual-path evidence reasoning.

J Comparison with Broader RAG and Memory-Centric Methods

We additionally compare *EHR-RAG* with broader RAG and memory-centric methods to examine whether general-purpose long-context memory mechanisms are sufficient for long-horizon structured EHR prediction. In addition to Direct Generation, Vanilla RAG, and ReAct RAG, we evaluate MemGPT (Packer et al., 2023), HippoRAG (Gutiérrez et al., 2024), HippoRAG-2 (Gutiérrez et al., 2025), and LightMem (Fang et al., 2025) on the *Long Length of Stay* task using *GPT-5-nano*. These methods are representative of memory-augmented or graph-inspired retrieval systems designed to maintain and retrieve information over long contexts.

As shown in Table 9, *EHR-RAG* achieves the strongest overall performance, reaching 57.08% Accuracy and 55.30% Macro-F1. MemGPT improves over Vanilla RAG, suggesting that explicit memory management can help in long-horizon settings. However, it still remains clearly below *EHR-RAG*, indicating that general-purpose memory mechanisms alone are insufficient for structured longitudinal EHR prediction. HippoRAG and HippoRAG-2 consume a large number of tokens, especially output tokens, but still underperform *EHR-RAG*. This suggests that higher token usage by itself does

Table 9: Comparison with broader RAG and memory-centric methods on the *Long Length of Stay* task using *GPT-5-nano*. We report Accuracy (%), Macro-F1 (%), and average token usage per prediction.

Method	Accuracy	Macro-F1	Avg Input Tokens	Avg Output Tokens
Direct Generation	43.36	43.34	70,959.6	115.0
Vanilla RAG	42.04	42.03	15,450.5	108.4
ReAct RAG	33.63	33.05	10,431.9	112.8
MemGPT	48.67	46.84	58,817.2	471.7
HippoRAG	37.50	36.51	81,117.1	9,304.3
HippoRAG-2	42.86	41.67	84,503.8	10,226.6
LightMem	28.57	27.08	27,090.7	1,432.0
EHR-RAG (Ours)	**57.08**	**55.30**	**96,426.6**	**585.9**

Table 10: Comparison with self-consistency variants of memory-centric baselines on the *Long Length of Stay* task. Inference time denotes the relative number of repeated inference runs used for majority voting.

Method	Accuracy	Macro-F1	Inference Time
MemGPT	48.67	46.84	1
MemGPT	51.33	50.96	3
LightMem	28.57	27.08	1
LightMem	35.29	34.39	4
EHR-RAG (Ours)	**57.08**	**55.30**	**1**

not guarantee stronger clinical prediction performance. A likely reason is that graph-based associative memory methods are primarily designed for general textual knowledge and multi-hop passage retrieval, whereas structured EHR data consists of heterogeneous clinical event types, sparse temporal signals, and irregular longitudinal trajectories. Such data provides fewer explicit relational structures for constructing reliable general-purpose knowledge graphs, making these methods less well matched to the EHR setting.

LightMem is designed to improve the efficiency of memory-augmented generation through lightweight memory compression and consolidation. While this design reduces input token usage, it also leads to a clear performance drop on long-horizon EHR prediction. This suggests that aggressive memory compression may discard fine-grained clinical signals that are essential for accurate prediction, especially when relevant evidence is temporally dispersed across a long patient history. In contrast, *EHR-RAG* explicitly models the structure of longitudinal EHR through event- and time-aware retrieval, expands incomplete evidence through adaptive iterative retrieval, and improves robustness through factual and counterfactual evidence reasoning.

To further account for inference-time budget differences, we also evaluate MemGPT and LightMem with self-consistency, increasing their inference cost through repeated runs and majority voting. As shown in Table 10, even under this stronger inference setting, MemGPT improves from 46.84 to 50.96 Macro-F1 and LightMem improves from 27.08 to 34.39 Macro-F1, but both remain below *EHR-RAG*'s 55.30 Macro-F1. These results indicate that the gains of *EHR-RAG* are not simply due to using more inference-time computation. Instead, the advantage comes from retrieval mechanisms specifically tailored to long-horizon structured EHR data. Overall, these comparisons show that *EHR-RAG* is better aligned with the temporal, heterogeneous, and evidence-sparse nature of clinical records than general-purpose RAG or memory-centric baselines.

K Variance Analysis for Stochastic GPT-based Experiments

Since GPT-based APIs may produce stochastic outputs, we further evaluate the robustness of *EHR-RAG* under repeated runs. Specifically, we independently run *EHR-RAG* and all LLM-based baselines with *GPT-5-nano* five times on the *Long Length of Stay* task, and report the mean and standard deviation in Table 11.

As shown in Table 11, *EHR-RAG* consistently outperforms all baselines across repeated runs. The performance margin between *EHR-RAG* and the strongest baseline is substantially larger than the ob-

Table 11: Variance analysis on the *Long Length of Stay* task using *GPT-5-nano*. We report mean ± standard deviation over five independent runs.

Method	Accuracy	Macro-F1	F1 (Short)	F1 (Long)
Direct Generation	43.36 ± 1.77	43.34 ± 1.79	42.81 ± 2.81	43.87 ± 0.77
Vanilla RAG	42.04 ± 0.45	42.03 ± 0.43	41.26 ± 0.19	42.80 ± 0.69
Uniform RAG	38.50 ± 1.33	38.40 ± 1.31	40.85 ± 1.53	35.95 ± 1.09
Rule-based RAG	39.82 ± 3.54	39.77 ± 3.49	41.30 ± 4.46	38.23 ± 2.52
ReAct RAG	33.63 ± 0.88	33.05 ± 1.21	27.10 ± 3.10	39.01 ± 0.67
EHR-RAG (Ours)	$\mathbf{57.08 \pm 1.33}$	$\mathbf{55.30 \pm 1.14}$	$\mathbf{64.19 \pm 1.50}$	$\mathbf{46.42 \pm 0.77}$

served run-to-run standard deviation. For example, *EHR-RAG* improves Macro-F1 over the strongest baseline by more than 11 points, while its standard deviation is only 1.14. This suggests that the observed improvement is robust and is unlikely to be caused by stochastic variation in GPT-based generation.

L CASE STUDY

We present a detailed representative case study to illustrate the effectiveness of *EHR-RAG* in clinical prediction, with a particular focus on its dual-path reasoning capability for balancing contradictory clinical evidence. This case demonstrates how *EHR-RAG* avoids false alarms by systematically integrating both risk factors and stabilizing clinical signals, leading to more calibrated and reliable predictions. These results are generated using *GPT-5* on the *Long Length of Stay* task.

Ethical and privacy note. This case study is based on de-identified EHRSHOT data and does not include any personally identifiable information. It is intended solely to qualitatively illustrate retrieval and reasoning behavior, rather than to provide clinical guidance or diagnostic recommendations.

L.1 CLINICAL SCENARIO

Patient: Female, 56 years old; BMI 43.0
Admission Date: 2014-12-05 20:00 (postoperative admission after intracranial aneurysm clipping)
Prediction Time: 2014-12-05 23:59
Ground Truth Outcome: LOS = 1 day (discharged 2014-12-06)
Task: Predict whether long LOS ($\geq$ 7 days)

Clinical Context. Initially admitted on 2014-11-27 for subarachnoid hemorrhage (SAH) due to ruptured intracranial aneurysm; discharged 2014-11-28 after conservative management. Readmitted on 2014-12-04 for elective intracranial aneurysm clipping. The microsurgical procedure was completed successfully, followed by routine postoperative monitoring in the neurological intensive care unit.

Baseline Comorbidities. Type 2 diabetes mellitus / Hypertension / Hyperlipidemia / Class III obesity (BMI 43.0) / Tobacco use

Model Predictions and Outcomes.

- ***EHR-RAG*:** Prediction = 0 (**Correct**). *Rationale:* Balances surgical risk factors with stabilizing postoperative signals, including rapid neurologic recovery and early clinical stability.
- **Direct Generation:** Prediction = 1 (**Incorrect**). *Rationale:* Overemphasizes surgical complexity and baseline comorbidities, overestimating prolonged hospitalization risk.
- **Vanilla RAG:** Prediction = 1 (**Incorrect**). *Rationale:* Retrieved evidence is dominated by high-risk features without incorporating countervailing stabilization indicators, yielding a biased risk assessment.

L.2 EVENT- AND TIME-AWARE HYBRID EHR RETRIEVAL IN *EHR-RAG*

The *Event- and Time-aware Hybrid EHR Retrieval* (ETHER) module first performs knowledge-guided coarse retrieval to collect candidate clinical indicators, then applies LLM-based reranking to select task-relevant indicators for downstream reasoning.

L.2.1 COARSE RETRIEVAL

Using medical priors for postoperative neurologic and respiratory monitoring, ETHER retrieves candidates spanning: neurologic status (e.g., GCS), respiratory parameters (e.g., PEEP, FiO_2, airway pressure), arterial blood gases (pH, $PaCO_2$, PaO_2), hemodynamics (blood pressure), and coagulation markers (INR, aPTT), as well as additional candidates including vasopressor use, ventilation settings, pain, and sedation metrics.

L.2.2 LLM-BASED INDICATOR RERANKING

The LLM reranks these candidates and selects the top-10 indicators:

1. Glasgow Coma Scale (GCS)
2. Arterial blood gas measures (pH, $PaCO_2$, PaO_2)

3. Fraction of inspired oxygen (FiO_2)
4. Mean airway pressure and PEEP
5. Central venous pressure (CVP)
6. Diastolic blood pressure (DBP)
7. International Normalized Ratio (INR)

L.2.3 Retrieved Clinical Evidence

Hemodynamics and Vital Signs. Stable SBP 98–130 mmHg, DBP 57–64 mmHg, and HR 81–89 bpm throughout the postoperative period; no vasopressor support required. Mean arterial pressure is consistently maintained within a normal range.

Neurologic Status. Rapid neurologic recovery: transient post-anesthetic decline (GCS 14) improves to full responsiveness (GCS 15). Pain decreases from 6/10 to 0/10 within two hours postoperatively; spontaneous respiration is preserved.

Respiratory and Blood Gas Trends. Expected postoperative disturbances with rapid normalization: mild acidosis (pH 7.24) and hypercapnia ($PaCO_2$ 52 mmHg) improve within two hours (pH 7.34, $PaCO_2$ 46.5 mmHg). Oxygenation remains adequate; FiO_2 is weaned from 60% to 50%, indicating improving respiratory status with minimal ventilatory support.

Coagulation and Metabolic Status. Normal coagulation (INR 1.1; aPTT 21.9 s); preserved renal function (creatinine 0.9 mg/dL); electrolytes stable. Transient hyperglycemia (consistent with postoperative stress) is observed.

Monitoring and Support. Standard postoperative monitoring is in place, including arterial and central venous lines and temporary airway protection, without evidence of escalating support requirements.

Summary. Overall, the retrieved evidence indicates rapid postoperative stabilization across neurologic, respiratory, and hemodynamic domains, providing strong signals against prolonged hospitalization despite high baseline risk factors.

L.3 Adaptive Information Retrieval (AIR) in *EHR-RAG*

AIR refines retrieval by adaptively generating positive and negative queries to gather evidence supporting both hypotheses.

L.3.1 Positive Reasoning Path (Evidence for Long Stay)

Iteration 0: Query = "Cerebral vasospasm following SAH"

- Retrieved: Vasospasm monitoring protocols, CBF measurements, transcranial doppler assessments (5 documents).
- Rationale: Vasospasm is a major SAH complication requiring prolonged ICU stay; early monitoring is critical.

Iteration 1: Query = "Hydrocephalus following intracranial hemorrhage"

- Retrieved: Ventriculomegaly assessments, intracranial pressure trends, EVD indications (6 documents).
- Rationale: SAH frequently causes hydrocephalus; EVD insertion commonly prolongs ICU stay.

Iteration 2: Query = "External ventricular drain management post-aneurysm"

- Retrieved: EVD care protocols, CSF characteristics, EVD removal criteria, duration of drainage (3 documents).
- Rationale: EVD presence strongly predicts extended ICU stay (typically 5–7+ days).

L.3.2 NEGATIVE REASONING PATH (EVIDENCE FOR SHORT STAY)

Iteration 0: Query = "Discharge to home post-neurosurgery without complications"

- Retrieved: Home discharge orders, absence of postoperative complications, successful emergence from anesthesia (4 documents).
- Rationale: Home discharge after neurosurgery indicates rapid, uncomplicated recovery.

Iteration 1: Query = "Social work assessment and discharge planning after ICU"

- Retrieved: Discharge readiness criteria, caregiver assessments, skilled nursing vs. home discharge (5 documents).
- Rationale: Early discharge planning and lack of complications facilitate short stay.

Iteration 2: Query = "Physical therapy: independent ambulation post-operative"

- Retrieved: PT clearance for discharge, functional independence measures, ICU mobility protocols (3 documents).
- Rationale: Independent ambulation signals discharge readiness.

L.4 DUAL EVIDENCE RETRIEVAL AND REASONING (DER) IN *EHR-RAG*

DER evaluates evidence for long- vs. short-stay hypotheses independently and then fuses contradictory evidence.

L.4.1 POSITIVE REASONING (SUPPORTING LONG STAY $\geq$ 7 DAYS)

Severe Diagnosis and Procedure Complexity:

- Intracranial aneurysm clipping is a high-risk, complex microsurgical procedure.
- Requires precise dissection of carotid circulation and careful handling of fragile vessel.
- Typical post-op course involves 5–7 day minimum ICU stay for neuro monitoring.
- SAH increases risk of post-operative complications.

High-Intensity ICU Care Requirements:

- Mechanical ventilation for airway protection post-op.
- Continuous neuro monitoring (GCS, pupil reactivity, motor/sensory).
- Arterial line for beat-to-beat BP monitoring (neurosurgery standard).
- Central venous line for volume assessment and medication administration.
- Thermoregulation management (hypothermia induction protocol common in SAH).
- Vasoactive medications (nitroprusside documented) for hemodynamic optimization.

Acute Physiologic Derangements:

- Early metabolic acidosis (pH 7.24) indicating post-operative stress.
- Hypercapnia ($PaCO_2$ 52 mmHg) suggesting ventilation/perfusion mismatch.
- Stress hyperglycemia (glucose 235 mg/dL) indicating systemic metabolic stress.
- These derangements typically persist for 3–5 days, prolonging ICU stay.

High-Risk Comorbidity Profile:

- Type 2 diabetes increases post-operative infection risk and impairs healing.
- Hypertension complicates post-op BP management in neurosurgery.
- Obesity increases anesthetic risk and mobility limitations.
- Tobacco use associated with respiratory complications.
- Combination of comorbidities predicts longer recovery trajectory.

Rehabilitation and Disposition Planning:

- Documented rehabilitation facility admission suggests anticipated prolonged stay.
- Comorbidities and mobility limitations often necessitate post-acute care facility.
- Home discharge typically requires minimal comorbidities and full functional independence.

L.4.2 NEGATIVE REASONING (SUPPORTING SHORT STAY $<$ 7 DAYS)

Rapid Anesthetic Emergence and Neurologic Recovery:

- By 20:42 (42 minutes post-op): MAC 0–0.1 indicates patient nearly fully awake.
- By 21:15 (75 minutes post-op): GCS 15 with full consciousness and appropriate responses.
- No requirement for prolonged sedation or neuromuscular blockade.
- Rapid emergence suggests excellent anesthetic recovery and minimal post-op CNS depression.

Stable Hemodynamics Without Vasopressor Dependence:

- SBP 98–130 mmHg without vasopressor infusions; mean BP 70–90s sufficient for cerebral perfusion.
- No hypotensive episodes requiring intervention.
- No documented inotropic support (dopamine, epinephrine).
- Hemodynamic stability by 1–2 hours post-op is a strong negative predictor.

Preserved Neurologic Integrity:

- GCS 15 at 21:15 demonstrates full consciousness.
- Temporary GCS 14 at 22:00 is mild and transient (likely opioid-related).
- No focal neurologic deficits; no stroke or intraoperative ischemic complications.
- Neurologic preservation suggests successful repair without complications.

Excellent Pain Control and Comfort:

- Pain reduced from 6/10 to 0/10 within 45 minutes.
- Adequate analgesia supports early mobilization and reduced ICU stay.

Laboratory Normalization and No Coagulopathy:

- Creatinine 0.9 mg/dL (normal renal function); electrolytes within normal limits.
- INR 1.1 and aPTT 21.9 indicate normal coagulation; no transfusions documented.
- Lack of metabolic/coagulation complications reduces ICU stay duration.

Physiologic Improvement and ABG Trend:

- pH improves from 7.24 to 7.34 within 2 hours.
- $PaCO_2$ decreases from 52 to 46.5 mmHg; base deficit improves from 5 to 1 mmol/L.
- Rapid recovery suggests minimal complications and improving discharge readiness.

Absence of Documented Post-operative Complications:

- No rebleeding, hematoma expansion, new stroke/ischemia, ventilator failure, or respiratory distress.
- No infection/fever/sepsis; no escalation to vasopressors/ECMO.

Prior Rapid Recovery Pattern:

- Prior SAH admission (11/27) discharged within 1 day (11/28), indicating rapid recovery.
- Readmission on 12/04 for planned elective repair (not emergency).
- Planned procedures typically have better outcomes than emergency cases.
- Patient's prior trajectory supports expectation of rapid current recovery.

L.4.3 EVIDENCE FUSION

The DER fusion module synthesizes both positive and negative evidence:

> *"While the procedure itself is complex and carries inherent risks typical of intracranial aneurysm repair, this particular patient demonstrated exceptionally rapid post-operative recovery. Within 2 hours of ICU admission, she was fully conscious (GCS 15), hemodynamically stable without vasopressor support, and showing rapid physiologic improvement (ABG normalization, pain control achieved).*
> *The presence of comorbidities (diabetes, obesity, HTN) and historical SAH are concerning for prolonged stay. However, the complete absence of any documented post-operative complication, combined with the patient's prior history of rapid discharge after SAH and the elective (not emergency) nature of this procedure suggest excellent surgical repair with minimal morbidity.*
> *By prediction time (23:59, 4 hours post-op), all indicators point to an uncomplicated trajectory: no vasospasm signs, no hydrocephalus indicators, no coagulopathy, normal renal function, and preserved neurologic status. This contrasts with the typical 5–7 day post-op course.*
> ***Conclusion:** The balance of evidence strongly favors shorter length of stay (<7 days). The rapid recovery trajectory and absence of complications outweigh theoretical risks from procedure complexity and comorbidities."*

Final Prediction: 0 (Length of Stay < 7 days) **Confidence:** High
Ground Truth: Discharged 2014-12-06 (1 day post-op) **Outcome:** ✓ **Correct**

L.5 FAILURE ANALYSIS: WHY DIRECT GENERATION AND VANILLA RAG FAILED

L.5.1 DIRECT GENERATION FAILURE

Reasoning Pattern:

> *"Complex intracranial aneurysm clipping surgery in a patient with multiple comorbidities (diabetes, obesity, hypertension) typically requires a lengthy ICU stay for monitoring of post-operative complications such as vasospasm and hydrocephalus. The patient's initial presentation with SAH suggests high-risk recovery. Expected LOS $\geq$ 7 days."*

Failure Mode – Confirmation Bias:

- **Received Data:** Accessed identical clinical data (GCS, ABGs, vital signs, etc.).
- **Reasoning Error:** Anchored on surgical complexity and comorbidities, underweighting recovery trajectory.
- **Bias:** Overweighted worst-case complications (vasospasm, hydrocephalus) without evidence.
- **Missed Signal:** Failed to treat rapid neurologic recovery (GCS 15 within hours) as countervailing evidence.
- **Static Thinking:** Evaluated a single timepoint rather than dynamic improvement.

L.5.2 VANILLA RAG FAILURE

Reasoning Pattern:

> *"Retrieval augmented generation returned articles on post-operative complications of intracranial aneurysm surgery, typical ICU stay durations (5–7 days), and management of vasospasm and hydrocephalus. Based on retrieved context, predicted LOS $\geq$ 7 days."*

Failure Mode – Lack of Systematic Evaluation:

- **Retrieval Bias:** Query biased toward complications and prolonged monitoring.
- **Missing Dual Perspective:** Retrieved risk-confirming documents without systematic search for rapid recovery evidence.
- **No Negative Evidence:** No structured mechanism to identify/weight evidence against high-risk conclusion.

- **Static Aggregation:** Treated typical 5–7 day LOS as a rule, ignoring patient-specific trajectory.
- **Confirmation Loop:** Retrieval reinforced initial impression without critical reappraisal.

L.5.3 EHR-RAG SUCCESS

Key Advantages:

- **Systematic Dual Evaluation:** Explicitly considers both positive and negative evidence.
- **Balanced Retrieval:** AIR issues positive queries (vasospasm, hydrocephalus) and negative queries (recovery, discharge planning).
- **Trajectory Analysis:** DER emphasizes dynamic trends (e.g., ABG improvement) rather than static risk factors.
- **Evidence Fusion:** Appropriately weights rapid improvement against theoretical risks.
- **False-alarm Reduction:** Prevents overconfident high-risk classification via structured dual reasoning.

L.6 CLINICAL SIGNIFICANCE AND LEARNING POINTS

Key Insight. Complex surgery does not automatically imply prolonged stay:

1. **Procedure complexity $\neq$ complication inevitability:** Skilled repair can avoid major post-op complications.
2. **Rapid recovery in high-risk patients:** Despite obesity/diabetes/HTN, full neurologic recovery occurs within hours.
3. **Trajectory over static risk:** Improving vitals/labs/neuro status within 2 hours is a strong predictor of short stay.
4. **No early complications:** Absence of complications by 4 hours post-op increases short-stay likelihood.

Clinical Impact of Accurate Prediction:

- **ICU resource allocation:** Enables earlier step-down and frees beds for critical patients.
- **Discharge planning:** Supports planning for rapid discharge vs. prolonged facility stay.
- **Patient/family expectations:** Reduces unnecessary anxiety about prolonged recovery.
- **Cost reduction:** Avoids unnecessary ICU days (at ~$5000/day).
- **Mobility/infection prevention:** Earlier step-down supports mobilization and reduces ICU-related risks.

L.7 SUMMARY

The structured dual reasoning of EHR-RAG mitigates false-alarm bias by preventing worst-case anchoring without proportional evidence. By enforcing balanced evaluation of contradictory evidence, EHR-RAG maintains sensitivity to high-risk cases while reducing excessive false positives.

M METADATA FOR CLINICAL PREDICTION TASKS

Table 12 presents the metadata of the EHRSHOT clinical prediction tasks used in our experiments, including core task definitions, basic task queries, and task-specific instructions. This metadata is provided to the LLMs under a unified prompt framework to support clinical prediction across different tasks.

Table 12: Metadata of EHRSHOT clinical prediction tasks used in our experiments. Each column corresponds to a task, and each row reports a specific metadata field.

Field	**guo_los**	**guo_readmission**	**new_acutemi**	**lab_anemia**
Task name	Long Length of Stay	30-day Readmission	Acute Myocardial Infarction	Anemia
Category	Operational Outcomes	Operational Outcomes	Assignment of New Diagnoses	Anticipating Lab Test Results
Description	Predict whether a patient will have a long hospital stay ($\geq$7 days) based on their EHR data.	Predict whether a patient will be readmitted within 30 days after hospital discharge based on their EHR data.	Predict whether a patient will receive a new acute myocardial infarction diagnosis within 1 year after discharge based on their EHR data.	Predict the severity category of the next anemia-related laboratory result based on the patient's prior EHR data.
Factual query	clinical factors and events associated with prolonged hospital stay	clinical factors and events associated with 30-day hospital readmission	clinical risk factors and events associated with acute myocardial infarction	clinical factors and events relevant to predicting anemia severity
Counterfactual query	clinical factors and events associated with short hospital stay	clinical factors and events associated with no readmission	clinical risk factors and events indicating absence of acute myocardial infarction	clinical factors and events indicating no anemia
Label type	binary	binary	binary	multiclass_4
Label values	$\{0, 1\}$	$\{0, 1\}$	$\{0, 1\}$	$\{0, 1, 2, 3\}$
Label description	$\{0, 1\}$ aka {<7 days, $\geq$7 days}	$\{0, 1\}$ aka {no readmission, readmission}	$\{0, 1\}$ aka {no diagnosis, diagnosis}	$\{0, 1, 2, 3\}$ aka {low, medium, high, abnormal}
Task-specific instructions	No instructions provided.	30-day readmission is uncommon. Default to 0 unless there is clear, strong, patient-specific evidence; do not predict 1 from vague risk factors.	Be sensitive to positives: if there is any reasonable, patient-specific evidence suggesting acute MI, lean toward 1; if uncertain, prefer 1.	Choose among {0,1,2,3} with calibrated preference for mild-to-moderate (1 or 2) when uncertain; use 0 only with strong evidence of no anemia; use 3 only with clear severe/abnormal anemia evidence.

N PROMPT DESIGN

Figures 3 present the prompt template used by the LLM-based clinical prediction baselines. Figures 4 and 5 illustrate the prompts employed in the *Event- and Time-Aware Hybrid EHR Retrieval (ETHER)* and *Adaptive Iterative Retrieval (AIR)* components of *EHR-RAG*, respectively. Figures 6, 7, and 8 show the prompt templates used for factual reasoning, counterfactual reasoning, and evidence fusion in the *Dual-Path Evidence Retrieval and Reasoning (DER)* component.

Prompt of Clinical Prediction Baselines

```
# Task
You are a reliable clinical expert.
Based on the information below and the patient medical history,
answer the clinical question for the task of "{task_name}", which belongs to the "{task_category}".

# Information
Clinical Question: {description}
Prediction time: {prediction_time}
Label Definition: {label_description}

Patient History:
{patient_history}

# Instructions
1. You must provide a clinical prediction based on the provided patient history at the end of your response.
2. If the evidence is incomplete or uncertain, make the best-supported prediction based on the available
information.
3. This is a {label_type} classification task, so your final answer must be an integer chosen from
{label_values}.
{task_specific_instructions}

# Response Format
Please think step-by-step, showing your reasoning, and then provide the final answer strictly in this format
(each on its own line):

Reasoning: <your reasoning here>
Answer: <your final answer of classification result>

Now, give your response:
```

Figure 3: Prompt template used for LLM-based clinical prediction baselines. Blue text indicates input variables.

Prompt of Clinical Indicator Selection (ETHER)

```
# Task
You are a clinical expert selecting the most informative clinical indicators to support a downstream prediction
task.
The clinical question corresponds to the task "{task_name}", which belongs to the category "{task_category}".

# Information
Clinical Question: {description}
Candidate Clinical Indicators: {indicator_list}

# Instructions
1. Select the top {topk} indicators that are most relevant and informative for answering the clinical question.
2. Focus on indicators that are clinically meaningful, discriminative, and directly related to the task.
3. Do NOT explain your choice.

# Response Format
Return ONLY a JSON list (no extra text).
Each element must be the indicator name EXACTLY as it appears above.

Example:
["indicator A", "indicator B", "indicator C"]
```

Figure 4: Prompt template used for clinical indicator selection in the *Event- and Time-Aware Hybrid EHR Retrieval* component of *EHR-RAG*. Blue text denotes input variables.

Prompt of Query Refinement (AIR)

```
# Task
You are refining a retrieval query to find missing EHR evidence that supports the objective.

# Information
Task: {task_name}
Clinical question: {description}
Objective: {objective}
Previously used queries: {other_queries}

Evidence already retrieved:
{evidence_snippets}

# Instructions
1. Only output ENOUGH if the retrieved evidence already covers all major clinical aspects that are typically
required to answer the clinical question.
2. Otherwise, write ONE refined query to retrieve an important missing piece of clinical evidence.
3. Assume the retrieval source does NOT contain direct answers to the clinical question.
4. The refined query must:
  - Be concise (<= 15 words)
  - Focus on a single clinical aspect (no lists, no multiple features)
  - Target evidence NOT already present above
  - Avoid dates, timestamps, and NOT mention prediction targets
  - Be different from all previous queries

# Response Format
Return ONLY one of the following:
- ENOUGH
- The refined query text
```

Figure 5: Prompt template used for query refinement in the *Adaptive Iterative Retrieval* component of *EHR-RAG*. Blue text denotes input variables.

Prompt of Factual Reasoning (DER)

```
# Task
You are a reliable clinical expert.
Based on the information below and the patient medical history,
answer the clinical question for the task of "{task_name}", which belongs to the "{task_category}".

# Information
Clinical Question: {description}
Prediction time: {prediction_time}
Label Definition: {label_description}

Patient History:
{patient_history}

indicator_information:
{indicator_information}

# Instructions
1. Assume the outcome is positive.
2. Your role is to identify and reason over evidence that SUPPORTS the positive outcome (label =
{positive_label}).
3. Focus ONLY on strong, direct clinical evidence that would justify predicting label = {positive_label}.
4. Do NOT consider evidence that argues against a positive outcome.
5. If no strong supporting evidence exists, explicitly state that the evidence for label = {positive_label} is weak.

Now, think step-by-step, showing your reasoning:
```

Figure 6: Prompt template used for factual reasoning in the *Dual-Path Evidence Retrieval and Reasoning* component of *EHR-RAG*. Blue text denotes input variables.

Prompt of Counterfactual Reasoning (DER)

 # Task
You are a reliable clinical expert.
Based on the information below and the patient medical history,
answer the clinical question for the task of "{task_name}", which belongs to the "{task_category}".

 # Information
Clinical Question: {description}
Prediction time: {prediction_time}
Label Definition: {label_description}

Patient History:
{patient_history}

indicator_information:
{indicator_information}

 # Instructions
1. Assume the outcome is negative.
2. Your role is to identify and reason over evidence that SUPPORTS the negative outcome (label = {negative_label}).
3. Focus ONLY on strong, direct clinical evidence that would justify predicting label = {negative_label}.
4. Do NOT consider evidence that argues against a negative outcome.
5. If no strong supporting evidence exists, explicitly state that the evidence for label = {negative_label} is weak.

Now, think step-by-step, showing your reasoning:

Figure 7: Prompt template used for counterfactual reasoning in the *Dual-Path Evidence Retrieval and Reasoning* component of *EHR-RAG*. Blue text denotes input variables.

Prompt of Evidence Fusion Reasoning (DER)

 # Task
You are a reliable clinical expert.
Using the patient history and two complementary reasoning summaries
(one arguing for the positive outcome and one for the negative outcome),
answer the clinical question for the task "{task_name}" in the category "{task_category}".

 # Information
Clinical Question: {description}
Prediction time: {prediction_time}
Label Definition: {label_description}

Patient History:
{patient_history}

indicator_information:
{indicator_information}

Reasoning in support of the positive outcome:
{factual_reasoning}

Reasoning in support of the negative outcome:
{counterfactual_reasoning}

 # Instructions
1. Compare the strength, directness, and relevance of evidence on both sides. Prefer strong, direct evidence over weak or speculative evidence.
2. Choose the label that is most consistent with the available evidence, and state why.
3. If the evidence on both sides is roughly balanced in strength, or if both sides are weak, DEFAULT to the NEGATIVE outcome (0),
since the default assumption is that the patient has no condition unless there is strong evidence to the contrary.
4. This is a {label_type} classification task. Your final answer must be an integer chosen from {label_values}.
{task_specific_instructions}

 # Response Format
Please think step-by-step, showing your reasoning, and then provide the final answer strictly in this format (each on its own line):

Reasoning: <your reasoning here>
Answer: <your final answer of classification result>

Now, give your response:

Figure 8: Prompt template used for evidence fusion reasoning in the *Dual-Path Evidence Retrieval and Reasoning* component of *EHR-RAG*. Blue text denotes input variables.