EHR-RAGp: PROTOTYPE-GUIDED RETRIEVAL OF LONGITUDINAL ELECTRONIC HEALTH RECORDS FOR CLINICAL PREDICTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic Health Records (EHR) contain rich longitudinal patient information and are widely used in predictive modeling applications. However, effectively leveraging historical data remains challenging due to long trajectories, heterogeneous events, temporal irregularity, and the varying relevance of past clinical context. Existing approaches often rely on fixed windows or uniform aggregation, which can obscure clinically important signals. In this work, we introduce EHR-RAGp, a retrieval-based framework that dynamically integrates the most relevant patient history consisting of diverse clinical event types. We propose a prototype-guided retrieval module that acts as an alignment mechanism and estimates the relevance of retrieved historical chunks with respect to a given prediction task, guiding the model towards the most informative context. Across multiple clinical prediction tasks and two benchmark datasets, EHR-RAGp consistently outperforms state-of-the-art EHR-based and transformer-based baselines. Furthermore, EHR-RAGp is model-agnostic, as integrating it with various backbone models yields substantial performance gains. Overall, EHR-RAGp establishes a novel direction for modeling long-range clinical context to improve downstream performance via retrieval.

1 INTRODUCTION

Healthcare systems rely extensively on Electronic Health Record (EHR) databases to keep records of diverse clinical events that occur during patient encounters, such as demographics, diagnoses, vital signs, lab results, and medications (Nasarudin et al., 2024). Recent advances in machine learning have enabled the development of various clinical prediction models for EHR data, where patient trajectories are represented as chronologically ordered sequences of events, in a manner that mimics the structures of sentences in natural language (Li et al., 2024; Ren et al., 2025; Wornow et al., 2024). However, unlike natural language, clinical events are inherently irregular, and span multiple episodes of care that accumulate over extended periods (Zhong et al., 2025).

The current state of EHR modeling is shaped by specific design assumptions, including: (i) information present only within a predefined time window for a specific patient visit is relevant to predict the outcome of interest (Wornow et al., 2023; Elsharief et al., 2025); and (ii) a preselected subset of events or features, such as labs, vitals, diagnoses, or combinations of them, is sufficient to predict patient outcomes (Odgaard et al., 2024; Hur et al., 2023; Yang et al., 2023). Although these assumptions provide practical abstractions, they may restrict the model's ability to capture relevant information beyond predefined temporal windows or feature sets. Beyond these design assumptions, computational constraints and the difficulty of representing long-range, irregular patient trajectories present additional barriers to comprehensive EHR modeling (Wornow et al., 2024), which may also lead to the omission of clinically important historical events. From a clinical perspective, the patient's state is almost always shaped by their accumulated longitudinal history (Braitman & Davidoff, 1996; Duan et al., 2019). This includes slow-moving chronic patterns and cross-encounter interactions that clinicians routinely examine when making decisions. Such a central gap highlights a fundamental limitation in current EHR modeling paradigms, namely, the inability to integrate and reason over a patient's extended history within a unified representation learning framework.

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.

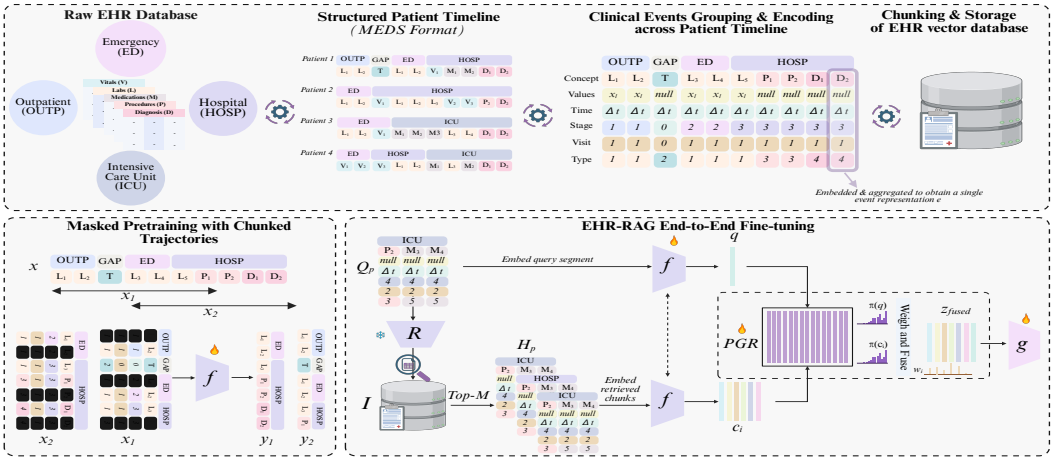


Figure 1: Overview of the EHR−RAGp, including chunking and indexing of patient trajectories (*top*), masked pre-training of the backbone encoder (*bottom left*), and end-to-end fine-tuning of the final architecture (*bottom right*).

Filling this gap requires a new outlook on EHR representation learning. Rather than limiting the patient's predictive context to a predefined window, a model must be equipped to access, retrieve, and integrate information present in the patient's complete trajectory. The central assumption motivating this work is that retrieval provides a scalable mechanism for incorporating long-range historical context without forcing the model to encode the entire trajectory within its fixed internal state. Given the remarkable success of Retrieval-Augmented Generation (Lewis et al., 2020) (RAG) techniques in Natural Language Processing (NLP) and their ability to leverage external knowledge beyond the training data of the base Large Language Model (LLM) (Yu et al., 2024; Zhao et al., 2024; Zheng et al., 2025), the same concept can be applied for retrieving and integrating historical EHR data. In other words, by considering the patient's historical events as retrievable units of information, an EHR model can dynamically condition on what matters the most for the patient's current state. This perspective shifts longitudinal modeling from a purely sequential problem into a retrieval-augmented one. Hence, in this work, we propose EHR−RAGp, a retrieval-based EHR modeling framework designed to reason over extended context of a patient's longitudinal EHR history. The main contributions of the work are summarized as follows:

1. We establish modeling of long EHR records as a retrieval problem, by constructing *a vector database of longitudinal EHR patient trajectories* using multi-granular chunking strategies, including event-based, time-based, visit-level, and care-stage, enabling flexible representation of heterogeneous events, temporal structure, and care progression.
2. We present EHR−RAGp, a *prototype-guided retrieval mechanism* that acts as an alignment operator, estimating the relevance of retrieved patient chunks with respect to the downstream task, guiding the model to integrate the most informative retrieved clinical context.
3. Extensive evaluation across *multiple clinical prediction tasks and two datasets* demonstrates that EHR−RAGp consistently improves performance over multiple clinical and general baselines.
4. We demonstrate the effectiveness of EHR−RAGp as a *model-agnostic framework*, by experimenting with multiple EHR backbone models, improving their performance beyond fixed context.
5. Existing EHR models vary in their selection and preprocessing of input features. By training all baselines using a *unified feature set and training pipeline*, we establish new benchmark results on four clinical prediction tasks and make our code publicly available at: link.

2 Related Work

EHR Modeling. Traditional EHR modeling methods process patient timelines as sequences of discrete events collected over time. Some models primarily consider the occurrence of certain events, while others process the occurrence of the clinical event and its measured value. Early techniques

utilized recurrent models such as Long-Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) or Gated Recurrent Unit (GRU) networks (Cho et al., 2014) to capture long-term dependencies (Che et al., 2018; Harutyunyan et al., 2019; Duan et al., 2019; Hayat et al., 2022).

The emergence of transformer-based LLMs (Devlin et al., 2019; Radford et al., 2019) paved the way for advancing EHR modeling. For example, encoder-based models such as BEHRT (Li et al., 2020), Hi-BEHRT (Li et al., 2022), CEHR-BERT (Pang et al., 2021), and Med-BERT (Rasmy et al., 2021) introduced large-scale pretraining over longitudinal sequences of medical events to enhance the contextualization of patient trajectories across visits. Other decoder-based transformer models such as ETHOS (Renc et al., 2025), CEHR-GPT (Pang et al., 2024), Curiosity (Waxler et al., 2025), and EHR2Path (Pellegrini et al., 2025) focused on the development of generative EHR foundation models for clinical prediction tasks. The common link between all these models is that they mimic standard language models by learning contextualized medical code vocabularies as tokens.

Another category of EHR foundation models shifts away from code-based modeling towards text-based modeling by treating each medical code as a textual description. Hence, the patient timeline constitutes a set of textual descriptions that can be processed directly by the LLM (Hur et al., 2022b; 2023; Lee et al., 2025). Despite the remarkable progress, these models continue to inherit core limitations from their language-model roots. They operate on truncated encounter windows, compress long and irregular patient histories into fixed-length sequences, and rely on restricted types of clinical events. While some methods have explored the use of longer context windows to accommodate extended trajectories (Waxler et al., 2025; Hur et al., 2022b; 2023), their vocabularies are heavily tokenized for medical codes and numerical values, or textualized and dependent on the base model's reasoning capabilities. As a result, substantial portions of a patient's longitudinal trajectory remain inaccessible in downstream prediction tasks. This emphasizes the need for scalable frameworks that are capable of integrating and reasoning over complete EHR trajectories.

Retrieval-Augmented Modeling. RAG has emerged as a powerful technique in NLP for extending large language models capabilities by allowing them to condition on external, non-parametric knowledge sources. The original RAG framework (Lewis et al., 2020) demonstrated how combining dense retrieval with generative transformers improves factual correctness and reduces hallucinations. The concept was later extended by models such as RETRO (Borgeaud et al., 2022), FLARE (Jiang et al., 2023b), and REPLUG (Shi et al., 2024). These architectures share a common principle: retrieval allows the model to incorporate long-range or domain-specific information without encoding the entire knowledge base within its internal parameters. The success of the retrieval paradigm has also influenced healthcare, particularly in clinical question answering (Saba et al., 2024; Sohn et al., 2025), evidence grounding (Lewis et al., 2025; Jia et al., 2025), biomedical knowledge integration (Feng et al., 2025b;a), phenotyping and cohort identification (Ziletti & D'Ambrosi, 2025), and many other applications related to medical textual data (Abo El-Enen et al., 2025; Amugongo et al., 2025).

In the context of EHR modeling, retrieval has been explored in work such as EMERGE (Zhu et al., 2024a), REALM (Zhu et al., 2024b), and RAM-EHR (Xu et al., 2024). The goal of these studies is to incorporate knowledge from external related sources into their modeling pipelines to enhance predictive accuracy. Hence, the retrieval objective of these studies is not aligned with our scope, *which specifically pertains to retrieval over the patient's history.* Furthermore, these models operate on multimodal data beyond EHR events only, like notes. To the best of our knowledge, REMED (Kim et al., 2023) is most closely related to our work. Its core idea is to retrieve the most relevant subset of medical codes for a given outcome prediction task, to overcome manual selection of clinical events or features. Overall, recent advancements highlight the potential of retrieval-augmented modeling for EHR, but they also reveal a clear unmet need for mechanisms that can dynamically access and integrate a patient's entire structured history.

3 METHODOLOGY

3.1 OVERVIEW

`EHR-RAGp` is a retrieval-augmented model for reasoning over longitudinal EHR trajectories (Figure 1). The framework encodes structured clinical events, organizes them into retrievable units, retrieves relevant historical segments through prototype-guided alignment, and integrates them with a task-specific query for downstream task. Here, we describe the core architectural components.

3.2 PRELIMINARIES

Let $\mathcal{D} = \{\mathcal{H}_p\}_{p=1}^{\mathcal{P}}$ represent a large-scale EHR dataset consisting of patient-level longitudinal records for a cohort of $\mathcal{P}$ patients. For a given patient p, the longitudinal medical history is represented as:

$$\mathcal{H}_p = \{\mathcal{V}_p^{(1)}, \mathcal{V}_p^{(2)}, \ldots, \mathcal{V}_p^{(N_p)}\},$$

where N_p denotes the total number of visits for the respective patient (see Figure S1). The n-th visit $\mathcal{V}_p^{(n)}$ consists of a set of clinical events:

$$\mathcal{V}_p^{(n)} = \{(c_{nk}, v_{nk}, t_{nk})\}_{k=1}^{K_n},$$

where c_{nk} denotes a clinical event such as a diagnosis, procedure, laboratory test, vital sign, or medication administration, v_{nk} is an optional associated value such as a lab result, medication dosage, or vital sign measurement, and t_{nk} is the exact timestamp at which the event occurred. Events within a visit may vary in type, granularity, and frequency, reflecting the inherently irregular and heterogeneous nature of structured EHR data.

For a given index visit $\mathcal{V}_p^{(\tau)}$, the goal is to predict a target patient outcome with true label $y_p^{(\tau)}$ by retrieving the most relevant historical information. We define a **query segment** $\mathcal{Q}_p^{(\tau)} \subseteq \mathcal{V}_p^{(\tau)}$ as the portion of the current visit used as the model's immediate input. Since the full visit may exceed the model's maximum input length, the query is restricted to a fixed-size segment within its context window. All preceding events, whether from the same visit or earlier visits, form the **historical context**:

$$\mathcal{H}_p^{(<\tau)} = \{\mathcal{V}_p^{(1)}, \ldots, \mathcal{V}_p^{(\tau-1)}, \mathcal{V}_p^{(\tau<q)}\},$$

where $\mathcal{V}_p^{(\tau<q)}$ denotes the prefix of the last visit preceding the query (see Figure S2). Hence, the model aims to learn a mapping function:

$$\gamma : \left(\mathcal{Q}_p^{(\tau)}, \mathcal{H}_p^{(<\tau)}\right) \rightarrow \hat{y}_p^{(\tau)}.$$

A naive formulation of this problem would require the model to encode the full patient history in its fixed internal state, which is a significant challenge due to computational constraints. Hence, the goal of our proposed framework is to enable the model to retrieve and integrate the most relevant information from the patient's history $\mathcal{H}_p^{(<\tau)}$ based on the query segment $\mathcal{Q}_p^{(\tau)}$.

3.3 CHUNKING

After representing individual events, `EHR-RAGp` partitions the patient timeline into coherent chunks that serve as retrievable units. We propose four chunking methods, namely, *event-based*, *time-based*, *visit-level*, and *care-stage* chunking (Figure S3). Each chunking method captures a different aspect of longitudinal patient structure, enabling retrieval over heterogeneous temporal granularities and clinical contexts. Event-based chunking groups a fixed number of consecutive events, time-based chunking partitions the timeline into fixed temporal windows, visit-level chunking organizes events according to hospital admissions, and care-stage chunking separates events based on transitions between clinical care settings, such as outpatient, emergency department, hospital ward, and ICU.

3.4 PROTOTYPE-GUIDED RETRIEVAL

Given a query segment $\mathcal{Q}_p^{(\tau)}$, `EHR-RAGp` retrieves relevant historical segments from the patient's longitudinal history $\mathcal{H}_p^{(<\tau)}$ using a two-stage procedure: (1) coarse-grained semantic retrieval to identify candidate chunks, followed by (2) prototype-guided alignment that estimates the relevance of each retrieved chunk via agreement in the learned prototype assignment space, to enable weighted integration of context for the downstream task.

Candidate Retrieval. The query segment $\mathcal{Q}_p^{(\tau)}$ is first encoded using a frozen pretrained retriever R, producing a fixed-dimensional query representation:

$$q' = R(\mathcal{Q}_p^{(\tau)}) \in \mathbb{R}^d.$$

The retriever, R, is pretrained using masked language modeling. Further details are provided in Appendix C. Using the same retriever R, each historical chunk $h_i \in \mathcal{H}_p^{(<\tau)}$ is embedded as

$$c_i' = R(h_i) \in \mathbb{R}^d.$$

Candidate retrieval is then performed by computing **cosine** similarity between the query embedding q' and historical chunk embeddings $\{c_i'\}$ stored in precomputed similarity index $\mathcal{I}$. We then retrieve the top-M most similar chunks:

$$\{h_1, h_2, \ldots, h_M\} \subset \mathcal{H}_p^{(<\tau)}.$$

Query and Candidate Encoding. Upon retrieval, the query segment $\mathcal{Q}_p^{(\tau)}$ and the top-M chunks $\{h_1, h_2, \ldots, h_M\}$ are passed to a shared backbone encoder f to obtain q and $\{c_i\}$, respectively.

$$q = f(\mathcal{Q}_p^{(\tau)}) \in \mathbb{R}^d, \qquad c_i = f(h_i) \in \mathbb{R}^d.$$

Using the same encoder f for queries and history ensures that both are represented in a common space, enabling consistent comparison and alignment between the query and retrieved context.

Prototype-Guided Alignment. `EHR-RAGp` maintains a learnable set of L prototypes, which serve as latent clinical centroids that organize a patient's longitudinal history into distinct semantic modes:

$$P = \{p_1, p_2, \ldots, p_L\}, \qquad p_l \in \mathbb{R}^d,$$

These prototypes serve as latent anchors that organize the representation space into distinct modes, enabling structured comparison between query and historical chunks. Rather than being fixed, they are learned end-to-end, allowing the model to capture recurring patterns in clinical trajectories and facilitate alignment between the query and retrieved context with respect to the downstream task.

For a chunk embedding $x \in \mathbb{R}^d$, the model computes a probability distribution over the L prototypes:

$$\pi(x) = \mathrm{softmax}\left(\frac{xP^\top}{T_\pi}\right) \in \mathbb{R}^L,$$

where x corresponds to either to the **query** q or a **candidate chunk** c_i, and T_π is a temperature parameter that controls the sharpness of the distribution (Assran et al., 2022). In practice, we use different temperatures for the query and history branches, denoted by T_q and T_h, respectively, with $T_q < T_h$. This asymmetric design reflects the roles of the two branches: the query must identify a focused subset of relevant patterns for the prediction task, while historical chunks represent diverse clinical contexts that may only partially match the query. Accordingly, we obtain:

$$\pi_q = \pi(q), \qquad \pi_i = \pi(c_i),\ i = 1, \ldots, M.$$

π_q represents the prototype assignment distribution of the query, while each π_i represents the assignment of candidate chunk c_i to the set of learned prototypes L. To measure how well a candidate history chunk aligns with the query in the prototype space, `EHR-RAGp` measures the agreement between their prototype distributions using a cross-entropy-based alignment score:

$$\alpha_i = -\sum_{l=1}^{L} \pi_q^{(l)} \log \pi_i^{(l)},$$

Where $\alpha \in [0, \infty)$. This formulation measures how well the candidate distribution π_i explains the query distribution π_q, treating the query as a reference. Lower negative cross-entropy values indicate stronger alignment, meaning that the candidate assigns high probability mass to the prototypes emphasized by the query, whereas higher values indicate weaker alignment. In addition, this distributional alignment enables smooth and differentiable matching in the prototype space without requiring hard assignment to a single prototype.

Prototype-Guided Weighting. Given the prototype alignment scores $\{\alpha_i\}_{i=1}^M$, `EHR-RAGp` converts them into normalized importance weights using a temperature-controlled softmax:

$$w_i = \mathrm{softmax}\left(\frac{-\alpha_i}{T_s}\right),$$

where T_s is a temperature parameter that controls the smoothness of the weighting distribution.

`EHR-RAGp` adopts a fully differentiable soft weighting mechanism, where all retrieved chunks contribute to the final prediction with varying degrees of importance. This design allows the model

to preserve distributed clinical evidence across multiple chunks, which is particularly important in EHR data where relevant signals are often weak and spread over time. The resulting weights $\{w_i\}$ are used to modulate the contribution of each retrieved chunk during the subsequent fusion stage.

Retrieval-Augmented Fusion. The query representation q and the retrieved historical chunks $\mathcal{R}_p^{(<\tau)} = \{c_1, \ldots, c_n\}$ are combined through a prototype-guided weighting mechanism. Each retrieved chunk is modulated by its corresponding importance weight w_i:

$$z_{\text{fused}} = \big[q,\ w_1 c_1,\ w_2 c_2,\ \ldots,\ w_n c_n\big].$$

This sequence is then processed by a transformer encoder $g(\cdot)$ followed by a a linear classifier $h(\cdot)$ to yield the final prediction:

$$\hat{y}_p^{(\tau)} = h\left(g(z_{\text{fused}})\right).$$

Training Objective. EHR-RAGp is trained end-to-end using a supervised binary cross-entropy loss associated with the target prediction task, while keeping the retrieval encoder, R, frozen and updating the backbone encoder, f, the prototype parameters, and the fusion module. The overall objective combines the task loss with a prototype usage regularization term that encourages balanced utilization of the prototype space across both query and history representations:

$$\mathcal{L} = l\big(y_p^{(\tau)}, \hat{y}_p^{(\tau)}\big) - \lambda_{\text{u}} \left(\mathcal{H}(\bar{\pi}_q) + \mathcal{H}(\bar{\pi}_h)\right),$$

where $\bar{\pi}_q = \frac{1}{B}\sum_{b=1}^{B} \pi_q^{(b)}$ and $\bar{\pi}_h = \frac{1}{\sum_b K_b}\sum_{b,k} \pi_{h_k}^{(b)}$ denote the average prototype assignment distributions over a batch for the query and historical chunks, respectively, $\mathcal{H}(\cdot)$ is the entropy function, and $\lambda_{\text{u}} > 0$ determines the strength of the regularization term. This regularization encourages the model to avoid degenerate solutions in which only a small subset of prototypes are used.

4 Experiments

Datasets & Tasks. We evaluate our proposed framework on multiple tasks using the large-scale real-world MIMIC-IV dataset (V3.1) (Johnson et al., 2020), and EHRSHOT (Wornow et al., 2023) dataset. MIMIC-IV consists of heterogeneous EHR data gathered from patients admitted to the Beth Israel Deaconess Medical Center between 2008-2022. It includes clinical data collected from 364K patients associated with 546K hospital admissions and 94K ICU stays (Johnson et al., 2023). EHRSHOT is a longitudinal dataset for performance benchmarking of foundation models for clinical prediction tasks. EHRSHOT covers data from 6,739 patients, 921K visits, and 15 tasks.

For the MIMIC-IV dataset, we assess EHR-RAGp on four risk prediction tasks, including: in-hospital mortality, ICU-readmission within 30 days, long length of stay (7 days or more), and 1-year mortality post-discharge. The selected tasks span both short-term and long-term clinical outcomes and reflect common benchmarks in EHR-based modeling. For EHRSHOT dataset, we consider the assignment of new diagnosis task family, including: Acute MI (MI), Lupus (LP), Hyperlipidemia (HL), Hypertension (HT), Celiac (CEL), and Pancreatic Cancer (PC). We also consider operational outcome task family including: ICU admission (ICU), long length of stay $>$7 days (LLOS), and 30-day Readmission (RA). Together, these tasks assess the model's ability to capture acute risk, longitudinal disease progression, and healthcare utilization by leveraging structured patient histories. Detailed information pertaining to the tasks and data preprocessing are in Appendix C.2.

Implementation Details. We use RoBERTa *base* (Liu et al., 2019) as the backbone encoder for EHR-RAGp, for the MIMIC-IV experiments. In the first stage, we pretrain the encoder using an MLM objective. For fairness of comparison, all code-based baselines in our evaluation are pretrained using the same objective and data splits. The pretrained EHR-RAGp encoder is subsequently reused as the embedding model for constructing and querying the similarity index. In the second stage, we fine-tune the full EHR-RAGp model end-to-end for the downstream prediction tasks in a supervised setting. For the EHRSHOT dataset, we integrate CLMBR (Wornow et al., 2023), a pretrained model, published with the dataset, in the EHR-RAGp pipeline as a backbone. We conduct extensive hyperparameter tuning using Bayesian optimization for all baselines and EHR-RAGp models. Detailed optimization settings and training schedules are provided in Appendix C. We evaluate MIMIC-IV experiments using the Area Under the Operator Receiver Curve (AUROC) and the Area Under the Precession-Recall (AUPRC). We also report statistical significance based on paired

Table 1: Performance of EHR-RAGp and baselines on the MIMIC-IV dataset. Best results are marked in **bold**. Second-best results are underlined (ties allowed). †denotes p$<$ 0.05 compared to second-best.

	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
Model	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
DescEmb I	0.661	0.078	0.904	0.618	0.822	0.405	0.754	0.376
DescEmb II	0.606	0.071	0.920	0.648	0.838	0.441	0.805	0.471
GenHPF	0.697	0.128	0.915	0.641	0.843	0.456	0.781	0.417
Med-BERT	0.664	0.081	0.893	0.579	0.813	0.412	0.729	0.338
REMed	0.561	0.060	0.865	0.501	0.786	0.348	0.680	0.280
CEHR-BERT	0.604	0.061	0.929	0.676	0.840	0.435	0.806	0.473
BEHRT	0.668	0.077	0.902	0.609	0.824	0.444	0.767	0.382
Hi-BEHRT	0.620	0.069	0.886	0.542	0.798	0.392	0.703	0.298
EHRMamba	0.665	0.082	0.929	0.650	0.829	0.424	0.796	0.429
ModernBERT	0.604	0.068	0.920	0.640	0.827	0.440	0.765	0.381
LongFormer	0.645	0.095	0.910	0.577	0.821	0.423	0.760	0.377
BigBird	0.719	0.118	0.913	0.643	0.830	0.444	0.787	0.424
RoFormer	0.699	0.114	0.927	0.672	0.846	0.461	0.806	0.460
RoBERTa	0.721	0.125	0.925	0.671	0.841	0.465	0.805	0.466
EHR-RAGp (Ours)	**0.750**†	**0.141**	**0.950**†	**0.736**†	**0.867**†	**0.509**†	**0.853**†	**0.564**†

Table 2: Performance of EHR-RAGp and baselines on the EHRSHOT dataset in terms of AUROC. Best results are marked in **bold**. Second-best results are underlined (ties allowed). Δ represents performance difference compared to the base backbone CLMBR.

	Operational tasks			Diagnostic tasks					
Model	**ICU**	**LLOS**	**RA**	**MI**	**LP**	**HL**	**HT**	**CEL**	**PC**
LLaMA	0.845	0.810	**0.823**	0.744	0.724	0.691	0.722	0.783	0.805
GPT-3	0.859	0.813	0.814	0.735	0.739	0.687	0.722	0.706	0.840
Hyena	0.738	0.766	0.762	0.629	0.602	0.659	0.647	0.733	0.740
Mamba	0.855	0.827	0.816	0.745	0.784	0.698	0.721	0.751	0.866
CLMBR	0.848	0.814	0.810	0.729	0.747	0.675	0.718	0.557	0.813
EHR-RAGp	**0.881**	**0.845**	0.816	**0.761**	**0.863**	**0.721**	**0.749**	**0.785**	**0.873**
Δ	+3.89%	+3.80%	+0.74%	+4.39%	+15.53%	+6.81%	+4.31%	+41.0%	+7.38%

bootstrap testing. We use AUROC only to evaluate EHRSHOT experiments, based on existing results in previous work (Wornow et al., 2024).

Baselines. Our main goal is to compare fixed context length EHR models to our proposed retrieval approach. Hence, we consider a specific set of baselines spanning clinical EHR models, and general transformer models. We also examine how LLMs perform in a zero-shot fashion using fixed context. Clinical baselines include existing EHR-specific models such as DescEmb (Hur et al., 2022a), GenHPF (Hur et al., 2023), Med-BERT (Rasmy et al., 2021), CEHR-BERT (Pang et al., 2021), BEHRT (Li et al., 2020), Hi-BEHRT (Li et al., 2022), EHRMamba (Fallahpour et al., 2024), and REMed (Kim et al., 2023). We denote DescEmb I as the CLS-FT variant, and DescEmb II as the BERT-FT variant. Since these models were originally trained on different subsets of clinical features, we adapt all baselines using a shared vocabulary and EHR embedding layer to ensure unified input representations and fair comparison. general transformer models include RoBERTa (Liu et al., 2019), Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), RoFormer (Su et al., 2024), and ModernBERT (Warner et al., 2025). Finally, we evaluate both general-purpose (3 models) and medically adapted LLMs (3 models) For EHRSHOT experiments. We also compare EHRSHOT results with long context baselines including LLaMA (Grattafiori et al., 2024), GPT-3 (Brown et al., 2020), Hyena (Poli et al., 2023), Mamba (Gu & Dao, 2024) with contexts length of $4K$, $4K$, $16K$, $16K$, respectively. Additional details are provided in Appendix D.

Table 3: Performance gains of existing EHR baselines with EHR−RAGp. Best results are shown in **bold**. †denotes p< 0.05 compared to the base model.

Model	EHR−RAGp	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
		AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Med-BERT	×	0.664	0.081	0.893	0.579	0.813	0.412	0.729	0.338
	√	**0.715**†	**0.117**†	**0.938**†	**0.702**†	**0.843**†	**0.455**†	**0.823**†	**0.476**†
CEHR-BERT	×	0.604	0.061	0.929	0.676	0.840	0.435	0.806	0.473
	√	**0.698**†	**0.102**†	**0.941**†	**0.715**†	**0.848**†	**0.452**	**0.821**†	**0.495**†
BEHRT	×	0.668	0.077	0.902	0.609	0.824	0.444	0.767	0.382
	√	**0.723**†	**0.126**†	**0.941**†	**0.711**†	**0.849**†	**0.465**†	**0.826**†	**0.490**†
Hi-BEHRT	×	0.620	0.069	0.886	0.542	0.798	0.392	0.703	0.298
	√	**0.684**†	**0.100**†	**0.934**†	**0.691**†	**0.835**†	**0.443**†	**0.786**†	**0.401**†

5 Main Results

MIMIC-IV Results. Table 1 reports the performance of EHR−RAGp and the evaluated baselines on the four MIMIC-IV dataset tasks. EHR−RAGp achieves the highest AUROC and AUPRC across all tasks. Relative to the strongest baseline for each metric, it improves AUROC by 0.029 for ICU readmission, 0.021 for in-hospital mortality, 0.021 for long LOS, and 0.047 for 1-year mortality. The corresponding AUPRC improvements are 0.013, 0.060, 0.044, and 0.091, respectively. Paired-bootstrap tests indicate that the differences from the strongest baseline are statistically significant for all metrics except ICU-readmission AUPRC, for which the improvement over GenHPF is not significant. Beyond the strongest baselines, EHR−RAGp consistently achieves higher performance gain across both medical and transformer based baselines. Table S5 shows that LLM baselines perform substantially worse across all tasks. Overall, these results suggest that retrieving relevant information from the longitudinal patient history can complement fixed-context representations across various clinical outcomes.

EHRSHOT Results. Table 2 compares EHR−RAGp with CLMBR and the long-context models on the EHRSHOT tasks. Relative to its CLMBR backbone, EHR−RAGp yields positive AUROC differences across all nine tasks. Among the reported results, EHR−RAGp obtains the highest point estimate on eight of all tasks. Overall, EHR−RAGp provides substantially larger gains on the new-diagnosis assignment tasks than on the operational outcomes. Moreover, EHR−RAGp matches or exceeds the reported long-context models on eight of the nine EHRSHOT tasks and remains competitive on the remaining task. These findings suggest that retrieval can preserve, and frequently improve upon, the performance of long-context modeling without relying on an increasingly large context window.

EHR−RAGp is Model-Agnostic. Table 3 demonstrates the benefit of integrating EHR−RAGp with other model backbones. Across the four models, tasks, and evaluation metrics, incorporating EHR−RAGp improves performance, yielding positive differences in all 32 model/task/metric comparisons. The AUROC gains range from 0.008 to 0.094, while the AUPRC gains range from 0.017 to 0.149. Particularly large improvements are observed for Med-BERT on 1-year mortality, where AUROC increases from 0.729 to 0.823 and AUPRC from 0.338 to 0.476, and for Hi-BEHRT on in-hospital mortality, where AUPRC increases from 0.542 to 0.691. Paired-bootstrap testing indicates that 31 of the 32 improvements are statistically significant; the only exception is the AUPRC improvement of CEHR-BERT on long LOS. These results, along with CLMBR results on EHRSHOT, demonstrate that the gains provided by EHR−RAGp are not restricted to a single backbone and that retrieving relevant longitudinal history consistently strengthens a range of existing EHR models.

6 Ablations

Effect of Prototypes. Table 4 separates the contributions of retrieving longitudinal history and prototype-guided alignment. Relative to the RoBERTa model without retrieval, incorporating semantically retrieved history without prototypes improves AUROC and AUPRC by 0.018 and 0.026, respectively, for long-LOS prediction, and by 0.034 and 0.050 for 1-year

Table 4: Ablation of retrieval mechanisms. †denotes p< 0.05 compared to EHR−RAGp.

Config.	Long LOS 7d		1YR Mortality	
	AUROC	AUPRC	AUROC	AUPRC
w/o Retrieval	0.841†	0.465†	0.805†	0.466†
w/o Prototypes	0.859†	0.491†	0.839†	0.516†
EHR−RAG$_p$	**0.867**	**0.509**	**0.853**	**0.564**

mortality. Introducing prototype-guided alignment provides further improvements of 0.008 AUROC and 0.018 AUPRC for long LOS, and 0.014 AUROC and 0.048 AUPRC for 1-year mortality. Differences between the full model and the ablated variants are statistically significant under paired-bootstrap testing. These results suggest that access to longitudinal history provides a consistent benefit, while prototypes alignment contributes additional improvements, particularly in AUPRC.

Effect of Regularization. Table 5 evaluates the impact of prototype regularization (λ_u) on both representation behavior and downstream performance. Without regularization, prototype assignments are highly peaked for queries (e.g., $P_{\max} = 0.887$ for Long LOS), indicating collapse to a few dominant prototypes, while history assignments remain sparse and less informative. Enabling regularization significantly increases entropy and reduces $P_{\max}$ for both query and history, leading to more distributed and balanced prototype usage. This results in consistent performance gains across tasks, with Long LOS AUPRC improving from 0.472 to 0.509, and ICU readmission AUPRC from 0.118 to 0.141. Overall, the results show that regularization prevents prototype collapse, promotes better utilization of the prototype space, and improves performance. See Appendix E.4 for further analysis on the usage regularization λ_u.

Table 5: Ablation of the usage of regularization.

λ_u	Query		History		Performance	
	p Max	Entropy	p Max	Entropy	AUROC	AUPRC
Long LOS 7d						
×	0.887	0.663	0.094	4.048	0.843	0.472
✓	0.084	4.096	0.027	4.156	**0.867**	**0.509**
ICU Readmission 30d						
×	0.716	1.578	0.030	6.099	0.715	0.118
✓	0.014	6.198	0.004	6.232	**0.750**	**0.141**

Effect of Chunking. Table 6 compares the four chunking strategies on ICU readmission and in-hospital mortality. Performance varies only slightly across the strategies. Event-based chunking achieves the highest ICU-readmission AUROC and AUPRC (0.750 and 0.141) and the highest in-hospital mortality AUROC (0.950), while care-stage chunking obtains the highest in-hospital mortality AUPRC (0.739). Time- and visit-based chunking remain competitive across both tasks. Overall, these results indicate limited sensitivity to the choice of chunking strategy, with modest task-dependent differences.

Table 6: Impact of the chunking strategy.

Model	ICU Readmission 30d		In-Hospital Mortality	
	AUROC	AUPRC	AUROC	AUPRC
Event-Based	**0.750**	**0.141**	**0.950**	0.736
Time-Based	0.741	0.129	0.946	0.738
Visit-Level	0.748	0.140	0.948	0.738
Care-Stage	0.747	0.138	0.947	**0.739**

Additional results and ablations are provided in Appendix E.

7 Discussion & Conclusion

In this work, we introduced `EHR-RAGp`, a retrieval-augmented framework that incorporates extended longitudinal patient history through semantic retrieval and prototype-guided alignment. Across two benchmarks and thirteen prediction tasks, `EHR-RAGp` achieves the strongest results across all MIMIC-IV metrics, improves upon its CLMBR backbone on all EHRSHOT tasks, and matches or exceeds the reported long-context models on eight of nine tasks. Integrating the framework with four additional EHR encoders also improves all 32 model/task/metric combinations, with 31 reaching statistical significance. Controlled ablations indicate that both retrieved history and prototype-guided alignment contribute to performance, while prototype regularization promotes more balanced prototype usage. Performance remains relatively consistent across the four chunking strategies, suggesting limited sensitivity to how patient histories are segmented. Overall, the results support retrieval as a useful complement to fixed-context EHR representations, particularly for outcomes that depend on information distributed across the longitudinal record.

Limitations. First, although we evaluate on two EHR benchmarks, additional institutions, patient populations, and coding systems are needed to assess broader generalizability. Second, `EHR-RAGp` focuses on structured EHR events, while patient records include other modalities such as clinical notes, images, and physiological waveforms. Future work could extend retrieval across these data sources. Finally, candidate retrieval relies on a fixed pretrained encoder and precomputed similarity index. Jointly learned or clinically informed retrieval represents a promising direction for improving task-specific adaptation.

ETHICS STATEMENT

This study uses previously collected, de-identified EHR data from MIMIC-IV and EHRSHOT and involves no direct patient interaction or clinical intervention. No additional IRB approval was required for this secondary analysis. Dataset access was obtained in accordance with the applicable data-use agreements and required CITI training, and no patient-level data will be redistributed. Because these datasets may reflect biases in their source institutions and patient populations, the reported results should not be assumed to generalize to other healthcare settings without external validation. `EHR-RAGp` is intended for research purposes and should not be used for clinical decision-making without prospective evaluation and appropriate assessment of safety, fairness, and calibration.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the `EHR-RAGp` architecture, retrieval and prototype-alignment mechanisms, and training objectives in Section 3. The appendix documents dataset preprocessing, task definitions, cohort construction, data splits, baseline adaptations, optimization settings, and hyperparameter configurations. We also report confidence intervals and paired-bootstrap testing procedures. MIMIC-IV and EHRSHOT are available through their respective controlled-access procedures and data-use agreements. An anonymized implementation containing the model code, preprocessing pipeline, experiment configurations, and evaluation scripts is provided in the contribution part of the introduction.

AI USE STATEMENT

Generative AI tools were used to support language editing and proofreading, assist with code for scientific visualizations, clarify literature reviewed by the authors, and brainstorm experimental analyses and manuscript presentation. AI-generated text was not used as a source of references or citations. All AI-assisted text was reviewed and revised by the authors, and AI-assisted code and visualizations were inspected and tested for correctness. The authors made all final research decisions and take responsibility for the complete content of this work.

REFERENCES

Mohamed Abo El-Enen, Sally Saad, and Taymoor Nazmy. A survey on retrieval-augmentation generation (rag) models for healthcare applications. *Neural Computing and Applications*, 37(33): 28191–28267, 2025.

Khalid Alghatani, Nariman Ammar, Abdelmounaam Rezgui, and Arash Shaban-Nejad. Predicting intensive care unit length of stay and mortality using patient vital signs: machine learning model development and validation. *JMIR medical informatics*, 9(5):e21347, 2021.

Lameck Mbangula Amugongo, Pietro Mascheroni, Steven Brooks, Stefan Doering, and Jan Seidel. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6):e0000877, 2025.

Malaikannan Sankarasubbu Ankit Pal. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`, 2024.

Bert Arnrich, Edward Choi, Jason Alan Fries, Matthew BA McDermott, Jungwoo Oh, Tom Pollard, Nigam Shah, Ethan Steinberg, Michael Wornow, and Robin van de Water. Medical event data standard (meds): Facilitating machine learning for health. In *ICLR 2024 Workshop on Learning from Time Series For Health*, pp. 03–08, 2024.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pp. 456–473. Springer, 2022.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Leonard E Braitman and Frank Davidoff. Predicting clinical states in individual patients. *Annals of Internal Medicine*, 125(5):406–412, 1996.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

Talen Chen, Samaneh Madanian, David Airehrour, and Marianne Cherrington. Machine learning methods for hospital readmission prediction: systematic analysis of literature. *Journal of Reliable Intelligent Environments*, 8(1):49–66, 2022.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Alex GC de Sá, Daniel Gould, Anna Fedyukova, Mitchell Nicholas, Lucy Dockrell, Calvin Fletcher, David Pilcher, Daniel Capurro, David B Ascher, Khaled El-Khawas, et al. Explainable machine learning for icu readmission prediction. *arXiv preprint arXiv:2309.13781*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Huilong Duan, Zhoujian Sun, Wei Dong, Kunlun He, and Zhengxing Huang. On clinical event prediction in patient treatment trajectory using longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 24(7):2053–2063, 2019.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, L Julián Lechuga López, and Farah E Shamout. Medmod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans. *Proceedings of Machine Learning Research*, 287:1–23, 2025.

Adibvafa Fallahpour, Mahshid Alinoori, Wenqian Ye, Xu Cao, Arash Afkanpour, and Amrit Krishnan. Ehrmamba: Towards generalizable and scalable foundation models for electronic health records. *arXiv preprint arXiv:2405.14567*, 2024.

Hui Feng, Yuntzu Yin, Emiliano Reynares, and Jay Nanavati. Ontologyrag: better and faster biomedical code mapping with retrieval-augmented generation (rag) leveraging ontology knowledge graphs and large language models. In *International Workshop on Knowledge-Enhanced Information Retrieval*, pp. 71–86. Springer, 2025a.

Yichun Feng, Jiawei Wang, Ruikun He, Lu Zhou, and Yixue Li. A retrieval-augmented knowledge mining method with deep thinking llms for biomedical research and clinical support. *GigaScience*, 14:giaf109, 2025b.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 75–84, 2014.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pp. 479–503. PMLR, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Huiling Hu, Jiashuai Li, Hui Ge, Bilin Wu, Tingting Feng, Xue Wu, and Xuanna Wu. Prognostic models for unplanned intensive care unit readmission risk prediction: A systematic review and meta-analysis based on hsroc model. *Nursing in critical care*, 30(2):e13306, 2025.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. Unifying heterogeneous electronic health records systems via text-based code embedding. In *Conference on Health, Inference, and Learning*, pp. 183–203. PMLR, 2022a.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, and Edward Choi. Unihpf: Universal healthcare predictive framework with zero domain knowledge. *arXiv preprint arXiv:2211.08082*, 2022b.

Kyunghoon Hur, Jungwoo Oh, Junu Kim, Jiyoun Kim, Min Jae Lee, Eunbyeol Cho, Seong-Eun Moon, Young-Hak Kim, Louis Atallah, and Edward Choi. Genhpf: General healthcare predictive framework for multi-task multi-source learning. *IEEE Journal of Biomedical and Health Informatics*, 28(1):502–513, 2023.

Shuyue Jia, Subhrangshu Bit, Varuna H Jasodanand, Yi Liu, and Vijaya B Kolachalama. Agentic memory-augmented retrieval and evidence grounding for medical question-answering tasks. *medRxiv*, pp. 2025–08, 2025.

Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825*, 10: 3, 2023a.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023b.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pp. 49–55, 2020.

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.

Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

Junu Kim, Chaeeun Shim, Bosco Seong Kyu Yang, Chami Im, Sung Yoon Lim, Han-Gil Jeong, and Edward Choi. General-purpose retrieval-enhanced medical prediction model using near-infinite history. *arXiv preprint arXiv:2310.20204*, 2023.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

Simon A Lee, Sujay Jain, Alex Chen, Kyoka Ono, Arabdha Biswas, Ákos Rudas, Jennifer Fang, and Jeffrey N Chiang. Clinical decision support using pseudo-notes from multiple streams of ehr data. *npj Digital Medicine*, 8(1):394, 2025.

Matthew Lewis, Samuel Thio, Richard JB Dobson, and Spiros Denaxas. Grounding large language models in clinical evidence: A retrieval-augmented generation system for querying uk nice clinical guidelines. *arXiv preprint arXiv:2510.02967*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*, 2024.

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

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Matthew BA McDermott, Justin Xu, Teya S Bergamaschi, Hyewon Jeong, Simon A Lee, Nassim Oufattole, Patrick Rockenschaub, Kamilė Stankevičiūtė, Ethan Steinberg, Jimeng Sun, et al. Meds: Building models and tools in a reproducible health ai ecosystem. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6243–6244, 2025.

Hong Meng, Liang Guo, Yucheng Pan, Bin Kong, Wei Shuai, and He Huang. Machine learning based clinical prediction model for 1-year mortality in sepsis patients with atrial fibrillation. *Heliyon*, 10(21), 2024.

Nurul Athirah Nasarudin, Fatma Al Jasmi, Richard O Sinnott, Nazar Zaki, Hany Al Ashwal, Elfadil A Mohamed, and Mohd Saberi Mohamad. A review of deep learning models and online healthcare databases for electronic health records and their use for health prediction. *Artificial Intelligence Review*, 57(9):249, 2024.

Beatriz Nistal-Nuño. Developing machine learning models for prediction of mortality in the medical intensive care unit. *Computer Methods and Programs in Biomedicine*, 216:106663, 2022.

Mikkel Odgaard, Kiril Vadimovic Klein, Sanne Møller Thysen, Espen Jimenez-Solem, Martin Sillesen, and Mads Nielsen. Core-behrt: A carefully optimized and rigorously evaluated behrt. *arXiv preprint arXiv:2404.15201*, 2024.

Chao Pang, Xinzhuo Jiang, Krishna S Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In *Machine Learning for Health*, pp. 239–260. PMLR, 2021.

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S Kalluri, Elise L Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, et al. Cehr-gpt: Generating electronic health records with chronological patient timelines. *arXiv preprint arXiv:2402.04400*, 2024.

Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Matthias Keicher, and Nassir Navab. From ehrs to patient pathways: Scalable modeling of longitudinal health trajectories with llms. *arXiv preprint arXiv:2506.04831*, 2025.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

Weijieying Ren, Jingxi Zhu, Zehao Liu, Tianxiang Zhao, and Vasant Honavar. A comprehensive survey of electronic health record modeling: From deep learning approaches to large language models. *arXiv preprint arXiv:2507.12774*, 2025.

Pawel Renc, Michal K Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew BA McDermott, Jaroslaw Was, Anthony E Samir, Jonathan W Cunningham, et al. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14: giaf107, 2025.

Walid Saba, Suzanne Wendelken, and James Shanahan. Question-answering based summarization of electronic health records using retrieval augmented generation. *arXiv preprint arXiv:2401.01469*, 2024.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, 2024.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 464–472. IEEE, 2017.

Jiwoong Sohn, Yein Park, Chanwoong Yoon, Sihyeon Park, Hyeon Hwang, Mujeen Sung, Hyunjae Kim, and Jaewoo Kang. Rationale-guided retrieval augmented generation for medical question answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12739–12753, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, 2025.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, et al. Generative medical event models improve with scale. *arXiv preprint arXiv:2508.12104*, 2025.

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.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, 2024.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 754–765, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature communications*, 14(1):7857, 2023.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pp. 102–120. Springer, 2024.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Min Zhang and Tsung-Ting Kuo. Early prediction of long hospital stay for intensive care units readmission patients using medication information. *Computers in biology and medicine*, 174: 108451, 2024.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*, 2024.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*, 2025.

Shaojie Zhong, Li Rong Wang, Zhuoxuan Zhan, Yih Yng Ng, and Xiuyi Fan. A hybrid approach for irregular-time series prediction using electronic health records: an intensive care unit mortality case study. *ACM Transactions on Computing for Healthcare*, 6(4):1–33, 2025.

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.

Angelo Ziletti and Leonardo D'Ambrosi. Generating patient cohorts from electronic health records using two-step retrieval-augmented text-to-sql generation. *arXiv preprint arXiv:2502.21107*, 2025.

A Appendix

B Data Preprocessing

We describe the preprocessing details performed on MIMIC-IV dataset in this section from raw tables to training ready data. For EHRSHOT dataset, the maintainer provides a training ready dataset along with splits and a labels. For further details regarding EHRSHOT, we refer the reader to the original work (Wornow et al., 2023).

B.1 Cohort construction

Pretraining Cohort: We first convert the MIMIC-IV dataset into the Medical Event Data Standard (MEDS) format (Arnrich et al., 2024; McDermott et al., 2025), representing each patient history as a chronologically ordered sequence of medical events covering all patient visits within a single sequence. MEDS extraction pipelines of MIMIC-IV fix the splits percentages upon extraction to 80% train, 10% tuning, and 10% held-out based on the *subject_id*. We keep the splits percentages as is across the whole preprocessing pipelines for ease of data reproducibility. Following that, we use the training split as the pretraining cohort in our work and exclude both tuning and held-out splits to avoid any possible *leakage*. The resulting cohort consists of $291,702$ patients with $707,716,112$ medical events. Finally, we hold out 5% of the training patients for pretraining quality monitoring.

Downstream Cohort: For downstream tasks, we construct the dataset using ICU stays only and exclude patients with hospital admissions but no ICU stays. We follow the inclusion and exclusion criteria proposed in previous work (Harutyunyan et al., 2019) to define the downstream cohort. First, we include all adult patients with an age at admission of $\geq$ 18 years. Second, we retain patients with a single ICU stay within the same hospital admission. Third, we include only patients whose admission and discharge ICU units are identical. Finally, we examine the length of each valid ICU stay and retain only those stays lasting at least 24 hours. This process results in a downstream cohort comprising 49,839 patients with 61,163 ICU stays.

B.2 Data Preprocessing

We perform preprocessing to reduce irregularities and transform the data into a standardized format. First, we exclude HCPCS billing codes from the patient timelines, as they appear in the charted data in the patient timeline rather than at discharge time. This situation may introduce temporal leakage of billing codes, so we safely remove them. Second, we observed that vital signs are present in the *OMR.csv* table as text values rather than numerical values. We convert them into numerical values and store them in the *numeric_value* column instead of the *text_value* column. Third, we handle outliers in numeric values using the *MEDS-transform* Python package by computing dataset-wide statistics over the *train* split for each numerical event and excluding values that lie beyond three standard deviations (3σ) from the mean. Fourth, we filter out codes that appear rarely in the dataset, with an appearance frequency of < 4 events. We also exclude events with no valid timestamp, except for static events that are intentionally without timestamps. Finally, we normalize all remaining numeric values using *MEDS-transform* by recomputing normalization statistics after outlier removal.

B.3 Patient Timeline Construction

Figure S1 depicts a sample patient timeline segmented into distinct care stages and covering all relevant clinical information. To construct a patient timeline, we first examine each patient sequence and identify hospital admissions (HOSP) using the *hadm_id* field, as well as ICU stays using the *icustay_id* field, which are the only care stages in MIMIC-IV associated with unique identifiers.

During this process, we observe that certain medical events, appear in the patient timeline without an explicit reference to any hospital admission. To handle these events without excluding them, we assess their temporal proximity to admissions present in the patient timeline. Events occurring after *ED_registration*, and events up to *HOSPITAL_ADMISSION*, are labeled as emergency department

(ED) events, while events that occur between two consecutive hospital stays outside the admission/emergency boundaries are treated as outpatient (OUTP) events.

To represent time gaps between consecutive visits, we insert artificial time tokens following prior work (Pang et al., 2021; Fallahpour et al., 2024), denoted as GAP, as illustrated in Figure S1. These tokens encode temporal intervals spanning, days, weeks, months, or years, depending on the elapsed time between visits.

A standard patient timeline primarily consists of hospital events including laboratory measurements, medications, ICD procedure codes, ICD diagnosis codes, and DRG codes, as well as ICU events such as chart events, infusions, procedures, and fluid outputs.

Each patient sequence is prepended with two events indicating the gender and birth date. While gender is a static event with no timestamp, birth date is a timestamped event represented as *MEDS_BIRTH*. If a patient dies during or after discharge, a special timestamped token, `MEDS_DEATH`, is appended to the sequence. We also include special tokens required for language modeling objectives, such as `[PAD]`, `[MASK]`, `[CLS]`, and `[UNK]`, where applicable.

For events associated with numeric values, such as laboratory tests and ICU chart events, we attach the corresponding numeric measurements, normalized. For events without numeric values, including `PROCEDURE-ICD` and `DIAGNOSIS-ICD`, we associate a learnable *null* parameter. To encode temporal information, we compute the time difference between consecutive events in minutes and attach this value to each event. Because time gaps may vary substantially both within and across visits, we scale time deltas using the transformation defined in Equation 1, which maps values to the range $[0, 1]$ and prevents extreme magnitudes.

$$\Delta t_i^{'} = \frac{log(1 + \Delta t_i)}{log(\Delta T_{max})} \tag{1}$$

where Δt_i is the consecutive time difference at event i, and ΔT_{max} is the maximum time difference present in the dataset. Temporal representations are then processed using Time2Vec (Kazemi et al., 2019) during training.

To further contextualize each event within the patient timeline, we introduce three additional categorical representations, care stage, visit order, and event type, as illustrated in the bottom rows of Figure S1. The care stage embedding explicitly encodes the clinical context in which each event occurs, distinguishing between outpatient (OUTP), emergency department (ED), inpatient hospitalization (HOSP), ICU stay (ICU), and artificial gap (GAP) events. This representation enables the model to differentiate identical medical events occurring under different clinical settings, which often carry distinct semantic meanings.

In addition, we assign a visit order index to each event, indicating the chronological visit number to which it belongs. The visit order is incremented across care episodes and reset for time gap tokens, allowing the model to reason over longitudinal disease progression across multiple encounters rather than treating the timeline as a flat sequence. Finally, the event type embedding encodes the high-level category of each event, such as laboratory tests, medications, diagnoses, procedures, administrative markers, or special tokens. Below, we provide a formal definition of each component that define a complete event (c_{nk}, v_{nk}, t_{nk}):

1. **Concept Embedding:** Each clinical concept c_{nk} (e.g., diagnosis, lab test, procedure, medication) is associated with a learnable semantic embedding: $\mathbf{u}_{c_{nk}} \in \mathbb{R}^d$.
2. **Value Embedding:** For events with a numerical measurement v_{nk}, the value is normalized and projected, such that $\mathbf{v}_{nk} = \theta(v_{nk}) \in \mathbb{R}^d$, and $\theta(\cdot)$ is a multi-layer perceptron. Events without values are assigned a learnable (null-value) parameter.
3. **Time Encoding:** Each timestamp t_{nk} is mapped into a temporal embedding representing the local temporal information as time deltas between consecutive events, such that $\mathbf{t}_{nk} \in \mathbb{R}^d$.
4. **Visit Order Embedding:** To capture patient-level longitudinal ordering, each event inherits a visit index embedding: $\mathbf{r}_{\text{visit}(n)} \in \mathbb{R}^d$, where the indices are encoded as continuous learnable positional vectors. This allows the model to distinguish between early and late episodes in the patient trajectory (global temporal representation).

	Visit 1											Visit 2									
	OUTP		GAP	ED		HOSP					GAP	ED		HOSP				ICU			HOSP	
Concept	L_1	L_2	T	L_1	L_2	L_3	P_1	P_2	D_1	D_2	T	L_1	L_2	P_1	M_1	M_2	P_2	M_3	M_4	D_1	D_2
Value	x_l	x_l	*null*	x_l	x_l	x_l	*null*	*null*	*null*	*null*	*null*	x_l	x_l	*null*	*null*	*null*	*null*	*null*	*null*	*null*	*null*
Time	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt	Δt
Stage	*1*	*1*	*0*	*2*	*2*	*3*	*3*	*3*	*3*	*3*	*0*	*2*	*2*	*3*	*3*	*3*	*4*	*4*	*4*	*3*	*3*
Visit	*1*	*1*	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*	*0*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*
Type	*1*	*1*	*2*	*1*	*1*	*1*	*3*	*3*	*4*	*4*	*2*	*1*	*1*	*3*	*5*	*5*	*3*	*5*	*5*	*4*	*4*

Figure S1: Sample patient timeline consisting of two visits. We illustrate the main components constituting the patient timeline in our implementation including medical events, numeric values, time values, care stage representation, visit order representation, and event type representation. Abbreviations, OUTP: Outpatient, ED: Emergency Department, HOSP: Hospital Admission, ICU: Incentive Care Unit, L: Labs, T: TIME-GAP, P: Procedure, M Medication, D: Diagnosis, x_a numeric value. This is a sample timeline created randomly for illustration purposes and does not reflect real data.

5. **Care Stage Embedding:** Each visit belongs to a clinical stage (e.g., outpatient, emergency, inpatient, ICU). We assign a learned embedding: $\mathbf{s}_{\text{stage}(n)} \in \mathbb{R}^d$ to enable the model to incorporate care-context information, e.g., ICU vs. emergency events carry different clinical semantics.
6. **Type Embedding:** Each event also receives a higher-level event type identifier that encodes the data source from which it originates (ICU chart event, ICU fluid output, etc.): $\mathbf{w}_{\text{type}(nk)} \in \mathbb{R}^d$. This provides local structural information that allows the model to differentiate heterogeneous event types.

To process these events within a unified architecture, we map each event and its associated metadata into a single representation $e_{nk} \in \mathbb{R}^d$, where d is the vector dimension. This embedding considers six key components: clinical concept $\mathbf{u}_{c_{nk}}$, numeric value $\mathbf{v}_{nk}$, time $\mathbf{t}_{nk}$, visit-order $\mathbf{r}_{\text{visit}(n)}$, care stage $\mathbf{s}_{\text{stage}(n)}$, and event-type $\mathbf{w}_{\text{type}(nk)}$ embeddings as shown below. The unified event representation is then constructed by summation following (Fallahpour et al., 2024; Li et al., 2020):

$$e_{nk} = (\mathbf{u}_{c_{nk}} + \mathbf{v}_{nk} + \mathbf{t}_{nk} + \mathbf{r}_{\text{visit}(n)} + \mathbf{s}_{\text{stage}(n)} + \mathbf{w}_{\text{type}(nk)}) \in \mathbb{R}^d$$

By jointly modeling care stage, visit order, and event type, the patient timeline representation preserves structural, temporal, and semantic distinctions across events, facilitating more expressive and context-aware representation learning over heterogeneous EHR sequences.

B.4 Query/History Definition

For each prediction instance in the downstream cohort, we decompose the patient timeline into a *query* segment and a corresponding *history* segment to enable retrieval-based conditioning. The query represents the patient's current clinical state exactly just before the time of prediction and consists of events observed within a task-specific observation window. For details on the observation windows used for each task, we refer the reader to Section C.2 and Table S1. All events occurring strictly before the query window constitute the patient history and are segmented into fixed-length units that serve as retrievable candidates. Figure S2 illustrates the query/history definitions for the various prediction windows considered in this work. This formulation mirrors real-world clinical reasoning, in which decisions are made based on the present encounter while selectively consulting relevant past visits. By explicitly separating query and history, the model is encouraged to retrieve and integrate the most relevant historical context, rather than uniformly encoding the entire past, enabling scalable reasoning over long and heterogeneous patient trajectories.

B.5 Leakage Prevention

Due to the sensitivity of this query/history definition with respect to temporal leakage, we take strict measures to avoid any leakage. (1) We first index each patient timeline of n events from $0 - n$. (2) We then specify the query boundaries on the indexed patient timeline. (3) We examine each query

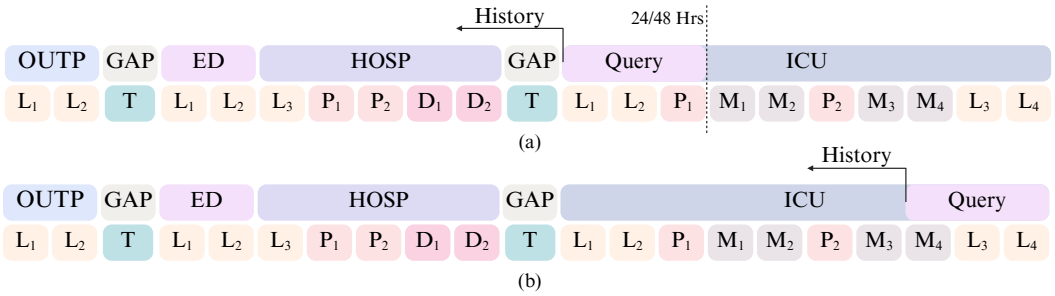


Figure S2: Illustration of query and history definition (a): Query/history boundaries for tasks with time-bounded prediction window. (b) Query/history boundaries task with entire stay prediction window.

if its contents could include any billing codes that might leak labels. (4) We iteratively refine query boundaries within the task window boundaries to ensure absence of any ineligible information. The same process is applied to the history chunks, especially those chunks closer to the query boundaries.

B.6 Chunking

Structured EHR trajectories are inherently irregular, with clinical events occurring at variable temporal resolutions and across diverse care settings. To organize longitudinal patient history into retrievable units, `EHR-RAGp` partitions the timeline into clinically coherent chunks. We investigate four chunking strategies: *event-based*, *time-based*, *visit-level*, and *care-stage* chunking. These strategies provide complementary views of patient history by capturing different temporal and clinical granularities. Figure S3 visually illustrates these chunking strategies. Below, we formally outline our proposed chunking strategies:

- **Event-Based:** Events are segmented into fixed-size groups based on event count. Each chunk contains a predefined number of embedded clinical events. Consecutive chunks overlap by a fixed number of events to preserve continuity across chunk boundaries. We set the chunk overlap to 12.5% of the original chunk length.
- **Time-Based:** Events are segmented into fixed time windows, for example 6-hour, 12-hour, 24-hour blocks. Each chunk contains all events occurring in its temporal window.
- **Visit-level:** Each hospital visit, admission, or encounter is treated as a candidate chunk. This aligns with natural clinical boundaries: emergency department stays, inpatient admissions, and outpatient encounters.
- **Care-Stage:** Events are grouped by clinical stage (outpatient, emergency, inpatient, ICU), ensuring that each chunk reflects a coherent care stage.

In cases where the chunk length is shorter than the context window, we pad the sequence to match the context length. In cases where the chunk length is greater than the context length, we recursively impose *Event-Based Chunking* with proper metadata handling, without overlap.

C Experimental Setup

C.1 Data Splitting

We use a fixed train/validation/test partition shared across all downstream prediction tasks. Of the 61,163 eligible stays, 49,080 (80.2%) are assigned to training, 5,986 (9.8%) to validation, and 6,097 (10.0%) to testing. The validation set is used for model selection and hyperparameter tuning, while the test set is reserved for final evaluation. Table S2 summarizes the task-specific label distributions within each split. To prevent information leakage, all stays associated with the same patient are assigned to a single split. The splits are inherited from the initial MEDS extraction, for ease of reproducibility.

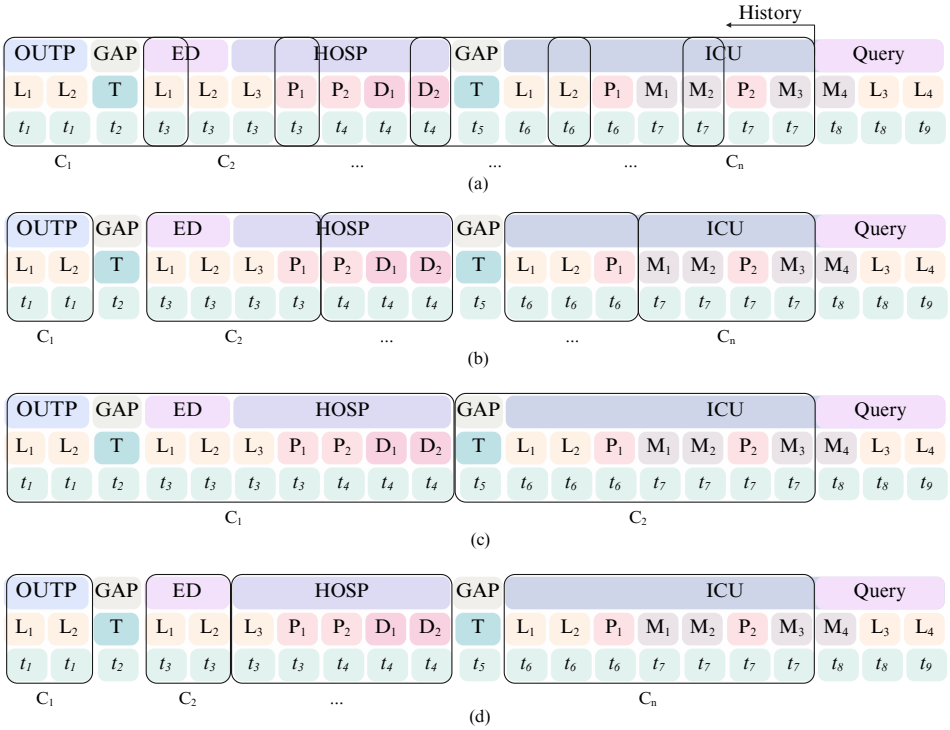


Figure S3: Visual illustration of the chunking strategies employed with EHR-RAGp. (a) Event-based chunking. (b) Time-based chunking. (c) Visit-level chunking. (d) Care-stage chunking.

C.2 Downstream Tasks

- **In-Hospital Mortality:** This task aims to predict whether a patient will die during the current hospital admission as a binary classification problem, using information available up to a predefined observation window. It is a widely used benchmark for evaluating clinical risk prediction models. In our setup, we condition the in-hospital mortality task on the first 48 hours spent in the ICU, following (Harutyunyan et al., 2019; Hayat et al., 2022; Elsharief et al., 2025).
- **1-Year Mortality Post-Discharge:** This task focuses on predicting whether a patient will die within one year following hospital discharge or no. It captures longer-term outcomes and reflects the model's ability to leverage both acute and chronic patterns in a patient's longitudinal history. We consider the entire ICU stay data as a valid prediction window (Meng et al., 2024; Nistal-Nuño, 2022; Ghassemi et al., 2014). For patients with multiple ICU stay and positive label of mortality within 1-year, we only assign the positive label to all stays that are within one year from the mortality date.
- **Long Length of Stay:** This task predicts whether a patient's hospital stay will exceed a predefined duration threshold. It serves as a proxy for resource utilization and clinical complexity and is commonly used to assess a model's ability to identify patients at risk of prolonged hospitalization. We set the length-of-stay threshold to 7 days following (Wornow et al., 2024; 2023; Kim et al., 2023) and define the prediction window as the first 24 hours spent in the ICU, following (Zhang & Kuo, 2024; Alghatani et al., 2021).
- **ICU Readmission:** This task aims to predict whether a patient will be readmitted to the intensive care unit within a specified time window after ICU discharge. It is an important indicator of care quality and patient instability and requires effective modeling of prior clinical trajectories. We set the readmission window to 30 days post-discharge following (Hu et al., 2025; Wornow et al., 2023; Chen et al., 2022) and use the entire ICU stay as the prediction window, following (Wornow et al., 2023; de Sá et al., 2023).

Table S1: Overview of the downstream tasks and the query/ history definition for each task

Task	Label type	Prediction Window	Reference point	Query Window	History Chunk
In-hospital mortality	Binary	48 hours since ICU admission	ICU admission	Last L event $\leq$ ICU admission + 48 hours	All events $<$ query start
Long Length of stay-7 Days	Binary	24 hours since ICU admission	ICU admission	Last L event $\leq$ ICU admission + 24 hours	All events $<$ query start
ICU Readmission-30 Days	Binary	Entire ICU stay	ICU discharge	Last L event $\leq$ ICU discharge	All events $<$ query start
1 Year Mortality	Binary	Entire ICU stay	ICU discharge	Last L event $\leq$ ICU discharge	All events $<$ query start

Table S2: Label distribution across data splits for the downstream prediction tasks.

Task	Train (N = 49,080 stays)		Validation (N = 5,986 stays)		Test (N = 6,097 stays)	
	Pos. (%)	Neg. (%)	Pos. (%)	Neg. (%)	Pos. (%)	Neg. (%)
In-Hospital Mortality	5,004 (10.196)	44,076 (89.804)	642 (10.725)	5,344 (89.275)	621 (10.185)	5,476 (89.815)
Long LOS ($\geq$ 7 Days)	6,678 (13.606)	42,402 (86.394)	840 (14.033)	5,146 (85.967)	763 (12.514)	5,334 (87.486)
1-Year Mortality Post Discharge	8,271 (16.852)	40,809 (83.148)	1,029 (17.190)	4,957 (82.810)	1,029 (16.877)	5,068 (83.123)
ICU Readmission within 30 Days	2,093 (4.264)	46,987 (95.736)	227 (3.792)	5,759 (96.208)	273 (4.478)	5,824 (95.522)

C.3 EHR-RAGp Architecture

EHR-RAGp uses a RoBERTa-base encoder (Liu et al., 2019) with 12 transformer layers, 12 attention heads, a hidden size of 768, and learned absolute positional embeddings. The encoder supports sequences of up to 512 tokens and produces contextualized token representations, which are aggregated into sequence-level representations using either [CLS] or mean pooling.

Each event token is represented as the sum of multiple embedding components. In addition to the core event embedding, we incorporate categorical embeddings for care stage, visit order, and event type, all projected to the same hidden dimension (768) and learned jointly with the model. Numerical values associated with events are normalized and projected through a small MLP before being fused with the categorical embeddings. In the absence of a numeric value, we use a learnable null embedding of the same dimensionality. Temporal information is encoded using Time2Vec (Kazemi et al., 2019), applied to scaled time-delta features and added to the token representation.

For retrieval, the pretrained backbone encoder is used to embed historical timeline chunks, which are stored in a vector database. During training, the query representation is used to retrieve the top-M most similar history chunks based on cosine similarity in the embedding space. Retrieval is non-parametric and does not involve gradient updates during the lookup stage; however, the encoder EHR-RAGp itself is fine-tuned end-to-end during downstream training.

The prototype module consists of a learnable set of prototype vectors defined in the same embedding space as the query and history representations. Both query and retrieved history embeddings are projected into the prototype space, where relevance scores are computed via similarity between history embeddings and prototypes, conditioned on the query representation. These scores are used to weight or filter retrieved history chunks, enabling selective integration of long-range context.

The fusion module combines the query representation with the filtered history representations using a shallow transformer with 2 layers and 4 attention heads. The fused representation is then passed to a task-specific prediction head implemented as a lightweight feedforward network. For all downstream tasks, the prediction head outputs a single logit optimized using binary cross-entropy (BCE) loss. All architectural components beyond retrieval are differentiable and optimized jointly during supervised fine-tuning.

C.4 Pretraining Details

We adopt masked language modeling (MLM) objective applied to structured EHR event sequences for EHR-RAGp pretraining. During pretraining, patient timelines are tokenized into fixed-length sequences. We consider the full patient timeline during pretraining and pass it to the encoder backbone as overlapped chunks of 512 tokens with 64 tokens of overlap. Hence, we produce dense sequences, whereas padding occurs only in the last chunk to fulfill the sequence length. During pretraining, a subset of 15% of the chunk events is randomly selected for masking following the RoBERTa (Liu et al., 2019) masking strategy. Specifically, masked positions are replaced with a special [MASK] token with probability 80%, substituted with a random event token with probability 10%, or left un-

changed with probability 10%. The model is trained to predict the original event identity at masked positions using a cross-entropy loss over the entire vocabulary of $48{,}506$ tokens.

All embedding components, including event embeddings and categorical embeddings, are trained jointly during this stage. No temporal, or numeric value components are used during pretraining; our objective at this stage is solely to learn contextualized representations of structured EHR sequences. Pretraining is conducted on the full pretraining cohort described in Section B.1.

We use *AdamW* (Loshchilov & Hutter, 2017) optimizer with an initial learning rate specified using cyclical learning rates following (Smith, 2017), a weight decay of 1×10^{-2}, a batch size of 64, and a cosine annealing learning scheduler. Training is performed for 100 epochs over the pretraining corpus. We monitor pretraining quality using accuracy on a held-out validation set, as described in Section B.1. Early stopping is implemented with a patience of three epochs if validation accuracy does not improve. The resulting pretrained backbone is subsequently reused for downstream fine-tuning and retrieval embedding generation. All pretraining experiments are conducted using a single NVIDIA A100 GPUs.

C.5 Vector Database Setup

To enable retrieval-augmented conditioning, we construct a patient-specific vector database that stores embeddings of historical visit segments derived from the pretrained `EHR-RAGp` backbone. We use Facebook AI Similarity Search (FAISS[1]) as the vector index for `EHR-RAGp`. For each patient in the downstream cohort, the history portion of the timeline (as defined in Section B.4) is segmented into fixed-length chunks, where each chunk corresponds according to the various chunking strategies proposed in our work. Each chunk is independently encoded using the pretrained backbone encoder, and the resulting pooled representation is stored as a dense vector.

We use mean pooling over the sequence representations and normalize the resulting embeddings prior to storage in the index. The vector index is constructed offline before downstream training and remains fixed during supervised fine-tuning. This design avoids repeated recomputation of historical embeddings and ensures scalable retrieval over long patient histories. During retrieval, we perform similarity search based on cosine similarity between the query embedding and historical chunk embeddings. It is worthy to note as well that we do not store chunked data along with their encoded vector, instead; we keep track of the chunk indices in the full chunked patient timeline. Upon retrieval, we perform online chunking during training to retrieve selected candidates by similarity search via their indices. This approach allow for faster retrieval and reduce the need for maintaining large database that accommodates the chunked data along with the vector index.

C.6 Downstream Training Details

Table S3 summarizes the hyperparameter settings and search spaces used for downstream training of `EHR-RAGp`. Core backbone parameters are fixed across all experiments, including the RoBERTa-base encoder with a hidden dimension of 768, 12 layers, and 12 attention heads.

Retrieval-related hyperparameters including query chunk size, history chunk size, history chunk overlap, number of retrieved chunks, and similarity metric. all retrieval hyperparameters are held constant with values of 1024, 256, 32, 24, *cosine similarity*, respectively.

For prototypes module, we use fixed prototype dimensionality of 768. For temperature, we experiment with temperature pairs, (T_q, T_h), for prototype assignment of $[(0.025, 0.1), (0.05, 0.2), (0.02, 0.08)]$ and Weighing temperature T_s values of $\{0.1, 0.15, 0.2, 0.25\}$, whereas the number of prototypes is set to be optimized with discrete search space in $\{64\,128, 256, 512\}$.

For fusion, we set up a transformer encoder of 2 layers and 4 attention heads, and dropout of 0.1. We set the pooling strategy to be set during hyperparameter optimization with search space of $\{mean, query\}$.

[1] https://github.com/facebookresearch/faiss

Table S3: Hyperparameter settings and search spaces used in EHR-RAGp experiments.

Category	Hyperparameter	Fixed	Search Space
Backbone	Encoder architecture	RoBERTa-base	–
	Hidden dimension (d)	768	–
	Number of layers	12	–
	Number of attention heads	12	–
	pooling	–	{mean, CLS}
Retrieval	Query Chunk size	1024	–
	History Chunk size	256	–
	History Chunk overlap*	32	–
	Number of retrieved chunks (M)	24	–
	Similarity metric	Cosine	–
Prototypes	Number of prototypes	–	{64, 128, 256, 512}
	Prototype dimension	768	–
	Prototype temperature (T_q, T_h)	–	$\{(0.025, 0.1), (0.05, 0.2), (0.02, 0.08)\}$
	Weighing temperature T_s	–	$\{0.1, 0.15, 0.2, 0.25\}$
Fusion	Fusion layers	2	–
	Fusion attention heads	4	–
	pooling	–	{mean, query}
	Dropout	0.1	–
Training	Optimizer	SGD	–
	Usage penalty λ_u	–	$\{0.004, 0.005, 0.006, 0.007\}$
	Learning rate	–	$[1e{-}5, 5e{-}4]$
	Weight decay	–	$[1e{-}3, 1e{-}2]$
	Batch size	–	{8,12,16,20}
	Max epochs	75	–

* overlap is used only with event-based chunking strategy

Table S4: Best configurations selected for each downstream task after hyperparameter optimization.

Hyperparameter	ICU-Readmit 30d	In-hospital Mortality	Long LOS 7d	1YR Mortality
Chunking Strategy	Event-Based	Event-Based	Visit-Level	Care-Stage
Prototype count	64	256	64	512
(T_q, T_h)	$(0.025, 0.1)$	$(0.05, 0.2)$	$(0.025, 0.1)$	$(0.05, 0.2)$
Weighting temperature (T_s)	0.25	0.20	0.25	0.22
Usage penalty (λ_u)	0.005	0.005	0.006	0.006
Learning rate	2.592×10^{-4}	1.909×10^{-4}	1.119×10^{-4}	1.948×10^{-5}
Weight decay	1.454×10^{-3}	1.380×10^{-3}	1.961×10^{-3}	1.54×10^{-3}

For training hyperparameters, we experiment with Stochastic Gradient Descent (SGD) optimizer, learning rate in $[1e{-}5, 5e{-}4]$, Weight decay in $[1e{-}3, 1e{-}2]$ and cosine annealing learning rate scheduler. For usage penalty λ_u, we experiment with fixed values in $\{0.004, 0.005, 0.006, 0.007\}$.

For hyperparameters optimization, we run 25 trails for each prediction task using Bayesian search. All downstream evaluation experiments were conducted using a single NVIDIA A100 80GB GPUs. Maximum batch size with cane be used with such a GPU in our experiments is 24 query chunks of length 512 events along with 24 history chunk of length 256, while average epoch time given the max settings is 38 minutes. It is worthy to note that higher capacity GPUs can be also used with increase in the max settings mentioned earlier. With NVIDIA H200 141GB for instance, max batch size can reach 60, while the epoch time can reach as low as 13 minutes. Table S3 summarizes the the downstream tasks training hyperparameters for both fixed and optimized ones, whereas Table S4 reports the best hyperparameters obtained for each task.

D BASELINES

D.1 CLINICAL BASELINES

- **DescEmb** (Hur et al., 2022a): is a text-based framework that represents patient history as textual descriptions of medical events, covering the event type, numeric value (if any), along with its unit of measurement. These sequences are then processed via clinical or general-purpose pretrained language models. The main goal of DescEmb is to reduce reliance on predefined medical code vocabularies through a shared linguistic vocabulary that can be utilized across different EHR datasets. We consider two variants of DescEmb in our experiments, including *BERT Fine-Tune* and *CLS Fine-Tune*.
- **GenHPF** (Hur et al., 2023): is an extension of DescEmb that goes beyond the basic description of the medical event type, value, and unit to include all features associated with an event in the raw EHR tables, enhancing medical event representation. For instance, a medication event will cover type, name, frequency, dosage, route, etc. It shares the same objective as DescEmb of being agnostic to differing EHR schemas across institutions.

- **Med-BERT** (Rasmy et al., 2021): is a BERT-based foundation model that is pretrained on structured EHR records for disease prediction tasks. It relies solely on structured diagnosis ICD codes as pretraining and evaluation data. The pretraining approach typically follows masked language modeling as proposed in the BERT paper (Devlin et al., 2019).
- **BEHRT** (Li et al., 2020): is another BERT-based foundation model that learns contextual embeddings of the patient's structured records in a similar fashion to Med-BERT via masked language modeling using diagnosis codes. However, it differs from Med-BERT by incorporating patient age as a temporal signal.
- **Hi-BEHRT** (Li et al., 2022): is a hierarchical extension of BEHRT that aims to handle longer, more comprehensive patient records. It captures short- and long-term patient patterns from hierarchically segmented medical sequences and is pretrained with the self-supervised BYOL framework (Grill et al., 2020).
- **CEHR-BERT** (Pang et al., 2021): is another BERT-based model that incorporates temporal and visit information to improve disease prediction. It represents patient visits as sequences of medical codes, augmented with artificial time tokens, age embeddings, and time embeddings.
- **REMed** (Kim et al., 2023): is a retrieval-based foundation model that processes textual descriptions in a similar fashion to both DescEmb and GenHPF. It operates by randomly sampling events from patient history and ranking them to retrieve the most relevant ones for different downstream tasks. In contrast to our proposed model, REMed operates at the medical event level, whereas `EHR-RAGp` operates at historical segment level to cover broad history beyond limited set of events, 128 in case of REMed.
- **EHRMamba** (Fallahpour et al., 2024): typically follows the embedding structure of CEHR-BERT but uses the state-space architecture MAMBA (Gu & Dao, 2024) instead of a BERT encoder. The use of a state-space model aims to enable efficient processing of relatively long EHR sequences compared to transformer models.

D.2 Transformer-based Baselines

- **RoBERTa** (Liu et al., 2019): is a transformer-based language model that builds upon BERT by removing the next-sentence prediction objective and adopting dynamic masking and larger pretraining corpora. RoBERTa proves that careful optimization of the pretraining pipleines significantly improves contextual representation learning, which makes it a strong general-purpose encoder baseline.
- **LongFormer** (Beltagy et al., 2020): is a transformer variant designed to handle long input sequences efficiently through a sparse attention mechanism that combines local windowed attention with task-specific global attention. This design reduces the quadratic complexity of standard self-attention, which in turn enables modeling of substantially longer contexts compared to standard transformers.
- **BigBird** (Zaheer et al., 2020): is another long-context transformer architecture that replaces full self-attention with a sparse attention pattern composed of random, local, and global attention. BigBird is theoretically proven to be a universal approximator while enabling scalable processing of long sequences, which making it suitable for tasks that involve extended contextual dependencies.
- **RoFormer** (Su et al., 2024): is a transformer model that introduces rotary positional embeddings (RoPE) to encode relative positional information directly into the self-attention mechanism. This approach improves extrapolation to longer sequences and better captures relative token relationships, which makes it effective for sequence modeling tasks with variable-length inputs.
- **ModernBERT** (Warner et al., 2025): is a re-engineered BERT-style encoder that incorporates recent architectural and training improvements, including optimized attention implementations, improved normalization, and enhanced efficiency on modern hardware. ModernBERT aims to retain the strengths of encoder-based transformers while achieving better performance and scalability.

D.3 Large Language Model Baselines

- **Qwen2.5-7B-Instruct** (Yang et al., 2024): is a general-purpose instruction-tuned large language model designed for conversational reasoning and text understanding across diverse domains.

- **Mistral-7B-Instruct** (Jiang et al., 2023a): is an instruction-following variant of the Mistral family of open-source large language models optimized for efficient inference and general reasoning capabilities.

- **Llama-3.1-8B-Instruct**: (Grattafiori et al., 2024) is a general-purpose instruction-tuned large language model designed for conversational interaction, reasoning, and text understanding across diverse domains.

- **MedGemma-1.5-4B-it** (Sellergren et al., 2026) is a medically adapted instruction-tuned language model developed for healthcare and biomedical applications. It is designed to support clinical reasoning and medical text understanding, and is evaluated here as a medical LLM baseline.

- **BioMistral-7B** (Labrak et al., 2024): is a biomedical domain-adapted variant of Mistral trained on large-scale biomedical corpora to improve performance on medical and scientific language tasks.

- **Llama3-OpenBioLLM-8B** (Ankit Pal, 2024): is a biomedical instruction-tuned large language model adapted from the Llama 3 architecture to improve medical knowledge, clinical reasoning, and biomedical text understanding.

D.4 LLM Evaluation Protocol

We evaluated both general-domain and medical-domain large language models using a zero-shot next-token prediction framework for clinical risk prediction. For each downstream task, patient timelines were converted into sequential textual event representations and truncated according to predefined prediction windows. To ensure fair comparison with encoder-based baselines, the same temporal slicing strategy and prediction horizons were used across all models. A fixed precomputed slices 1024 clinical events along with their numeric values events used.

Each patient sequence was formatted as a chronological list of clinical events and passed directly to the language model without supervised fine-tuning. We used causal language modeling inference and extracted the logits corresponding to the "Yes" and "No" tokens from the next-token distribution. All other token logits were ignored, and probabilities were normalized over the binary "Yes/No" subset using a softmax operation. The normalized probability assigned to the "Yes" token was used as the prediction score. We present the prompt used for each task in The prompts used for zero-shot evaluation are shown in Box D.4.

Performance was evaluated using AUROC and AUPRC. Confidence intervals were estimated using nonparametric bootstrap resampling with 1000 iterations.

LLM Prompt Templates

In-Hospital Mortality

```
You are an expert clinical risk prediction model using electronic health records.

--- PATIENT DATA ---
Electronic Health Records:
{ehr_text}

--- TASK ---
Will this patient die during this hospital admission?

Answer only using one word: Yes or No.

Answer:
```

1-Year Mortality

```
You are an expert clinical risk prediction model using electronic health records.

--- PATIENT DATA ---
Electronic Health Records:
{ehr_text}

--- TASK ---
Will this patient die within 1 year after hospital discharge?

Answer only using one word: Yes or No.

Answer:
```

ICU Readmission within 30 Days

```
You are an expert clinical risk prediction model using electronic health records.

--- PATIENT DATA ---
Electronic Health Records:
{ehr_text}

--- TASK ---
Will this patient be readmitted to the ICU within 30 days after ICU discharge?

Answer only using one word: Yes or No.

Answer:
```

Long Length of Stay ¿ 7 Days

```
You are an expert clinical risk prediction model using electronic health records.

--- PATIENT DATA ---
Electronic Health Records:
{ehr_text}

--- TASK ---
Will this patient have a hospital length of stay longer than 7 days?

Answer only using one word: Yes or No.

Answer:
```

E Additional Results

E.1 LLM Baseline Results

Table S5 reports the zero-shot performance of general-purpose and clinically adapted LLMs. Performance varies across tasks: MedGemma performs best on ICU readmission, BioMistral on long LOS, and Qwen2.5 on both mortality outcomes. Clinically adapted LLMs do not consistently outperform their general-purpose counterparts. As expected under a zero-shot setting, these models perform below the task-specific `EHR-RAGp` model; therefore, we include them as reference results rather than as directly matched baselines.

Table S5: Performance results zero-shot evaluation across LLM baselines

Model	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)
General-purpose LLMs								
Qwen2.5-7B-Instruct	0.484 (0.451, 0.520)	0.043 (0.038, 0.052)	**0.729 (0.707, 0.750)**	**0.309 (0.273, 0.345)**	0.504 (0.483, 0.525)	0.130 (0.119, 0.144)	**0.576 (0.557, 0.594)**	**0.201 (0.187, 0.219)**
Mistral-7B-Instruct	0.470 (0.432, 0.504)	0.044 (0.037, 0.054)	0.700 (0.678, 0.724)	0.249 (0.219, 0.284)	0.529 (0.507, 0.552)	0.145 (0.131, 0.163)	0.533 (0.513, 0.553)	0.188 (0.173, 0.207)
Llama-3.1-8B-Instruct	0.498 (0.463, 0.530)	0.044 (0.038, 0.051)	0.659 (0.636, 0.681)	0.184 (0.162, 0.209)	0.452 (0.430, 0.474)	0.113 (0.105, 0.124)	0.520 (0.500, 0.537)	0.177 (0.165, 0.192)
Clinical LLMs								
medgemma-1.5-4b-it	**0.515 (0.481, 0.546)**	**0.047 (0.040, 0.056)**	0.531 (0.504, 0.555)	0.117 (0.105, 0.132)	0.531 (0.511, 0.552)	0.142 (0.130, 0.159)	0.496 (0.477, 0.516)	0.169 (0.158, 0.183)
BioMistral-7B	0.466 (0.432, 0.498)	0.041 (0.036, 0.048)	0.484 (0.461, 0.506)	0.098 (0.089, 0.108)	**0.554 (0.532, 0.576)**	**0.146 (0.134, 0.159)**	0.473 (0.453, 0.493)	0.159 (0.148, 0.171)
Llama3-OpenBioLLM-8B	0.498 (0.463, 0.528)	0.043 (0.037, 0.050)	0.455 (0.430, 0.479)	0.107 (0.094, 0.127)	0.495 (0.472, 0.517)	0.127 (0.116, 0.142)	0.483 (0.465, 0.501)	0.164 (0.151, 0.178)

E.2 Latent Representation Structure

Figure S4 visualizes the relationship between history chunk embeddings and learned prototypes across the four tasks. Overall, the embedding spaces exhibit partially separated care-stage structure, while the prototypes follow the geometry of the learned manifolds: they are concentrated in high-density regions, extend across distinct components, and rarely appear as isolated points. Their allocation is non-uniform and differs across tasks, ranging from broad coverage of the embedding space to greater concentration within particular regions. Prototypes also occur in both care-stage-dominant and mixed-stage regions, indicating that their organization is not reducible to care-stage labels alone. These projections provide qualitative evidence that the prototypes act as distributed reference points within the representation space, although UMAP geometry and prototype locations should not be interpreted as establishing the clinical meaning of individual prototypes.

E.3 Qualitative Analysis of Prototype-Guided Retrieval

This section provides a qualitative analysis of the prototype-guided retrieval mechanism in `EHR-RAGp`. To illustrate the behavior of the model, we present two representative case studies from the in-hospital mortality task, corresponding to positive and negative prediction examples. For each case, we analyze prototype usage, query-history alignment, chunk-level relevance weighting, and the contribution of visits and care stages to the final prediction. These examples provide insight into how `EHR-RAGp` organizes longitudinal patient history and integrates clinically relevant context during inference.

E.3.1 Case of Positive Example

Figure S5 presents a representative positive example from the in-hospital mortality task, corresponding to a correctly predicted positive case (`label=1`, `pred=1`) with high confidence (`prob=0.901`) for a query segment originating from Visit V10 in the patient timeline.

E.3.2 Case of Negative Example

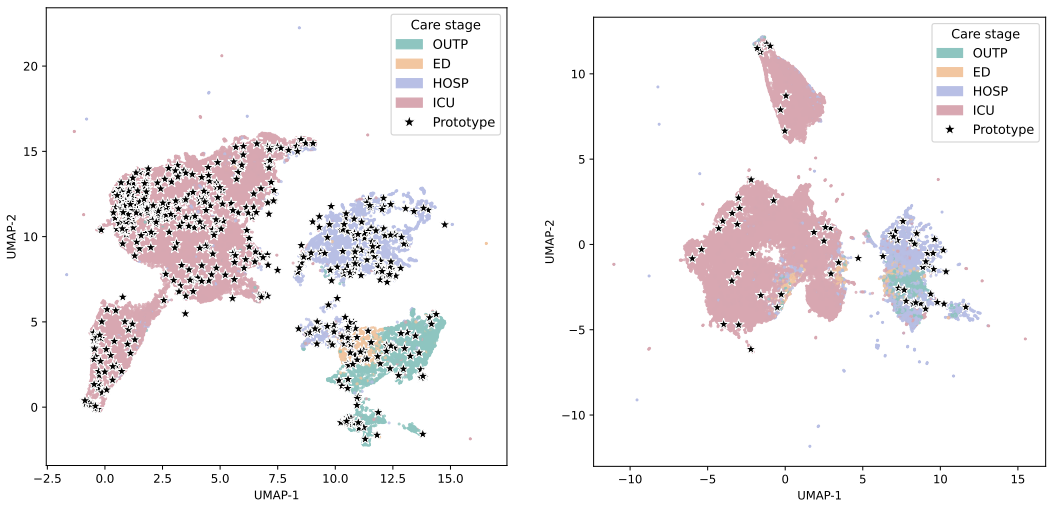
Figure S6 presents a representative negative example from the in-hospital mortality task, corresponding to a correctly predicted negative case (`label=0`, `pred=0`) with low predicted probability (`prob=0.000`) for a query segment originating from Visit V14 in the patient timeline.

E.4 Analysis of Prototype Regularization

This section analyzes the effect of prototype regularization on the behavior of `EHR-RAGp`. Specifically, we compare the model with and without regularization from two perspectives: (1) global prototype utilization patterns, illustrating how regularization affects the distribution and diversity of prototype assignments, and (2) qualitative case studies, showing how regularization influences prototype-guided retrieval and chunk weighting for individual patient trajectories. Together, these analyses highlight the role of regularization in preventing prototype collapse and encouraging more balanced use of the latent prototype space.

E.4.1 Global Prototypes Usage

Figure S7 illustrates the effect of prototype regularization on global prototype utilization. Without regularization as in Figure S7a, query assignments exhibit highly imbalanced usage patterns, where



(a) 1-Year Mortality Post-Discharge

(b) ICU Readmission within 30 Days

(c) In-Hospital Mortality

(d) Long LOS ($\geq$ 7 Days)

Figure S4: UMAP visualizations of history-chunk embeddings and learned prototypes across four clinical prediction tasks: (a) 1-year mortality post-discharge, (b) ICU readmission within 30 days, (c) in-hospital mortality, and (d) long length of stay ($\geq$ 7 days).

a small subset of prototypes receive disproportionately large assignment probabilities while many prototypes remain minimally utilized. A similar, although less severe, imbalance is also observed for history chunk assignments. This behavior indicates partial prototype collapse, where the model concentrates representation learning around a limited number of latent prototypes. In turn, such a situation makes agreement between history and query rather difficult due to the sharper distribution of the query as compared to the history assignments. In contrast, enabling regularization, as in Figure S7b, produces a substantially more uniform assignment distribution across prototypes for both queries and history chunks. The resulting increase in prototype diversity suggests that regularization encourages broader utilization of the latent prototype space, leading to more balanced representation learning and improved retrieval behavior.

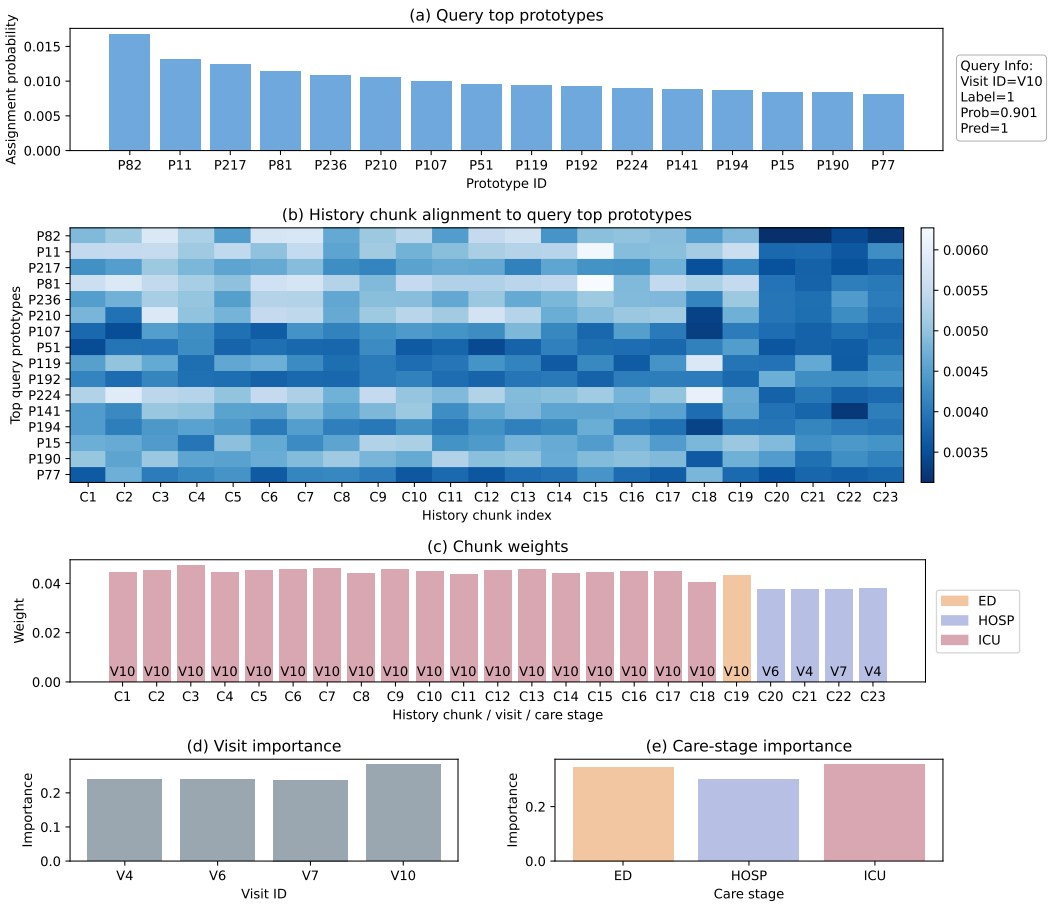


Figure S5: Qualitative analysis of prototype-guided retrieval for an in-hospital mortality positive example. (a) Top prototype assignments for the query segment. (b) Alignment between retrieved history chunks and the query prototypes. (c) Prototype-guided relevance weights assigned to retrieved chunks. (d) Aggregated contribution of retrieved visits. (e) Aggregated importance across care stages.

E.4.2 Qualitative Analysis of Prototypes Usage

To further understand the role of usage regularization λ_u, we examine the model behavior with regularization, as in in Figure S8, and without regularization, as in Figure S9. In the settings where usage regularization is activated, prototype assignments are distributed across multiple latent prototypes, resulting in relatively balanced alignment scores and chunk weights across retrieved history segments. This behavior indicates that the model integrates information from several relevant historical chunks rather than relying on a single dominant prototype or retrieval path. On the other hand, without usage regularization, the query representation collapses strongly toward a single prototype (P0), which leads to highly concentrated retrieval behavior where few retrieved chunk receives the majority of the final weight while the remaining chunks contribute minimally. With respect to the model performance, it can be observed that despite the collapse the model can still provide correct prediction but less confident. Overall, these examples qualitatively demonstrate how prototype regularization stabilizes retrieval dynamics, encourages broader utilization of patient history, and prevents highly concentrated dependence on a narrow subset of latent prototypes, which in turn deteriorates the retrieval quality.

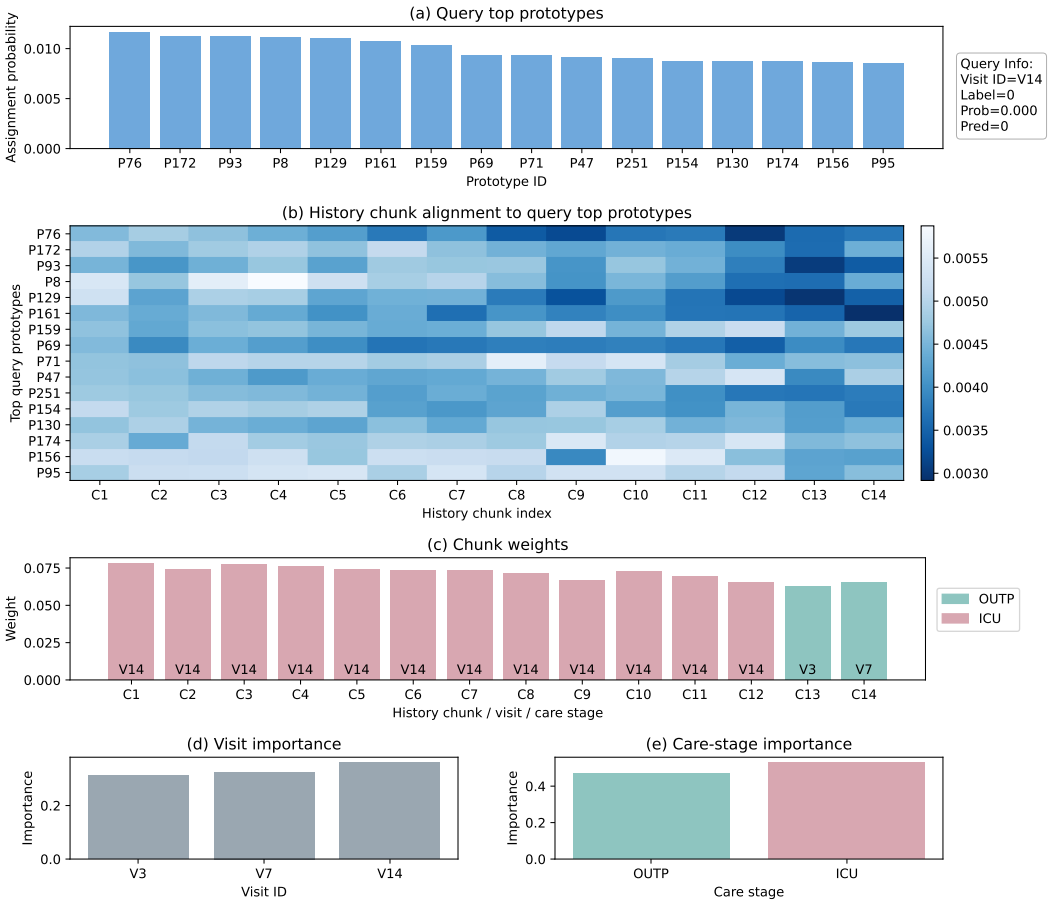
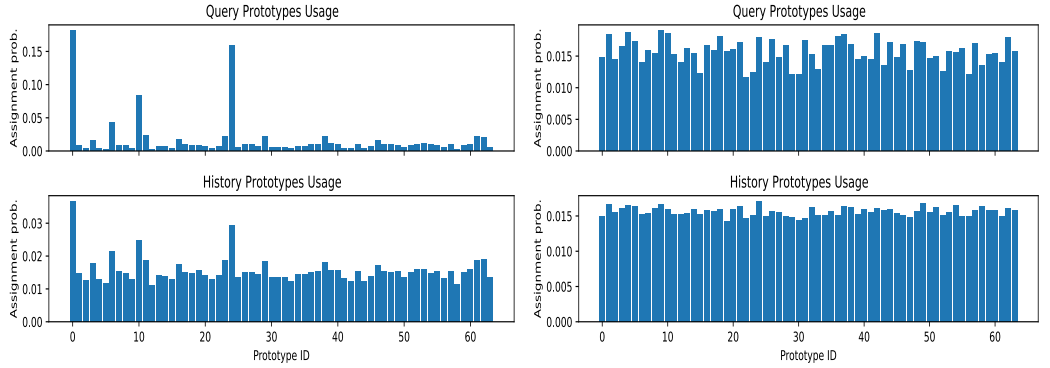


Figure S6: Qualitative analysis of prototype-guided retrieval for an in-hospital mortality negative example. (a) Top prototype assignments for the query segment. (b) Alignment between retrieved history chunks and the query prototypes. (c) Prototype-guided relevance weights assigned to retrieved chunks. (d) Aggregated contribution of retrieved visits. (e) Aggregated importance across care stages.



(a) (b)

Figure S7: Global prototype utilization patterns. (a) Without Regularization. (b) With regularization.

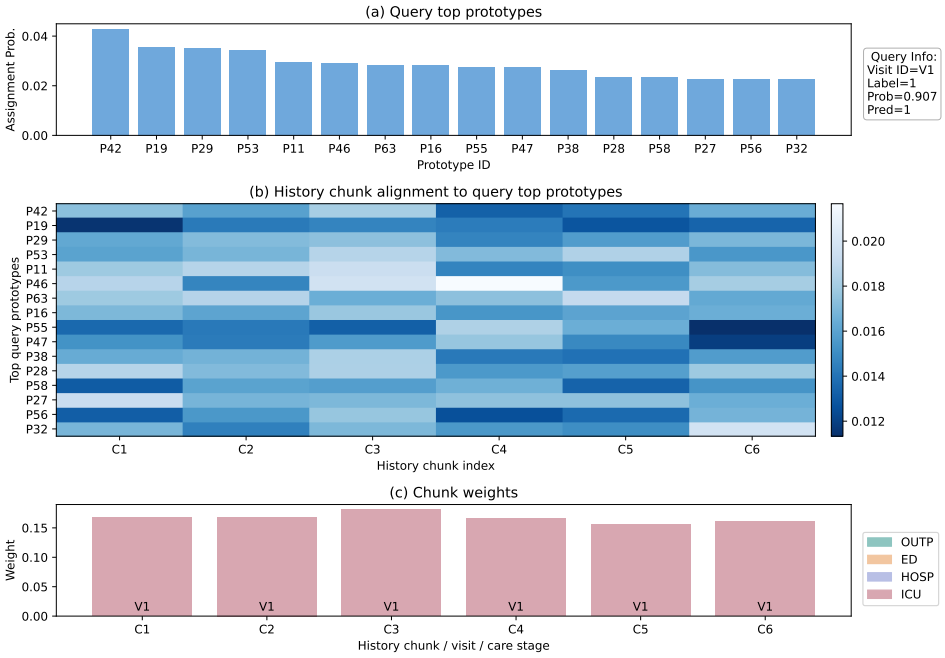


Figure S8: Qualitative analysis of sample prediction for long length of stay (7 days) task with regularization (a) Top prototype assignments for the query segment. (b) Alignment between retrieved history chunks and the query prototypes. (c) Prototype-guided relevance weights assigned to retrieved chunks.

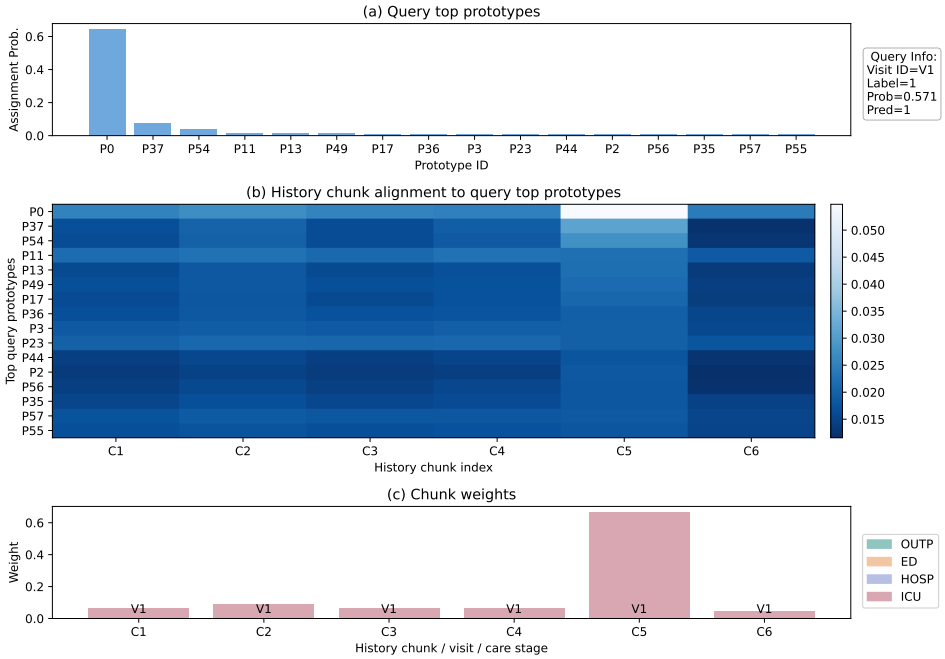


Figure S9: Qualitative analysis of sample prediction for long length of stay (7 days) task without regularization (a) Top prototype assignments for the query segment. (b) Alignment between retrieved history chunks and the query prototypes. (c) Prototype-guided relevance weights assigned to retrieved chunks.

E.5 EHRSHOT BENCHMARK RESULTS

Comparisons in Table S6 are descriptive and based on the reported AUROC point estimates. Although all methods are evaluated using the predefined EHRSHOT splits, we do not claim that differences between `EHR-RAGp` and the long-context baselines are statistically significant, as paired predictions for these baselines are unavailable. The confidence intervals and significance indicators reported for the long-context models reflect their original comparisons with CLMBR-*t*-base and are included only for completeness.

E.6 ADDITIONAL RESULTS ON MIMIC-IV DATASET

Table S7 reports the bootstrapped confidence intervals computed for all baselines models reported in Table 1, whereas Table S8 reports the confidence interval for all models reported in Table 3. Table S9 complements the results presented in table 6 and report the confidence intervals for each reported result. Table S10 report the p-values computed for all baselines used in our experiments.

E.7 ADDITIONAL ABLATIONS ON MIMIC-IV DATASET

Table S11 complements the results provided in Table 4 by comparing prototype effect for ICU-Readmit 30d and In-hospital Mortality tasks. Table S12 presents ablation with number of prototypes, while Table S13 shows the effect of softmax temperature on the performance.

Table S6: AUROC differences relative to CLMBR-*t*-base on operational-outcome and new-diagnosis assignment tasks. For each long-context baseline, we report the maximum context length evaluated by Wornow et al. (Wornow et al., 2024). The best and second-best results for each task are shown in **bold** and underlined, respectively. A checkmark indicates that the reported 95% confidence interval excludes zero.

Context	Model	Task	Δ over CLMBR-*t*-base	95% CI	Sig.
4096	LLaMA	ICU Admission	-0.003	(-0.026, 0.021)	
		Long LOS	-0.004	(-0.018, 0.010)	
		30-day Readmission	**0.013**	(0.002, 0.026)	✓
		Acute MI	0.015	(-0.008, 0.038)	
		Celiac	0.226	(0.097, 0.365)	✓
		Hyperlipidemia	0.016	(-0.002, 0.036)	
		Hypertension	0.004	(-0.013, 0.021)	
		Lupus	-0.023	(-0.097, 0.049)	
		Pancreatic Cancer	-0.008	(-0.056, 0.033)	
4096	GPT	ICU Admission	0.011	(-0.021, 0.044)	
		Long LOS	-0.001	(-0.014, 0.014)	
		30-day Readmission	0.004	(-0.009, 0.015)	
		Acute MI	0.006	(-0.022, 0.033)	
		Celiac	0.149	(0.041, 0.278)	✓
		Hyperlipidemia	0.012	(-0.018, 0.043)	
		Hypertension	0.004	(-0.015, 0.024)	
		Lupus	-0.008	(-0.095, 0.088)	
		Pancreatic Cancer	0.027	(-0.008, 0.062)	
16384	Hyena	ICU Admission	-0.110	(-0.147, -0.075)	✓
		Long LOS	-0.048	(-0.068, -0.029)	✓
		30-day Readmission	-0.048	(-0.067, -0.026)	✓
		Acute MI	-0.100	(-0.145, -0.053)	✓
		Celiac	0.176	(0.029, 0.318)	✓
		Hyperlipidemia	-0.016	(-0.069, 0.034)	
		Hypertension	-0.071	(-0.125, -0.023)	✓
		Lupus	-0.145	(-0.268, -0.017)	✓
		Pancreatic Cancer	-0.073	(-0.148, 0.006)	
16384	Mamba	ICU Admission	0.007	(-0.028, 0.040)	
		Long LOS	0.013	(-0.005, 0.029)	
		30-day Readmission	0.005	(-0.008, 0.017)	
		Acute MI	0.016	(-0.005, 0.036)	
		Celiac	0.194	(0.108, 0.333)	✓
		Hyperlipidemia	0.023	(-0.013, 0.058)	
		Hypertension	0.003	(-0.018, 0.023)	
		Lupus	0.037	(-0.056, 0.132)	
		Pancreatic Cancer	0.053	(0.024, 0.087)	✓
$512 + (24 \times 256)$	EHR-RAGp	ICU Admission	**0.033**	–	–
		Long LOS	**0.031**	–	–
		30-day Readmission	0.006	–	–
		Acute MI	**0.032**	–	–
		Celiac	**0.228**	–	–
		Hyperlipidemia	**0.046**	–	–
		Hypertension	**0.031**	–	–
		Lupus	**0.116**	–	–
		Pancreatic Cancer	**0.060**	–	–

Table S7: Performance comparison between EHR−RAGp and all baselines. Best results are marked in **bold**. Second best results are underlined.

Model	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)
DescEmb I	0.661 (0.630, 0.690)	0.078 (0.064, 0.095)	0.904 (0.890, 0.915)	0.618 (0.578, 0.656)	0.822 (0.806, 0.836)	0.405 (0.369, 0.440)	0.754 (0.738, 0.770)	0.376 (0.347, 0.406)
DescEmb II	0.606 (0.573, 0.638)	0.071 (0.058, 0.091)	0.920 (0.909, 0.931)	0.648 (0.607, 0.686)	0.838 (0.824, 0.852)	0.441 (0.405, 0.480)	0.805 (0.791, 0.820)	0.471 (0.440, 0.504)
GenHPF	0.697 (0.667, 0.729)	0.128 (0.099, 0.169)	0.915 (0.904, 0.926)	0.641 (0.604, 0.681)	0.843 (0.828, 0.856)	0.456 (0.419, 0.494)	0.781 (0.767, 0.795)	0.417 (0.388, 0.448)
REMed	0.561 (0.528, 0.595)	0.060 (0.049, 0.079)	0.865 (0.851, 0.880)	0.501 (0.462, 0.542)	0.786 (0.770, 0.803)	0.348 (0.317, 0.383)	0.680 (0.662, 0.697)	0.280 (0.257, 0.303)
Med-BERT	0.664 (0.633, 0.695)	0.081 (0.067, 0.105)	0.893 (0.879, 0.907)	0.579 (0.540, 0.621)	0.813 (0.798, 0.829)	0.412 (0.378, 0.446)	0.729 (0.714, 0.745)	0.338 (0.312, 0.366)
CEHR-BERT	0.604 (0.573, 0.637)	0.061 (0.051, 0.075)	0.929 (0.919, 0.938)	0.676 (0.641, 0.714)	0.840 (0.825, 0.854)	0.435 (0.396, 0.473)	0.806 (0.791, 0.820)	0.473 (0.440, 0.506)
BEHRT	0.668 (0.638, 0.698)	0.077 (0.065, 0.094)	0.902 (0.889, 0.913)	0.609 (0.569, 0.650)	0.824 (0.809, 0.840)	0.444 (0.408, 0.481)	0.767 (0.751, 0.782)	0.382 (0.353, 0.413)
Hi-BEHRT	0.620 (0.587, 0.652)	0.069 (0.058, 0.087)	0.886 (0.873, 0.899)	0.542 (0.501, 0.585)	0.798 (0.780, 0.816)	0.392 (0.356, 0.430)	0.703 (0.687, 0.721)	0.298 (0.275, 0.325)
EHRMamba	0.665 (0.633, 0.695)	0.082 (0.067, 0.100)	0.929 (0.918, 0.938)	0.650 (0.611, 0.690)	0.829 (0.813, 0.844)	0.424 (0.385, 0.464)	0.796 (0.782, 0.809)	0.429 (0.400, 0.462)
ModernBERT	0.604 (0.571, 0.635)	0.068 (0.056, 0.087)	0.920 (0.910, 0.930)	0.640 (0.604, 0.679)	0.827 (0.811, 0.842)	0.440 (0.404, 0.475)	0.765 (0.750, 0.780)	0.381 (0.354, 0.414)
LongFormer	0.645 (0.613, 0.679)	0.095 (0.074, 0.129)	0.910 (0.900, 0.921)	0.577 (0.536, 0.620)	0.821 (0.806, 0.837)	0.423 (0.385, 0.463)	0.760 (0.744, 0.775)	0.377 (0.349, 0.409)
BigBird	0.719 (0.691, 0.746)	0.118 (0.094, 0.152)	0.913 (0.903, 0.924)	0.643 (0.606, 0.680)	0.830 (0.813, 0.844)	0.444 (0.407, 0.481)	0.787 (0.773, 0.803)	0.424 (0.393, 0.455)
RoFormer	0.699 (0.670, 0.728)	0.114 (0.088, 0.149)	0.927 (0.917, 0.937)	0.672 (0.637, 0.706)	0.846 (0.830, 0.859)	0.461 (0.424, 0.502)	0.806 (0.791, 0.820)	0.460 (0.427, 0.495)
RoBERTa	0.721 (0.690, 0.750)	0.125 (0.099, 0.162)	0.925 (0.916, 0.934)	0.671 (0.637, 0.705)	0.841 (0.827, 0.856)	0.465 (0.430, 0.506)	0.805 (0.791, 0.818)	0.466 (0.436, 0.498)
EHR−RAGp (Ours)	**0.750 (0.723, 0.774)**	**0.141 (0.115, 0.184)**	**0.950 (0.941, 0.956)**	**0.736 (0.701, 0.771)**	**0.867 (0.853, 0.878)**	**0.509 (0.472, 0.545)**	**0.853 (0.841, 0.864)**	**0.564 (0.531, 0.595)**

Table S8: Performance gains of existing EHR baseline with EHR−RAGp.

Model	EHR−RAGp	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
		AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)
Med-BERT	×	0.664 (0.633, 0.695)	0.081 (0.067, 0.105)	0.893 (0.879, 0.907)	0.579 (0.540, 0.621)	0.813 (0.798, 0.829)	0.412 (0.378, 0.446)	0.729 (0.714, 0.745)	0.338 (0.312, 0.366)
	√	**0.715 (0.687, 0.742)**	**0.117 (0.093, 0.152)**	**0.938 (0.929, 0.947)**	**0.702 (0.663, 0.736)**	**0.843 (0.829, 0.858)**	**0.455 (0.421, 0.494)**	**0.823 (0.810, 0.836)**	**0.476 (0.445, 0.510)**
CEHR-BERT	×	0.604 (0.573, 0.637)	0.061 (0.051, 0.075)	0.929 (0.919, 0.938)	0.676 (0.641, 0.714)	0.840 (0.825, 0.854)	0.435 (0.396, 0.473)	0.806 (0.791, 0.820)	0.473 (0.440, 0.506)
	√	**0.698 (0.669, 0.727)**	**0.102 (0.082, 0.130)**	**0.941 (0.932, 0.949)**	**0.715 (0.676, 0.751)**	**0.848 (0.835, 0.862)**	**0.452 (0.417, 0.491)**	**0.821 (0.807, 0.834)**	**0.495 (0.465, 0.527)**
BEHRT	×	0.668 (0.638, 0.698)	0.077 (0.065, 0.094)	0.902 (0.889, 0.913)	0.609 (0.569, 0.650)	0.824 (0.809, 0.840)	0.444 (0.408, 0.481)	0.767 (0.751, 0.782)	0.382 (0.353, 0.413)
	√	**0.723 (0.692, 0.751)**	**0.126 (0.098, 0.161)**	**0.941 (0.933, 0.949)**	**0.711 (0.677, 0.748)**	**0.849 (0.836, 0.863)**	**0.465 (0.427, 0.503)**	**0.826 (0.813, 0.838)**	**0.490 (0.458, 0.524)**
Hi-BEHRT	×	0.620 (0.587, 0.652)	0.069 (0.058, 0.087)	0.886 (0.873, 0.899)	0.542 (0.501, 0.585)	0.798 (0.780, 0.816)	0.392 (0.356, 0.430)	0.703 (0.687, 0.721)	0.298 (0.275, 0.325)
	√	**0.684 (0.652, 0.716)**	**0.100 (0.080, 0.128)**	**0.934 (0.924, 0.942)**	**0.691 (0.655, 0.728)**	**0.835 (0.821, 0.848)**	**0.443 (0.408, 0.478)**	**0.786 (0.772, 0.800)**	**0.401 (0.37, 0.435)**

Table S9: Impact of choice of chunking strategy on model performance across all tasks.

Model	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)	AUROC (CI)	AUPRC (CI)
Event-Based	**0.750 (0.723, 0.774)**	**0.141 (0.115, 0.184)**	**0.950 (0.941, 0.956)**	0.736 (0.701, 0.771)	**0.867 (0.854, 0.878)**	0.506 (0.470, 0.542)	0.851 (0.838, 0.864)	0.553 (0.521, 0.588)
Time-Based	0.741 (0.712, 0.767)	0.129 (0.103, 0.163)	0.946 (0.938, 0.954)	0.738 (0.701, 0.766)	0.862 (0.848, 0.875)	0.507 (0.469, 0.541)	0.847 (0.834, 0.859)	0.538 (0.506, 0.573)
Visit-Level	0.748 (0.720, 0.778)	0.140 (0.110, 0.181)	0.948 (0.940, 0.955)	0.738 (0.704, 0.772)	**0.867 (0.853, 0.878)**	**0.509 (0.472, 0.545)**	0.852 (0.839, 0.864)	0.553 (0.522, 0.583)
Care-Stage	0.747 (0.719, 0.774)	0.138 (0.111, 0.177)	0.947 (0.938, 0.955)	**0.739 (0.704, 0.771)**	0.865 (0.851, 0.878)	0.506 (0.470, 0.544)	**0.853 (0.841, 0.864)**	**0.564 (0.531, 0.595)**

Table S10: Paired-bootstrap significance of the performance differences between EHR−RAGp and each baseline.

Model	ICU-Readmit 30d		In-hospital Mortality		Long LOS 7d		1YR Mortality	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
DescEmb I	< 0.0001	0.0014	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
DescEmb II	< 0.0001	0.0002	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
GenHPF	0.0004	0.2395	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
REMed	< 0.0001	0.0004	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Med-BERT	< 0.0001	0.0006	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
CEHR-BERT	< 0.0001	0.0006	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
BEHRT	< 0.0001	0.0011	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Hi-BEHRT	< 0.0001	0.0018	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
EHRMamba	< 0.0001	0.0016	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
ModernBERT	< 0.0001	0.0008	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Longformer	< 0.0001	0.0009	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
BigBird	0.0098	0.0355	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
RoFormer	< 0.0001	0.0163	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
RoBERTa	0.0013	0.0246	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table S11: Ablation of retrieval mechanisms. †denotes p< 0.05 compared to EHR−RAGp.

Configuration	ICU Readmission 30d		1-Year Mortality	
	AUROC	AUPRC	AUROC	AUPRC
w/o Retrieval	$0.721^{\dagger}$	$0.125^{\dagger}$	$0.925^{\dagger}$	$0.671^{\dagger}$
w/o Prototypes	$0.735^{\dagger}$	0.134	$0.941^{\dagger}$	$0.713^{\dagger}$
w/ Prototypes	0.750	0.141	0.950	0.736

Table S12: Ablation of the number of prototypes.

# Prototypes	ICU Readmission 30d		1-Year Mortality	
	AUROC	AUPRC	AUROC	AUPRC
64	0.744	**0.139**	0.850	0.551
128	**0.745**	0.137	0.844	0.536
256	0.742	0.137	0.845	0.536
512	0.727	0.129	**0.851**	**0.555**

Table S13: Effect of the softmax temperature on model performance.

T_s	1-Year Mortality		In-Hospital Mortality	
	AUROC	AUPRC	AUROC	AUPRC
0.10	0.845	0.540	0.947	0.728
0.15	0.852	0.552	0.948	0.730
0.20	**0.853**	**0.564**	0.948	0.732
0.25	0.851	0.553	**0.950**	**0.736**