Risk Horizons: Structured Hypothesis Spaces for Longitudinal Clinical Prediction

Zhan Qu
TU Dresden and ScaDS.AI
Dresden, Germany
zhan.qu@tu-dresden.de

Michael Färber
TU Dresden and ScaDS.AI
Dresden, Germany
michael.faerber@tu-dresden.de

Abstract

Predicting future clinical events from longitudinal electronic health records (EHRs) requires selecting plausible outcomes from a large and structured event space under sparse observations. While clinical coding systems provide hierarchical organization of events, cross-modal and temporal relationships are not explicitly specified and must instead be inferred from data, making prediction difficult for weakly observed longitudinal transitions. We introduce **Risk Horizons**, a geometry-aware framework for constructing patient-specific candidate spaces for multi-modal next-visit prediction. Risk Horizons combines deterministic coding hierarchies with data-driven lagged cross-modal associations, embeds the resulting clinical graph in hyperbolic space, and retrieves candidate futures using directional risk cones. This reframes longitudinal prediction as ranking within a compact, clinically coherent hypothesis space rather than scoring an unconstrained vocabulary. Experiments on MIMIC-IV and eICU demonstrate competitive next-visit prediction performance, with consistently improved hierarchy consistency and candidate-space coverage across clinical modalities. Further analysis suggests that hyperbolic structured candidate retrieval is the primary driver of performance, while LLMs are effective as constrained inference-time rerankers operating over clinically grounded candidate sets.

1 Introduction

Longitudinal electronic health records (EHRs) are central to clinical decision support, enabling models of disease progression, treatment response, and adverse events over time [36, 16, 9]. A standard formulation is next-visit prediction, where the goal is to forecast a patient's future diagnoses, procedures, and medications from prior visits. Existing approaches commonly represent patient histories as temporal sequences and apply recurrent or transformer-based architectures to predict future clinical events [10, 21]. More recently, large language models (LLMs) have been explored for clinical reasoning due to their strong semantic representations and few-shot generalization capabilities [40, 30]. However, longitudinal EHR prediction remains difficult because the space of plausible future events is large, sparse, and strongly structured.

A key difficulty is that EHR trajectories are sparse, irregularly sampled, and multi-modal [22, 6]. Within a single admission, information is recorded as multiple unordered sets of entities drawn from different coding systems [19], such as diagnoses in ICD and medications in NDC/ATC, rather than as a coherent sequence of clinically salient states. This fragmentation makes it unclear which elements within a visit best summarize the patient's condition, and it leaves many clinically meaningful relationships implicit, such as which diagnoses motivate which treatments.

Clinical knowledge resources provide partial structure for this problem, but not the full hypothesis space required for longitudinal prediction. External biomedical knowledge bases such as UMLS

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

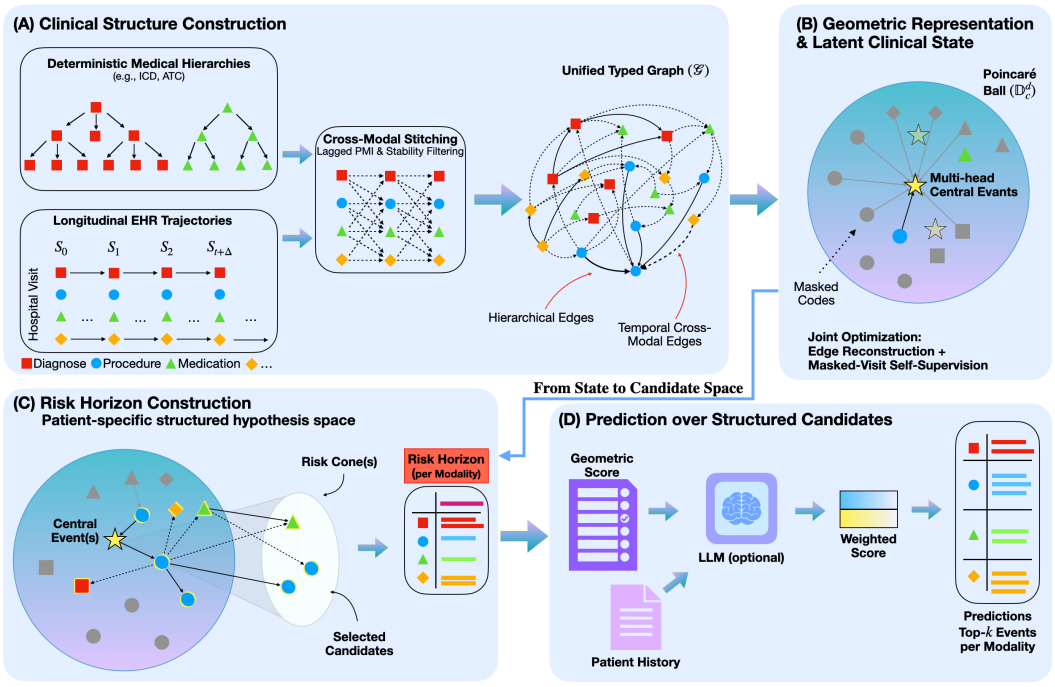


Figure 1: **Overview of Risk Horizons.** (A) Clinical structure is constructed by combining coding hierarchies with data-driven cross-modal associations. (B) Concepts are embedded in hyperbolic space, and visits are summarized as multi-head Central Events. (C) Structure-aware retrieval constructs a patient-specific *Risk Horizon*, a set of plausible future events. (D) Prediction ranks candidates within the Risk Horizon using geometric scores, optionally refined by an LLM.

[3] provide broad relations and hierarchies, but can be difficult to operationalize at scale due to heterogeneous vocabularies and weakly specified relations. In contrast, coding systems such as ICD-10 [44] and the Anatomical Therapeutic Chemical classification [45] provide reliable hierarchical organization of clinical events. However, these hierarchies do not explicitly encode cross-modal or temporal relationships among diagnoses, procedures, and medications. Such dependencies must therefore be inferred from longitudinal data, where they are often sparse or weakly observed.

This motivates a different view of next-visit prediction. Rather than treating the task only as reasoning over patient history, we view it as constructing a patient-specific hypothesis space of plausible futures. In clinical practice, decision-making does not proceed by scoring over all possible diagnoses or treatments. Instead, clinicians operate within a context-dependent space of plausible alternatives constrained by prior knowledge, hierarchical relationships, and disease progression. This suggests that models should first recover clinically coherent candidate futures and only then rank predictions within this space. This is particularly important for sparse and weakly supported transitions, where full-vocabulary scoring or unstructured retrieval often yields diffuse or clinically inconsistent predictions.

We introduce **Risk Horizons**, a geometry-aware framework for constructing patient-specific structured candidate spaces for longitudinal EHR prediction. As shown in Figure 1, Risk Horizons builds a unified clinical graph by combining deterministic coding-system hierarchies with data-driven lagged cross-modal associations. The graph is embedded in hyperbolic space, whose geometry is well suited to hierarchical abstraction and asymmetric relationships [29, 13]. Patient visits are summarized as probabilistic *Central Events*, which denoise sparse multi-modal observations into latent clinical states. At inference time, directional geometric *risk cones* retrieve compact sets of clinically plausible future events aligned with hierarchy and inferred temporal progression.

This formulation turns next-visit prediction into a two-stage procedure: first construct a structured Risk Horizon, then rank candidates within it. LLMs are used only as optional inference-time rerankers over retrieved candidates, not as unconstrained generators. This design preserves the semantic strengths of LLMs while restricting their output space to clinically plausible candidates,

reducing the likelihood of clinically implausible generations [17, 53, 34]. Our experiments suggest that geometry-aware candidate construction contributes substantially to performance, while LLM reranking provides additional gains when constrained to the Risk Horizon.

Our contributions are: (1) we reframe longitudinal EHR prediction as structured hypothesis space construction rather than only sequence modeling or unconstrained reasoning; **(2)** we introduce Risk Horizons, a geometry-aware framework that combines deterministic coding hierarchies with data-driven lagged cross-modal associations; **(3)** we propose directional retrieval via geometric risk cones to construct compact, patient-specific candidate spaces; and **(4)** we provide extensive evaluation on MIMIC-IV and eICU, demonstrating strong next-visit prediction performance and improved hierarchy consistency under partially observed longitudinal structure. Our code is publicly available at: https://github.com/ZhanQu945/Risk-Horizons.

2 Related Work

Deep Sequential Models for Longitudinal EHRs. Longitudinal EHRs are commonly modeled as temporal sequences. Early recurrent models demonstrated the feasibility of next-visit prediction from diagnosis codes [9, 32], while RETAIN introduced reverse-time attention for interpretability [10] and later work incorporated uncertainty-aware and survival objectives [1]. Transformer-based models treat medical codes as tokens to capture long-range dependencies [21, 37], with recent extensions exploring long-context, encoder–decoder, and foundation-model pretraining for irregular clinical trajectories [46, 51, 41]. Autoregressive models such as HALO further model high-dimensional EHR sequences [42]. While effective at modeling temporal patterns, these approaches largely treat prediction as sequence modeling over flattened event sets, without explicitly constructing structured candidate spaces for plausible future events.

Ontology-Aware and Graph-Based Clinical Representation Learning. Ontology-aware methods incorporate medical hierarchies such as ICD and SNOMED-CT. GRAM propagates information from ancestor concepts via attention [11], while KAME integrates external knowledge graphs into memory networks [26]. Graph-based approaches construct patient–code or code–code graphs and apply GNNs for clinical representation learning [39, 27], with recent work extending this direction to temporal heterogeneous graphs using time-aware message passing or graph transformers [50, 7]. These methods embed relational and ontological structure into learned representations, but primarily use this structure for representation learning or downstream prediction rather than explicitly constructing constrained candidate spaces for future events.

Hyperbolic Geometry for Medical Representation Learning. Hyperbolic geometry provides low-distortion embeddings for hierarchical and tree-structured data due to exponential volume growth [38, 29]. Hyperbolic entailment cones and hyperbolic GNNs extend these ideas to non-Euclidean neural models [12, 13, 5, 23]. In medicine, hyperbolic embeddings have been used to capture ICD hierarchies and improve mortality, readmission, diagnosis prediction, and ICD coding tasks [2, 25, 24, 28, 4, 48]. Prior work primarily uses hyperbolic geometry for representation learning or classification, and does not exploit its geometric structure to construct structured hypothesis spaces for downstream prediction.

RAG and LLM-Based Clinical Reasoning. LLMs show strong performance on medical reasoning benchmarks [40, 30], but hallucination remains a concern in clinical settings [17]. Retrieval-augmented generation grounds predictions in retrieved evidence [20, 15], and clinical RAG studies show that retrieval strategy and corpus choice affect performance [49]. Beyond text retrieval, knowledge graph–augmented prompting retrieves entities, relations, or subgraphs [31, 52], and graph-based retrieval has been integrated into medical RAG pipelines [47]. Recent systems for longitudinal EHR reasoning, such as TRACE, maintain structured patient states over time [35]. Most retrieval mechanisms, however, rely on lexical similarity, nearest-neighbor retrieval, or linearized structured evidence, limiting their ability to model hierarchical abstraction and directional clinical progression.

3 Methodology

We present **Risk Horizons**, a geometry-aware framework for constructing structured candidate spaces for longitudinal EHR prediction. The central challenge in next-visit prediction is that the space of possible future clinical events is only partially structured: while coding systems define hierarchical

relationships, many clinically meaningful temporal and cross-modal dependencies remain implicit in longitudinal data. This yields a large and weakly constrained prediction space that is difficult to reason over directly. Risk Horizons addresses this challenge by constructing patient-specific candidate spaces that recover missing structure and constrain prediction to clinically plausible futures before final ranking. Figure 1 summarizes the framework, and Algorithm 1 in Appendix A.1 provides the end-to-end procedure.

3.1 Problem Setup and Notation

A patient trajectory is a sequence of visits indexed by time $t \in \{1, \dots, T\}$. Each visit is a set-valued, multi-modal observation:

$$S_t = \left\{ S_t^{(\mathrm{dx})},\ S_t^{(\mathrm{proc})},\ S_t^{(\mathrm{med})} \right\}, \qquad S_t^{(m)} \subset \mathcal{V}^{(m)}. \tag{1}$$

Given history $S_{1:t}$, the goal is to predict next-visit events $S_{t+1}^{(m)}$ for each modality m. We denote the unified vocabulary by

$$\mathcal{V} = \bigcup_m \mathcal{V}^{(m)} \cup \mathcal{V}^{(\mathrm{anc})}, \tag{2}$$

where $\mathcal{V}^{(\mathrm{anc})}$ contains deterministic ancestor nodes induced by coding hierarchies. We formulate prediction as ranking events within a patient-specific structured candidate space $\mathcal{R}_T \subset \mathcal{V}$, referred to as the *Risk Horizon*. The Risk Horizon is intended to capture clinically plausible future events before final modality-wise ranking.

3.2 Risk Horizons: Structured Candidate Space Construction

Unified Clinical Graph. We construct a directed typed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with two edge families: (i) deterministic hierarchical edges and (ii) data-driven temporal cross-modal edges. Edges are partitioned by relation types $r \in \mathcal{R}$:

$$\mathcal{E} = \mathcal{E}_{\mathrm{hier}} \cup \mathcal{E}_{\mathrm{cross}}, \qquad \mathcal{E} = \bigcup_{r \in \mathcal{R}} \mathcal{E}_r. \tag{3}$$

Hierarchical edges. Within each modality, we build deterministic parent–child edges from coding structure (e.g., ICD prefixes, ATC levels). We direct edges from abstract parents to specific children:

$$\mathcal{E}_{\mathrm{hier}} = \left\{ (p \to v)\ :\ p = \mathrm{parent}(v),\ v \in \mathcal{V}^{(m)} \right\}. \tag{4}$$

This directionality ensures that specialization corresponds to moving "downstream" in the hierarchy.

Temporal Association Stitching via Lagged PMI. To connect clinical concepts beyond deterministic coding hierarchies, we infer statistical associations from longitudinal EHR trajectories, including both intra-modal and cross-modal relationships. For $a \in \mathcal{V}^{(m)}$ and $b \in \mathcal{V}^{(m')}$, where m and m' may be identical or different, we define lagged co-occurrence as

$$P_\Delta(a, b) = \Pr\left(a \in S_t^{(m)},\ b \in S_{t+\Delta}^{(m')} \right), \qquad \Delta \in \{0, \dots, \Delta_{\max}\}. \tag{5}$$

Here, Δ denotes visit-order lag rather than elapsed calendar time. In particular, $\Delta = 0$ captures within-visit associations, whereas $\Delta > 0$ captures associations across visits. We compute lagged pointwise mutual information (PMI):

$$\mathrm{PMI}_\Delta(a \to b) = \log \frac{P_\Delta(a, b)}{P(a)P(b)}. \tag{6}$$

We retain an association edge $a \to b$ if

$$\mathrm{PMI}_\Delta(a \to b) > \tau_{m,m',\Delta} \quad \wedge \quad \mathrm{cnt}_\Delta(a, b) \geq \kappa, \tag{7}$$

where κ is a minimum support threshold. Each retained edge is assigned a relation type

$$r = \mathrm{type}(a \to b) = (m \to m', \Delta), \tag{8}$$

yielding $\mathcal{E}_{\mathrm{cross}} = \bigcup_r \mathcal{E}_r$. Edge direction reflects observed temporal ordering for $\Delta > 0$ and does not imply causality; associations may also reflect clinical documentation or workflow patterns.

Stability Filtering. To reduce dataset-specific artifacts, we apply bootstrap stability filtering to PMI-derived edges. Let $\mathcal{E}^{(b)}_{\text{cross}}$ denote the cross-modal edge set recovered on bootstrap resample $b \in \{1, \ldots, B\}$. We keep edge e iff

$$\text{stab}(e) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}\Big[e \in \mathcal{E}^{(b)}_{\text{cross}}\Big] \geq q, \tag{9}$$

where $q \in (0, 1]$ controls the stability requirement.

Hyperbolic Representation. We embed nodes into the d-dimensional Poincaré ball of curvature $-c$:

$$\mathbb{D}^d_c = \left\{x \in \mathbb{R}^d : \|x\| < 1/\sqrt{c}\right\}, \qquad z_v \in \mathbb{D}^d_c \ \ \forall v \in \mathcal{V}. \tag{10}$$

For any two embeddings $x, y \in \mathbb{D}^d_c$, we use the hyperbolic distance $d_{\mathbb{D}}(x, y)$ (Appendix A.2). For an observed edge $(u \to v)$ of type r, we define a compatibility score

$$s_r(u, v) = -d_{\mathbb{D}}(z_u, z_v). \tag{11}$$

Although hyperbolic distance is symmetric, directionality is incorporated through the directed and typed clinical graph, relation-specific edge reconstruction, and directional risk-cone retrieval. Thus, the compatibility score itself does not explicitly parameterize asymmetric clinical transitions.

Embeddings are trained via typed edge reconstruction with negative sampling, using the objective detailed in Appendix A.3. Optimization employs Riemannian SGD/Adam with projection to enforce $\|z_v\| < 1/\sqrt{c}$.

Latent Clinical State via Probabilistic Central Events. Let $\mathcal{C}_t = \bigcup_m S^{(m)}_t$ be the set of all codes observed at visit t. We compute a visit-level hyperbolic barycenter as a closed-form approximation of the Fréchet mean using log–exp maps at the origin:

$$\mu_t = \exp^c_0\left(\frac{1}{\sum_{c_i \in \mathcal{C}_t} w_{t,i}} \sum_{c_i \in \mathcal{C}_t} w_{t,i} \log^c_0(z_{c_i})\right), \tag{12}$$

where $w_{t,i} \geq 0$ are code weights (uniform by default).

Rather than snapping μ_t to a single nearest node, we define a soft assignment over candidate nodes:

$$p_t(v) = \frac{\exp(-\beta\, d_{\mathbb{D}}(\mu_t, z_v))}{\sum_{v' \in \mathcal{V}_t} \exp(-\beta\, d_{\mathbb{D}}(\mu_t, z_{v'}))}, \tag{13}$$

where $\mathcal{V}_t$ is a restricted candidate pool (e.g., same-modality nodes and ancestors) for efficiency. The Central Event is $\text{CE}_t = (\mu_t, p_t(\cdot))$.

Multi-head Central Events. A single visit barycenter may collapse heterogeneous clinical signals, e.g., diagnoses, procedures, and medications, into one overly coarse state that obscures distinct clinical factors. We therefore represent each visit by H head-specific Central Events. For each code $c_i \in \mathcal{C}_t$, let $\tilde{z}_i = \log^c_0(z_{c_i})$. Each head applies a learned tangent-space projection $A_h \in \mathbb{R}^{d \times d}$ and computes

$$\bar{u}^{(h)}_t = \frac{\sum_{c_i \in \mathcal{C}_t} w_{t,i} A_h \tilde{z}_i}{\sum_{c_i \in \mathcal{C}_t} w_{t,i}}, \qquad \mu^{(h)}_t = \exp^c_0(\bar{u}^{(h)}_t). \tag{14}$$

Each head induces a soft Central Event distribution

$$p^{(h)}_t(v) = \frac{\exp(-\beta d_{\mathbb{D}}(\mu^{(h)}_t, z_v))}{\sum_{v' \in \mathcal{V}_t} \exp(-\beta d_{\mathbb{D}}(\mu^{(h)}_t, z_{v'}))}. \tag{15}$$

The head projections are trained through masked-visit reconstruction, allowing different heads to specialize in distinct clinical factors. When a single visit summary is required, we use the mixture barycenter

$$\mu_t = \exp^c_0\left(\sum_{h=1}^{H} \pi_h \log^c_0(\mu^{(h)}_t)\right), \qquad \sum_{h=1}^{H} \pi_h = 1. \tag{16}$$

This reduces to the single-head formulation when $H = 1$.

Masked-Visit Reconstruction for Self-Supervised Denoising. To train Central Events to preserve predictable clinical signal while filtering noise, we randomly mask a subset of visit codes $\mathcal{M}_t \subset \mathcal{C}_t$ and compute μ_t from $\mathcal{C}_t \setminus \mathcal{M}_t$, where μ_t is the mixture barycenter of the multi-head Central Events, which is used as a unified representation for reconstruction. We then score a masked code $c \in \mathcal{M}_t$ by hyperbolic distance:

$$p(c \mid \mu_t) \propto \exp(-d_{\mathbb{D}}(\mu_t, z_c)/\tau)\,. \tag{17}$$

Training minimizes the sampled-softmax / contrastive reconstruction loss

$$\mathcal{L}_{\text{mask}} = -\sum_{c \in \mathcal{M}_t} \log p(c \mid \mu_t), \tag{18}$$

with an efficient negative-sampling normalization. The complete sampled-softmax formulation and negative sampling strategy are detailed in Appendix A.4, Eq. (30).

Structure-Aware Retrieval via Multi-Cone Risk Horizons. Given the most recent Central Event at time T, Risk Horizons performs *structure-aware retrieval* to identify a compact set of candidate next-visit events. We refer to this retrieved, modality-typed candidate set as the *Risk Horizon*. Unlike standard nearest-neighbor retrieval in embedding space, our approach explicitly incorporates hierarchical structure, temporal associations, and directional progression. The goal is to favor candidates representing plausible downstream specialization or progression rather than merely nearby semantic similarity.

For each head h, we select a representative concept $v_T^{(h)} = \arg\min_{v \in \mathcal{V}} d_{\mathbb{D}}(z_v, \mu_T^{(h)})$ and construct a structured candidate pool from a bounded local neighborhood around $v_T^{(h)}$. This neighborhood is obtained through outgoing graph traversal and incident PMI associations, including hierarchy and association edges as well as incoming and outgoing same-visit and lagged links. The candidate pool is the union of these head-specific neighborhoods, restricted to eligible prediction codes.

To favor clinically plausible downstream progression, each head defines a tangent-space risk cone for each modality m. Let r_m denote the corresponding modality root. The downstream direction is

$$d_{T,m}^{(h)} = \log_0^c(\mu_T^{(h)}) - \log_0^c(z_{r_m}). \tag{19}$$

Directional compatibility is defined by

$$a_{T,m}^{(h)}(u) = \frac{\langle \log_0^c(z_u), d_{T,m}^{(h)} \rangle}{\| \log_0^c(z_u) \| \| d_{T,m}^{(h)} \|}, \tag{20}$$

with numerical stabilization for near-zero norms. A candidate satisfies the cone condition when $a_{T,m}^{(h)}(u) > \phi_m$, where ϕ_m is a modality-specific cosine similarity threshold parameterizing the cone aperture; larger values correspond to narrower cones.

The cone constraint supports two retrieval modes: *hard filtering*, which penalizes candidates violating the cone condition for any head with positive weight using a large negative score, and *soft scoring*, which retains the structured candidate pool and rewards cone membership. In hard-filtering mode, an empty-cone fallback may waive a head's constraint when no candidates satisfy it. For soft scoring, each candidate receives a head-specific geometric score

$$s_{\text{geo}}^{(h)}(u; T) = -d_{\mathbb{D}}(z_u, \mu_T^{(h)}) + \eta_m \cdot \mathbb{I}\Big[a_{T,m}^{(h)}(u) > \phi_m\Big]\,, \tag{21}$$

where $\eta_m \geq 0$ controls the modality-specific directional preference. In hard-filtering mode, the cone-membership bonus is omitted, and candidates violating an active head's cone condition receive a large negative penalty rather than being explicitly removed. The final score aggregates across heads using normalized weights:

$$s_{\text{geo}}(u; T) = \sum_{h=1}^{H} \pi_h \, s_{\text{geo}}^{(h)}(u; T). \tag{22}$$

The typed Risk Horizon $\mathcal{R}_T$ is formed by selecting up to the top-K eligible candidates per modality.

3.3 Prediction over Structured Candidate Spaces

Given the Risk Horizon $\mathcal{R}_T$, prediction is performed by ranking candidates within each modality. By default, we use the geometric score $s_{\text{geo}}(u;T)$ from the Risk Horizon construction. Optionally, retrieved candidates can be reranked using a constrained LLM-based scoring module. Let $\text{Desc}(\cdot)$ map codes to short text descriptions. The LLM receives (i) a compressed history from Central Events and (ii) candidates grouped by modality. For each candidate $u \in \mathcal{R}_T^{(m)}$, we obtain an LLM score $s_{\text{llm}}(u;T)$ and combine it with the geometric score:

$$s(u;T) = \lambda\, s_{\text{llm}}(u;T) + (1-\lambda)\, s_{\text{geo}}(u;T). \tag{23}$$

When no LLM reranker is used, we set $s(u;T) = s_{\text{geo}}(u;T)$. The LLM is used strictly at inference time and is not allowed to introduce new candidates; we do not backpropagate through the LLM. Final top-k predictions are obtained per modality:

$$\widehat{S}_{T+1}^{(m)} = \text{Top-}k\{s(u;T) : u \in \mathcal{R}_T^{(m)}\}. \tag{24}$$

3.4 Training Objective and Optimization

We jointly train node embeddings and Central Event denoising using

$$\mathcal{L} = \mathcal{L}_{\text{edge}} + \alpha\, \mathcal{L}_{\text{mask}}, \tag{25}$$

where $\mathcal{L}_{\text{edge}}$ is the typed edge reconstruction loss (Appendix A.3, Eq. (28)) and $\mathcal{L}_{\text{mask}}$ is the masked-visit reconstruction loss (Appendix A.4, Eq. (30)). We optimize on the Poincaré ball using Riemannian SGD/Adam with projection to maintain $z_v \in \mathbb{D}_c^d$.

4 Experiments

4.1 Experimental Setup

We evaluate Risk Horizons on MIMIC-IV [19], with additional cross-institution validation on eICU [33]. The task is **multi-modal next-visit prediction**: given patient history $S_{1:t}$, the model ranks diagnosis, procedure, and medication events in the next visit S_{t+1}.

Dataset. After filtering for trajectories with at least two visits, MIMIC-IV contains 34,775 patients, 97,943 visits, and 63,168 next-visit prediction pairs. The unified clinical graph contains 62,574 nodes and 713,717 edges spanning diagnoses, procedures, medications, and hierarchy ancestors. Visits are dense and highly multi-modal, averaging 32.81 events per visit (16.49 diagnoses, 3.23 procedures, 13.08 medications).

We use patient-level 80/10/10 train/validation/test splits, ensuring that visits from the same patient never appear across splits. All data-driven graph statistics, including lagged PMI associations, are computed exclusively from training data to prevent information leakage. Additional preprocessing details, dataset statistics, and split coverage analyses are provided in Appendix B.

Baselines. We compare Risk Horizons against three classes of methods. (i) *Sequential models*, including Copy Last Visit, RETAIN [10], Transformer [43], Med-BERT [37], and HALO [42], which model longitudinal trajectories as Euclidean token sequences. (ii) *Retrieval-based baselines*, including Euclidean Risk Horizons. Euclidean Risk Horizons uses the same graph topology and retrieval pipeline as the proposed method but replaces hyperbolic geometry with Euclidean embeddings, isolating the effect of geometry-aware retrieval. (iii) *Hyperbolic baselines*, including Hyperbolic Linear Probe and Hyperbolic NN Global, which operate on the same frozen hyperbolic representations learned by Risk Horizons but replace structured candidate-space construction with lightweight classification or global nearest-neighbor retrieval. Full baseline details are provided in Appendix C.

Evaluation. We report Recall@10 and nDCG@10, macro-averaged across visits and modalities. To evaluate the quality of retrieved hypothesis spaces, we report Candidate Recall, Risk-Horizon Recall@K, and Oracle Upper Recall@K. These metrics separate candidate-space coverage from final ranking quality and allow analysis of retrieval versus reranking behavior. We further evaluate structural consistency using hierarchy-aware metrics, including hierarchy distance and Ancestor Match@2. Additional metrics, implementation details, and extended results are provided in Appendices D–G.

4.2 Main Results

Table 1 reports multi-modal next-visit prediction results on MIMIC-IV. Risk Horizons achieves strong overall performance across diagnoses, procedures, and medications. Importantly, substantial performance already emerges from geometry-aware candidate-space construction before constrained reranking is introduced. Geometry-only Risk Horizons achieves performance comparable to strong sequential baselines, with 0.357 overall R@10 and 0.447 overall nDCG@10, despite relying solely on geometric retrieval and ranking within the Risk Horizon. While RETAIN and Transformer remain competitive on aggregate metrics, Risk Horizons achieves comparable performance through structured hypothesis-space construction without requiring a sequential prediction model. Adding constrained reranking further improves performance to 0.425 R@10 and 0.516 nDCG@10 overall with Llama-3.3-70B-Instruct [14], while Meditron-70B [8] achieves the strongest procedure performance (0.389 / 0.365). These results suggest that constructing a sufficiently complete and clinically coherent hypothesis space is an important component of longitudinal EHR prediction, enabling effective ranking of plausible future events without unconstrained generation over the full event vocabulary.

Table 1: Multi-modal next-visit prediction on MIMIC-IV (mean ± std over 5 seeds).

	Dx		Proc		Med		Overall	
Model	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
Copy Last	0.190 ± 0.002	0.303 ± 0.001	0.166 ± 0.030	0.152 ± 0.024	0.452 ± 0.031	0.619 ± 0.012	0.269 ± 0.029	0.358 ± 0.012
RETAIN	0.220 ± 0.013	0.423 ± 0.019	0.336 ± 0.016	0.268 ± 0.017	0.532 ± 0.030	0.723 ± 0.034	0.363 ± 0.021	0.471 ± 0.023
Transformer	0.219 ± 0.015	0.423 ± 0.018	0.327 ± 0.015	0.263 ± 0.014	0.539 ± 0.025	0.734 ± 0.030	0.362 ± 0.018	0.473 ± 0.021
Med-BERT	0.149 ± 0.013	0.293 ± 0.023	0.280 ± 0.021	0.221 ± 0.010	0.505 ± 0.012	0.694 ± 0.011	0.311 ± 0.015	0.402 ± 0.018
HALO	0.128 ± 0.056	0.251 ± 0.043	0.244 ± 0.037	0.185 ± 0.060	0.495 ± 0.074	0.683 ± 0.075	0.289 ± 0.056	0.373 ± 0.059
Euclidean RH	0.161 ± 0.013	0.337 ± 0.016	0.224 ± 0.040	0.179 ± 0.032	0.426 ± 0.051	0.572 ± 0.065	0.270 ± 0.004	0.362 ± 0.009
Hyper. Linear Probe	0.129 ± 0.030	0.249 ± 0.023	0.249 ± 0.017	0.179 ± 0.030	0.489 ± 0.034	0.678 ± 0.035	0.289 ± 0.027	0.369 ± 0.029
Hyper. NN Global	0.125 ± 0.063	0.243 ± 0.028	0.166 ± 0.021	0.125 ± 0.050	0.362 ± 0.035	0.514 ± 0.025	0.218 ± 0.040	0.294 ± 0.034
RH (w/o LLM)	0.195 ± 0.005	0.392 ± 0.012	0.298 ± 0.009	0.243 ± 0.009	0.579 ± 0.012	0.705 ± 0.014	0.357 ± 0.013	0.447 ± 0.017
RH+Llama-3.3-70B-I.	0.306 ± 0.032	0.496 ± 0.064	0.368 ± 0.005	0.360 ± 0.008	0.602 ± 0.016	0.693 ± 0.014	0.425 ± 0.018	0.516 ± 0.029
RH+Meditron-70B	0.297 ± 0.011	0.417 ± 0.033	0.389 ± 0.023	0.365 ± 0.009	0.587 ± 0.058	0.601 ± 0.027	0.424 ± 0.031	0.461 ± 0.023

Geometry-aware retrieval is the primary source of improvement. Euclidean Risk Horizons performs worse than its hyperbolic counterpart despite using the same graph topology and retrieval pipeline. For example, Euclidean Risk Horizons achieves 0.161 diagnosis R@10, compared to 0.195 for geometry-only Risk Horizons, with overall R@10 increasing from 0.270 to 0.357. This gap suggests that the gain is not due to retrieval alone, but also to geometry-aware modeling of hierarchical abstraction and cross-modal temporal structure. The additional hyperbolic baselines further isolate this effect. Despite using hyperbolic representations, both the linear probe and global nearest-neighbor retrieval remain below full Risk Horizons, particularly on diagnoses and procedures. This suggests that hyperbolic representation learning alone is insufficient; the main benefit arises from explicitly constructing patient-specific structured candidate spaces.

Table 2: Hierarchy consistency and candidate-space quality on MIMIC-IV (mean ± std over 5 seeds).

Model	**Type**	Hier. Dist ↓	Anc. Match ↑	Cand. Rec. ↑	RH@10 ↑	RH@40 ↑	Oracle@10 ↑
Hyper. NN Global	Dx	1.900 ± 0.021	0.797 ± 0.006	0.356 ± 0.012	0.125 ± 0.008	0.248 ± 0.011	0.333 ± 0.014
	Proc	1.910 ± 0.026	0.759 ± 0.009	0.391 ± 0.015	0.166 ± 0.010	0.272 ± 0.013	0.391 ± 0.017
	Med	1.250 ± 0.018	1.000 ± 0.001	0.981 ± 0.004	0.362 ± 0.013	0.706 ± 0.016	0.771 ± 0.011
Risk Horizons	Dx	1.859 ± 0.019	0.797 ± 0.005	0.413 ± 0.010	0.113 ± 0.007	0.270 ± 0.009	0.376 ± 0.012
	Proc	1.508 ± 0.023	0.835 ± 0.007	0.578 ± 0.013	0.265 ± 0.011	0.450 ± 0.014	0.578 ± 0.010
	Med	1.260 ± 0.016	1.000 ± 0.001	0.994 ± 0.002	0.489 ± 0.014	0.896 ± 0.012	0.774 ± 0.009

Candidate-space analysis reveals a retrieval–ranking decomposition. Table 2 analyzes the quality of the retrieved hypothesis space. Across modalities, Risk Horizons consistently improves Candidate Recall, RH@40, and Oracle@10 over global hyperbolic nearest-neighbor retrieval. The largest improvements occur for procedures, where Candidate Recall increases from 0.391 to 0.578 and RH@40 improves from 0.272 to 0.450. Similar trends appear for medications, where RH@40 reaches 0.896. These metrics reveal an important decomposition. Candidate Recall measures whether the correct future event is contained anywhere in the retrieved Risk Horizon before final ranking, while Oracle@10 estimates the best achievable performance under perfect reranking. The Oracle@10 values indicate that many correct future events are already present within the retrieved candidate space, particularly for procedures and medications. The remaining gap between Oracle@10 and final Recall@10 highlights opportunities for improving ranking, while incomplete candidate coverage, especially for diagnoses, remains an additional limitation.

Table 3: Cross-dataset validation on eICU.

Model	R@10	nDCG@10
RETAIN	0.381 ± 0.019	0.424 ± 0.010
HALO	0.338 ± 0.012	0.373 ± 0.013
Risk Horizons	0.339 ± 0.015	0.377 ± 0.004
RH+Llama-3.3-70B-I.	**0.401** ± 0.027	**0.433** ± 0.030

Table 4: Ablation studies on MIMIC-IV.

Method	R@10	nDCG@10
RH (w/o LLM)	**0.357** ± 0.013	**0.447** ± 0.017
RH w/o stitching	0.005 ± 0.001	0.004 ± 0.001
RH w/o cones	0.180 ± 0.019	0.206 ± 0.011
RH w/o denoising	0.115 ± 0.017	0.124 ± 0.018

Constrained reranking is effective once retrieval becomes structured. This decomposition helps explain why constrained reranking improves overall performance. Geometry-only Risk Horizons already retrieves compact and clinically grounded candidate spaces, reducing the effective output space from the full medical vocabulary to a smaller set of plausible futures. Constrained reranking then refines ordering within this space, improving overall R@10 from 0.357 to 0.425 with Llama-3.3-70B-Instruct. This configuration also achieves the highest MRR (0.704; Table 7 in Appendix F), indicating that relevant events tend to be ranked earlier once prediction is restricted to structured candidate sets.

Risk Horizons improves structural consistency and generalization. Risk Horizons also improves several hierarchy-aware consistency metrics (Table 2). For procedures, hierarchy distance decreases from 1.910 to 1.508 while Ancestor Match@2 increases from 0.759 to 0.835, indicating that retrieved futures become more structurally coherent within the clinical ontology.

Table 3 further evaluates the framework on an independent EHR dataset. On eICU, Risk Horizons + Llama-3.3-70B-Instruct achieves the best overall performance (0.401 R@10, 0.433 nDCG@10), while geometry-only Risk Horizons remains competitive with several baselines, although it trails RETAIN. These findings suggest that structure-aware retrieval can provide a useful inductive bias across datasets with different coding systems and event distributions.

Ablation Studies. Table 4 shows that all major components of Risk Horizons contribute to constructing high-quality candidate spaces. Removing lagged PMI stitching causes a near-complete collapse in performance (R@10: 0.357 $\rightarrow$ 0.005), indicating that deterministic hierarchies alone are insufficient without inferred temporal and cross-modal associations. Replacing cone-constrained retrieval with distance-only retrieval substantially degrades performance (0.357 $\rightarrow$ 0.180), suggesting that directional retrieval constraints are important for capturing temporal associations encoded in the graph. Finally, removing masked-visit denoising reduces R@10 from 0.357 to 0.115, showing that denoised Central Events improve robustness under sparse and noisy trajectories.

4.3 Qualitative Analysis and Failure Modes

Figure 2 illustrates a central pattern underlying the behavior of Risk Horizons: the framework is particularly effective when future events emerge from structured longitudinal progression, but becomes less reliable for genuinely novel or externally driven events.

In the left example, most future diagnoses, procedures, and medications are strongly supported by prior clinical history. The patient trajectory reflects a coherent progression involving chemotherapy treatment, cardiovascular comorbidities, and supportive care. In this regime, Risk Horizons successfully retrieves and ranks repeated or evolutionarily related events, producing high Recall@10 across all modalities. Importantly, several non-exact predictions remain clinically plausible. For example, the prediction of gastro-esophageal reflux disease is clinically plausible given the patient's gastrointestinal medications and chemotherapy-related treatment context, indicating that the model captures meaningful cross-modal clinical associations rather than relying solely on exact code repetition.

In contrast, the right example demonstrates a substantially harder setting where the next visit is dominated by unexpected external events. Although the patient's history strongly suggests a typical cardio-renal progression, the true outcome is instead driven by traumatic injury following a fall. As a result, many ground-truth diagnoses and procedures correspond to novel events with weak historical support. In this setting, the model continues to generate clinically coherent predictions centered around renal and cardiovascular deterioration, but fails to anticipate the traumatic transition.

This distinction is consistent with the quantitative trends observed throughout the paper. Risk Horizons is strongest when future events can be recovered through structured retrieval over clinically coherent trajectories, particularly for repeated or progression-related events. Performance degrades primarily

Qualitative examples: modeling structured progression vs. unexpected events

Structured progression (high performance)	**Unexpected external event (reasonable failure)**
Clinical context. Patient undergoing chemotherapy with stable cardiovascular comorbidities.	**Clinical context.** Patient with chronic kidney disease and cancer history.
Ground-truth next visit. *Diagnoses:* Atherosclerotic heart disease, Essential hypertension, Long-term anticoagulant use, Adverse effect of chemotherapy, Functional decline (DNR status). *Procedures:* Continuation of chemotherapy infusion. *Medications:* Blood-related therapies, Gastrointestinal treatments.	**Ground-truth next visit.** *Diagnoses:* Facial fracture due to fall at home, Hypovolemic shock, Acute cardiac complications, Chronic kidney disease. *Procedures:* Emergency facial repair surgery. *Medications:* Supportive therapies, Acute-care medications.
Predictions. *Diagnoses:* Atherosclerotic heart disease, Essential hypertension, Long-term anticoagulant use, Adverse effect of chemotherapy, Gastroesophageal reflux disease. *Procedures:* Chemotherapy infusion, Additional vascular access procedures. *Medications:* Blood-related therapies, Gastrointestinal treatments, Additional unrelated medications.	**Predictions.** *Diagnoses:* Acute kidney failure, Hypertensive chronic kidney disease, Obesity, Atrial fibrillation. *Procedures:* Routine vascular and monitoring procedures. *Medications:* Supportive therapies, Chronic-condition medications.
Recall@10: dx = 0.50, proc = 1.00, med = 1.00; avg = 0.83.	**Recall@10:** dx = 0.00, proc = 0.00, med = 0.25; avg = 0.083.

Figure 2: Structured progression vs. unexpected events. The model performs well when future events follow longitudinal clinical structure (left), but struggles on externally driven transitions (right). **Colors:** history-supported truth; novel truth; correct/plausible prediction; incorrect prediction.

for highly novel events weakly grounded in prior history, including trauma, acute accidents, and unexpected interventions, many of which are weakly supported or unrecoverable from longitudinal clinical structure alone. Even under failure cases, predictions typically remain structurally and clinically related to the underlying patient state, suggesting that errors arise more from intrinsic uncertainty in longitudinal forecasting than from arbitrary behavior. This pattern is consistent with the repeated-versus-novel event statistics reported in Appendix F.3, where retrieval-based performance is substantially stronger for history-supported events than for highly novel transitions.

5 Conclusion

We presented Risk Horizons, a framework for multi-modal next-visit prediction that combines hierarchical clinical structure, temporal associations, hyperbolic representation learning, and geometry-aware retrieval to construct clinically coherent candidate spaces for future events. Across MIMIC-IV and eICU, Risk Horizons achieves strong predictive accuracy and improved hierarchy consistency, while analysis shows that most gains arise from structured retrieval rather than unconstrained generation. Our results further suggest that LLMs are most effective as inference-time rerankers operating over clinically grounded hypothesis spaces. More broadly, this work highlights the importance of structure-aware retrieval for integrating geometric representations and LLMs in longitudinal EHR reasoning.

Acknowledgments and Disclosure of Funding

The authors acknowledge the financial support by the Federal Ministry of Research, Technology, and Space of Germany and by the Saxon State Ministry for Science, Culture, and Tourism, in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig“, project identification number: ScaDS.AI.

The authors acknowledge the computing time made available to them on the high-performance computer at the NHR Center of TU Dresden. This center is jointly supported by the Federal Ministry of Research, Technology and Space of Germany and the state governments participating in the NHR (www.nhr-verein.de/unsere-partner).

The authors declare no competing interests.

References

[1] Ahmed M Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. *Advances in neural information processing systems*, 32, 2019.

[2] Brett K Beaulieu-Jones, Isaac S Kohane, and Andrew L Beam. Learning contextual hierarchical structure of medical concepts with poincairé embeddings to clarify phenotypes. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 24, page 8, 2019.

[3] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[4] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. Hypercore: Hyperbolic and co-graph representation for automatic icd coding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3105–3114, 2020.

[5] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

[6] 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.

[7] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. Predictive modeling with temporal graphical representation on electronic health records. In *IJCAI: proceedings of the conference*, volume 2024, page 5763, 2024.

[8] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[9] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

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

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

[12] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*, pages 1646–1655. PMLR, 2018.

[13] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

[14] 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.

[15] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[16] 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.

[17] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[18] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[19] 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.

[20] 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.

[21] 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.

[22] Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

[23] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[24] Chang Lu, Chandan K Reddy, and Yue Ning. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4): 2124–2136, 2021.

[25] Qiuhao Lu, Nisansa De Silva, Sabin Kafle, Jiazhen Cao, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Brent Hailpern, Berthold Reinwald, and Yunyao Li. Learning electronic health records through hyperbolic embedding of medical ontologies. In *Proceedings of the 10th acm international conference on bioinformatics, computational biology and health informatics*, pages 338–346, 2019.

[26] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752, 2018.

[27] Chengsheng Mao, Liang Yao, and Yuan Luo. Medgcn: Medication recommendation and lab test imputation via graph convolutional networks. *Journal of Biomedical Informatics*, 127:104000, 2022.

[28] Usman Naseem, Surendrabikram Thapa, Qi Zhang, Shoujin Wang, Junaid Rashid, Liang Hu, and Amir Hussain. Graph learning with label attention and hyperbolic embedding for temporal event prediction in healthcare. *Neurocomputing*, 592:127736, 2024.

[29] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[30] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[31] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

[32] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 30–41. Springer, 2016.

[33] Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. eICU Collaborative Research Database. *PhysioNet*, April 2019. doi: 10.13026/C2WM1R. URL `https://doi.org/10.13026/C2WM1R`. Version 2.0.

[34] Zhan Qu and Michael Färber. Medieval: A unified medical benchmark for patient-contextual and knowledge-grounded reasoning in llms. *arXiv preprint arXiv:2512.20822*, 2025.

[35] Zhan Qu and Michael Färber. Trace: Temporal reasoning via agentic context evolution for streaming electronic health records (ehrs). *arXiv preprint arXiv:2602.12833*, 2026.

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

[37] 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.

[38] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International symposium on graph drawing*, pages 355–366. Springer, 2011.

[39] Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346*, 2019.

[40] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[41] Ethan Steinberg, Jason Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. *arXiv preprint arXiv:2301.03150*, 2023.

[42] Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature communications*, 14(1):5305, 2023.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] WHO. *International Statistical Classification of Diseases and related health problems: Alphabetical index*, volume 3. World Health Organization, 2004.

[45] WHO. Anatomical therapeutic chemical (atc) classification, 2024.

[46] 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.

[47] Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph rag: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, 2025.

[48] Yuzhou Wu, Xuechen Chen, Xin Yao, Yongang Yu, and Zhigang Chen. Hyperbolic graph convolutional neural network with contrastive learning for automated icd coding. *Computers in Biology and Medicine*, 168:107797, 2024.

[49] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, 2024.

[50] Yuyang Xu, Haochao Ying, Siyi Qian, Fuzhen Zhuang, Xiao Zhang, Deqing Wang, Jian Wu, and Hui Xiong. Time-aware context-gated graph attention network for clinical risk prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7557–7568, 2022.

[51] 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.

[52] Zefan Zeng, Qing Cheng, Xingchen Hu, Yan Zhuang, Xinwang Liu, Kunlun He, and Zhong Liu. Kosel: Knowledge subgraph enhanced large language model for medical question answering. *Knowledge-Based Systems*, 309:112837, 2025.

[53] Yuxuan Zhou, Xien Liu, Chenwei Yan, Chen Ning, Xiao Zhang, Boxun Li, Xiangling Fu, Shijin Wang, Guoping Hu, Yu Wang, and Ji Wu. Evaluating LLMs across multi-cognitive levels: From medical knowledge mastery to scenario-based problem solving. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=sgrJs7dbWC`.

A Additional Method Details

A.1 End-to-End Procedure

Algorithm 1 Risk Horizons: Geometry-Aware Structured Inference

Require: Longitudinal EHR trajectories $\{S_{1:T}^{(i)}\}_{i=1}^{N}$
Ensure: Next-visit prediction $\widehat{S}_{T+1}$
Construct hierarchical edges $\mathcal{E}_{\text{hier}}$ and lagged PMI edges $\mathcal{E}_{\text{cross}}$
Apply stability filtering to obtain graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
Learn node embeddings $\{z_v\} \subset \mathbb{D}_c^d$ via typed edge reconstruction loss $\mathcal{L}_{\text{edge}}$
Train Central Event representations via masked-visit reconstruction loss $\mathcal{L}_{\text{mask}}$
Compute multi-head Central Events $\{\mu_T^{(h)}\}_{h=1}^{H}$ for the current visit
for each head h **do**
    Select representative concept $v_T^{(h)}$
    Construct candidate pool $\mathcal{C}_{T,h}^{\text{cand}}$
    Apply cone constraint and compute scores $s_{\text{geo}}^{(h)}$
end for
Aggregate head-specific scores: $s_{\text{geo}}(u;T) = \sum_h \pi_h s_{\text{geo}}^{(h)}(u;T)$
Select top-K candidates per modality to form Risk Horizon $\mathcal{R}_T$
Rank retrieved candidates using geometric scores $s_{\text{geo}}(u;T)$
(Optional) Apply constrained LLM reranking over $\mathcal{R}_T$
Return top-k predictions per modality as $\widehat{S}_{T+1}$

A.2 Hyperbolic Geometry

We embed nodes in the d-dimensional Poincaré ball of curvature $-c$,

$$\mathbb{D}_c^d = \{x \in \mathbb{R}^d : \|x\| < 1/\sqrt{c}\}.$$

The hyperbolic distance between $x, y \in \mathbb{D}_c^d$ is

$$d_{\mathbb{D}}(x,y) = \frac{2}{\sqrt{c}} \operatorname{arctanh}\big(\sqrt{c}\,\|(-x) \oplus_c y\|\big), \tag{26}$$

where $\oplus_c$ denotes Möbius addition.

We use the logarithmic and exponential maps at the origin:

$$\log_0^c(x) = \frac{2}{\sqrt{c}} \operatorname{arctanh}(\sqrt{c}\|x\|)\frac{x}{\|x\|}, \qquad \exp_0^c(v) = \tanh\left(\frac{\sqrt{c}\|v\|}{2}\right)\frac{v}{\sqrt{c}\|v\|}. \tag{27}$$

After each gradient update, embeddings are projected to ensure $\|z_v\| < 1/\sqrt{c}$.

A.3 Typed Edge Reconstruction Objective

For a directed edge $(u \to v)$ of type r, the compatibility score is defined as

$$s_r(u,v) = -d_{\mathbb{D}}(z_u, z_v).$$

We optimize the negative-sampling objective

$$\mathcal{L}_{\text{edge}} = -\sum_{r\in\mathcal{R}} \sum_{(u\to v)\in\mathcal{E}_r} \left[\log\sigma(\gamma_r + s_r(u,v)) + \sum_{v^-\sim\mathcal{N}_r(v)} \log\sigma(\gamma_r - s_r(u,v^-))\right], \tag{28}$$

where $\mathcal{N}_r(v)$ samples negatives from the same target modality and γ_r is a learned type-specific bias.

A.4 Masked-Visit Reconstruction Loss

Given a masked visit $\mathcal{M}_t \subset \mathcal{C}_t$ and Central Event μ_t, we define the probability of recovering a masked code c as

$$p(c \mid \mu_t) = \frac{\exp(-d_{\mathbb{D}}(\mu_t, z_c)/\tau)}{\exp(-d_{\mathbb{D}}(\mu_t, z_c)/\tau) + \sum_{c^- \sim \mathcal{N}(c)} \exp(-d_{\mathbb{D}}(\mu_t, z_{c^-})/\tau)}. \tag{29}$$

The reconstruction loss is

$$\mathcal{L}_{\text{mask}} = -\sum_{c \in \mathcal{M}_t} \log p(c \mid \mu_t), \tag{30}$$

where negatives are sampled from the same modality as c.

B Datasets and Cohort Details

B.1 Task Definition and Evaluation Protocol

We consider multi-modal next-visit prediction on longitudinal EHR trajectories. Given patient history $S_{1:t}$, the task is to predict the (potentially multi-label) set of events $S_{t+1}^{(m)}$ for each modality $m \in \{\text{dx}, \text{proc}, \text{med}\}$.

The problem is formulated as a multi-label ranking task, where models assign scores to possible events and produce a ranked list of predictions independently for each modality. These predictions are evaluated against the ground-truth next-visit events using ranking and classification metrics (Appendix D).

B.2 Cohort Selection and Preprocessing

We conduct experiments on the MIMIC-IV and eICU. We construct patient trajectories as time-ordered sequences of visits, where each visit is a multi-modal event set containing diagnoses (ICD-9/10), procedures (ICD-9/10), and medications (ATC).

B.3 Sequence Length.

We analyze the effect of the minimum trajectory length $T_{\min}$ on model performance. Increasing $T_{\min}$ improves temporal context but reduces the number of available training samples. We compare $T_{\min} \in \{2, 3, 4\}$ and observe that $T_{\min} = 2$ provides the best trade-off between data coverage and predictive performance, while larger values lead to reduced sample size and no significant performance gains. Results in this table use the single-head variant of Risk Horizons to isolate the effect of minimum trajectory length independently of the final multi-head retrieval configuration used in the main experiments.

Table 5: Effect of minimum sequence length $T_{\min}$ on sample size and performance using the single-head Risk Horizons variant.

		MIMIC-IV			eICU		
$T_{\min}$	**Method**	#Pairs	R@10	nDCG@10	#Pairs	R@10	nDCG@10
2	Risk Horizons (geometry-only, single head)	63,168	0.282	0.380	15,214	0.249	0.271
3	Risk Horizons (geometry-only, single head)	42,406	0.289	0.388	7,039	0.290	0.295
4	Risk Horizons (geometry-only, single head)	27,452	0.294	0.399	3,697	0.220	0.228

B.4 Vocabulary and Hierarchy Construction.

For each retained leaf code, we include deterministic hierarchical ancestors derived from standard coding systems (ICD for diagnoses, ICD procedure hierarchies for procedures, and ATC for medications). The final vocabulary $\mathcal{V}$ is the union of retained leaf nodes and all their ancestors.

Diagnosis nodes. Diagnosis nodes follow an ICD hierarchy with up to four levels in our graph (*modality root* $\rightarrow$ coarse prefix $\rightarrow$ 3-character category $\rightarrow$ leaf code). For example, `I50.23` [acute on chronic systolic heart failure] expands as `I` [circulatory system diseases] $\rightarrow$ `I50` [heart failure] $\rightarrow$ `I50.23` [acute on chronic systolic heart failure].

Procedure nodes. Procedure nodes follow ICD procedure structure and are constructed using prefix ancestors derived from the code string. For ICD-10-PCS, a 7-character code induces a prefix chain; for example, `0FT44ZZ` [resection of gallbladder, percutaneous endoscopic approach] expands as `0` $\rightarrow$ `0F` $\rightarrow$ `0FT` $\rightarrow$ `0FT4` $\rightarrow$ `0FT44` $\rightarrow$ `0FT44Z` $\rightarrow$ `0FT44ZZ`. Including the modality root gives up to eight levels. ICD-9 procedure codes are shallower and therefore induce fewer levels.

In practice, procedure hierarchy depth is data-dependent. Many ICD-9-style procedure codes induce shallow chains (typically a 2-digit prefix plus leaf), while ICD-10-PCS-style codes induce longer prefix chains. Therefore, procedure hierarchy depth is variable and does not always correspond to a full 7-level semantic ontology in the processed graph.

Medication nodes. Medication nodes (MIMIC-IV) follow ATC prefixes. With ATC4 granularity in our pipeline, this yields up to five levels including root (root $\rightarrow$ level-1 anatomical group $\rightarrow$ level-2 therapeutic subgroup $\rightarrow$ level-3 pharmacological subgroup $\rightarrow$ level-4 chemical subgroup). For example, `A10A` [insulins and analogues] expands as `A` [alimentary tract and metabolism] $\rightarrow$ `A10` [drugs used in diabetes] $\rightarrow$ `A10A` [insulins and analogues] (and with modality root: `med:ROOT` $\rightarrow$ `A` $\rightarrow$ `A10` $\rightarrow$ `A10A`).

Dataset-specific statistics (MIMIC-IV). For MIMIC-IV, on the `min_visits` ≥ 2 subset, diagnosis codes are mostly ICD-10 (62.64%), with ICD-9 accounting for 37.29% and a negligible remainder (0.07%). Procedure codes are mixed: 51.08% ICD-9 and 48.92% ICD-10-PCS, indicating that both coding eras are substantially represented.

For the root-to-leaf path length, diagnosis paths are highly consistent (min 3, max 4, mean 3.95, median 4), while procedure paths are more variable (min 3, max 8, mean 5.37, median 3), reflecting the coexistence of shallow ICD-9-style and deeper ICD-10-PCS-style prefix chains. Medication paths in MIMIC-IV are fixed in this preprocessing pipeline (all length 4).

Dataset-specific statistics (eICU). For eICU (`min_visits` ≥ 2), diagnoses are overwhelmingly ICD-9 (98.20%), with only 1.80% ICD-10. Procedures are not stored using standard ICD procedure ontologies in the processed tables; instead, they are represented by synthetic `PROC_*` identifiers. Medications are represented by normalized medication identifiers/text fields rather than a clean ATC hierarchy. These source-table differences lead to substantially shallower procedure hierarchies, while medication hierarchies remain relatively deeper after preprocessing.

On the `min_visits` ≥ 2 subset of eICU, diagnosis path lengths remain compact (min 3, max 4, mean 3.94, median 4). Procedure path lengths are mostly shallow (min 3, max 8, mean 3.03, median 3). Medication paths are deeper and nearly uniform (min 5, max 6, mean 6.00, median 6).

Event Representation Modes. We support three event representation modes: `code`, `text`, and `hybrid`. In `code` mode, events are represented by raw codes only. In `text` mode, events are represented by mapped short descriptions. In `hybrid` mode, events are represented as `text (code)`.

Text descriptions are derived from code-to-text mappings constructed during preprocessing, primarily from official ICD terminology tables for diagnoses and procedures, with analogous medication mappings when available.

Because graph construction and hierarchy stitching are based on canonical code hierarchies rather than free-text semantics, geometry-side performance is effectively invariant across display modes under matched preprocessing and splits. For consistency with LLM-related evaluation, we conduct experiments in `text` mode.

B.5 Dataset Statistics

MIMIC-IV. After preprocessing (minimum 2 visits per patient, without minimum frequency filtering), the cohort contains 34,775 patients and 97,943 visits, with 63,168 next-visit prediction pairs for $T_{\min} = 2$. On average across 5 seeds (patient-level 80/10/10 split), train/val/test contain 27,820/3,477/3,478 patients and 78,364/9,807/9,772 visits. Per-visit event counts are: diagnosis (mean 16.49, median 15, min 1, max 40, std 8.50), procedure (mean 3.23, median 2, min 1, max 35, std 2.82), and medication (mean 13.08, median 12, min 1, max 67, std 7.07). The combined events per visit have mean 32.81, median 31, min 3, max 119, and std 13.67. The final vocabulary has 62,574 nodes (dx: 25,711; proc: 36,617; med: 246).

eICU. Using the same preprocessing protocol, the cohort contains 10,766 patients and 25,980 visits, with 15,214 next-visit prediction pairs for $T_{\min} = 2$. Averaged over 5 seeds, train/val/test contain 8,612/1,076/1,078 patients and 20,768/2,618/2,594 visits. Per-visit event counts are: diagnosis (mean 3.58, median 2, min 1, max 58, std 3.62), procedure (mean 7.62, median 4, min 1, max 148, std 9.42), and medication (mean 18.85, median 17, min 1, max 109, std 10.90). The combined events per visit have mean 30.05, median 25, min 3, max 274, and std 18.36. The final vocabulary has 5,912 nodes (dx: 1,290; proc: 2,638; med: 1,984).

Data Splits and Leakage Prevention. We split the dataset *by patient* into training (80%), validation (10%), and test (10%) sets, ensuring patient-level leakage control (no visits from the same patient across splits). Crucially, all data-driven graph statistics used for temporal stitching, especially lagged PMI, are computed *strictly on the training split* to prevent label leakage. No information from validation or test splits is used in graph construction, model training, or hyperparameter selection.

Split Coverage. We report overlap of unique event nodes and lag-1 transition edges between training and held-out splits (averaged over 5 seeds: 13, 21, 42, 87, 123). For **MIMIC-IV**, validation/test node overlap with training is high (val: 91.97%, test: 92.17%), with only 8.03%/7.83% unseen nodes, respectively. Transition-edge overlap is lower but still substantial (val: 71.17%, test: 71.50%), leaving 28.83%/28.50% unseen edges. For **eICU**, node overlap is even higher (val: 97.35%, test: 97.28%), with 2.65%/2.72% unseen nodes; transition-edge overlap is 80.70% (val) and 80.64% (test), with 19.30%/19.36% unseen edges. Overall, most validation/test nodes and edges are observed during training, while a non-trivial unseen fraction remains, reflecting a realistic distributional shift in longitudinal EHR data. As a result, models must generalize beyond memorized co-occurrence patterns and reason over partially observed transition structure, which is particularly challenging for rare or weakly supported cross-modal dependencies.

C Baseline Implementations

We compare Risk Horizons against sequential, retrieval-based, hyperbolic, and LLM-based baselines under a unified protocol. All baselines use patient-level 80/10/10 splits, minimum visits $= 2$, and minimum code frequency $= 0$.

Sequential Baselines.

- **Copy Last.** A deterministic persistence baseline that predicts events from the most recent observed visit.
- **Transformer [43].** A visit-sequence Transformer over flattened visit event sets, using hidden dimension 256, four layers, eight attention heads, an FFN dimension of 1024, and a BCE loss.
- **RETAIN [10].** A reverse-time dual-RNN attention model with visit-level and variable-level attention, trained with BCE loss.
- **GRAM [11].** An ontology-aware model that aggregates code and ancestor representations using hierarchy-derived attention.
- **BEHRT [21].** A BERT-style longitudinal EHR encoder with visit-position and modality-type embeddings.
- **Med-BERT [37].** A medical BERT-style longitudinal encoder over event tokens with visit-position embeddings.
- **HALO [42].** An autoregressive EHR sequence model adapted to multi-label next-visit prediction.

Geometry and Retrieval Baselines.

- **Euclidean Risk Horizons.** Risk Horizons with the same graph construction and objectives, but replacing hyperbolic embeddings and distances with Euclidean counterparts.
- **Euclidean Risk Horizons +LLM variants.** LLM reranking applied to Euclidean candidate sets using the same constrained scoring protocol as Risk Horizons.

Hyperbolic Baselines.

- **Hyperbolic Linear Probe.** A lightweight multi-label linear classifier trained on frozen Risk Horizons embeddings and visit representations.
- **Hyperbolic NN Global.** A non-parametric nearest-neighbor retrieval baseline using frozen hyperbolic embeddings, without patient-specific Risk Horizon construction.

LLM Baselines.

- **Frequency-candidate reranking.** LLMs rerank candidate pools constructed from training-set frequency statistics using the same constrained candidate-scoring protocol as Risk Horizons. Candidate size is set to $K = 80$ by default. The evaluated LLM backends include `meta-llama/Llama-3.1-8B-Instruct` [14], `meta-llama/Llama-3.3-70B-Instruct` [14], `epfl-llm/meditron-70b` [8], and `mistralai/Mixtral-8x7B-Instruct-v0.1` [18], all accessed through Hugging Face Transformers.

D Evaluation Metrics

We evaluate ranking quality, hierarchy consistency, retrieval quality, and efficiency. Main MIMIC-IV results are reported in Table 6, extended metrics in Table 7, and eICU results in Table 11.

Ranking Metrics. We report Recall@k and nDCG@k ($k \in \{5, 10, 20\}$), computed per visit and macro-averaged across modalities. We additionally report MRR and Micro-F1 after threshold tuning on the validation split.

Hierarchy and Retrieval Diagnostics. To evaluate structural consistency and candidate-space quality beyond exact matching, we report:

- **Hierarchy Distance (Tree Dist, ↓):** Average shortest-path distance between predicted and ground-truth events in the hierarchy.
- **Ancestor Match@2 (↑):** Fraction of predictions sharing an ancestor with the target within two hierarchy levels.
- **Candidate Recall (↑):** Fraction of ground-truth events covered by the retrieved candidate pool.
- **Risk-Horizon Recall@K (↑):** Fraction of ground-truth events retained in the top-K retrieved candidates.
- **Oracle Recall@K (↑):** Upper-bound recall achievable under perfect reranking of the retrieved candidate set.

Efficiency. We report average evaluation latency per visit step.

E Implementation Details

All models are implemented in PyTorch. Hyperbolic optimization is implemented with `geoopt`. Experiments are conducted on a single NVIDIA H100 GPU.

General Training Configuration. Unless otherwise noted, Risk Horizons uses:

- embedding dimension $d = 64$,
- Poincaré ball manifold with initial curvature $c = 1.0$,
- Riemannian Adam optimizer,
- learning rate 10^{-2},
- batch size 256,

- maximum training epochs 50,
- 10 negative samples per positive edge,
- edge-loss margin 1.0,
- masked-visit loss weight $\alpha = 1.0$,
- masking ratio $\rho = 0.30$,
- soft assignment temperature $\beta = 5.0$.

Graph Construction and Retrieval. The main Risk Horizons variant uses:

- lag horizon $L = 1$,
- PMI threshold $\tau = 0.0$,
- minimum support $\kappa = 3$,
- bootstrap stability filtering threshold $q = 0.5$,
- intra-visit and intra-modality stitching enabled,
- multi-head Central Events with $H = 4$ heads,
- multi-cone retrieval with cone aperture $\phi = 0.5$.

Validation and Model Selection. Early stopping is performed using validation Recall@10 with patience 5. Threshold tuning for Micro-F1 is performed on the validation split prior to final evaluation.

Computational resources. Experiments were conducted using one NVIDIA H100 GPU, eight CPU cores, and 128 GB of memory. The longest experimental run (HALO) required at most 54.4 hours.

LLM Usage. For LLM-augmented variants, the LLM is used strictly as an inference-time reranker over candidates retrieved by Risk Horizons. The LLM cannot introduce new candidates, and no gradients are propagated through the language model.

E.1 LLM Prompting and Scoring Details

Figure 3 shows the prompt template used for constrained LLM reranking. The LLM receives a linearized patient history together with modality-specific candidate sets retrieved by Risk Horizons and returns ranked predictions restricted to the provided candidates.

We use a listwise reranking formulation in which diagnoses, procedures, and medications are grouped by modality. The prompt instructs the LLM to consider temporal progression and cross-modal consistency while avoiding unsupported candidate expansion. Candidates not selected by the LLM receive the lowest reranking score. LLM outputs are combined with geometric retrieval scores using Eq. (23).

To control context length and inference cost, we retain only the most recent visit in the linearized patient history. The retrieved candidate pool size is determined by the Risk Horizon retrieval configuration and is typically around 100 candidates per modality in our main experiments. Typical histories contain an average of 177 diagnosis tokens, 46 procedure tokens, and 182 medication tokens. The resulting prompts contain approximately 3,500 input tokens on average, with a maximum context budget of 4,096 tokens.

Generation is performed with greedy decoding (temperature $T = 0$) for deterministic inference. Candidate expansion is explicitly disabled by constraining the LLM to rank only events appearing in the retrieved candidate list. Outputs are restricted to structured JSON responses with a maximum generation budget of 4,096 tokens. No in-context examples are provided, and no gradients are propagated through the language model.

All LLM experiments are executed through Hugging Face using identical prompts and decoding configurations across datasets and evaluation splits.

Prompt Template for Constrained Candidate Reranking

You are a clinical event prediction assistant.

Task:
Given a patient's longitudinal visit history and a candidate list, select and rank only the events from the provided candidate list that are most likely to occur in the NEXT visit.

Modalities:
- dx: diagnosis
- proc: procedure
- med: medication

Instructions:
- Consider temporal progression (earlier → later visits).
- Consider cross-modality consistency (diagnoses, procedures, medications).
- Favor clinically plausible developments over unrelated or overly generic events.
- Select likely events only from the provided candidate list.
- Do not introduce, infer, or output any event that is not present in the candidate list.
- Keep outputs concise and clinically grounded.

Output:
Return JSON only in the following format:
`{"dx": [{"code": "...", "text": "...", "rank": 1}], "proc": [...], "med": [...]}`
Use objects of the form `{"code": "...", "text": "...", "rank": ...}` for each selected event.
Do not output any other text.

Patient history (oldest → newest):
`{history_text}`

Candidate list:
```
dx: {dx_candidates}
proc: {proc_candidates}
med: {med_candidates}
```

Candidate context:
- Candidates are produced by upstream structured retrieval.
- The candidate list is intended to constrain prediction to plausible next-visit events.

Ranked next-visit events (JSON):

Figure 3: Prompt template used for constrained LLM reranking. The LLM selects and ranks events only from the provided candidate list and is not allowed to expand the candidate set.

E.2 Reproducibility

The public project repository linked in the Introduction contains preprocessing scripts, graph construction pipelines, model implementations, training configurations, and evaluation code required to reproduce the experiments after obtaining dataset access.

All data splits are generated at the patient level using fixed random seeds. To avoid redistributing restricted dataset-derived information, we release the split-generation procedure rather than patient-level split files. All results are averaged over five random seeds, and all data-dependent statistics (e.g., lagged PMI edges) are computed strictly on the training split to prevent information leakage.

Data Accessibility. All datasets used in this work are publicly available through PhysioNet[1]. Access to MIMIC-IV and eICU requires credentialed approval and completion of the required data-use training procedures specified by PhysioNet.

[1] `https://physionet.org/`

F Additional Results on MIMIC-IV

F.1 Full Evaluation Results on MIMIC-IV

Tables 6 and 7 provide the full MIMIC-IV evaluation, including baselines and metrics omitted from the main text for space. These results complement the main analysis by separating three effects: whether a method can recover plausible future events, whether it can rank them well, and whether downstream reranking can compensate for weak candidate construction.

First, the complete table confirms that the advantage of Risk Horizons is not explained by model capacity alone. Sequential baselines remain competitive for frequent and persistent events, especially medications, but their performance varies substantially across modalities. Transformer, for example, achieves the highest medication nDCG@10 among sequential baselines, while RETAIN achieves the strongest overall R@10 in this group; however, neither provides consistently strong diagnosis, procedure, and medication prediction. This supports the main-text observation that longitudinal EHR prediction is not only a temporal modeling problem, but also a structured output-space problem.

Second, LLM-only baselines perform poorly when restricted to frequent-event candidate sets. Even large models such as Llama-3.3-70B-Instruct and Meditron-70B remain far below sequential and Risk Horizons variants. This result should be interpreted as a limitation of unconstrained or weakly constrained prediction rather than as a limitation of language models themselves: next-visit prediction requires selecting sparse, multi-label clinical futures from a large and structured vocabulary, and frequent-event candidate sets provide insufficient patient-specific structure.

Third, Euclidean retrieval and Euclidean Risk Horizons remain weak even when paired with LLM reranking. In several cases, adding an LLM to Euclidean candidates does not improve performance and may reduce it. This indicates that downstream reasoning cannot reliably repair a poorly constructed candidate space. By contrast, hyperbolic Risk Horizons without LLM reranking already provides a substantially stronger starting point, suggesting that the main benefit comes from constructing a structured and patient-specific Risk Horizon before final ranking.

Table 6: Multi-modal next-visit prediction on MIMIC-IV (mean ± std over 5 seeds).

Model	Dx		Proc		Med		Overall	
	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
Sequential Baselines								
Copy Last	0.190 ± 0.002	0.303 ± 0.001	0.166 ± 0.030	0.152 ± 0.024	0.452 ± 0.031	0.619 ± 0.012	0.269 ± 0.029	0.358 ± 0.012
Transformer	0.219 ± 0.015	0.423 ± 0.018	0.327 ± 0.015	0.263 ± 0.014	0.539 ± 0.025	0.734 ± 0.030	0.362 ± 0.018	0.473 ± 0.021
BEHRT	0.141 ± 0.007	0.271 ± 0.002	0.258 ± 0.005	0.201 ± 0.006	0.499 ± 0.051	0.687 ± 0.010	0.299 ± 0.038	0.386 ± 0.007
RETAIN	0.220 ± 0.013	0.423 ± 0.019	0.336 ± 0.016	0.268 ± 0.017	0.532 ± 0.030	0.723 ± 0.034	0.363 ± 0.021	0.471 ± 0.023
GRAM	0.093 ± 0.003	0.178 ± 0.003	0.170 ± 0.008	0.119 ± 0.010	0.479 ± 0.030	0.661 ± 0.021	0.247 ± 0.014	0.319 ± 0.011
Med-BERT	0.149 ± 0.013	0.293 ± 0.023	0.280 ± 0.021	0.221 ± 0.010	0.505 ± 0.012	0.694 ± 0.011	0.311 ± 0.015	0.402 ± 0.018
HALO	0.128 ± 0.056	0.251 ± 0.043	0.244 ± 0.037	0.185 ± 0.060	0.495 ± 0.074	0.683 ± 0.075	0.289 ± 0.056	0.373 ± 0.059
LLMs (Top80 Freq. Events)								
Llama-3.1-8B-Instruct	0.112 ± 0.001	0.135 ± 0.001	0.096 ± 0.001	0.124 ± 0.004	0.205 ± 0.001	0.276 ± 0.002	0.138 ± 0.001	0.178 ± 0.002
Llama-3.3-70B-Instruct	0.196 ± 0.021	0.216 ± 0.034	0.077 ± 0.019	0.074 ± 0.012	0.220 ± 0.009	0.382 ± 0.010	0.164 ± 0.016	0.224 ± 0.019
Mixtral-8x7B-Instruct-v0.1	0.187 ± 0.039	0.148 ± 0.035	0.068 ± 0.045	0.067 ± 0.041	0.209 ± 0.071	0.397 ± 0.060	0.155 ± 0.052	0.204 ± 0.045
Meditron-70B	0.201 ± 0.030	0.205 ± 0.017	0.129 ± 0.009	0.146 ± 0.013	0.274 ± 0.009	0.391 ± 0.004	0.201 ± 0.016	0.247 ± 0.011
Graph & Retrieval								
Euclidean RH	0.161 ± 0.013	0.337 ± 0.016	0.224 ± 0.040	0.179 ± 0.032	0.426 ± 0.051	0.572 ± 0.065	0.270 ± 0.004	0.362 ± 0.009
Euc.RH+Llama-3.1-8B-I.	0.182 ± 0.010	0.368 ± 0.023	0.222 ± 0.032	0.150 ± 0.039	0.396 ± 0.031	0.496 ± 0.020	0.267 ± 0.021	0.338 ± 0.023
Hyperbolic Baselines								
Hyperbolic Linear Probe	0.129 ± 0.030	0.249 ± 0.023	0.249 ± 0.017	0.179 ± 0.030	0.489 ± 0.034	0.678 ± 0.035	0.289 ± 0.027	0.369 ± 0.029
Hyperbolic NN Global	0.125 ± 0.063	0.243 ± 0.028	0.166 ± 0.021	0.125 ± 0.050	0.362 ± 0.035	0.514 ± 0.025	0.218 ± 0.040	0.294 ± 0.034
Ours								
Risk Horizons (w/o LLM)	0.195 ± 0.005	0.392 ± 0.012	0.298 ± 0.009	0.243 ± 0.009	0.579 ± 0.012	0.705 ± 0.014	0.357 ± 0.013	0.447 ± 0.017
RH+Llama-3.1-8B-I.	0.213 ± 0.050	0.453 ± 0.012	0.298 ± 0.009	0.243 ± 0.010	0.579 ± 0.031	0.705 ± 0.047	0.363 ± 0.030	0.467 ± 0.023
RH+Llama-3.3-70B-I.	0.306 ± 0.032	0.496 ± 0.064	0.368 ± 0.005	0.360 ± 0.008	0.602 ± 0.016	0.693 ± 0.014	0.425 ± 0.018	0.516 ± 0.029
RH+Meditron-70B	0.297 ± 0.011	0.417 ± 0.033	0.389 ± 0.023	0.365 ± 0.009	0.587 ± 0.058	0.601 ± 0.027	0.424 ± 0.031	0.461 ± 0.023

The extended metrics in Table 7 further support this retrieval–ranking decomposition. Risk Horizons + Llama-3.3-70B-Instruct achieves the strongest Recall@5, Recall@20, Micro-F1, and MRR, indicating that gains persist beyond the top-10 cutoff. The improvement in MRR is particularly informative: geometry-only Risk Horizons already retrieves relevant events effectively, while constrained reranking moves those events earlier in the ranked list. This is consistent with the main-text interpretation that Risk Horizons addresses the candidate-space coverage problem, while LLM reranking primarily improves local ordering within that space.

Finally, the latency results show the expected trade-off between retrieval-only and reranked inference. Geometry-only Risk Horizons is substantially faster, while LLM-based reranking adds inference

overhead. This cost is bounded by the compact Risk Horizon, since the LLM scores a restricted candidate set rather than the full vocabulary.

Table 7: Extended evaluation metrics on MIMIC-IV (overall, macro; mean ± std over 5 seeds).

Model	R@5	R@20	nDCG@5	nDCG@20	Micro-F1	MRR	Latency (s/step)
Sequential Baselines							
Copy Last	0.177 ± 0.013	0.337 ± 0.007	0.361 ± 0.009	0.347 ± 0.002	0.325 ± 0.005	0.481 ± 0.002	0.001
Transformer (Eucl.)	0.242 ± 0.006	0.494 ± 0.020	0.500 ± 0.014	0.475 ± 0.015	0.326 ± 0.012	0.651 ± 0.010	0.589
BEHRT	0.194 ± 0.005	0.424 ± 0.006	0.405 ± 0.010	0.395 ± 0.004	0.270 ± 0.009	0.565 ± 0.007	0.193
RETAIN	0.240 ± 0.005	0.499 ± 0.023	0.496 ± 0.005	0.475 ± 0.012	0.319 ± 0.003	0.652 ± 0.005	0.176
GRAM	0.152 ± 0.009	0.365 ± 0.014	0.334 ± 0.013	0.332 ± 0.018	0.229 ± 0.006	0.479 ± 0.007	0.172
Med-BERT	0.204 ± 0.011	0.439 ± 0.023	0.422 ± 0.021	0.410 ± 0.025	0.278 ± 0.026	0.583 ± 0.036	0.165
HALO	0.189 ± 0.045	0.409 ± 0.081	0.391 ± 0.057	0.381 ± 0.034	0.261 ± 0.080	0.541 ± 0.083	0.031
LLMs (Top 80 Freq. Events)							
Llama-3.1-8B-Instruct	0.076 ± 0.002	0.250 ± 0.004	0.177 ± 0.003	0.202 ± 0.003	0.217 ± 0.002	0.296 ± 0.003	5.211
Llama-3.3-70B-Instruct	0.092 ± 0.018	0.276 ± 0.021	0.208 ± 0.025	0.232 ± 0.020	0.224 ± 0.012	0.332 ± 0.021	5.461
Mixtral-8x7B-Instruct-v0.1	0.090 ± 0.034	0.268 ± 0.043	0.203 ± 0.041	0.227 ± 0.038	0.226 ± 0.030	0.326 ± 0.039	5.667
Meditron-70B	0.116 ± 0.012	0.312 ± 0.015	0.254 ± 0.014	0.276 ± 0.013	0.235 ± 0.009	0.384 ± 0.014	5.893
Graph & Retrieval							
Euclidean Risk Horizons	0.183 ± 0.006	0.364 ± 0.012	0.392 ± 0.009	0.356 ± 0.009	0.284 ± 0.006	0.548 ± 0.008	0.656
Euc.RH+Llama-3.1-8B	0.185 ± 0.018	0.352 ± 0.027	0.385 ± 0.026	0.321 ± 0.024	0.280 ± 0.018	0.536 ± 0.026	6.687
Hyperbolic Baselines							
Hyperbolic Linear	0.197 ± 0.017	0.367 ± 0.016	0.352 ± 0.010	0.445 ± 0.014	0.284 ± 0.012	0.537 ± 0.013	0.077
Hyperbolic NN Global	0.145 ± 0.021	0.303 ± 0.020	0.265 ± 0.018	0.358 ± 0.019	0.232 ± 0.016	0.501 ± 0.014	0.073
Ours							
RH (w/o LLM backend)	0.223 ± 0.009	0.460 ± 0.003	0.465 ± 0.007	0.442 ± 0.006	0.320 ± 0.006	0.613 ± 0.010	0.319
RH+Llama-3.1-8B-I.	0.228 ± 0.030	0.465 ± 0.037	0.454 ± 0.022	0.431 ± 0.025	0.307 ± 0.019	0.622 ± 0.025	5.967
RH+Llama-3.3-70B-I.	0.279 ± 0.011	0.566 ± 0.018	0.519 ± 0.034	0.503 ± 0.029	0.338 ± 0.012	0.704 ± 0.028	46.781
RH+Meditron-70B	0.274 ± 0.023	0.558 ± 0.032	0.532 ± 0.031	0.512 ± 0.027	0.327 ± 0.020	0.686 ± 0.031	46.923

F.2 Additional Ablation and Sensitivity Studies

Table 4 analyzes the contribution of the main components of Risk Horizons. All components contribute substantially to performance, supporting the central hypothesis that accurate next-visit prediction depends primarily on constructing structured and temporally coherent candidate spaces before final ranking.

Cross-modal temporal stitching is essential. Removing lagged PMI stitching causes a near-complete collapse in performance, reducing overall R@10 from 0.357 to 0.005 and nDCG@10 from 0.447 to 0.004. This indicates that deterministic coding hierarchies alone are insufficient for constructing clinically meaningful future hypothesis spaces. Most predictive signal instead arises from inferred cross-modal and temporal dependencies between diagnoses, procedures, and medications.

Directional retrieval improves hypothesis-space quality. Replacing cone-constrained retrieval with distance-only retrieval reduces overall R@10 from 0.357 to 0.180 and nDCG@10 from 0.447 to 0.206. This suggests that directional constraints are important for modeling asymmetric clinical progression. Without directional retrieval, candidate spaces remain semantically related but become substantially less predictive of future transitions.

Effect of graph directionality. We further examine whether preserving edge direction contributes to prediction performance by comparing the original directed clinical graph with undirected and reversed-edge variants under identical embedding, retrieval, and ranking settings. On MIMIC-IV, the directed configuration achieves an overall R@10 of 0.357 and nDCG@10 of 0.447, compared with 0.272/0.381 for undirected edges and 0.291/0.396 for reversed edges. These results demonstrate that the orientation of hierarchical and temporal associations contributes substantially to predictive performance. Importantly, this effect arises from the directed graph construction and typed edge modeling rather than from asymmetry in hyperbolic distance itself.

Masked-visit denoising stabilizes latent clinical states. Removing masked-visit reconstruction reduces overall R@10 from 0.357 to 0.115 and nDCG@10 from 0.447 to 0.124. This indicates that denoising-based Central Event training improves robustness under sparse and noisy multi-modal trajectories by encouraging visit representations to preserve predictable clinical structure rather than memorizing individual observations.

Central-event weighting. The central-event weighting results further suggest that more complex weighting schemes do not automatically improve the latent visit representation. IDF, attention-based, and modality-level weighting all underperform uniform weighting. One possible explanation is that

Table 8: Central Event Weighting Strategies.

Method	R@10	nDCG@10
Central Events Weighting		
idf weighting	0.194	0.224
attention weighting	0.189	0.217
modality weighting	0.192	0.220

the downstream Risk Horizon already imposes strong structural constraints, so additional weighting can overemphasize rare or locally salient codes at the expense of stable patient-state representation.

Sensitivity Analysis. The sensitivity analysis shows that Risk Horizons is robust to moderate hyperparameter variation but sensitive to choices that damage candidate-space quality. Increasing the number of Central Event heads improves performance up to a moderate value and then saturates, suggesting that multiple heads help separate heterogeneous clinical factors but excessive capacity provides limited benefit. The PMI threshold has a relatively mild effect, while aggressive stability filtering substantially degrades performance, likely because useful but weakly observed temporal associations are removed. Cone aperture exhibits a similar trade-off: moderate apertures preserve directional selectivity, whereas overly broad cones admit semantically related but temporally less plausible candidates. Overall, the sensitivity results reinforce the main conclusion that performance depends on constructing candidate spaces that are broad enough to cover plausible futures but constrained enough to remain clinically coherent.

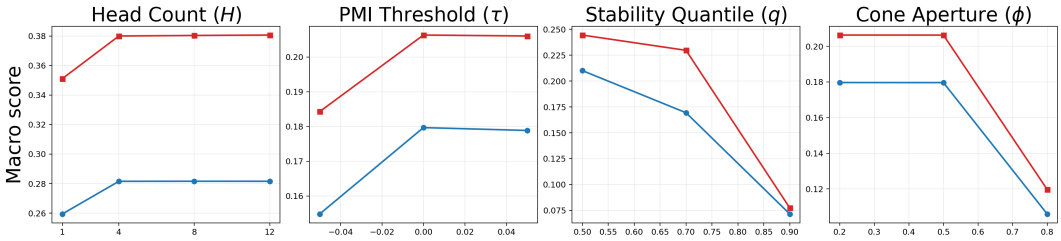


Figure 4: Sensitivity analysis: Recall@10 (blue) and nDCG@10 (red). From left to right: (a) number of heads H, (b) PMI Threshold τ, (c) Stability q, (d) cone aperture ϕ.

Statistical significance. We assess performance differences using paired patient-cluster bootstrap resampling with 10,000 replicates. Patients are sampled with replacement, retaining all evaluated transitions from each sampled patient to account for within-patient dependence. We report differences in overall R@10 and nDCG@10, defined as Risk Horizons minus the comparator, together with 95% bootstrap confidence intervals and two-sided bootstrap p-values.

Compared with RETAIN, geometry-only Risk Horizons shows no statistically significant difference in R@10 ($\Delta = -0.006$, 95% CI $[-0.013, 0.002]$, $p = 0.11$) or nDCG@10 ($\Delta = -0.024$, 95% CI $[-0.066, 0.004]$, $p = 0.10$). In contrast, it significantly outperforms MedBERT in both R@10 ($\Delta = +0.046$, 95% CI $[0.031, 0.059]$, $p < 0.001$) and nDCG@10 ($\Delta = +0.045$, 95% CI $[0.032, 0.057]$, $p < 0.001$). These results indicate that geometry-only Risk Horizons is competitive with RETAIN under this test while significantly outperforming MedBERT.

F.3 Repeated vs. Novel Event Analysis

To better understand the behavior of Risk Horizons, we analyze prediction performance separately for repeated events (events already observed in the patient history) and novel events (events not previously observed in the history). This distinction is particularly important in longitudinal EHR prediction because many future events correspond either to persistent chronic conditions or to clinically coherent progression patterns, while others arise from weakly observed, exogenous, or highly sparse transitions.

Table 9 reports Recall separately for repeated and novel events on MIMIC-IV using the geometry-only Risk Horizons variant. Across all modalities, performance is substantially higher for repeated events

than for novel events. This gap is especially pronounced for diagnoses, where repeated-event Recall reaches 0.646 while novel-event Recall is only 0.084.

Table 9: Recall on repeated vs. novel events on MIMIC-IV. Repeated events are events already present in the patient history; novel events are newly introduced events not previously observed in the trajectory.

	Event Distribution		**Recall**		**Weakly Supported**
Modality	Repeated Share	Novel Share	Repeated	Novel	Novel Events
Dx	28.2%	71.8%	0.646	0.084	2.2%
Proc	13.4%	86.6%	0.707	0.171	6.9%
Med	66.1%	33.9%	0.854	0.339	0.0%

These findings are consistent with the structured retrieval formulation of Risk Horizons. The framework is most effective when future events follow coherent longitudinal progression and can be recovered through structured candidate construction. In such settings, the retrieved Risk Horizon successfully captures persistent conditions, evolutionarily related diagnoses, and temporally associated medications and procedures.

In contrast, novel-event prediction remains substantially harder, particularly for diagnoses and procedures. Many novel events correspond to traumatic injuries, acute complications, unexpected interventions, or weakly observed transitions with limited historical support. Such events are intrinsically difficult to anticipate from prior longitudinal structure alone. Importantly, even when exact novel events are missed, predictions often remain clinically and hierarchically related to the underlying patient state, consistent with the qualitative examples discussed in Section 4.3.

The final column of Table 9 reports the fraction of novel events that are weakly supported in the analyzed trajectories, meaning that they never appear in the historical portion of any patient trajectory within the analyzed samples. Although these events account for only a minority of novel targets (2.2% for diagnoses and 6.9% for procedures), they are almost never recovered by the model. Many correspond to rare acute conditions, traumatic injuries, or unexpected interventions with limited recoverable longitudinal structure.

Overall, these findings suggest that the primary strength of Risk Horizons lies in recovering clinically plausible progression trajectories through structured retrieval, while highly novel or exogenous transitions remain a fundamental challenge for longitudinal EHR forecasting more broadly.

Generalization across transition support. We additionally stratify held-out target events according to their maximum directed lag-1 transition support in the training data, measured over events in the current visit before PMI filtering or graph pruning. Frequent transitions (support > 5) achieve R@10 of 0.389 and candidate Recall@100 of 0.421, whereas rare transitions (support 1–5) achieve 0.113 and 0.114, respectively. Unseen transitions (support 0) achieve zero on both metrics because their target associations are absent from the learned graph. The analysis covers 18,687 transition–modality cases; groups are defined at the target-event level and may overlap within a case. These results identify limited training support as a major bottleneck for candidate coverage, particularly for rare and unseen transitions.

Effect of inter-visit time intervals. Although Risk Horizons models temporal relationships through visit-order lag rather than elapsed calendar time, the interval between visits may influence prediction difficulty. We therefore stratify MIMIC-IV test transitions by the elapsed time between consecutive visits. For intervals of 0–30 days (1,655 transitions), overall R@10, nDCG@10, and Candidate Recall@100 are 0.3764, 0.4964, and 0.4674, respectively. These values decrease to 0.3482, 0.4678, and 0.4411 for intervals of 31–180 days (1,646 transitions), and further to 0.2773, 0.3505, and 0.3680 for intervals exceeding 180 days (2,928 transitions). The consistent decline suggests that longer intervals weaken the predictive relevance of the most recent clinical state. The accompanying decrease in candidate recall indicates that reduced candidate coverage, rather than ranking alone, contributes to this difficulty.

F.4 Geometry of the Learned Retrieval Space

We analyze the geometry induced by Risk Horizons to better understand how hierarchical structure and cross-modal temporal associations shape the retrieved hypothesis space.

Global hierarchy and local retrieval structure. Figure 5 compares embeddings learned using hierarchy edges alone versus the full graph with cross-modal PMI stitching. The hierarchy-only embedding forms a predominantly tree-like radial organization characteristic of hyperbolic representations. After introducing PMI edges, the embedding develops localized multimodal neighborhoods while still preserving large-scale hierarchical structure. Rather than collapsing the hierarchy, cross-modal stitching selectively deforms local regions of the space, bringing temporally associated diagnoses, procedures, and medications into closer proximity. This supports the central retrieval objective of Risk Horizons: preserving global semantic organization while constructing clinically coherent local candidate regions.

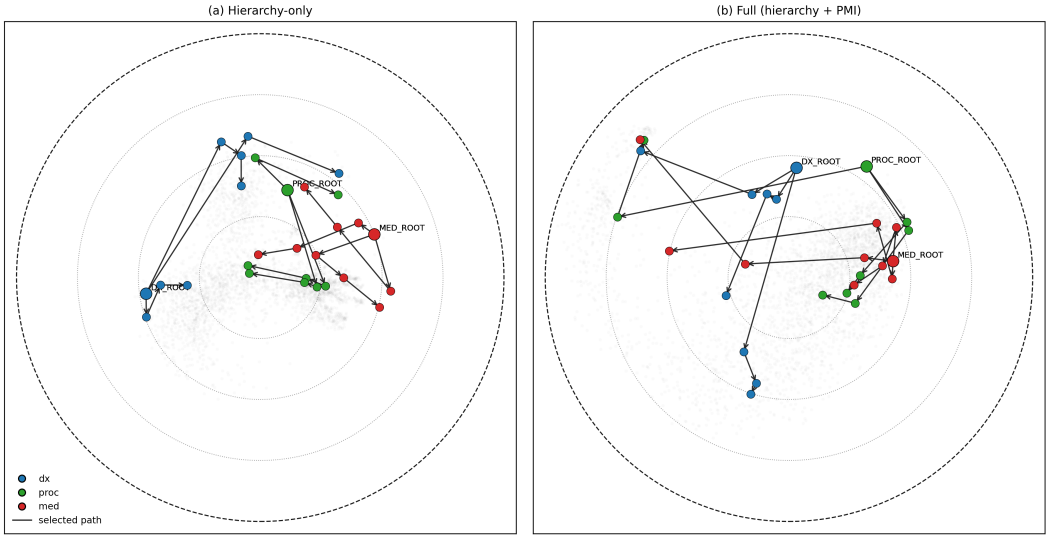


Figure 5: Hierarchy-only vs. full embedding (with PMI edges).

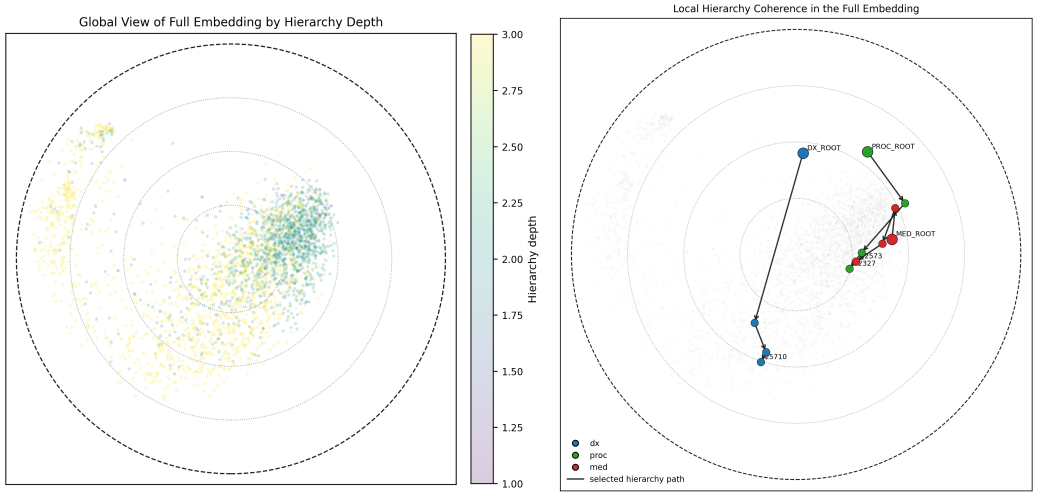


Figure 6: Left: Full embedding colored by hierarchy depth. Right: Selected hierarchical paths in the full embedding.

Figure 6 (left) further shows that hierarchy depth remains approximately organized along the radial direction of the Poincaré disk, with deeper concepts generally positioned closer to the boundary.

Although PMI stitching introduces local deviations from strict tree geometry, the embedding retains substantial hierarchical organization even after multimodal integration.

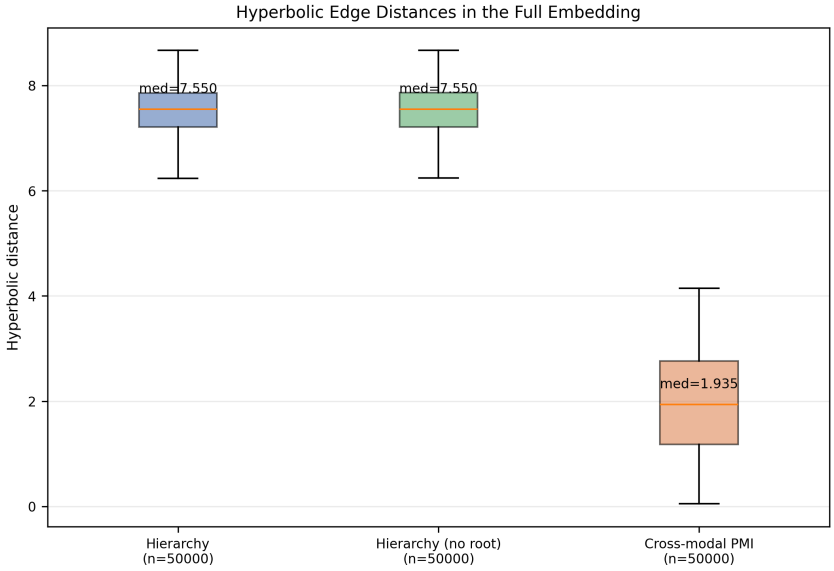


Figure 7: Hyperbolic distance distributions for hierarchy and PMI edges.

Edge-type separation and local hierarchy preservation. Figure 7 shows a strong geometric separation between hierarchy and PMI edges. Hierarchical relations occupy substantially larger hyperbolic distances (median ≈ 7.55), while PMI edges remain highly localized (median ≈ 1.93). This indicates that cross-modal temporal associations primarily act as local retrieval shortcuts rather than globally distorting the hierarchy.

Figure 6 (right) visualizes representative root-to-leaf paths in the full embedding. Despite local deformation induced by cross-modal stitching, hierarchical trajectories remain approximately radial and locally coherent. Together, these observations suggest that Risk Horizons preserves enough global hierarchy for semantic organization while introducing sufficient local flexibility to support clinically meaningful retrieval neighborhoods.

Clinical interpretation of PMI edges. Table 10 presents representative associations retained after PMI-based graph construction and stability filtering. Several edges correspond to established clinical relationships: end-stage renal disease is associated with haemodialysis as renal replacement therapy, acute respiratory failure may require endotracheal intubation, and dopaminergic agents are commonly used to treat Parkinson disease. The association between chemotherapy encounters and chemotherapy infusion reflects a direct clinical workflow relationship. In contrast, the association between hypertension and analgesics/antipyretics is less clinically specific and may reflect common medication use or documentation patterns rather than a direct therapeutic relationship. Lag 0 indicates within-visit associations, whereas lag 1 indicates associations between consecutive visits rather than a fixed elapsed time. Support denotes the number of observed source–target pairs, and stability represents the frequency with which an edge is retained across bootstrap samples. All five examples have a stability of 1.000, indicating consistent selection under the applied filtering procedure. These results illustrate that stability filtering retains both clinically interpretable and less specific associations; the learned edges capture statistical relationships and should not be interpreted as evidence of causality.

G Evaluation Results on eICU

Table 11 reports multi-modal next-visit prediction results on eICU. Overall, the trends remain consistent with MIMIC-IV: structure-aware retrieval substantially improves next-visit prediction,

Table 10: Representative stability-filtered PMI edges in the clinical event graph.

Source → Target	**Lag**	**Support**	**Stability**
ESRD → Haemodialysis	1	1,399	1.000
Acute respiratory failure → Endotracheal intubation	0	1,362	1.000
Dopaminergic agents → Parkinson disease	1	135	1.000
Antineoplastic chemotherapy encounter → Chemotherapy infusion	1	412	1.000
Hypertension → Analgesics/antipyretics	0	14,226	1.000

Table 11: Multi-modal next-visit prediction on eICU (mean ± std over 5 seeds).

Model	**Dx**		**Proc**		**Med**		**Overall**	
	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10	R@10	nDCG@10
Sequential Baselines								
Copy Last	0.273 ± 0.002	0.263 ± 0.003	0.170 ± 0.018	0.175 ± 0.014	0.249 ± 0.021	0.421 ± 0.011	0.231 ± 0.012	0.286 ± 0.009
Transformer (Eucl.)	0.499 ± 0.016	0.412 ± 0.014	0.338 ± 0.015	0.327 ± 0.013	0.295 ± 0.020	0.537 ± 0.016	0.378 ± 0.017	0.425 ± 0.014
BEHRT	0.386 ± 0.007	0.302 ± 0.006	0.257 ± 0.006	0.241 ± 0.008	0.241 ± 0.009	0.462 ± 0.007	0.295 ± 0.007	0.335 ± 0.007
RETAIN	0.505 ± 0.011	0.412 ± 0.012	0.346 ± 0.010	0.326 ± 0.009	0.292 ± 0.013	0.533 ± 0.010	0.381 ± 0.019	0.424 ± 0.010
GRAM	0.335 ± 0.008	0.241 ± 0.007	0.210 ± 0.009	0.194 ± 0.008	0.220 ± 0.012	0.441 ± 0.009	0.255 ± 0.008	0.292 ± 0.008
Med-BERT	0.426 ± 0.009	0.337 ± 0.010	0.288 ± 0.008	0.272 ± 0.009	0.252 ± 0.011	0.480 ± 0.010	0.322 ± 0.009	0.363 ± 0.009
HALO	0.443 ± 0.013	0.353 ± 0.014	0.310 ± 0.011	0.282 ± 0.012	0.261 ± 0.014	0.485 ± 0.012	0.338 ± 0.012	0.373 ± 0.013
LLMs (Top 80 Freq. Events)								
Llama-3.1-8B-Instruct	0.195 ± 0.041	0.289 ± 0.043	0.169 ± 0.032	0.173 ± 0.030	0.241 ± 0.012	0.310 ± 0.024	0.202 ± 0.029	0.257 ± 0.032
Llama-3.3-70B-Instruct	0.291 ± 0.035	0.311 ± 0.051	0.247 ± 0.034	0.211 ± 0.050	0.239 ± 0.033	0.479 ± 0.039	0.259 ± 0.034	0.334 ± 0.044
Mixtral-8x7B-Instruct-v0.1	0.283 ± 0.027	0.297 ± 0.032	0.231 ± 0.025	0.152 ± 0.021	0.199 ± 0.030	0.269 ± 0.026	0.238 ± 0.026	0.239 ± 0.025
Meditron-70B	0.297 ± 0.032	0.312 ± 0.039	0.249 ± 0.039	0.198 ± 0.024	0.247 ± 0.033	0.451 ± 0.020	0.264 ± 0.033	0.320 ± 0.027
Graph & Retrieval								
Euclidean Risk Horizons	0.313 ± 0.006	0.240 ± 0.009	0.183 ± 0.006	0.182 ± 0.010	0.144 ± 0.013	0.258 ± 0.021	0.213 ± 0.008	0.227 ± 0.013
Euc.RH+Llama-3.1-8B	0.187 ± 0.027	0.246 ± 0.014	0.154 ± 0.030	0.201 ± 0.024	0.258 ± 0.022	0.274 ± 0.019	0.200 ± 0.025	0.240 ± 0.019
Hyperbolic Baselines								
Hyperbolic Linear Probe	0.460 ± 0.017	0.367 ± 0.011	0.307 ± 0.009	0.291 ± 0.010	0.259 ± 0.012	0.484 ± 0.014	0.342 ± 0.012	0.381 ± 0.008
Hyperbolic NN Global	0.407 ± 0.008	0.312 ± 0.009	0.241 ± 0.007	0.241 ± 0.018	0.249 ± 0.011	0.429 ± 0.012	0.299 ± 0.008	0.327 ± 0.011
Ours								
Risk Horizons (w/o LLM)	0.462 ± 0.007	0.378 ± 0.007	0.282 ± 0.014	0.271 ± 0.006	0.274 ± 0.005	0.483 ± 0.009	0.339 ± 0.015	0.377 ± 0.004
RH+Llama-3.1-8B-I.	0.469 ± 0.049	0.363 ± 0.023	0.285 ± 0.022	0.287 ± 0.031	0.276 ± 0.036	0.489 ± 0.043	0.343 ± 0.030	0.380 ± 0.034
RH+Llama-3.3-70B-I.	0.519 ± 0.029	0.463 ± 0.019	0.388 ± 0.022	0.333 ± 0.045	0.296 ± 0.031	0.502 ± 0.039	0.401 ± 0.027	0.433 ± 0.030
RH+Meditron-70B	0.506 ± 0.044	0.419 ± 0.043	0.396 ± 0.020	0.327 ± 0.034	0.292 ± 0.036	0.497 ± 0.043	0.398 ± 0.033	0.414 ± 0.039

while constrained reranking provides additional gains when the retrieved candidate space is sufficiently coherent.

Compared to MIMIC-IV, however, the relative gap between geometry-only and reranked variants is smaller. Geometry-only Risk Horizons already performs strongly on diagnoses (0.462 R@10, 0.378 nDCG@10), outperforming both hyperbolic baselines, although remaining below RETAIN and Transformer. Adding constrained reranking still improves overall performance, reaching 0.401 overall R@10 and 0.433 overall nDCG@10 with Llama-3.3-70B-Instruct, but the gains are noticeably smaller than those observed on MIMIC-IV. This suggests that, on eICU, candidate retrieval remains the dominant factor, while downstream reranking becomes comparatively less important.

The behavior of the Euclidean and hyperbolic baselines further supports this interpretation. Euclidean retrieval methods remain substantially weaker, even when paired with LLM reranking, indicating that downstream reasoning alone cannot compensate for poorly structured candidate spaces. At the same time, the gap between Hyperbolic NN Global and full Risk Horizons is narrower than on MIMIC-IV, suggesting that eICU trajectories contain weaker hierarchical and cross-modal structure overall. In this setting, much of the predictive signal can already be recovered from local geometric proximity, reducing the relative advantage of more sophisticated horizon construction.

The sequential baselines also behave differently on eICU. RETAIN performs particularly strongly overall, while Copy Last remains competitive for medications, reflecting stronger short-term event persistence and lower effective vocabulary complexity. LLM-only baselines also improve relative to their MIMIC-IV performance, likely because the eICU prediction space is substantially smaller and less diverse. Nevertheless, they still remain well below structured retrieval methods, reinforcing that prediction with frequency-based rather than patient-specific candidate sets remains difficult even in reduced-vocabulary settings.

These differences are consistent with the underlying dataset characteristics. Compared to MIMIC-IV, eICU contains a substantially smaller and less standardized vocabulary, including synthetic procedure identifiers and non-ATC medication coding. Diagnosis trajectories are also sparser and more heterogeneous across institutions. As a result, the latent hierarchy is weaker, temporal transitions

are noisier, and clinically coherent future states are more difficult to organize into sharply separated Risk Horizons. Under these conditions, geometry-aware retrieval remains beneficial, but the additional gains from constrained reranking become smaller because the retrieved candidate space itself is less structured.

Table 12 further supports this interpretation. Risk Horizons consistently improves hierarchy distance, candidate recall, and Risk-Horizon Recall across all modalities relative to Hyperbolic NN Global, indicating that the retrieved horizons remain more structurally coherent even in the noisier eICU setting. However, the absolute gaps are smaller than on MIMIC-IV, especially for Ancestor Match and Oracle@10, again suggesting that eICU contains weaker recoverable hierarchy structure overall. Importantly, Oracle@10 remains substantially higher than the achieved Recall@10 across modalities, indicating that ranking within the retrieved horizon remains a partially unsolved problem even when candidate retrieval is strong.

Table 12: Hierarchy consistency and candidate-space quality on eICU (mean $\pm$ std over 5 seeds).

Model	**Type**	Hier. Dist $\downarrow$	Anc. Match $\uparrow$	Cand. Rec. $\uparrow$	RH@10 $\uparrow$	RH@40 $\uparrow$	Oracle@10 $\uparrow$
Hyper. NN Global	Dx	1.337 ± 0.014	0.978 ± 0.003	0.870 ± 0.008	0.407 ± 0.011	0.707 ± 0.010	0.863 ± 0.006
	Proc	1.537 ± 0.019	0.956 ± 0.005	0.660 ± 0.013	0.241 ± 0.010	0.458 ± 0.014	0.641 ± 0.012
	Med	1.902 ± 0.027	0.986 ± 0.002	0.740 ± 0.011	0.249 ± 0.009	0.519 ± 0.013	0.586 ± 0.010
Risk Horizons	Dx	1.295 ± 0.012	0.980 ± 0.002	0.887 ± 0.007	0.452 ± 0.010	0.724 ± 0.009	0.890 ± 0.005
	Proc	1.493 ± 0.017	0.959 ± 0.004	0.694 ± 0.012	0.269 ± 0.011	0.486 ± 0.013	0.667 ± 0.011
	Med	1.828 ± 0.024	0.984 ± 0.003	0.766 ± 0.010	0.271 ± 0.008	0.540 ± 0.012	0.599 ± 0.009

G.1 Hyperparameter Sensitivity on eICU.

Figure 8 reports single-seed sensitivity analyses on eICU, transferring the corresponding MIMIC-IV configurations without dataset-specific retuning. This setting evaluates the transferability of the hyperparameter choices across institutions and avoids optimizing the framework separately for each dataset. Performance is relatively stable across $H \in \{1, 4, 8, 12\}$, with R@10 ranging from 0.3147 to 0.3201 and nDCG@10 from 0.3619 to 0.3656, suggesting limited benefit from additional heads. In contrast, graph construction parameters exhibit greater sensitivity. PMI sensitivity is non-monotonic: $\tau = 0$ achieves R@10/nDCG@10 of 0.1288/0.1487, compared with 0.0983/0.1040 at $\tau = 0.05$. Within the 64-bootstrap stability sweep, increasing q from 0.5 to 0.9 improves performance from 0.0923/0.0931 to 0.1565/0.2037. Cone sensitivity is also non-monotonic, with $\phi = 0.8$ yielding the highest scores among the tested values (0.2876/0.3225). Here, ϕ parameterizes the cone aperture through a cosine similarity threshold, so larger values impose narrower directional constraints.

Compared with MIMIC-IV, these patterns suggest that eICU may be more sensitive to the selection of graph associations and directional constraints than to the expressiveness of multi-head visit representations. One possible explanation is that differences in coding distributions, visit structure, and cross-institutional documentation practices alter the density and reliability of inferred temporal associations. Under such conditions, stricter stability filtering or narrower cones may suppress weakly supported candidates, although this mechanism has not been directly verified. Importantly, the sweeps use different fixed configurations, so absolute scores should be compared within, rather than across, panels.

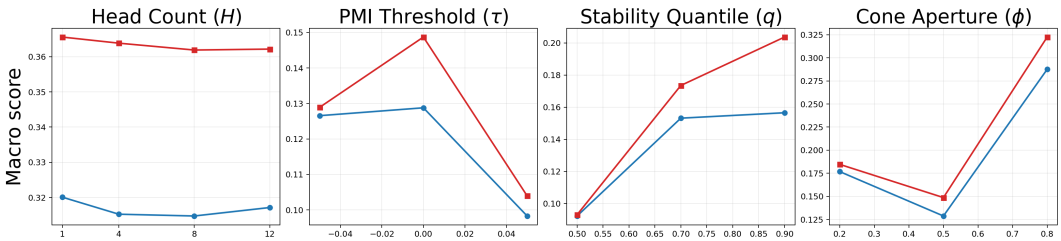


Figure 8: Hyperparameter sensitivity on eICU. Recall@10 (blue) and nDCG@10 (red). From left to right: (a) number of heads H, (b) PMI Threshold τ, (c) Stability q, (d) cone aperture ϕ.

G.2 Embedding Geometry on eICU

We visualize the learned hyperbolic embeddings on eICU to examine whether the geometric patterns observed on MIMIC-IV generalize across datasets with a weaker hierarchical structure.

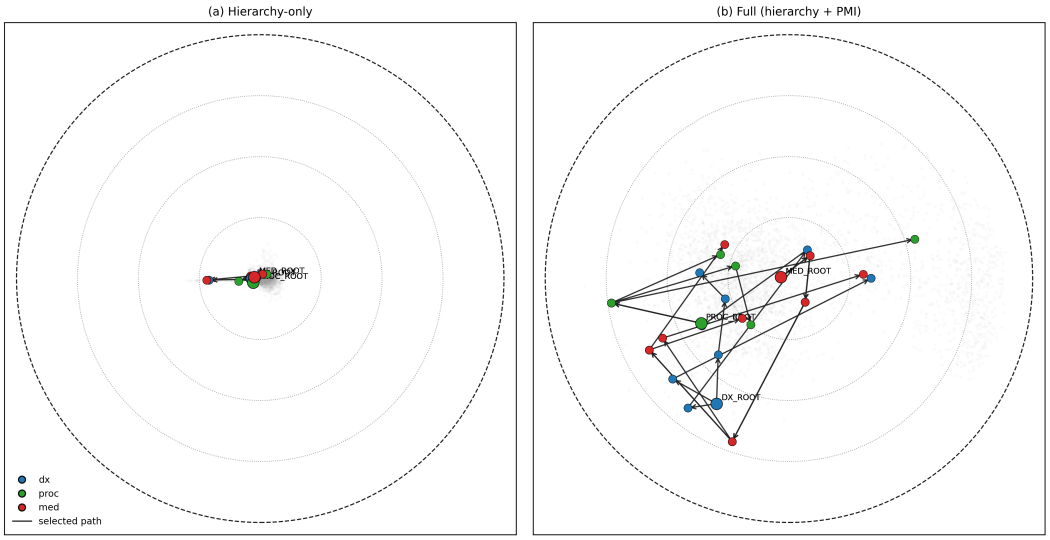


Figure 9: Hierarchy-only versus full hyperbolic embedding on eICU. Adding cross-modal PMI edges substantially improves geometric separation and local organization.

Figure 9 compares embeddings learned using hierarchy edges alone versus the full graph with cross-modal PMI stitching. Similar to MIMIC-IV, the hierarchy-only embedding exhibits substantial geometric collapse, with most nodes concentrated near the origin and limited angular separation between branches. Adding PMI-based cross-modal edges produces a substantially more dispersed organization, introducing local clustering and improving separation across modalities and clinical contexts.

The effect is more pronounced on eICU than on MIMIC-IV, consistent with the weaker ontology structure of eICU, including synthetic procedure identifiers and less standardized medication coding. In this setting, hierarchy alone provides insufficient geometric constraints, while cross-modal temporal associations become critical for recovering meaningful latent structure.

Figure 10 further shows that hierarchy and PMI edges remain geometrically separated in the learned space. Hierarchy edges occupy substantially larger hyperbolic distances than PMI edges, indicating that the embedding preserves a distinction between global hierarchical organization and local associative structure even under noisier coding systems.

H Limitations and Ethical Considerations

Limitations. Risk Horizons remains subject to several limitations inherent to longitudinal EHR modeling and structure-aware retrieval. First, clinical records are noisy, incomplete, and affected by coding variability across institutions and time, which can introduce inconsistencies in both hierarchical relations and temporal co-occurrence statistics.

Second, the framework depends on retrieval quality and candidate-space coverage. If clinically relevant future events are absent from the retrieved Risk Horizon, downstream reranking cannot recover them, regardless of the strength of the inference model. This dependence is particularly important under sparse supervision or weak ontology structure, as observed in eICU.

Third, the approach depends on predefined coding systems and hierarchical organization. While hyperbolic geometry effectively captures structured relations, the quality of the learned embedding is influenced by the consistency and granularity of the underlying ontology. Datasets with synthetic or weakly standardized coding systems may therefore provide weaker geometric supervision.

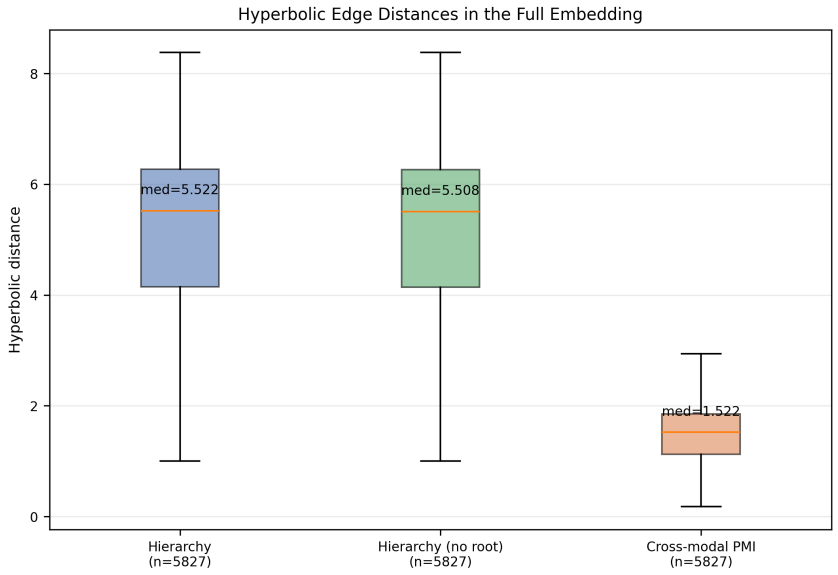


Figure 10: Hyperbolic distance distributions on eICU. PMI edges remain substantially more local than hierarchy edges in the learned space.

Fourth, although LLM-based reranking improves predictive performance, it also introduces additional computational cost and potential variability across model backends. The proposed framework is intended as a research system for structured clinical reasoning rather than a deployable clinical decision-making tool, and its outputs should not be interpreted as medical recommendations without expert oversight and external validation.

Finally, Risk Horizons also inherits limitations of candidate-space retrieval: highly novel events or events arising from coding drift may fall outside the retrieved horizon. Addressing these cases may require adaptive retrieval expansion or dynamically updated retrieval structures under distributional shift.

Ethical Considerations. The use of EHR data raises important ethical concerns, including patient privacy, demographic bias, and potential misuse of predictive models. Although we use publicly available de-identified datasets, biases present in the underlying healthcare systems may still be reflected in the learned representations and predictions. Performance disparities may therefore emerge across demographic groups, institutions, or rare clinical conditions. We do not conduct subgroup-specific fairness evaluations in this study, leaving the assessment of demographic performance disparities to future work.

Risk Horizons aims to improve transparency and reduce unconstrained hallucination through structure-aware retrieval and constrained candidate generation. However, these mechanisms do not eliminate bias, guarantee fairness, or ensure clinical correctness. Any real-world deployment would require rigorous external validation, subgroup fairness analysis, prospective evaluation, and compliance with clinical and regulatory standards.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by the methodological and experimental results presented throughout the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and ethical considerations are discussed in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not contain formal theoretical results or proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setup, datasets, baselines, implementation details, and evaluation protocols are described in Section 4 and Appendices B–E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A public code repository is linked in the Introduction, and reproduction instructions are provided with the released code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendix Implementation Details

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables report mean ± std over 5 seeds; described in Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational resources, including GPU hardware, training configurations, and inference latency, are documented in the experimental setup and appendix E. Table 7 reports inference latency for the evaluated models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Ethical considerations discussed in Appendix.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Potential societal impacts, risks, and ethical considerations are discussed in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No release of high-risk models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, pretrained models, and external assets are properly cited and used in accordance with their respective licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used as optional inference-time rerankers over structured candidate spaces. Their integration is described in Section 3.3, with implementation details and evaluation provided in the appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.