Discrete Neural Interlingua for Ontology-Agnostic EHR Modeling

Tong Ding[1,2†] **Caiwei Tian**[1,2†] **Dongmin Bang**[1,2] **Ming Y. Lu**[1,2] **Sophia J. Wagner**[1,2]
Andrew Zhang[1,2,3] **Rowland Pettit**[1,2] **Long Phi Le**[1,2] **Faisal Mahmood**[1,2]
[1]Mass General Brigham AI Institute [2]Harvard University
[3]Massachusetts Institute of Technology

Abstract

EHR foundation models promise scalable clinical prediction across health systems, yet their deployment is constrained by a tokenization bottleneck. The same clinical concept can appear as different ontology codes, local identifiers, synonyms, or free-text descriptions, which standard EHR tokenizers treat as unrelated symbols. A model trained on MIMIC-III with a closed-vocabulary tokenizer can use only 13% of eICU events. We introduce MEDIATOR, a discrete ontology-agnostic tokenizer that maps any clinical concept expressed as text into a shared vector-quantized codebook, turning EHR harmonization into open-vocabulary tokenization without manual mappings or a fixed vocabulary. Trained with synonym supervision and guided by enriched concept descriptions, code co-occurrence, and relation graphs, MEDIATOR collapses equivalent expressions while organizing distinct concepts in a medically meaningful discrete space. Used as the token interface of an EHR Transformer, MEDIATOR improves mean AUPRC over the strongest prior tokenizer by 4.0 points (8% relative) on 20 MIMIC-III tasks and by 8.6 points (29% relative) in zero-shot transfer to 14 eICU tasks, where it uses all events. The transfer gain combines coverage and representation, as MEDIATOR remains the best tokenizer under identical inputs (+1.0 and +1.9 points on operational and subcohort-mortality tasks). Same-encoder controls show that the text encoder drives in-domain accuracy, while the learned objectives of MEDIATOR drive transfer robustness and entity-linking quality. MEDIATOR also enables transferable attribution across hospitals. Code is available at `https://github.com/mahmoodlab/MEDIATOR`.

1 Introduction

Electronic health records (EHRs) are central to clinical prediction, real-world evidence generation, cohort construction, and decision support. Transformer-based EHR models have achieved strong results by learning from longitudinal sequences of diagnoses, procedures, medications, laboratory events, and other clinical observations [Shmatko et al., 2025, Waxler et al., 2025, Li et al., 2020, Wornow et al., 2023a, Steinberg et al., 2021]. These models promise to support multi-institutional learning and deployment across health systems. Yet this promise is constrained by a representation problem. Clinical concepts are recorded differently across coding systems, ontology versions, local institutional practice, and free-text documentation. A model trained on MIMIC-III [Johnson et al., 2016a,b] with a closed-vocabulary tokenizer can use only 13% of eICU events [Pollard et al., 2018, 2019], since the rest use other codes or free text. As a result, many EHR models learn from an input space that is fragmented, redundant, and difficult to transfer across settings (Figure 1).

Most existing EHR foundation models tokenize structured data by treating each medical code as an independent symbol with a learned embedding [Wornow et al., 2023b]. This design scales poorly to

[†]Equal contribution.

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

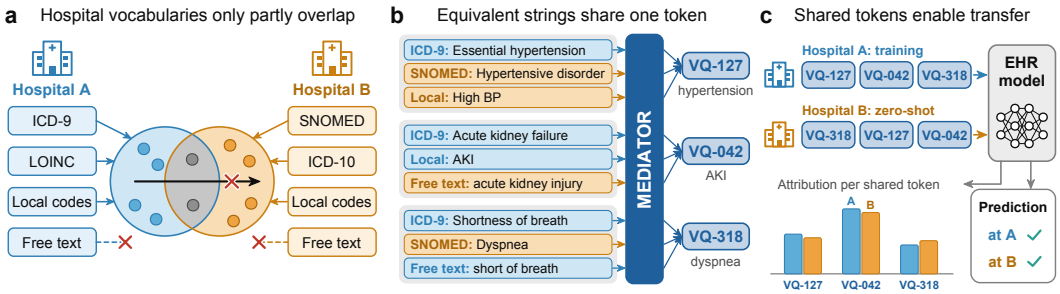


Figure 1: **Motivation for MEDIATOR. a**. The vocabularies of Hospital A (light blue) and Hospital B (orange) only partly overlap, so models trained at A fail on B-specific codes and drop free text. **b**. MEDIATOR maps equivalent surface forms from both hospitals, colored by source hospital, to one shared vector-quantized (VQ) token (dark blue). **c**. Shared tokens let one EHR model transfer across hospitals, with predictions and attributions defined over the same tokens.

large, heavy-tailed clinical vocabularies and fails to share statistical strength across related concepts [Su et al., 2025]. Code-level tokens also discard informative structure. Each code is defined by a textual description, belongs to an ontology hierarchy, and is linked to many synonyms and near-synonyms across other coding systems [Johnson et al., 2026]. Ignoring these semantics leads to inefficient vocabularies and fragmented representations, especially for rare or institution-specific codes. This brittleness is most visible across institutions. Hospitals differ in the ontologies they use, which versions they deploy, and how they encode local variants of common concepts. When a model is trained with one hospital's code vocabulary, deploying it elsewhere typically requires manual code mapping, vocabulary alignment, and coverage audits. Such mapping is expensive, error-prone, and often incomplete in the long tail. As a result, most models avoid clinical text entirely [Wornow et al., 2023b] and restrict modeling to a narrow subset of structured data such as diagnostic codes, labs, and medications, sacrificing clinical signal for modeling convenience.

Previous studies have attempted to address this heterogeneity from several complementary angles. Ontology-normalization approaches map events to a common vocabulary such as the Unified Medical Language System (UMLS) [Bodenreider, 2004, Wu et al., 2026a]; representation-learning approaches enrich fixed code vocabularies with text descriptions and medical knowledge graphs, and LLM-generated definitions [Su et al., 2025, Kim et al., 2026]; text-based embedding methods represent codes directly in continuous semantic space [Hur et al., 2022]; and multi-institutional harmonization frameworks align site-specific vocabularies using EHR summary statistics, biomedical graphs, and language-model semantics [Zhou et al., 2026]. These efforts improve semantic coverage and reduce reliance on manual pairwise code mapping, but they do not fully address the modeling interface required for scalable EHR sequence modeling. First, UMLS defines a massive and evolving vocabulary. It aggregates multiple source vocabularies, introducing redundancy and fine-grained distinctions that are often unnecessary for downstream prediction. Consequently, using UMLS identifiers directly as tokens reintroduces the problems of scale, sparsity, and redundancy inherent in raw code vocabularies [Su et al., 2025]. Fixed-vocabulary code embeddings remain closed over a prespecified set of structured codes, and site-alignment methods generally must be rerun when a new hospital, ontology, or local code system is introduced. Continuous embeddings are more flexible, but they do not provide the stable discrete interface used by standard masked-token and autoregressive EHR Transformers, which also complicate event-level inspection, attribution, caching, and generative trajectory modeling. A useful EHR tokenizer should therefore be both discrete and open-vocabulary. It should accept any medical concept expressed as text and map it into a shared, reusable code space.

To this end, we introduce MEDical InterlinguA TokenizeR (MEDIATOR), a discrete and ontology-agnostic concept space that serves as a neural interlingua. MEDIATOR takes as input any clinical concept expressible as text (including ontology code descriptions, local code descriptions, and uncoded text entries) and maps it to a discrete index from a fixed-size vector-quantized (VQ) codebook, in which each code is a semantic prototype with an embedding. Leveraging strong pretrained text encoders, we learn this mapping from large-scale cross-ontology synonym supervision, treating each underlying concept as an equivalence class of surface forms. To better shape the representation geometry, we further guide codebook learning with three complementary sources of medical knowledge including enriched concept descriptions, empirical code co-occurrence, and curated relation-graph structure. The resulting codebook defines a compact, reusable clinical vocabulary that

can be used as a drop-in token interface for standard EHR Transformers, improving interoperability and interpretability while avoiding fragile ontology-specific tokenization and manual code mapping.

We show that MEDIATOR is effective in settings that stress current EHR modeling. As the token interface of an EHR Transformer, MEDIATOR improves mean AUPRC over the strongest prior tokenizer by 4.0 points (8% relative) on 20 MIMIC-III tasks and by 8.6 points (29% relative) in zero-shot transfer to 14 eICU tasks without manual vocabulary alignment, and remains the best tokenizer when input coverage is matched to closed-vocabulary baselines. Controlled analyses decompose these gains. The transfer gain combines broader input coverage with better representations, the pretrained text encoder drives in-domain accuracy, and the learned objectives of MEDIATOR drive transfer robustness and entity-linking quality. Finally, the shared discrete space supports transferable interpretability across hospitals through common concept tokens.

2 Related work

EHR models. Transformer-based EHR models adapt the language-modeling paradigm to longitudinal sequences of structured clinical events. Bidirectional embedding models learn patient representations with masked-event or contextual prediction objectives, starting from diagnosis-focused token sequences and later expanding to medications, procedures, temporal features, and discretized numeric observations [Li et al., 2020, Rasmy et al., 2021, Pang et al., 2021, Wornow et al., 2023a], and graph-based variants model intra-visit structure [Poulain and Beheshti, 2024]. Autoregressive and encoder–decoder models use generative objectives to model patient trajectories and support forecasting, time-to-event prediction, and generative medical-event modeling [Steinberg et al., 2021, 2024, Kraljevic et al., 2024, Shmatko et al., 2025, Waxler et al., 2025, Yang et al., 2023], and toolkits such as PyHealth 2.0 improve their reproducibility [Wu et al., 2026b]. These studies underline the importance of tokenized EHR data, but concept embeddings learned from scratch or initialized from external resources remain tied to a closed vocabulary and transfer poorly across institutions.

Cross-hospital transfer. Improving transfer across hospitals requires resolving vocabulary mismatch from different ontologies, local code systems, and documentation practices. One strategy is to normalize structured events to a shared vocabulary before model training, such as UMLS [Bodenreider, 2004, Wu et al., 2026a] or the standard concepts of the OMOP common data model maintained by OHDSI [Overhage et al., 2012, Hripcsak et al., 2015], which are largely mapped by hand. This inherits the scale, redundancy, and incomplete coverage of the target vocabulary and requires ongoing curation. A second strategy learns continuous code representations from textual descriptions, or enriches a fixed structured vocabulary with biomedical knowledge graphs and LLM-generated concept descriptions [Hur et al., 2022, Su et al., 2025, Kim et al., 2026]. These representations improve semantic sharing among known codes, but they remain tied to a prespecified vocabulary or yield continuous event vectors that lack a discrete token interface. A third strategy aligns vocabularies across institutions using site-level summary statistics, biomedical graphs, and language-model semantics [Zhou et al., 2026]. However, the alignment is specific to the participating institutions and must be recomputed for new hospitals or unseen code systems. MEDIATOR learns an open-vocabulary text-to-token mapping that complements manual OMOP mapping. Any clinical concept expressible as text can be mapped into a shared discrete codebook and then reused by standard EHR sequence models.

Vector-quantized tokenization. Vector-quantized autoencoders such as VQ-VAE [van den Oord et al., 2017] learn a finite codebook of latent prototypes and discretize inputs by nearest-neighbor assignment. Such tokenizers now link perceptual encoders and sequence models in vision, audio, and multimodal learning [Esser et al., 2021, Zeghidour et al., 2022, Du et al., 2024]. In healthcare, VQ-based discretization has been used for time-series tokenization and ontology-aware token learning over structured medical codes [Lee et al., 2024, Su et al., 2025], but these tokenizers are trained on a closed set of observed codes. MEDIATOR provides an open-vocabulary text-to-token interface that maps unseen code descriptions or free-text concept expressions into the same discrete interlingua.

3 Methods

3.1 Problem definition

Let $\mathcal{C}$ be a universe of underlying clinical concepts. Each concept $c \in \mathcal{C}$ can be expressed by multiple textual descriptions, including ontology descriptions, local code descriptions, synonyms, and uncoded

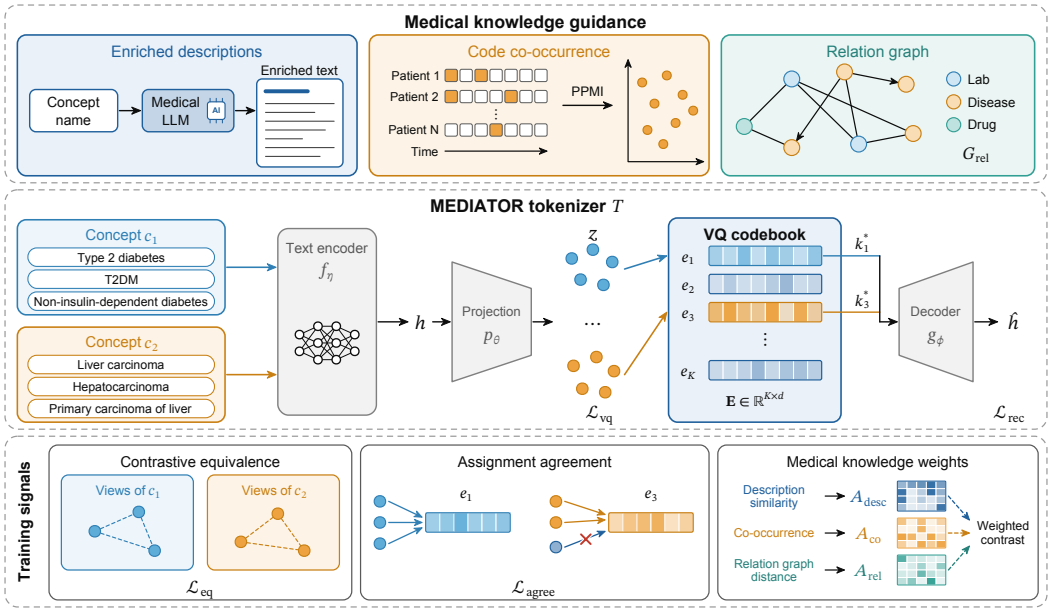


Figure 2: **MEDIATOR architecture.** A text encoder maps any clinical concept description to a dense embedding, a trainable adaptor projects it into the quantization space, and the result is assigned to a learned VQ codebook. Training combines same-concept contrastive objectives before and after quantization, multi-view assignment agreement, semantic teachers from enriched descriptions, code co-occurrence, and relation graphs, and reconstruction to the encoder embedding space.

text entries. We model these views as $\mathcal{X}_c = \{x_{c,1}, \ldots, x_{c,|\mathcal{X}_c|}\} \subset \mathcal{X}$, where $\mathcal{X}$ is the space of text strings. Our goal is to learn a discrete tokenizer $T : \mathcal{X} \to [K] = \{1, \ldots, K\}$ such that descriptions of the same concept map to the same discrete code, while different concepts remain separated as much as possible. Figure 2 gives an overview of the tokenizer (middle panel), its training signals (bottom panel), and the medical knowledge that guides training (top panel).

3.2 Vector-quantized tokenization

Given $x \in \mathcal{X}$, a text encoder produces a semantic embedding $h = f_\eta(x) \in \mathbb{R}^{d_e}$, and a projection head maps it to the quantization space $z = p_\theta(h) \in \mathbb{R}^d$. We maintain a codebook $\mathbf{E} = \{e_k\}_{k=1}^K$ with $e_k \in \mathbb{R}^d$. Using cosine similarity, the assigned code is

$$k^\star(z) = \arg\max_{k \in [K]} \cos(z, e_k), \qquad q(z) = e_{k^\star(z)}. \tag{1}$$

The tokenizer output is the discrete index $T(x) = k^\star(p_\theta(f_\eta(x)))$.

We update codebook statistics using exponential moving averages (EMA), following standard VQ practice [van den Oord et al., 2017, Razavi et al., 2019]. Let n_k be the number of assignments to code k in a minibatch and $m_k = \sum_{i:k^\star(z_i)=k} z_i$ be the sum of assigned vectors. We maintain EMA statistics (N_k, M_k) and update

$$N_k \leftarrow \gamma N_k + (1-\gamma) n_k, \qquad M_k \leftarrow \gamma M_k + (1-\gamma) m_k, \qquad e_k \leftarrow M_k / (N_k + \varepsilon), \tag{2}$$

with decay $\gamma \in (0,1)$ and small $\varepsilon > 0$. We use the commitment objective $\mathcal{L}_{\text{vq}} = \frac{1}{N} \sum_{i=1}^N \|z_i - \text{sg}(q(z_i))\|_2^2$, where $\text{sg}(\cdot)$ stops gradients.

To avoid permanent codebook death, we periodically reinitialize under-used codes from the empirical distribution of current encoder outputs, following dead-code revival in neural audio tokenization [Zeghidour et al., 2022]. For losses applied after quantization, we use the rotation trick [Fifty et al., 2025]. Let $\hat{z}$ denote the surrogate quantized representation used for backpropagation. Its forward value is $q(z)$, while its approximate Jacobian rotates gradients as

$$\frac{\partial \hat{z}}{\partial z} \approx \alpha R(z, q(z)), \qquad \nabla_z \mathcal{L} \approx \alpha R(z, q(z))^\top \nabla_{\hat{z}} \mathcal{L}, \tag{3}$$

where $R(\cdot,\cdot)$ is an orthonormal map from the direction of z to $q(z)$ and $\alpha > 0$ rescales magnitudes.

Finally, a decoder g_ϕ reconstructs the encoder embedding with $\mathcal{L}_{\text{rec}} = 1 - \frac{1}{N}\sum_{i=1}^{N}\cos(g_\phi(\hat{z}_i), h_i)$ to preserve clinically relevant semantics while discouraging code collapse [Qu et al., 2025].

3.3 Equivalence supervision from multiple views

The primary supervision signal is concept equivalence. All views of the same underlying concept should be interchangeable at the tokenizer interface. Given a minibatch of N views with concept labels c_i, let $\mathcal{P}(i) = \{j \neq i : c_j = c_i\}$ be the set of other views of the same concept. For a representation r_i, define

$$\pi_{ij}(r) = \frac{\exp(\cos(r_i, r_j)/\tau)}{\sum_{\ell \neq i}\exp(\cos(r_i, r_\ell)/\tau)}. \tag{4}$$

We use a multi-positive contrastive loss [Khosla et al., 2020]

$$\mathcal{L}_{\text{eq}}(r) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\log\sum_{j\in\mathcal{P}(i)}\pi_{ij}(r), \tag{5}$$

where $\mathcal{I} = \{i : |\mathcal{P}(i)| > 0\}$. This loss is applied before VQ to organize the continuous space, $\mathcal{L}_{\text{eq}}^{\text{pre}} = \mathcal{L}_{\text{eq}}(\psi_{\text{pre}}(z))$, and after quantization to act directly on the discrete interface, $\mathcal{L}_{\text{eq}}^{\text{post}} = \mathcal{L}_{\text{eq}}(\psi_{\text{post}}(\hat{z}))$.

We further align the assignment distributions of views from the same concept. Let $p_i(k) = \exp(\cos(z_i, e_k)/\tau_a)/\sum_{\ell=1}^{K}\exp(\cos(z_i, e_\ell)/\tau_a)$ be the soft code-assignment distribution. For concept c in a minibatch, let $\mathcal{V}_c = \{i : c_i = c\}$ and $\bar{p}_c = \frac{1}{|\mathcal{V}_c|}\sum_{i\in\mathcal{V}_c} p_i$. The assignment-agreement loss, a generalized Jensen–Shannon objective over code assignments, is

$$\mathcal{L}_{\text{agree}} = \frac{1}{|\mathcal{C}_B|}\sum_{c\in\mathcal{C}_B}\frac{1}{|\mathcal{V}_c|}\sum_{i\in\mathcal{V}_c}\text{KL}(p_i \,\|\, \bar{p}_c), \tag{6}$$

where $\mathcal{C}_B$ is the set of concepts represented in the minibatch.

3.4 Medical knowledge as semantic teachers

The discrete tokenizer is the desired interface, but learning a high-quality codebook benefits from a well-structured representation space before the VQ bottleneck. We therefore use medical knowledge to define soft semantic teachers over non-equivalent concepts.

We use three categories of medical knowledge. First, *LLM-generated enriched descriptions* provide definitional knowledge about what a concept means. Let y_c denote an enriched description for concept c and let $u_c = \rho(y_c)/\|\rho(y_c)\|_2$ be its normalized embedding. We define

$$A_{\text{desc}}(c_i, c_j) = \exp\left(\frac{\max(0, u_{c_i}^\top u_{c_j})}{\tau_{\text{desc}}}\right) - 1. \tag{7}$$

Second, *code co-occurrence* provides real-world evidence about which concepts appear in similar clinical contexts. We define co-occurrence over a set of contexts, where each context corresponds to a patient and a fixed-width time window. Let $B \in \{0,1\}^{D\times|\mathcal{C}|}$ be the binary context–concept matrix, where $B_{tc} = 1$ if concept c appears in context t. The sparse co-occurrence matrix is $C = B^\top B$, and the marginal context count for concept i is $n_i = C_{ii}$. We form a raw positive pointwise mutual information (PPMI) matrix [Beam et al., 2019]

$$M_{ij} = \begin{cases}\max\left(\log_2\frac{D\,C_{ij}}{n_i n_j},\, 0\right), & C_{ij} > 0,\\ 0, & C_{ij} = 0.\end{cases} \tag{8}$$

We define $A_{\text{co}}(c_i, c_j) = \log(1 + M_{c_i c_j})/\tau_{\text{co}}$, which preserves empirical usage evidence while damping very large PPMI values. Third, the *relation graph* captures curated medical relations. Given a graph G_{rel} over concepts and shortest-path distance d_{rel}, we define

$$A_{\text{rel}}(c_i, c_j) = \begin{cases}\exp(-d_{\text{rel}}(c_i, c_j)/\tau_{\text{rel}}), & d_{\text{rel}}(c_i, c_j) \leq R,\\ 0, & \text{otherwise.}\end{cases} \tag{9}$$

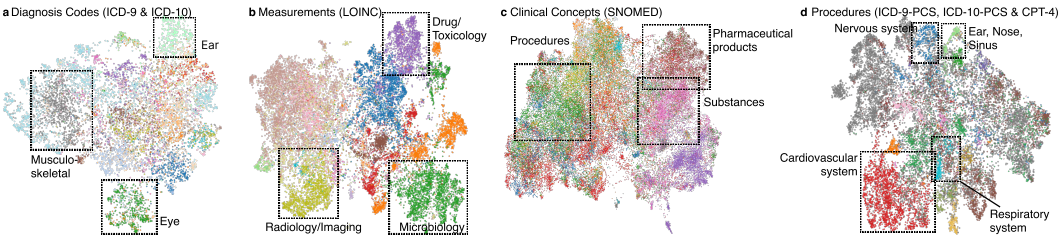


Figure 3: **UMAP visualization of the MEDIATOR codebook.** Each point is a source concept, shown by its assigned VQ code embedding projected with UMAP. Colors indicate independent ontology-derived categories for (a) diagnosis codes, (b) measurements, (c) SNOMED clinical concepts, and (d) procedures. Dashed boxes mark example categories. Full category legends are in Appendix E.

Let $\mathcal{M} = \{\text{desc}, \text{co}, \text{rel}\}$. For each teacher $m \in \mathcal{M}$, we convert affinities in a minibatch into a row-normalized target distribution over non-equivalent concepts

$$\tilde{w}_{ij}^{(m)} = \frac{\mathbf{1}[c_i \neq c_j] A_m(c_i, c_j)}{\sum_{\ell \neq i} \mathbf{1}[c_i \neq c_\ell] A_m(c_i, c_\ell)}. \tag{10}$$

Rows with zero total affinity are omitted, and sparse teachers keep only the largest affinities per row. With $\pi_{ij}^{\text{pre}} = \pi_{ij}(\psi_{\text{pre}}(z))$, the teacher loss is

$$\mathcal{L}_m = -\frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} \sum_{j \neq i} \tilde{w}_{ij}^{(m)} \log \pi_{ij}^{\text{pre}}. \tag{11}$$

Same-concept objectives determine what must collapse, and the semantic teachers determine how distinct concepts should be arranged before quantization.

3.5 Training objective and inference

The final objective, combining vector quantization, semantic reconstruction, same-concept equivalence, assignment agreement, and the semantic teachers, is

$$\mathcal{L} = \lambda_{\text{pre}} \mathcal{L}_{\text{eq}}^{\text{pre}} + \lambda_{\text{post}} \mathcal{L}_{\text{eq}}^{\text{post}} + \lambda_{\text{vq}} \mathcal{L}_{\text{vq}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{agree}} \mathcal{L}_{\text{agree}} + \sum_{m \in \mathcal{M}} \lambda_m \mathcal{L}_m. \tag{12}$$

At inference time, MEDIATOR is applied as an offline preprocessing step to unique clinical entity strings, and the resulting discrete token IDs $T(x)$ are cached and reused for all patient sequences. After this one-time tokenization step, downstream EHR models operate on discrete token IDs in the same way as standard code-embedding models. Because the codebook is fixed, new code systems such as ICD-11 can be tokenized without retraining, as in our eICU and entity-linking experiments (Appendix J); novel clinical distinctions remain untested.

4 Experiments

MEDIATOR pretraining dataset. We pretrain MEDIATOR on a concept-view corpus derived from UMLS 2025AB [Bodenreider, 2004]. We first merge CUIs linked by synonym relations into concept groups. All English, unsuppressed names and source descriptions within a group are positive views of the same concept. The corpus spans the International Classification of Diseases (ICD-9 [World Health Organization, 1988] and ICD-10 [World Health Organization, 2016]), SNOMED CT US [SNOMED International, 2025], LOINC [Forrey et al., 1996], RxNorm [National Library of Medicine, 2025], CPT [American Medical Association, 2026], HCPCS [Centers for Medicare & Medicaid Services, 2026], CVX [Centers for Disease Control and Prevention, 2026], ATC [WHO Collaborating Centre for Drug Statistics Methodology, 2026], and MeSH [National Library of Medicine, 2026]. We also align LLM-enriched descriptions, a PPMI matrix from EHR code co-occurrence, and a medical relation KG to the same concept groups. Details are in Appendix B.1.

Biomedical entity linking (BEL) benchmarks. We use three curated BEL datasets, NCBI Disease [Doğan et al., 2014], MedMentions [Mohan and Li, 2019], and COMETA [Basaldella et al., 2020],

Table 1: EHR prediction performance on MIMIC-III, reported as mean (std) AUPRC across folds. Across the 20 tasks, MEDIATOR improves over MedTok (one-sided Wilcoxon signed-rank test, $p = 9.5 \times 10^{-7}$). Patient-level bootstrap tests (5,000 resamples) show significant gains ($p < 0.05$) on 7/20 tasks over MedTok and 17/20 tasks over Code Emb.

Task	Code Emb	ClinVec	MedTok	$\text{MedTok}_{\text{text}}$-top5	$\text{MedTok}_{\text{text}}$-top1	MEDIATOR
Operational tasks						
LOS <7	0.819(0.012)	0.807(0.004)	0.827(0.007)	**0.830(0.005)**	0.762(0.011)	0.829(0.007)
ICU Mortality	0.571(0.015)	0.484(0.006)	0.556(0.008)	0.563(0.007)	0.418(0.014)	**0.573(0.010)**
Readmission	0.336(0.026)	0.342(0.018)	0.366(0.021)	0.347(0.021)	0.336(0.022)	**0.374(0.023)**
Mean	0.576	0.544	0.583	0.580	0.505	**0.592**
Subcohort mortality tasks						
Acute Kidney Injury	0.563(0.032)	0.513(0.025)	0.573(0.014)	0.572(0.024)	0.436(0.013)	**0.587(0.013)**
Acute MI	0.376(0.071)	0.397(0.037)	0.412(0.021)	0.404(0.045)	0.337(0.035)	**0.467(0.056)**
Acute Pancreatitis	0.263(0.053)	0.307(0.084)	0.359(0.074)	0.311(0.054)	0.234(0.061)	**0.396(0.082)**
Atrial Fibrillation	0.489(0.025)	0.416(0.027)	0.473(0.041)	0.482(0.023)	0.383(0.051)	**0.497(0.012)**
ARF/ARDS	0.542(0.028)	0.520(0.025)	0.585(0.005)	0.586(0.018)	0.497(0.027)	**0.603(0.030)**
Cirrhosis	0.291(0.072)	0.355(0.135)	0.358(0.150)	0.224(0.098)	0.195(0.044)	**0.444(0.120)**
Heart Failure	0.426(0.026)	0.362(0.031)	0.403(0.030)	0.430(0.026)	0.305(0.023)	**0.447(0.050)**
Liver Failure	0.564(0.055)	0.577(0.061)	0.599(0.029)	0.566(0.083)	0.470(0.073)	**0.615(0.024)**
GI Bleeding	0.350(0.052)	0.398(0.049)	0.470(0.085)	0.430(0.049)	0.298(0.035)	**0.479(0.032)**
Severe Sepsis/Septic Shock	0.581(0.035)	0.593(0.018)	0.640(0.012)	0.634(0.019)	0.514(0.029)	**0.676(0.034)**
Cerebrovascular Disease	0.514(0.049)	0.539(0.057)	0.577(0.014)	0.561(0.031)	0.525(0.017)	**0.631(0.038)**
Subarachnoid Hemorrhage	0.296(0.064)	0.317(0.135)	0.394(0.161)	0.360(0.070)	0.362(0.119)	**0.556(0.155)**
Mean	0.438	0.441	0.487	0.463	0.380	**0.533**
Survival tasks						
Liver Cancer	0.609(0.058)	0.611(0.077)	0.681(0.133)	0.625(0.052)	0.642(0.062)	**0.719(0.068)**
Prostate Cancer	0.460(0.179)	0.501(0.131)	0.557(0.185)	0.371(0.121)	0.472(0.181)	**0.584(0.149)**
Brain Cancer	0.607(0.055)	0.558(0.048)	0.550(0.081)	0.586(0.087)	0.533(0.101)	**0.650(0.109)**
Myeloid Leukemia	0.570(0.119)	0.586(0.095)	0.595(0.109)	0.589(0.085)	0.531(0.058)	**0.595(0.056)**
Lymphoid Leukemia	0.542(0.144)	0.547(0.089)	0.580(0.141)	0.503(0.114)	0.614(0.056)	**0.646(0.039)**
Mean	0.557	0.561	0.593	0.535	0.558	**0.639**

which map medical mentions from scientific text and social media to ontology concepts and are standard benchmarks for synonym alignment [Liu et al., 2021]. We tokenize each mention and report pairwise consistency, which measures whether synonymous strings collapse to the same code, code purity, which measures semantic coherence within each code, and their harmonic mean $\mathrm{F1} = 2\,\mathrm{Consistency} \cdot \mathrm{Purity}/(\mathrm{Consistency} + \mathrm{Purity})$.

EHR benchmark (in-domain). We evaluate whether MEDIATOR serves as a better tokenizer for downstream EHR sequence modeling on MIMIC-III [Johnson et al., 2016a,b]. We build patient timelines from diagnoses, procedures, medications, and other coded events, restricting inputs to events timestamped at or before the prediction time (diagnoses and procedures, which MIMIC-III records only at discharge, enter only discharge-time tasks; Appendix D.1), train the same BEHRT-style Transformer [Li et al., 2020] from scratch for each task and tokenizer, and report five-fold mean AUPRC on operational, disease-specific mortality, and survival tasks.

Zero-shot cross-dataset generalization. We apply models trained on MIMIC-III directly to the eICU Collaborative Research Database [Pollard et al., 2018, 2019], whose ontologies differ and whose events are partly free text (Appendix D.2). MEDIATOR tokenizes both structured code descriptions and free-text event strings without manual vocabulary alignment.

Baselines. On BEL benchmarks, we compare with a k-means tokenizer on frozen Qwen3-Embedding-8B embeddings [Zhang et al., 2025], which clusters UMLS training embeddings into k=100,000 centroids and assigns each text to its nearest centroid, and with the text branch of MedTok [Su et al., 2025], a state-of-the-art medical tokenizer that learns a VQ codebook with separate text and knowledge-graph branches. The top-5 variant (MedTok's default) represents each input by its five nearest codes, and we treat each unique 5-code combination as a token ($\text{MedTok}_{\text{text}}$-top5). The top-1 variant uses only the nearest code ($\text{MedTok}_{\text{text}}$-top1), which makes it directly comparable to MEDIATOR. For EHR prediction, we additionally compare against Code Emb, which treats each medical code as a separate token, ClinVec [Johnson et al., 2026], a pretrained knowledge graph embedding, and full MedTok. Under zero-shot transfer, the closed-vocabulary baselines (Code Emb, ClinVec, full MedTok) drop free text and codes outside the source vocabulary.

Table 2: EHR prediction performance on eICU, reported as mean (std) AUPRC across folds. Across the 14 tasks, MEDIATOR improves over MedTok (one-sided Wilcoxon signed-rank test, $p = 6.1 \times 10^{-5}$). Patient-level bootstrap tests (5,000 resamples) show significant gains ($p < 0.05$) on 12/14 tasks over MedTok and 14/14 tasks over Code Emb.

Task	Code Emb	ClinVec	MedTok	$\text{MedTok}_{\text{text}}$-top5	$\text{MedTok}_{\text{text}}$-top1	MEDIATOR
Operational tasks						
LOS <7	0.933(0.003)	0.939(0.002)	0.934(0.001)	0.941(0.026)	0.939(0.001)	**0.942(0.004)**
ICU Mortality	0.259(0.018)	0.211(0.010)	0.282(0.009)	0.162(0.066)	0.145(0.048)	**0.349(0.005)**
Mean	0.596	0.575	0.608	0.552	0.542	**0.645**
Subcohort mortality tasks						
Acute Kidney Injury	0.318(0.032)	0.237(0.019)	0.347(0.013)	0.233(0.031)	0.213(0.002)	**0.408(0.011)**
Acute MI	0.241(0.032)	0.191(0.015)	0.227(0.049)	0.110(0.011)	0.119(0.006)	**0.325(0.024)**
Acute Pancreatitis	0.110(0.034)	0.121(0.015)	0.140(0.064)	0.073(0.015)	0.077(0.000)	**0.178(0.049)**
Atrial Fibrillation	0.200(0.027)	0.169(0.018)	0.214(0.019)	0.125(0.020)	0.095(0.002)	**0.290(0.012)**
ARF/ARDS	0.285(0.025)	0.253(0.014)	0.363(0.022)	0.250(0.034)	0.254(0.004)	**0.397(0.020)**
Cirrhosis	0.239(0.020)	0.284(0.073)	0.290(0.082)	0.208(0.011)	0.194(0.001)	**0.325(0.062)**
Heart Failure	0.177(0.016)	0.140(0.003)	0.196(0.011)	0.136(0.011)	0.125(0.003)	**0.247(0.013)**
Liver Failure	0.323(0.027)	0.372(0.022)	0.375(0.027)	0.326(0.001)	0.353(0.002)	**0.487(0.028)**
GI Bleeding	0.153(0.027)	0.164(0.009)	0.226(0.013)	0.134(0.038)	0.132(0.005)	**0.331(0.025)**
Severe Sepsis/Septic Shock	0.286(0.018)	0.300(0.014)	0.287(0.017)	0.279(0.054)	0.274(0.000)	**0.403(0.013)**
Cerebrovascular Disease	0.136(0.027)	0.147(0.033)	0.105(0.011)	0.123(0.021)	0.104(0.000)	**0.294(0.010)**
Subarachnoid Hemorrhage	0.191(0.027)	0.199(0.016)	0.189(0.021)	0.222(0.043)	0.199(0.004)	**0.401(0.099)**
Mean	0.222	0.215	0.247	0.185	0.178	**0.340**

Table 3: **Direct open-vocabulary tokenization evaluation.** Tests whether each tokenizer maps unseen mentions and their ground-truth concept descriptions to one code while preserving purity.

	NCBI			MedMentions			COMETA		
Model	Acc.	Pur.	F1	Acc.	Pur.	F1	Acc.	Pur.	F1
Qwen3-8B + k-means	0.487	0.928	0.639	0.450	0.543	0.492	0.648	0.807	0.718
$\text{MedTok}_{\text{text}}$-top5	0.195	0.986	0.326	0.045	0.885	0.085	0.231	0.913	0.368
$\text{MedTok}_{\text{text}}$-top1	0.269	0.360	0.308	0.131	0.082	0.101	0.384	0.101	0.160
MEDIATOR	0.524	0.907	**0.664**	0.471	0.516	**0.493**	0.682	0.794	**0.734**

Implementation details. We use frozen Qwen3-Embedding-8B [Zhang et al., 2025] with precomputed instruction-conditioned embeddings, a codebook of 100,000 entries (size ablation in Appendix F) in a 512-dimensional VQ space, and four views per concept during training. Enriched descriptions are generated with MedGemma-1.5-4B [Sellergren et al., 2026], PPMI uses Mass General Brigham records of 7.1 million patients (a public alternative is evaluated in Section 4.5), and the relation graph uses UMLS medical relations. Once trained, MEDIATOR is applied offline to unique clinical strings, and the cached token IDs feed the downstream models (Appendix C.1).

4.1 MEDIATOR is a discrete neural interlingua

We first test whether MEDIATOR functions as a discrete neural interlingua, where synonymous surface forms should collapse to the same token, while each token should remain semantically coherent. As shown in Table 3, MEDIATOR achieves the highest F1 on every dataset across domains.

The MEDIATOR codebook should also provide a meaningful embedding space for downstream EHR models. We assess this qualitatively by tokenizing concepts from major medical vocabularies, retrieving their assigned VQ code embeddings, and projecting them with UMAP [McInnes et al., 2018]. As shown in Figure 3, concepts cluster by ontology-derived categories that are not used as direct classification targets, including ICD disease chapters, LOINC measurement classes, SNOMED semantic groups, and procedure body-system categories. The resulting organization suggests that medical knowledge-guided training shapes the codebook into a clinically structured discrete space.

4.2 MEDIATOR improves in-domain EHR modeling

Table 1 reports AUPRC on 20 MIMIC-III tasks spanning operations, subcohort mortality, and survival. MEDIATOR achieves the strongest mean performance in all three task families, improving over

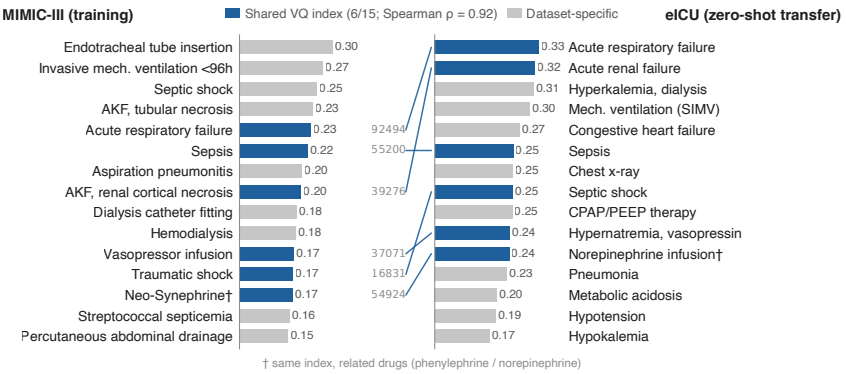


Figure 4: **Transferable attribution.** Top-15 tokens by IG attribution to cirrhosis mortality on MIMIC-III (left) and zero-shot eICU (right). Shared VQ indices are blue, others gray.

MedTok by 0.9, 4.6, and 4.6, respectively. Gains are most pronounced on clinically specific endpoints, including subarachnoid hemorrhage (+16.2), brain cancer (+10.0), cirrhosis (+8.6), and lymphoid leukemia (+6.6), suggesting that the shared semantic codebook improves statistical sharing across related codes. MEDIATOR also remains the best tokenizer with two stronger downstream backbones, GT-BEHRT [Poulain and Beheshti, 2024] and TransformEHR [Yang et al., 2023] (Appendix H).

4.3 MEDIATOR improves zero-shot transfer to eICU

Table 2 tests MIMIC-III-trained models directly on eICU. Of 3,578,036 eICU events, closed-vocabulary baselines retain only 466,408 (787 unique items), whereas MEDIATOR retains all of them, mapped to 4,102 distinct codes. MEDIATOR achieves the best transfer in both task families, improving over MedTok by 3.7 points on operational tasks and 9.3 points (38% relative) on subcohort mortality, with large gains on subarachnoid hemorrhage (+21.2), cerebrovascular disease (+18.9), severe sepsis/septic shock (+11.6), and liver failure (+11.2). The open-vocabulary tokenizer is thus more robust to ontology shift and free-text documentation.

Matched input coverage. Part of this gain is from coverage, the intended benefit of an open-vocabulary tokenizer. To isolate representation quality, we restrict all models to codes available to closed-vocabulary baselines (Table A.9, Appendix I). MEDIATOR remains the best on both task-family means, improving over MedTok by 1.0 (operational) and 1.9 (subcohort mortality). Coverage explains most of the full-input margin, and representation adds a consistent gain.

4.4 MEDIATOR enables transferable interpretability

Because MEDIATOR maps structured codes and free text into a shared token space, attributions are comparable across datasets. Figure 4 shows population-level integrated gradients (IG) [Sundararajan et al., 2017] for cirrhosis mortality trained on MIMIC-III and transferred to eICU. Across shared codes, attribution scores are strongly rank-correlated between datasets (Spearman $\rho = 0.92$, $p = 4 \times 10^{-26}$), so the importance profile transfers despite distribution shift. Six of the top 15 tokens appear in both datasets, against an expected 0.46 under a frequency-matched null over codes passing the attribution filter ($\geq$5% of patients; permutation $p = 5 \times 10^{-6}$), and overlap is also significant at top-10 and top-20. The shared tokens cover sepsis, shock, acute kidney and respiratory failure, and vasopressors, which match known mortality risk factors in decompensated cirrhosis. At K=100k, a token can also merge closely related concepts, such as the vasopressors phenylephrine and norepinephrine (token 54924), a granularity trade-off that code purity quantifies (Table 3). These attributions describe predictive associations only.

4.5 Ablation experiments

Table 4 reports leave-one-out ablations over all tasks, with per-family and additive results in Appendix G. Removing the agreement loss causes the largest overall drop, and removing PPMI a substantial drop, particularly on eICU. These two dominate downstream prediction and transfer, while enriched descriptions and relation-graph supervision give smaller, task-dependent gains. Against

Table 4: **Leave-one-out ablation.** Mean AUPRC ± std over all tasks.

Variant	MIMIC-III	eICU
$-\mathcal{L}_{\text{agree}}$	0.554±0.003	0.378±0.008
$-\mathcal{L}_{\text{co}}$ (PPMI)	0.562±0.008	0.370±0.011
$-\mathcal{L}_{\text{rel}}$	0.564±0.010	0.375±0.009
$-\mathcal{L}_{\text{desc}}$	0.561±0.018	0.380±0.006
MEDIATOR full	**0.568±0.011**	**0.384±0.007**

Table 5: **PPMI teacher source.** Mean AUPRC over all tasks.

PPMI teacher	MIMIC-III	eICU
Private (default)	**0.568**	**0.384**
Public (COHD + GoM)	0.565	**0.384**
None	0.562	0.370

the mean of the four ablated variants, the full objective improves AUPRC by 0.008 (95% CI [0.001, 0.014]; better on 24 of 34 tasks; Wilcoxon $p = 0.011$) and has the best mean rank (Friedman $p = 0.049$). In the additive ablation (Table A.6), the description and graph teachers mainly help in-domain, improving MIMIC-III operational AUPRC from 0.586 to 0.592 over the baseline with $\mathcal{L}_{\text{agree}}$ and PPMI, while eICU results are comparable.

To separate the frozen encoder from the tokenizer objectives, we compare MEDIATOR with same-encoder controls that keep the Qwen3-Embedding-8B embeddings, BEHRT model, and input coverage fixed and form tokens by k-means on the 4096-dimensional embeddings or on 512-dimensional PCA projections matching our quantization space (Table A.7 in Appendix G, which also varies the frozen encoder). While k-means achieves performance close to MEDIATOR on MIMIC-III, both controls trail it on zero-shot eICU, where the downstream model cannot relearn task-specific token mappings. Strong text embeddings and full coverage thus drive in-domain performance, while the objectives of MEDIATOR add transfer robustness and open-vocabulary mapping quality.

Finally, we rebuild the PPMI teacher from public sources. We compute PPMI matrices from Columbia Open Health Data (COHD) concept co-occurrence counts [Ta et al., 2018] and from Stanford Graph of Medicine (GoM) co-occurrence statistics [Finlayson et al., 2014], map both to our concept groups, and average them (Appendix B.1). The public teacher performs nearly identically to the private one (Table 5), so the reported gains do not depend on private co-occurrence data.

5 Conclusion

Ontology-specific token interfaces hinder the transfer of EHR models across institutions and exclude much clinically relevant text. MEDIATOR reframes EHR harmonization as open-vocabulary discrete tokenization, mapping any clinical concept expressible as text into a shared VQ codebook. Across codebook analysis, entity linking, in-domain prediction, zero-shot transfer, and attribution, MEDIATOR provides a compact, reusable token interface for standard EHR Transformers, and on eICU it uses every event where closed-vocabulary tokenizers use 13%. Open-vocabulary tokenization may also help other domains where many identifiers denote one entity, such as genes or financial securities.

Limitations. MEDIATOR inherits noise and omissions from its supervision sources, and a systematic error such as an incorrect synonym group affects every use of that concept, so clinically critical mappings should be validated before high-stakes use. The default co-occurrence teacher uses private data, though a public one performs nearly identically. Transfer gains partly reflect a coverage asymmetry, since MEDIATOR ingests events that closed-vocabulary baselines must discard. MEDIATOR tokenizes each concept description without context, so it cannot resolve short-form polysemy in ambiguous abbreviations, and it is trained on English sources only. Our downstream models are trained per task, and pretrained generative backbones and more hospitals and languages remain future work.

Acknowledgments and Disclosure of Funding

This work was funded by the Mass General Brigham AI Institute.

References

American Medical Association. CPT® code set overview. `https://www.ama-assn.org/practice-management/cpt/cpt-code-set-overview`, 2026. Accessed: 2026-01-25.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.253. URL https://aclanthology.org/2020.emnlp-main.253/.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. In *Pacific Symposium on Biocomputing 2020*, pages 295–306. World Scientific, November 2019. ISBN 978-981-12-1562-9. doi: 10.1142/9789811215636_0027.

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 2004.

Centers for Disease Control and Prevention. IIS: Code Sets: CVX – Vaccines Administered. https://www2a.cdc.gov/vaccines/iis/iisstandards/vaccines.asp?rpt=cvx, 2026. Accessed: 2026-05-06.

Centers for Medicare & Medicaid Services. Healthcare Common Procedure Coding System (HCPCS). https://www.cms.gov/medicare/coding-billing/healthcare-common-procedure-system, 2026. Accessed: 2026-10-16.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10, 2014. doi: 10.1016/j.jbi.2013.12.006.

Tianqi Du, Yifei Wang, and Yisen Wang. On the role of discrete tokenization in visual representation learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNLAkjUm19.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

Christopher Fifty, Ronald Guenther Junkins, Dennis Duan, Aniketh Iyengar, Jerry Weihong Liu, Ehsan Amid, Sebastian Thrun, and Christopher Ré. Restructuring vector quantization with the rotation trick. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GMwRl2e9Y1.

Samuel G. Finlayson, Paea LePendu, and Nigam H. Shah. Building the graph of medicine from millions of clinical narratives. *Scientific Data*, 1:140032, 2014. doi: 10.1038/sdata.2014.32.

Arden W Forrey, Clement J McDonald, Georges DeMoor, Stanley M Huff, Dennis Leavelle, Diane Leland, Tom Fiers, Linda Charles, Brian Griffin, Frank Stalling, et al. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry*, 42(1):81–90, 1996.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

George Hripcsak, Jon D. Duke, Nigam H. Shah, Christian G. Reich, Vojtech Huser, Martijn J. Schuemie, Marc A. Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R. Rijnbeek, et al. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574–578, 2015. doi: 10.3233/978-1-61499-564-7-574.

Kyunghoon Hur, Jiyoung Lee, Jungwoo Oh, Wesley Price, Younghak Kim, and Edward Choi. Unifying heterogeneous electronic health records systems via text-based code embedding. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 183–203. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/hur22a.html.

Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database. *PhysioNet*, September 2016a. doi: 10.13026/C2XW26. URL https://doi.org/10.13026/C2XW26. Version 1.4.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016b. doi: 10.1038/sdata.2016.35.

Ruth Johnson, Uri Gottlieb, Galit Shaham, Lihi Eisen, Jacob Waxman, Stav Devons-Sberro, Curtis R Ginder, Peter Hong, Raheel Sayeed, Xiaorui Su, et al. Embeddings of clinical codes enable knowledge-grounded AI in medicine. *npj Digital Medicine*, 9:786, 2026. doi: 10.1038/s41746-026-02664-9.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673, 2020.

Junmo Kim, Namkyeong Lee, Jiwon Kim, and Kwangsoo Kim. MedRep: medical concept representations for general electronic health record foundation models. *Journal of the American Medical Informatics Association*, 33(6):1089–1099, 2026. doi: 10.1093/jamia/ocag032.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Balston, Jack Ross, Esther Idowu, James T Teo, and Richard J B Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: A retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, April 2024. doi: 10.1016/S2589-7500(24)00025-6.

Yoonhyung Lee, Younhyung Chae, and Kyomin Jung. Leveraging VQ-VAE tokenization for autoregressive modeling of medical time series. *Artificial Intelligence in Medicine*, 154:102925, 2024.

Yikuan Li, Shishir Rao, Jose R. A. Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, 2020. doi: 10.1038/s41598-020-62922-y.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.334. URL https://aclanthology.org/2021.naacl-main.334/.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Sunil Mohan and Donghui Li. MedMentions: A large biomedical corpus annotated with UMLS concepts. In *Automated Knowledge Base Construction (AKBC)*, 2019.

National Library of Medicine. RxNorm. https://www.nlm.nih.gov/research/umls/rxnorm/, 2025.

National Library of Medicine. Medical Subject Headings. https://www.nlm.nih.gov/mesh/meshhome.html, 2026. Accessed: 2026-05-06.

J. Marc Overhage, Patrick B. Ryan, Christian G. Reich, Abraham G. Hartzema, and Paul E. Stang. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60, 2012. doi: 10.1136/amiajnl-2011-000376.

Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. In *Proceedings of Machine Learning for Health*, pages 239–260. PMLR, November 2021.

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.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178.

Raphael Poulain and Rahmatollah Beheshti. Graph transformers on EHRs: Better representation improves downstream performance. In *The Twelfth International Conference on Learning Representations*, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. TokenFlow: Unified image tokenizer for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2545–2555, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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):1–13, May 2021. doi: 10.1038/s41746-021-00455-y.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, 31(9):1844–1855, 2024. doi: 10.1093/jamia/ocae029.

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

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, November 2025. doi: 10.1038/s41586-025-09529-3.

SNOMED International. *SNOMED CT US Edition*, 2025. URL https://www.nlm.nih.gov/healthit/snomedct/us_edition.html.

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

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

Xiaorui Su, Shvat Messica, Yepeng Huang, Ruth Johnson, Lukas Fesser, Shanghua Gao, Faryad Sahneh, and Marinka Zitnik. Multimodal medical code tokenizer. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UaTrcei5Ba.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, 2017.

Cao N. Ta, Michel Dumontier, George Hripcsak, Nicholas P. Tatonetti, and Chunhua Weng. Columbia open health data, clinical concept prevalence and co-occurrence from electronic health records. *Scientific Data*, 5:180273, 2018. doi: 10.1038/sdata.2018.273.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale. *arXiv preprint arXiv:2508.12104*, 2025.

WHO Collaborating Centre for Drug Statistics Methodology. ATC Classification Index with DDDs, 2026. `https://atcddd.fhi.no/atc_ddd_index/`, 2026. Accessed: 2026-05-06.

World Health Organization. International classification of diseases—ninth revision (ICD-9). *Weekly Epidemiological Record = Relevé Épidémiologique Hebdomadaire*, 63(45):343–344, 1988.

World Health Organization. *International Statistical Classification of Diseases and Related Health Problems*. World Health Organization, Geneva, 10th edition, 2016. URL `https://icd.who.int/browse10/2016/en`.

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

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A. Pfeffer, Jason Fries, and Nigam H. Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, July 2023b. doi: 10.1038/s41746-023-00879-8.

Jialun Wu, Xin Mei, Rui Mao, Kai He, and Erik Cambria. TAKECare: A temporal-hierarchical framework with knowledge fusion for personalized clinical predictive modeling. *Information Fusion*, 126:103620, February 2026a. doi: 10.1016/j.inffus.2025.103620.

John Wu, Yongda Fan, Zhenbang Wu, Paul Landes, Eric Schrock, Sayeed Sajjad Razin, Arjun Chatterjee, Naveen Baskaran, Joshua Steier, Andrea Fitzpatrick, et al. PyHealth 2.0: A comprehensive open-source toolkit for accessible and reproducible clinical deep learning. *arXiv preprint arXiv:2601.16414*, 2026b.

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*, 14:7857, 2023.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Doudou Zhou, Han Tong, Linshanshan Wang, Suqi Liu, Xin Xiong, Ziming Gan, Griffier Romain, Boris P Hejblum, Yun-Chung Liu, Chuan Hong, et al. Representation learning to advance multi-institutional studies with electronic health record data from US and France. *Nature Communications*, 17:4800, 2026. doi: 10.1038/s41467-026-71152-1.

A Impact statement

This work proposes a discrete, ontology-agnostic tokenizer that aims to improve the interoperability of EHR-based machine learning models across coding systems, versions, and institutions, and to enable more structured use of information extracted from clinical notes. If adopted in clinical research or deployed in downstream decision-support tools, such representations could improve portability and reduce the engineering burden of manual code mapping, potentially accelerating model development and evaluation across sites. However, the approach inherits common risks of EHR modeling. Biases and missingness in clinical data may be preserved, performance may vary across demographic groups or care settings, and errors in concept normalization could propagate into downstream predictions. In particular, a systematic error in a supervision source, such as an incorrectly curated synonym group, affects every downstream use of that concept, so mappings for clinically critical concepts should be independently validated before deployment in high-stakes settings.

B Data processing

B.1 Training data preparation

We construct the MEDIATOR pretraining corpus as a set of concept groups, each associated with multiple textual views. We begin from UMLS 2025AB MRCONSO and retain English, unsuppressed entries from the target clinical and biomedical vocabularies. An LLM (Qwen3.5-35B-A3B [Qwen Team, 2026]) is used to filter the relations that are synonyms or medically related. We then merge CUIs connected by synonym-classified relations using union-find, where the resulting connected components define `CUI_GROUP` identifiers. All retained source strings from CUIs in the same `CUI_GROUP` are treated as positive textual views. We remove null and numeric-only strings, require at least two views per group before augmentation, and augment under-viewed groups so that each retained `CUI_GROUP` has at least five distinct descriptions.

LLM-generated text resources. For minimum-view augmentation, we use Qwen3.5-35B-A3B [Qwen Team, 2026] to generate additional equivalent names for concept groups with fewer than five descriptions. The system prompt instructs the model to act as a medical expert and generate only new alternative names that are concise, medically accurate, semantically equivalent, of similar granularity, distinct from all existing names, and returned as a JSON array without explanations or metadata. The user prompt is as follows.

> Representative concept name: {representative name}. Existing equivalent names already known: {existing names}. Generate {n} additional equivalent names that are distinct from all existing names. Output only a JSON array of strings.

For description-teacher supervision, the pipeline uses a CUI-level table with columns `CUI`, `STR`, and `enriched_description`. These enriched descriptions were generated offline with MedGemma-1.5-4B [Sellergren et al., 2026] and are aligned to `CUI_GROUP`; we release the generated descriptions and prompts with our code. The system prompt is as follows.

> You are a medical expert. Write only plain-text responses. Never use markdown, headers, bullet points, numbered lists, or bold text. Do not include thinking steps or generation notes – only the final description. Keep it concise.

and the user prompt is as follows.

> In plain-text sentences, describe what this concept is and its primary clinical relevance. Be informative and concise.

Embedding clinical strings and enriched descriptions. All tokenizer-input strings are embedded once with the frozen Qwen3-Embedding-8B encoder and cached before VQ training. We use instruction-aware formatting, `Instruct:` *Given a clinical concept string, produce an embedding that preserves exact biomedical concept identity across synonyms, ontology versions, abbreviations, and local hospital wording. Strings that refer to the same medical concept should be close. Strings that are clinically related but not equivalent should remain distinguishable.* `Query: {text}`. For enriched descriptions, we use the analogous "medical concept description" instruction.

Empirical co-occurrence resource. We build raw PPMI from structured EHR events of 7.1 million patients at Mass General Brigham (MGB). The MGB institutional review board approved the retrospective analysis of these records. The participants were not directly involved in or recruited for the study, and the requirement for informed consent to analyze the data was thus waived. The event sources are problem-list diagnoses, laboratory events, procedures, medications, and immunizations. Raw codes are normalized and mapped to UMLS using ICD10CM for diagnoses, LOINC for laboratories, CPT for procedures, RxNorm for medications, and CVX for immunizations. The UMLS code map is then collapsed to `CUI_GROUP`; codes that map ambiguously to multiple concept groups are excluded.

After mapping, we deduplicate events at the patient–date–`CUI_GROUP` level and define contexts by patient and 30-day calendar time bin. Let $B \in \{0,1\}^{D\times|\mathcal{C}|}$ be the sparse context–concept matrix over D patient-window contexts. We compute $C = B^\top B$, where C_{ij} is the number of contexts containing both concepts i and j, and $n_i = C_{ii}$ is the number of contexts containing concept i. The final sparse teacher is the positive pointwise mutual information (PPMI) matrix

$$M_{ij} = \max\left(0, \log_2 \frac{D\, C_{ij}}{n_i n_j}\right). \tag{13}$$

Public co-occurrence resource. To remove the dependence on private data, we also build a PPMI teacher from two public resources. Columbia Open Health Data (COHD) [Ta et al., 2018] provides concept prevalence and pairwise co-occurrence counts from the Columbia University Irving Medical Center EHR, encoded with OMOP standard concepts. The Stanford Graph of Medicine [Finlayson et al., 2014] provides co-occurrence counts of UMLS concepts mined from clinical narratives. We map COHD concepts to UMLS through their source vocabularies, collapse both resources to `CUI_GROUP`, compute a PPMI matrix from each resource's co-occurrence and marginal counts as above, and average the two matrices to form the public teacher.

Medical relation knowledge graph. The medical relation knowledge graph G_{rel} is built from UMLS relations classified as synonymous or medically related by Qwen3.5-35B-A3B [Qwen Team, 2026], after mapping both endpoints to `CUI_GROUP`. We remove self-loops after grouping, build an undirected graph over concept groups, and precompute two-hop neighborhoods for relation-based supervision during tokenizer training.

C Implementation details

C.1 MEDIATOR training

We initialize the codebook with k-means over the precomputed encoder embeddings projected into the quantization space, update codebook statistics with EMA, and periodically revive under-used codes from current encoder outputs. Post-quantization objectives use rotation-trick gradient transport. Semantic teachers from enriched descriptions, raw PPMI co-occurrence, and the medical relation knowledge graph G_{rel} are applied as separate weighted contrastive losses, and PPMI-neighbor sampling is used to increase the frequency of informative co-occurrence pairs in minibatches. Additional hyperparameters are listed in Table A.1. Training requires 16GB. All experiments were trained on a single NVIDIA A100 GPU.

MEDIATOR is applied once, offline, as a vocabulary-level preprocessing step over unique clinical-concept strings. Each unique concept string is passed once through the model to obtain a discrete integer token id; these ids are then cached. At training and inference time the downstream EHR model consumes only these cached integer ids, and the text encoder is never invoked again. Concretely, tokenizing the entire concept vocabulary takes ≈70s per dataset on a single NVIDIA RTX PRO 6000 GPU. The 12,533 MIMIC-III concepts take 71.2s (5.7 ms/concept) and the 10,779 eICU concepts take 69.9s (6.5 ms/concept). Tokenization cost therefore scales only with the number of unique concepts and is independent of dataset size and sequence length. The text encoder and adaptor are never loaded by the downstream model. The codebook ($100{,}000 \times 512$; 51.2M parameters, ≈102 MB in fp16) is used once to initialize the downstream embedding table (Appendix C.2).

Table A.1: Hyperparameters used in final MEDIATOR tokenizer training.

Hyperparameter	Value
Adaptor hidden dim.	1024
VQ embedding dim.	512
Codebook size	100,000
Views per concept	4
Batch size	1536
Training epochs	100
Adaptor learning rate	5×10^{-5}
Other learning rate	1×10^{-4}
Weight decay	1×10^{-5}
Warmup ratio	0.1
Max gradient norm	1.0
Contrastive temperature	0.07
Post-VQ temperature	0.07
Teacher temperature	0.15
EMA decay	0.99
EMA ϵ	1×10^{-5}
Revive threshold	2.0
Agreement warmup	10 epochs
PPMI batch fraction	0.083
λ_{vq}	0.25
λ_{pre}	1.0
λ_{post}	1.5
λ_{agree}	1000.0
λ_{rel}	0.1
λ_{co}	0.1
λ_{desc}	0.1
λ_{rec}	0.01

Table A.2: Hyperparameters used in downstream BEHRT-style EHR training.

Hyperparameter	Value
Hidden size	288
Number of hidden layers	6
Number of attention heads	12
Intermediate size	512
Dropout probability	0.1
Segmentation vocab size	3
Age vocab size	91
Sex vocab size	2
Batch size	256
Max epochs	100
Learning rate	5×10^{-4}
Warmup proportion	0.1
Weight decay	5×10^{-4}
Early-stop patience	10

C.2 Downstream EHR model

We train an end-to-end BEHRT [Li et al., 2020] model for every single task and integrate the tokens for time, visit, and patient information (sex, age) as in the original paper. Hyperparameters are listed in Table A.2.

For in-domain EHR prediction, the Code Emb baseline trains BEHRT with a standard learned embedding table over structured codes. For the full MedTok baseline, medical codes are passed to the original MedTok model and the resulting code embeddings are used as model inputs. ClinVec uses knowledge graph embeddings following the official setup [Johnson et al., 2026]. For text baselines, in MedTok$_{\text{text}}$-top5 and MedTok$_{\text{text}}$-top1 settings, the official description of each medical code is encoded using the MedTok text encoder, and the average of the top-5 returned embeddings or the top-1 embedding is used, respectively. For MEDIATOR, the input embedding table has one row per codebook entry ($100{,}000 \times 512$) and is initialized from the trained MEDIATOR codebook, followed by a linear projection from 512 to the 288-dimensional BEHRT hidden size. Both the embedding table and the projection are trainable during BEHRT training. Each event is represented by the projected embedding of its assigned VQ index. All models are trained with 5-fold data split.

For zero-shot transfer to eICU, we directly apply models trained on MIMIC-III without fine-tuning. All five fold-specific checkpoints are evaluated on eICU, and the final performance is reported as the average across folds. Code Emb, MedTok, and ClinVec must drop much of the text information, since eICU provides no official code mapping. To further ensure fair comparison, MedTok$_{\text{text}}$-top5, MedTok$_{\text{text}}$-top1 settings and MEDIATOR embed all texts and structured code descriptions. We allow eICU VQ indices that were not observed in MIMIC training; we therefore evaluate transfer with the full, unfiltered tokenizer output. Because the embedding table is initialized from the codebook, rows for indices never observed in MIMIC-III training receive no gradient updates and retain their codebook initialization (up to weight decay), so these indices are mapped through the learned projection to embeddings that reflect their position in the shared semantic space.

C.3 Integrated gradients analysis

We attribute predictions to MEDIATOR tokens using integrated gradients (IG) [Sundararajan et al., 2017] in embedding space. Let a patient sequence be represented by a sequence of VQ indices, and let the model input be the corresponding sequence of VQ embeddings. For IG, we treat each token embedding as the attribution unit and use the all-zero vector as the baseline embedding. We compute IG with respect to the positive-class (mortality) logit.

From per-patient IG scores, we compute population-level importance for each token in three steps. (i) We filter to tokens that appear in at least 5% of patients to avoid unstable estimates for rare tokens. (ii) Within each patient, if a token index appears multiple times, we apply max pooling over its IG scores so that each patient contributes at most one attribution value per token (capturing the strongest evidence for that concept in the trajectory). (iii) We then average pooled IG scores across patients (without taking absolute values) to obtain directional population-level importance, where positive scores indicate risk-associated concepts and negative scores indicate protective concepts.

D Task definition

The selection of downstream prediction tasks is guided by both clinical relevance and data availability across the MIMIC-III [Johnson et al., 2016b] and eICU [Pollard et al., 2018] datasets, both obtained from PhysioNet [Goldberger et al., 2000] under credentialed access. In particular, we prioritize tasks that are well-established in the clinical machine learning literature. For disease-specific mortality and cancer survival tasks, cohort selection is further constrained by the need to ensure sufficient statistical power. We ensure that all evaluated tasks are supported by sufficient data for reliable modeling and evaluation.

D.1 MIMIC-III

We use PyHealth 2.0 [Wu et al., 2026b] to construct the length-of-stay and readmission cohort. For other tasks, we follow a standard MIMIC-III cohort-construction pipeline. After filtering for age $\geq$ 18, ICU stay structure, and task-specific eligibility, we sample one admission per patient, so each patient contributes a single admission trajectory and all splits are patient-disjoint. We evaluate all downstream models with five-fold cross-validation. In each run, one fold serves as the validation set, one as the test set, and the remaining three as the training set, and the checkpoint with the best validation AUPRC is used for testing. eICU is never used for training, hyperparameter tuning, or model selection. Subcohort mortality is defined to predict ICU mortality for patients diagnosed with that specific disease within that stay. Survival tasks predict 1-year mortality after ICU discharge for patients diagnosed with a specific cancer. Cohort statistics are reproduced in Table A.3.

ICU length of stay, ICU mortality, and sub-cohort mortality tasks are predicted using data available within the first 24 hours of ICU admission. ICU readmission and survival tasks are predicted at discharge for patients discharged alive. ICU readmission is defined as a subsequent ICU admission within 30 days. For survival tasks, the prediction time is ICU discharge; patients who die on or before the ICU discharge date are excluded, and the label is positive if the recorded date of death falls within 365 days after ICU discharge. Patients without a recorded date of death are treated as survivors; we do not explicitly censor patients with incomplete follow-up.

MIMIC-III diagnosis and procedure codes (`DIAGNOSES_ICD` and `PROCEDURES_ICD`) carry no timestamps and are assigned at the end of the hospital stay, so we timestamp them at discharge time. For every task, model inputs are restricted to events timestamped at or before the prediction time.

Disease-specific mortality tasks are built from ICD-9 cohorts corresponding to 584 (acute kidney injury); 410 (acute myocardial infarction (MI)); 577.0 (acute pancreatitis); 427.31 (atrial fibrillation); 518.81 (acute respiratory failure/acute respiratory distress syndrome (ARF/ARDS)); 571.5, 571.6 (cirrhosis); 428 (heart failure); 570, 572.2, 572.8 (liver failure); 578 (gastrointestinal (GI) bleeding); 785.52, 995.92 (severe sepsis/septic shock); 430, 431, 433, 434, 436 (cerebrovascular disease, comprising subarachnoid and intracerebral hemorrhage, all subcodes of precerebral (433) and cerebral (434) artery occlusion with or without recorded infarction, and acute ill-defined cerebrovascular disease); and 430 (subarachnoid hemorrhage). Cancer survival tasks are built from ICD-9 codes 155 (liver cancer); 191 (brain cancer, i.e., malignant neoplasm of brain; histology not specified); 204 (lymphoid leukemia); 205 (myeloid leukemia); and 185 (prostate cancer).

Table A.3: Dataset statistics for prediction tasks on MIMIC-III and eICU. Total Stays is the number of ICU stays in each task cohort (one per patient); Events is the number of positive labels.

		MIMIC-III		eICU	
Group	Task	Total Stays	Events	Total Stays	Events
Hospital Operation	LOS $<$ 7 days	40757	18931	83166	21818
	ICU Mortality	31650	2954	83166	7003
	Readmission	9515	2613	–	–
Subcohort ICU Mortality	Acute Kidney Injury	6757	1370	8309	1367
	Acute MI	3681	439	5671	414
	Acute Pancreatitis	627	84	322	19
	Atrial Fibrillation	8179	984	7491	669
	ARF/ARDS	4732	1327	11170	2384
	Cirrhosis	755	133	497	92
	Heart Failure	7921	957	6806	708
	Liver Failure	1090	389	622	162
	GI Bleeding	1384	219	4984	403
	Severe Sepsis/Septic Shock	2468	818	6033	1279
	Cerebrovascular Disease	3014	525	7101	689
	Subarachnoid Hemorrhage	509	100	497	90
Cancer Survival	Liver Cancer	245	120	–	–
	Prostate Cancer	180	44	–	–
	Brain Cancer	143	66	–	–
	Lymphoid Leukemia	116	50	–	–
	Myeloid Leukemia	122	67	–	–

D.2 eICU

We likewise follow the same eICU cohort-construction protocol for diagnoses, medications, admission drugs, past medical history, and treatments, again keeping one ICU stay per patient. Because eICU lacks post-discharge follow-up, we do not evaluate readmission and survival tasks.

In addition to the ICD-9 label filtering rules applied to MIMIC-III, we define disease-specific subcohorts in eICU with the ICD-10 diagnosis codes N17 (acute kidney injury); I21, I22 (acute MI); K85 (acute pancreatitis); I48.0-I48.2, I48.91 (atrial fibrillation); J80 (ARF/ARDS); K74 (cirrhosis); I50 (heart failure); K72 (liver failure); K92.0, K92.1, and K92.2 (GI bleeding); R65.20, R65.21 (severe sepsis/septic shock); I60, I61, I62, I63, I64 (cerebrovascular disease); and I60 (subarachnoid hemorrhage). Because ARDS-specific coding is unavailable in ICD-9, the MIMIC-III ARF/ARDS cohort is defined by acute respiratory failure (518.81), whereas the eICU cohort uses ARDS (J80), a narrower subset; transfer results for this task should be interpreted with this definitional difference in mind. Cohort details are in Table A.3. eICU records are passed through the tokenizer without restricting predictions to VQ indices observed in the source dataset.

E Codebook visualization

Figure A.1 shows the UMAP projections of Figure 3 with the complete category legends and the number of concepts in each category.

F Codebook size ablations

We vary the codebook size K from 10k to 120k (Figure A.2, Table A.4). Codebook size affects entity linking and downstream prediction differently. Entity-linking quality generally improves with K. Mean F1 rises from 0.558 at 10k to 0.638 at 120k, monotonically on MedMentions and nearly so on COMETA. Downstream AUPRC on operational tasks is comparatively flat, ranging from 0.580 to 0.592 on MIMIC-III and from 0.635 to 0.649 on eICU, with no improvement beyond 100k on either dataset. From 100k to 120k, MIMIC-III declines from 0.592 to 0.586 and eICU stays at 0.645. We therefore select $K{=}100$k as a cost-performance operating point that attains the best MIMIC-III value, lies within 0.004 of the best eICU value, and keeps entity-linking quality close to that of the largest

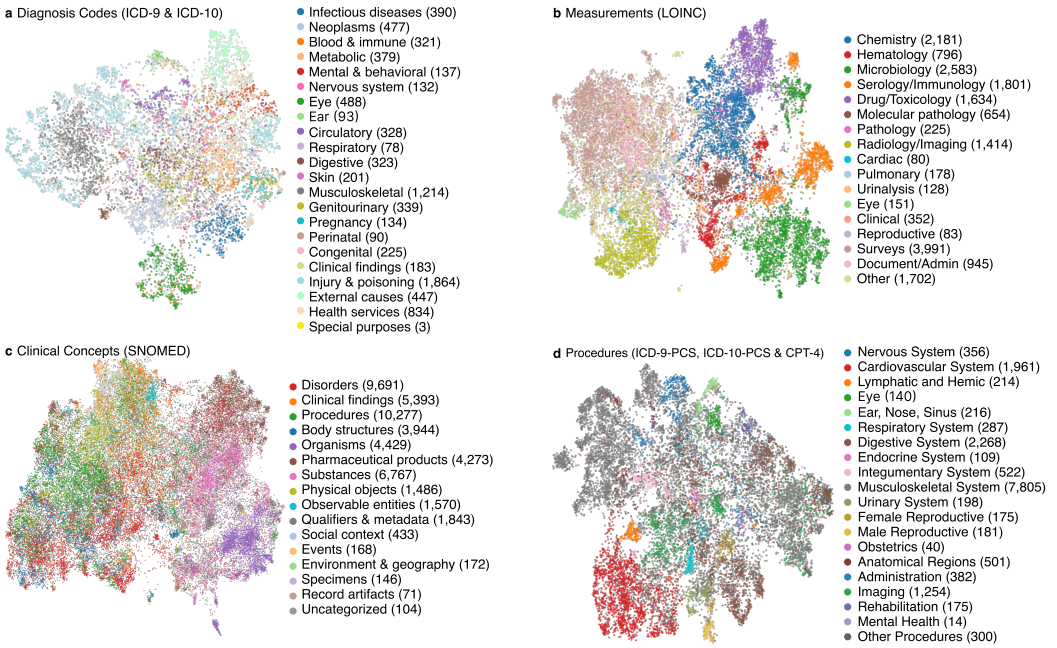


Figure A.1: **UMAP visualization of the MEDIATOR codebook with full category legends.** Each point denotes a source concept represented by its assigned VQ code embedding and projected to two dimensions with UMAP. Colors indicate ontology-derived categories, with the number of concepts per category in parentheses, for (a) diagnosis codes (ICD-9 and ICD-10), (b) measurement concepts (LOINC), (c) clinical concepts (SNOMED CT), and (d) procedure codes (ICD-9-PCS, ICD-10-PCS, and CPT-4).

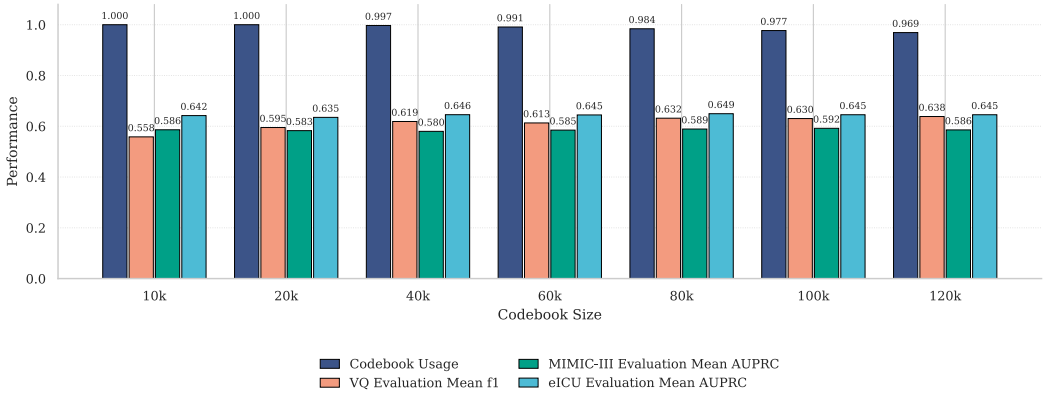


Figure A.2: **Effect of codebook size K.** We vary the VQ codebook size and report mean BEL (VQ Evaluation) F1 and operational-task mean AUPRC on MIMIC-III and zero-shot eICU. We additionally report codebook usage on the training corpus.

codebook while keeping the vocabulary compact. Across all K, codebook utilization on the training corpus stays above 96%, indicating that MEDIATOR's training stabilizes large codebooks and avoids the severe dead-code behavior commonly observed when scaling VQ models.

G Additional ablations

This section reports the full component and encoder ablations summarized in Section 4.5, including leave-one-out results by task family (Table A.5), the additive component ablation over operational tasks (Table A.6), and the encoder and same-encoder tokenizer controls (Table A.7). Panel A of

Table A.4: **Biomedical entity linking F1 across codebook sizes.**

Codebook size	NCBI	MedMentions	COMETA	Mean
10k	0.663	0.343	0.669	0.558
20k	0.670	0.409	0.707	0.595
40k	0.681	0.451	0.723	0.619
60k	0.637	0.469	0.732	0.613
80k	0.673	0.484	0.738	0.632
100k	0.664	0.493	0.734	0.630
120k	0.680	0.497	0.739	0.638

Table A.5: **Leave-one-out component ablation by task family.** Mean AUPRC ± standard deviation across folds for task-family means on MIMIC-III and zero-shot eICU transfer.

	MIMIC-III				eICU		
Method	Oper.	Mort.	Surv.	All	Oper.	Mort.	All
$-\mathcal{L}_{\text{agree}}$	0.582±0.010	0.524±0.010	0.610±0.023	0.554±0.003	0.642±0.006	0.334±0.010	0.378±0.008
$-\mathcal{L}_{\text{co}}$ (PPMI)	0.585±0.011	0.530±0.013	0.628±0.054	0.562±0.008	0.644±0.005	0.324±0.013	0.370±0.011
$-\mathcal{L}_{\text{rel}}$	0.590±0.007	**0.549**±0.006	0.585±0.045	0.564±0.010	0.644±0.005	0.330±0.010	0.375±0.009
$-\mathcal{L}_{\text{desc}}$	0.588±0.009	0.537±0.022	0.601±0.051	0.561±0.018	0.640±0.008	0.337±0.007	0.380±0.006
MEDIATOR full	**0.592**±0.007	0.533±0.012	**0.639**±0.037	**0.568**±0.011	**0.645**±0.003	**0.340**±0.009	**0.384**±0.007

Table A.7 replaces the frozen Qwen3-Embedding-8B encoder with BioLORD-2023 [Remy et al., 2024] or Qwen3-Embedding-0.6B while keeping the rest of the pipeline fixed. Both weaker encoders lower performance, although Qwen3-Embedding-0.6B retains most of the eICU transfer performance (0.372 vs. 0.384), indicating that the transfer gains come substantially from the learned mapping.

H Additional EHR architectures

To test whether the benefit of MEDIATOR depends on the specific downstream BEHRT-style architecture used in the main experiments, we evaluated the same tokenizer variants with two stronger EHR backbones, GT-BEHRT [Poulain and Beheshti, 2024] and TransformEHR [Yang et al., 2023]. For each architecture, the tokenizer is the only component varied, while the downstream training and evaluation protocol follows the main MIMIC-III in-domain and MIMIC-to-eICU zero-shot transfer setup. As shown in Table A.8, MEDIATOR remains the best tokenizer under both architectures and on both evaluation settings. This suggests that the gains from open-vocabulary discrete tokenization are not specific to the BEHRT-style model used in the main paper, but transfer to stronger EHR sequence modeling backbones. For GT-BEHRT, MIMIC-III and eICU in our benchmark contain a single admission trajectory per patient, so the original multi-visit, per-visit graph construction does not directly apply. We therefore use a single-visit adaptation with a per-patient star topology, where a learned readout token attends to and from all event tokens through dense masked attention, and edge typing is collapsed to code-category embeddings. This preserves the graph-attention mechanism of GT-BEHRT while matching the single-admission structure of our benchmark.

I Zero-shot comparison on the mappable subset

Table A.9 shows that matching coverage removes most, but not all, of the performance gap. Against the strongest baseline (MedTok), MEDIATOR's operational-mean margin falls from 3.7 to 1.0 points, and its subcohort-mortality margin falls from 9.3 to 1.9 points. MEDIATOR remains the best model on both task-family means, while MedTok leads the mean on neither.

J Worked example

Table A.10 traces three surface forms of one concept and one distinct concept through the MEDIATOR pipeline of text encoding, projection, and nearest-code assignment.

Table A.6: **Additive component ablation.** Mean AUPRC over operational tasks on MIMIC-III and zero-shot eICU transfer, adding components to a baseline trained with only the equivalence, VQ, and reconstruction losses.

Objective	MIMIC-III Oper.	eICU Oper.
Baseline	0.585	0.641
Baseline + $\mathcal{L}_{\text{desc}}$	0.586	0.642
Baseline + $\mathcal{L}_{\text{co}}$ (PPMI)	0.580	0.645
Baseline + $\mathcal{L}_{\text{rel}}$	0.584	0.635
Baseline + $\mathcal{L}_{\text{agree}}$	0.583	0.644
Baseline + $\mathcal{L}_{\text{agree}}$ + $\mathcal{L}_{\text{desc}}$	0.584	0.637
Baseline + $\mathcal{L}_{\text{agree}}$ + $\mathcal{L}_{\text{co}}$ (PPMI)	0.586	**0.651**
Baseline + $\mathcal{L}_{\text{agree}}$ + $\mathcal{L}_{\text{rel}}$	0.584	0.641
MEDIATOR full	**0.592**	0.645

Table A.7: **Encoder and tokenizer ablations.** Panel A varies the frozen text encoder used by MEDIATOR. Panel B keeps the same frozen Qwen3-Embedding-8B encoder fixed and compares MEDIATOR against direct clustering controls. All results report mean AUPRC on 20 MIMIC-III tasks and 14 zero-shot eICU transfer tasks.

Analysis	Variant	Dim.	MIMIC-III	eICU
A. Alternative frozen encoders for MEDIATOR				
Encoder	BioLORD-2023	768	0.551	0.312
Encoder	Qwen3-Embedding-0.6B	1024	0.550	0.372
Encoder	Qwen3-Embedding-8B	4096	**0.568**	**0.384**
B. Same-encoder clustering controls				
Tokenizer	Qwen embeddings + k-means	4096	**0.568**	0.378
Tokenizer	Qwen embeddings $\rightarrow$ PCA $\rightarrow$ k-means	512	0.558	0.379
Tokenizer	MEDIATOR	512	**0.568**	**0.384**

Table A.8: **Tokenizer comparison under stronger EHR architectures.** We evaluate MEDIATOR as the token interface for GT-BEHRT and TransformEHR. Results are mean AUPRC on 20 MIMIC-III tasks and 14 zero-shot eICU transfer tasks.

Tokenizer	MIMIC-III	eICU
GT-BEHRT		
Code Emb	0.572	0.313
ClinVec	0.535	0.280
MedTok	0.573	0.313
$\text{MedTok}_{\text{text}}$-top1	0.511	0.219
$\text{MedTok}_{\text{text}}$-top5	0.534	0.224
MEDIATOR	**0.602**	**0.414**
TransformEHR		
Code Emb	0.470	0.272
ClinVec	0.481	0.264
MedTok	0.512	0.299
$\text{MedTok}_{\text{text}}$-top1	0.434	0.227
$\text{MedTok}_{\text{text}}$-top5	0.499	0.218
MEDIATOR	**0.553**	**0.384**

Table A.9: **Matched-input zero-shot transfer on eICU.** All models are evaluated on the overlapping subset of eICU codes available to closed-vocabulary baselines. We report mean AUPRC with standard deviation across folds. Best results are bolded and second-best results are underlined.

Task	Code Emb	ClinVec	MedTok	$\text{MedTok}_{\text{text}}$-top5	$\text{MedTok}_{\text{text}}$-top1	MEDIATOR
Operational tasks						
LOS <7	0.933(0.003)	**0.939(0.002)**	0.934(0.001)	0.937(0.004)	0.926(0.003)	0.938(0.002)
ICU Mortality	0.259(0.018)	0.211(0.010)	0.282(0.009)	0.138(0.022)	0.130(0.010)	**0.298(0.021)**
Mean	0.596	0.575	0.608	0.537	0.528	**0.618**
Subcohort mortality tasks						
Acute Kidney Injury	0.318(0.032)	0.237(0.019)	**0.347(0.013)**	0.184(0.023)	0.185(0.011)	0.345(0.010)
Acute MI	**0.241(0.032)**	0.191(0.015)	0.227(0.049)	0.102(0.020)	0.108(0.010)	0.227(0.011)
Acute Pancreatitis	0.110(0.034)	0.121(0.015)	0.140(0.064)	0.087(0.022)	0.089(0.025)	**0.162(0.042)**
Atrial Fibrillation	0.200(0.027)	0.169(0.018)	0.214(0.019)	0.132(0.043)	0.084(0.009)	**0.242(0.015)**
ARF/ARDS	0.285(0.025)	0.253(0.014)	**0.363(0.022)**	0.231(0.014)	0.228(0.011)	0.340(0.022)
Cirrhosis	0.239(0.020)	0.284(0.073)	0.290(0.082)	0.196(0.020)	0.194(0.018)	**0.301(0.050)**
Heart Failure	0.177(0.016)	0.140(0.003)	0.196(0.011)	0.130(0.016)	0.112(0.008)	**0.201(0.006)**
Liver Failure	0.323(0.027)	0.372(0.022)	**0.375(0.027)**	0.272(0.046)	0.302(0.047)	0.357(0.048)
GI Bleeding	0.153(0.027)	0.164(0.009)	0.226(0.013)	0.108(0.015)	0.125(0.006)	**0.268(0.021)**
Severe Sepsis/Septic Shock	0.286(0.018)	**0.300(0.014)**	0.287(0.017)	0.227(0.035)	0.230(0.015)	0.299(0.017)
Cerebrovascular Disease	0.136(0.027)	0.147(0.033)	0.105(0.011)	0.117(0.013)	0.104(0.010)	**0.185(0.021)**
Subarachnoid Hemorrhage	0.191(0.027)	0.199(0.016)	0.189(0.021)	0.188(0.018)	0.183(0.022)	**0.268(0.034)**
Mean	0.222	0.215	0.247	0.165	0.162	**0.266**

Table A.10: **Worked example of MEDIATOR tokenization.** Each string is encoded by the frozen text encoder f_η, projected into the quantization space by p_θ, and assigned to its nearest codebook entry. Three surface forms of dyspnea, from a clinical phrase, a colloquial expression, and the SNOMED CT term, are assigned the same code i, whereas the distinct concept diabetes mellitus is assigned a different code j. Because the codebook is fixed after training, strings from new code systems are tokenized in the same way without retraining.

Input string x	$h = f_\eta(x)$	$z = p_\theta(h)$	$k^\star(z)$
"shortness of breath" (clinical phrase)	h_1	z_1	i
"can't breathe" (colloquial)	h_2	z_2	i
"dyspnea" (SNOMED CT term)	h_3	z_3	i
"diabetes mellitus" (distinct concept)	h_4	z_4	$j \neq i$

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 abstract and introduction state the contributions and their scope, including the coverage asymmetry and matched-input results.

   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: We have a section specifying the limitations of this work.

   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: This is not a theory paper.

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: We fully disclosed all the information needed.

   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: Code is available at `https://github.com/mahmoodlab/MEDIATOR`. MIMIC-III and eICU are available through PhysioNet credentialed access.

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: We specify all the details in paper and appendix.

   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: We conduct 5-fold analyses for all tasks and report the standard deviation across folds. We additionally report Wilcoxon signed-rank tests across tasks and patient-level bootstrap tests (Tables 1 and 2), and Wilcoxon and Friedman tests for the component ablation (Section 4.5).

   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: We specify all the information in appendix.

   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: We have reviewed the content and our work conforms with it.

   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: We provide an impact statement in 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: [Yes]

    Justification: Our model is a medical concept tokenizer with a low risk of misuse, and we provide clear instructions on how to use it properly.

    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: We cite all datasets, models, and resources we 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: [Yes]

    Justification: We release the MEDIATOR code, the generated concept descriptions, and the prompts at `https://github.com/mahmoodlab/MEDIATOR`, with documentation in the repository.

    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: We use no crowdsourcing and recruit no participants. All patient data are analyzed retrospectively.

    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: [Yes]

    Justification: The Mass General Brigham institutional review board approved the retrospective analysis of the records used for the co-occurrence teacher, with a waiver of informed consent (Appendix B.1). MIMIC-III and eICU are de-identified datasets used under the PhysioNet credentialed data use agreement.

    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: LLM is an essential part of our study design.

    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.