A Representation Learning Framework for Clinical Trajectories from Multimodal Longitudinal EHRs

Anonymous ACL submission

Abstract

Understanding heterogeneity in disease progression from longitudinal electronic health records (EHRs) remains challenging due to irregular temporal sampling, missing data, and unstructured clinical narratives. We present an empirical framework that integrates QA-based abstraction of clinical notes with multimodal temporal representation learning to study progression-aware patient representations. By aligning structured physiological measurements and interpretable semantic signals derived from clinical narratives, the framework induces admission-level trajectory representations under temporal irregularity. Using a sepsis cohort from MIMIC-III, we analyze latent trajectory groupings in the learned representation space and characterize their progression patterns, outcomes, and early predictability. Our findings suggest that QA-based abstraction of clinical text, when combined with temporal multimodal modeling, yields more coherent and clinically aligned trajectory representations than existing baselines. The code and data are available at https://github.com/anonymous-2344/clinical_trajectory_phenotype.

1 Introduction

In healthcare, the shift toward personalized and evidence-based care has underscored the importance of understanding heterogeneity in clinical trajectories among patients with the same diagnosis during hospitalization (Johnson et al., 2021; Wang et al., 2021). Characterizing such clinical trajectories is essential for capturing heterogeneity in disease progression, enabling early risk stratification, and supporting interpretable modeling of disease dynamics that can inform prognosis, treatment strategies, and timely intervention. The widespread adoption of electronic health records (EHRs) provides rich longitudinal data capturing patient evolution through both structured physiological measurements and unstructured clinical narratives (Cheng et al., 2016; Alaboud et al., 2023; Jana et al., 2022a). However, modeling disease progression from EHRs remains challenging due to irregular temporal sampling, pervasive missingness, and the complex interactions between numerical signals and free-text documentation (Kim et al., 2023). Effectively characterizing clinical trajectories therefore requires representation learning approaches that can integrate multimodal longitudinal data while explicitly preserving temporal dynamics and clinically aligned variation across patients.

Early studies on disease progression modeling predominantly adopted pathway-based and process-oriented frameworks to extract common clinical pathways Seoane et al., 2014; Zhang et al., 2015; Aisen et al., 2017; Arias et al., 2020. Subsequent work incorporated probabilistic modeling, visual analytics, and data-driven approaches to analyze progression dynamics, risk factors, and stage transitions across diverse diseases Kwon et al., 2020; Zhou et al., 2020; Vesga et al., 2021; Nenova and Shang, 2022; Nagamine et al., 2022. More recently, phenotype discovery methods have shifted toward representation learning, but have largely focused on stratifying patients by disease labels, rather than capturing heterogeneity in progression among patients with the same diagnosis. These approaches commonly rely on unsupervised factorization or topic modeling to identify latent patient groups from co-occurring diagnoses, procedures, and laboratory measurements (Afshar et al., 2020; Becker et al., 2023; Ahuja et al., 2022; Wang et al., 2024), with more recent extensions incorporating temporal information by clustering learned patient representations (Wang et al., 2025; De Freitas et al., 2021; Kauffman et al., 2025). Despite these advances, most existing methods rely predominantly on structured physiological parameters and underutilize unstructured clinical narratives that encode high-level semantic context critical for understanding disease

progression (Dey et al., 2023). Moreover, they often overlook irregular temporal sampling and pervasive missingness, both of which are inherent to real-world EHRs.

To address these limitations, we propose a unified framework for learning progression-aware representations from multimodal longitudinal EHR data and identify clinical trajectory phenotypes. Our main contributions are summarized as follows:

- We propose a QA-based abstraction of clinical narratives that transforms free-text notes into structured, interpretable semantic signals for longitudinal modeling.
- We develop a multimodal temporal representation learning framework that aligns QA-derived text abstractions with irregularly sampled physiological measurements.
- We empirically demonstrate that jointly modeling temporal dynamics and QA-based text abstractions yields more coherent and outcome-aligned trajectory representations than unimodal or non-temporal baselines.
- We present a case study on a sepsis cohort from MIMIC-III to analyze latent trajectory structure and assess early trajectory predictability.

We position this work as a empirical study focused on representation learning and trajectory structure, rather than on definitive clinical phenotype discovery.

2 Related Works

Temporal Representation Learning of EHRs
EHRs offer rich longitudinal signals for diagnosis, outcome prediction, and progression modeling, yet their high dimensionality, irregular sampling, and pervasive missingness challenge conventional models (Kim et al., 2023; Getzen et al., 2023). Early temporal EHR representations leveraged continuous-time formulations, such as latent ODEs and initial value problem–based models, to handle irregular sampling (Rubanova et al., 2019; Xiao et al., 2024). Subsequent approaches incorporated time-aware and personalized RNN to capture inter-visit intervals and patient-specific physiological trends (An et al., 2023; Al Olaimat et al., 2024). More recent studies have explored transformer-based architectures, introducing event-based encoders and temporal self-attention mechanisms to model irregular clinical trajectories without relying on dense imputation (Karami et al., 2024; Song et al., 2025; AlSaad et al., 2024). While these approaches advance temporal modeling of structured EHR data, most focus solely on numerical time series, with only a few recent works jointly modeling structured measurements and clinical narratives in a temporally aligned manner, leveraging multimodal encoders and cross-modal attention to enrich patient representations (Zhang et al., 2023; Ma et al., 2024).

Phenotype Discovery from EHRs
Prior works on EHR-based phenotype discovery relied on unsupervised matrix and tensor factorization to identify latent patient groups, modeling phenotypes as low-rank structures that capture co-occurrence patterns among diagnoses, procedures, and laboratory measurements (Afshar et al., 2020; Becker et al., 2023). Subsequent approaches adopted probabilistic topic models, representing phenotypes as latent topics over multimodal clinical features, with hierarchical extensions enabling the discovery of subphenotypes at multiple levels of granularity (Ahuja et al., 2022; Wang et al., 2024). More recent studies incorporate longitudinal information for learning latent patient representations in a self- or unsupervised manner to infer disease phenotypes, shifting the focus from static disease states to progression-oriented phenotypes (Estiri et al., 2021; Wang et al., 2025; De Freitas et al., 2021; Kauffman et al., 2025). Despite these advances, most existing approaches largely overlook unstructured clinical narratives and rich temporal dynamics, limiting their ability to discover interpretable, clinical trajectory-aware phenotypes from multimodal longitudinal EHRs.

3 Problem Formulation

In a multimodal EHR setting, each patient hospitalization h is observed as a heterogeneous longitudinal trajectory from admission to discharge. Specifically, an episode is represented by (i) static demographic features d_h, (ii) a sequence of structured physiological measurements $\mathcal{S}_h = \{(s_i, t_i)\}_{i=1}^{T_h}$ with $s_i \in \mathbb{R}^L$, and (iii) a temporally aligned sequence of free-text clinical notes $\mathcal{N}_h = \{(n_i, t_i)\}_{i=1}^{T_h}$. At each timestamp t_i, these modalities form a multimodal observation $m_i = (d_h, s_i, n_i, t_i)$, yielding a hospitalization

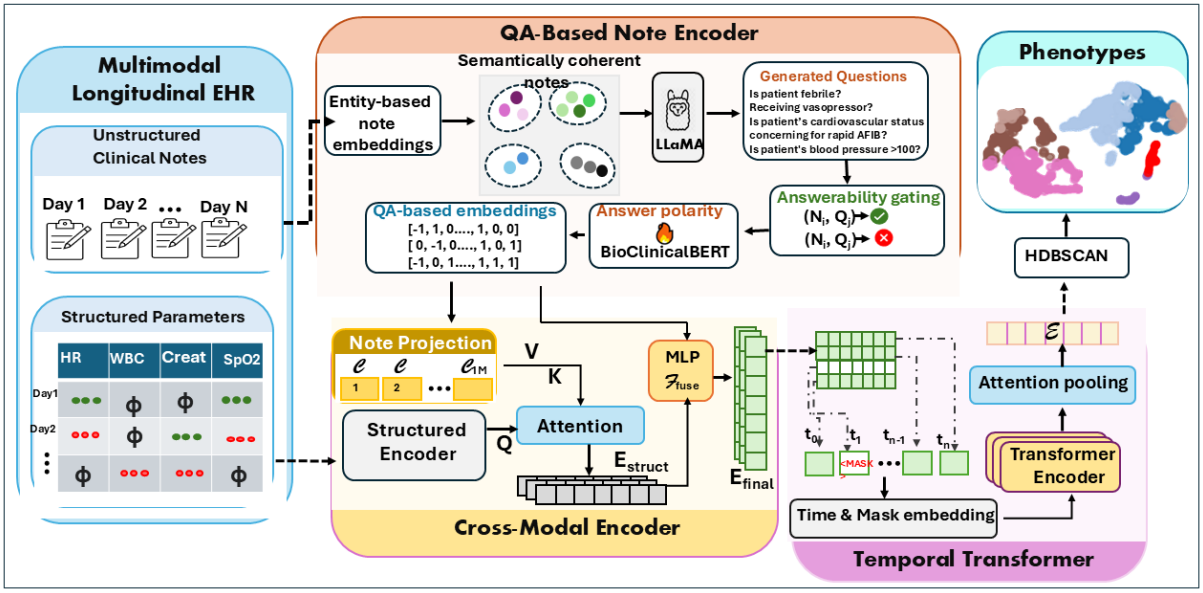

Figure 1: Schematic overview of proposed framework for learning progression-aware admission representation and clinical trajectory phenotypes discovery.

trajectory $\mathcal{M}_h = (m_1, \ldots, m_{T_h})$. We study cohorts sharing the same primary diagnosis so that observed variability primarily reflects differences in disease progression rather than etiology.

Our primary goal is to learn a progression-aware representation of each hospitalization trajectory. We assume that each trajectory is governed by one of K latent *clinical trajectory phenotypes*, denoted by $z_h \in \{1, \ldots, K\}$, corresponding to distinct patterns of temporal evolution in multimodal clinical signals. Using the learned trajectory embeddings, phenotype discovery is performed by grouping admissions with similar progression dynamics, inducing phenotype assignments z_h. As a downstream evaluation of representation quality, we additionally consider *early phenotype prediction*. Given only a prefix of the trajectory up to time t (e.g., the first 48 hours), $\mathcal{M}_{h,\leq t} = (m_1, \ldots, m_t)$, the objective is to predict the phenotype associated with the full hospitalization by estimating $f_\phi : \mathcal{M}_{h,\leq t} \rightarrow P_\phi(z_h \mid \mathcal{M}_{h,\leq t})$.

4 Methodology

We propose a multimodal representation learning framework (Figure 1) that models clinical trajectories from EHRs by jointly encoding unstructured clinical narratives and structured physiological measurements. At each timestamp, textual notes and laboratory or vital signals are aligned and fused into a unified clinical state representation, and aggregated with a Temporal Transformer to capture irregular progression dynamics, yielding progression-aware representations for latent clinical trajectory phenotype inference. The framework components are detailed in the following subsections.

4.1 Encoding Clinical Notes

Prior work on clinical outcome prediction typically encodes clinical notes using transformer-based models such as ClinicalBioBERT (Jana et al., 2022b,a). While effective at capturing lexical and contextual nuances (e.g., distinguishing *severe* from *mild* pain), the resulting dense embeddings are largely opaque and difficult to interpret, limiting clinical trust. In contrast, entity-based representations are interpretable but often sparse and brittle, failing to capture subtle yet clinically important semantic distinctions (Jana et al., 2025).

To illustrate this limitation, consider the following nursing note snippets:

> "Patient febrile, tachypneic, RR 28–30, warm to touch. Complains of chills and feeling weak. Maintained on 3L nasal cannula with SpO_2 91–93%. Alert but tired..."
>
> "Patient reports new abdominal discomfort. Mild tachypnea. On 3L nasal cannula with SpO_2 around 92–94%. Urine output normal..."
>
> "Patient appears lethargic and intermittently confused. On 3L nasal cannula, SpO_2 89–92%. Skin cool. BP trending low..."

Although all three patients receive similar oxygen support (*3L nasal cannula*), their underlying clinical states and progression risks differ

substantially. Simple entity extraction would produce nearly identical representations, obscuring differences in infection severity, hemodynamic instability, and disease evolution. In contrast, a targeted question such as *"Does the patient continue to exhibit fever despite receiving oxygen support?"* directly probes a clinically salient signal and distinguishes these trajectories. Motivated by this, we adopt a contrastive question–answering framework inspired by CQG-MBQA (Sun et al., 2024), in which a large language model(LLM) generates discriminative question–answer pairs to yield interpretable, fine-grained note embeddings for downstream trajectory modeling.

Contrastive Question Generation

Clinical notes are encoded via binary question–answer abstractions using a contrastive question generation framework. To induce contrastive supervision, each note is first mapped to a weakly structured, interpretable representation: a ternary entity vector $e \in \{-1, 0, 1\}^{|\mathcal{V}|}$ obtained by extracting clinically salient entities (e.g., diseases, symptoms, procedures, medications) with a LLM, where values indicate affirmative, negated, or absent mentions (Appendix B). These entity representations are clustered into semantically coherent groups that serve as anchors for question generation. For each cluster, we sample positive notes from clusters, hard negatives from neighboring clusters, and easy negatives from distant clusters. A LLM is then prompted to generate binary clinical questions that elicit consistent responses for positives and contrasting responses for negatives (Appendix C).

LLM-generated questions often exhibit semantic redundancy or inconsistent responses across positive samples. To remove near-duplicates, questions are embedded using the *all-mpnet-base-v2* sentence transformer and clustered via cosine similarity, retaining a single canonical representative per cluster. The remaining questions are scored for discriminative quality. Let $\mathcal{N}_p$ denote the positive set and $\mathcal{N}_h$, $\mathcal{N}_e$ the hard and easy negative sets. Given binary LLM answers $a(q, n) \in \{0, 1\}$, we define:

$$\hat{a}_p(q) = \mathbb{I}\big[\mathbb{E}_{n\sim\mathcal{N}_p}[a(q,n)] > 0.5\big]$$
$$C_{\text{intra}}(q) = \mathbb{E}_{n\sim\mathcal{N}_p}\big[\mathbb{I}\big[a(q,n) = \hat{a}_p(q)\big]\big]$$
$$C_{\text{inter}}(q) = \mathbb{E}_{n\sim\mathcal{N}_h\cup\mathcal{N}_e}\big[\mathbb{I}\big[a(q,n) \neq \hat{a}_p(q)\big]\big]$$

The final discriminative score is $S(q) = C_{\text{intra}}(q)\, C_{\text{inter}}(q)$, and questions with $S(q) > 0.8$ are retained.

Answerability-Aware Question Answering

Directly answering all note-question pairs with a LLM is computationally infeasible: in our setting, over 8,000 clinical notes and 2,000 questions yield ∼16M pairs. We therefore adopt a two-stage answerability-aware framework. An *answerability gating model* first filters unanswerable note-question pairs, followed by a *polarity model* that predicts binary answers for the remaining pairs, substantially reducing inference cost while preserving semantic fidelity.

Not all contrastive questions are answerable from every note, as clinical documentation may omit required evidence (e.g., a note without mention of *urine output* cannot answer *"Is urine output low?"*). To explicitly model such missing information, the gating model encodes the note and question using BioClinicalBERT and assesses answerability via cosine similarity, augmented with overlap between extracted clinical entities. A pair is deemed answerable if the similarity exceeds a threshold (e.g., 0.7) or if at least one relevant entity overlaps, filtering out approximately 20-30% of pairs. Full details are provided in Appendix D.

The polarity model is initialized from ClinicalBERT (Alsentzer et al., 2019) and fine-tuned as a three-class classifier with labels $\{1, -1, 0\}$ corresponding to *Yes*, *No*, and *Cannot Answer*. Despite gating, some pairs remain ambiguous, motivating the explicit *Cannot Answer* class. Training uses a curated subset of LLM-annotated pairs formatted as `[CLS] Question [SEP] Note [SEP]`, with the `[CLS]` representation fed to a classification head. The model is optimized with cross-entropy loss and achieves a macro-F1 of 80% on a held-out set. For more details on training and evaluation, see Appendix E. Finally, we infer answers for all answerable pairs to construct QA-based note embeddings. For a note n, the embedding over retained questions $\mathcal{Q}$ is defined as

$$E_n = \langle l(q) \rangle_{q\in\mathcal{Q}}, \quad l(q) \in \{-1, 0, 1\}$$

where $l(q)$ denotes the predicted polarity. This yields an interpretable note representation that encodes clinically aligned distinctions and serves as input to downstream modeling.

4.2 Cross-Modal Clinical State Representation

We represent each hospitalization as a sequence of timestamped observations $t = 1, \ldots, T$ consist-

ing of unstructured clinical notes and structured physiological measurements. The goal is to learn a unified, timestamp-level clinical state embedding that jointly captures both modalities under irregular sampling and missingness. At timestamp t, clinical notes are now encoded as QA-based embeddings $E_n^{(t)} \in \mathbb{R}^{D_n}$. Structured data comprise a set of clinical parameters $\mathcal{P}$, which are irregularly observed. To avoid heuristic imputation, each parameter $p \in \mathcal{P}$ is represented as a triplet $(v_{t,p}, m_{t,p}, \Delta t_{t,p})$, where $v_{t,p}$ is the observed value (or 0 if missing), $m_{t,p} \in \{0, 1\}$ indicates presence mask, and $\Delta t_{t,p}$ denotes time since last observation. Missing notes are handled analogously via a note-presence mask $m_n^{(t)}$.

Each structured parameter is encoded independently with a shared MLP,

$$s_{t,p} = f_{\text{struct}}(v_{t,p}, m_{t,p}, \Delta t_{t,p}) \in \mathbb{R}^{D_s}$$

yielding a parameter-level representation $S_t = [s_{t,1}, \ldots, s_{t,P}]$. In parallel, the note embedding is projected and decomposed into M latent context vectors $C_t = [c_{t,1}, \ldots, c_{t,M}] \in \mathbb{R}^{M \times D}$, capturing multiple semantic aspects of the narrative.

To align structured measurements with textual evidence, we apply cross-modal attention with structured parameters as queries and note contexts as keys and values: $\tilde{s}_{t,p} = \text{CrossAttn}(s_{t,p}, C_t)$, allowing each parameter to selectively retrieve relevant semantic information from the note. Parameter-wise attention weights $\alpha_{t,p}$ are then learned to aggregate note-conditioned parameters into a structured summary,

$$E_{\text{struct}}^{(t)} = \sum_{p=1}^{P} \alpha_{t,p}\, \tilde{s}_{t,p}$$

Finally, the note and structured summaries are fused to produce the cross-modal clinical state embedding,

$$E_{\text{final}}^{(t)} = f_{\text{fuse}}\big([E_n^{(t)}; E_{\text{struct}}^{(t)}]\big) \in \mathbb{R}^D$$

The model is trained with a self-supervised consistency objective that aligns the fused representation with available modalities:

$$\mathcal{L}_{\text{note}} = \sum_t m_n^{(t)} \big(1 - \text{sim}(E_{\text{final}}^{(t)}, E_n^{(t)})\big)$$

$$\mathcal{L}_{\text{struct}} = \sum_t m_s^{(t)} \big(1 - \text{sim}(E_{\text{final}}^{(t)}, E_{\text{struct}}^{(t)})\big)$$

$$\mathcal{L} = \mathcal{L}_{\text{note}} + \mathcal{L}_{\text{struct}}$$

where $m_s^{(t)}$ indicates the presence of any structured measurement at t and $\text{sim}(\cdot)$ denotes cosine similarity function.

4.3 Temporal Transformer for Disease Progression Modeling

Given an admission a with T_a clinical timestamps, each timestamp $t \in \{1, \ldots, T_a\}$ is represented by a cross-modal clinical state embedding $E_{\text{final},a}^{(t)} \in \mathbb{R}^D$. The admission trajectory is modeled as

$$\mathcal{X}_a = \big\{(E_{\text{final},a}^{(t)},\, m_a^{(t)},\, \Delta t_a^{(t)})\big\}_{t=1}^{T_a}$$

where $m_a^{(t)} \in \{0, 1\}$ denotes observation availability and $\Delta t_a^{(t)}$ is the elapsed time since the previous observed timestamp. Our goal is to aggregate these irregular, partially observed sequences into a fixed-dimensional progression representation. To explicitly encode temporal irregularity and missingness, each timestamp embedding is augmented with learnable time and mask embeddings,

$$z_a^{(t)} = E_{\text{final},a}^{(t)} + \phi(\Delta t_a^{(t)}) + \psi(m_a^{(t)})$$

where $\phi(\cdot)$ is a learnable time embedding and $\psi(\cdot)$ encodes observation presence. Given sequence $Z_a = \{z_a^{(t)}\}_{t=1}^{T_a}$, we apply a stack of Transformer encoder layers: $H_a = \text{TransformerEncoder}(Z_a, M_a)$, where M_a is a key-padding mask derived from $\{m_a^{(t)}\}$. Unlike standard Transformers, we omit absolute positional embeddings, relying on learned temporal encodings to model irregular sampling.

A learnable attention pooling mechanism aggregates the Transformer outputs $H_a = \{h_a^{(t)}\}$ into an admission-level progression embedding $\mathcal{E}_a$:

$$\alpha_a^{(t)} = \frac{\exp(h_a^{(t)\top} q)\, m_a^{(t)}}{\sum_{k=1}^{T_a} \exp(h_a^{(k)\top} q)\, m_a^{(k)}} \quad \mathcal{E}_a = \sum_{t=1}^{T_a} \alpha_a^{(t)} h_a^{(t)}$$

where $q \in \mathbb{R}^D$ is a learnable query vector.

The Temporal Transformer is trained in a fully self-supervised manner using two complementary contrastive objectives. First, a prefix-to-full trajectory loss encourages early clinical states to be predictive of overall disease progression. For a mini-batch of B admissions, let $\mathcal{E}_a^{(p)}$ denote the embedding of a prefix subsequence of admission a, the loss is defined as

$$\mathcal{L}_{\text{pf}} = -\frac{1}{B} \sum_{a=1}^{B} \log \frac{\exp(\text{sim}(\mathcal{E}_a^{(p)}, \mathcal{E}_a)/\tau)}{\sum_{b=1}^{B} \exp(\text{sim}(\mathcal{E}_a^{(p)}, \mathcal{E}_b)/\tau)}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity and τ is a temperature.

Second, to improve robustness to irregular sampling, we generate two stochastic temporal augmentations of each admission by randomly dropping timestamps and perturbing temporal gaps. Let $\mathcal{E}_a^{(1)}$ and $\mathcal{E}_a^{(2)}$ be the corresponding embeddings, the augmentation loss is defined by

$$\mathcal{L}_{\text{aug}} = -\frac{1}{B}\sum_{a=1}^{B}\log\frac{\exp(\mathrm{sim}(\mathcal{E}_a^{(1)},\mathcal{E}_a^{(2)})/\tau)}{\sum_{b=1}^{B}\exp(\mathrm{sim}(\mathcal{E}_a^{(1)},\mathcal{E}_b^{(2)})/\tau)}$$

The final training objective is $\mathcal{L} = \mathcal{L}_{\text{pf}} + \lambda_{\text{aug}}\mathcal{L}_{\text{aug}}$, where λ_{aug} balances trajectory predictiveness and robustness to temporal irregularity.

4.4 Trajectory Phenotype Inference

Latent clinical trajectory phenotypes are inferred by performing fully unsupervised clustering over admission-level progression embeddings $\mathcal{E}_a$ for all admissions $a \in \mathcal{A}$. Each phenotype corresponds to a distinct pattern of disease progression over hospitalization, integrating multimodal signals from clinical narratives and physiological measurements. Prior to clustering, embeddings are standardized via z-score normalization and projected using principal component analysis (PCA) to reduce noise and improve clustering stability. We apply HDBSCAN (McInnes et al., 2017) on projected embeddings to identify intrinsic trajectory phenotypes without pre-specifying their number. HDBSCAN discovers dense regions in the embedding space as phenotypes while assigning sparse or unstable points to a noise set, capturing rare trajectories. Formally, the admissions are partitioned into phenotype clusters $\{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$ and a set of atypical trajectories $\mathcal{P}_N$.

5 Experiments

Dataset We evaluate our framework on a cohort of 1,660 sepsis-related hospital admissions derived from MIMIC-III v1.4, a large publicly available clinical database (Johnson et al., 2016). Among clinical narratives, we utilize *nursing progress notes*, which are recorded at high temporal frequency and provide comprehensive documentation of patients' evolving clinical status. In addition, we extract 27 irregularly sampled vital sign and laboratory variables selected from established sepsis severity scores (SOFA, qSOFA, and SIRS), supplemented with commonly monitored inflammatory, metabolic, renal, hepatic, and hematologic markers recommended in sepsis management guidelines (see full list in Appendix I-Table 6). Hospitalizations shorter than 48 hours are excluded to ensure sufficient temporal context for progression modeling and early phenotype prediction. Summary statistics of the cohort are reported in Appendix A-Table 3. See Appendix H for detailed experimental setup. Next, we present a comprehensive evaluation of the proposed framework across three core aspects: (i) quality of learned representations, (ii) clinical characteristics of inferred trajectory phenotypes, and (iii) early phenotype prediction as a downstream probing task. Collectively, these analyses examine the model's representational robustness, clinical interpretability, and practical applicability.

Representation Quality Analysis Primarily, 30,360 nursing progress notes are extracted and consolidated into daily notes by concatenating multiple entries per hospital day. Using the proposed note encoding pipeline, we construct 2,424 discriminative contrastive questions spanning disease states, comorbidities, symptoms, interventions, and progression patterns (examples in Appendix J, Fig. 6). The answerability gate identifies approximately 10M answerable note–question pairs, of which 132K are annotated with LLM-generated labels to train the polarity classifier. The trained model infers answers for all answerable pairs, producing QA-based note embeddings (statistics in Appendix F, Table 4). These embeddings are fused with structured vital and laboratory measurements and aggregated by the temporal transformer to yield admission-level trajectory representations.

We assess representation quality prior to phenotype inference, focusing on structural organization, temporal coherence, and clinical relevance. Figure 2 presents UMAP visualizations of admission-level embeddings under different representation learning variants in ablation study. Embeddings are visualized using two complementary clinical annotations length-of-stay bins and in-hospital mortality. The results indicate that jointly modeling clinical text and structured signals with explicit temporal dynamics produces the most coherent and clinically aligned embedding structure, supporting the suitability of the learned representations for unsupervised clinical trajectory phenotype inference.

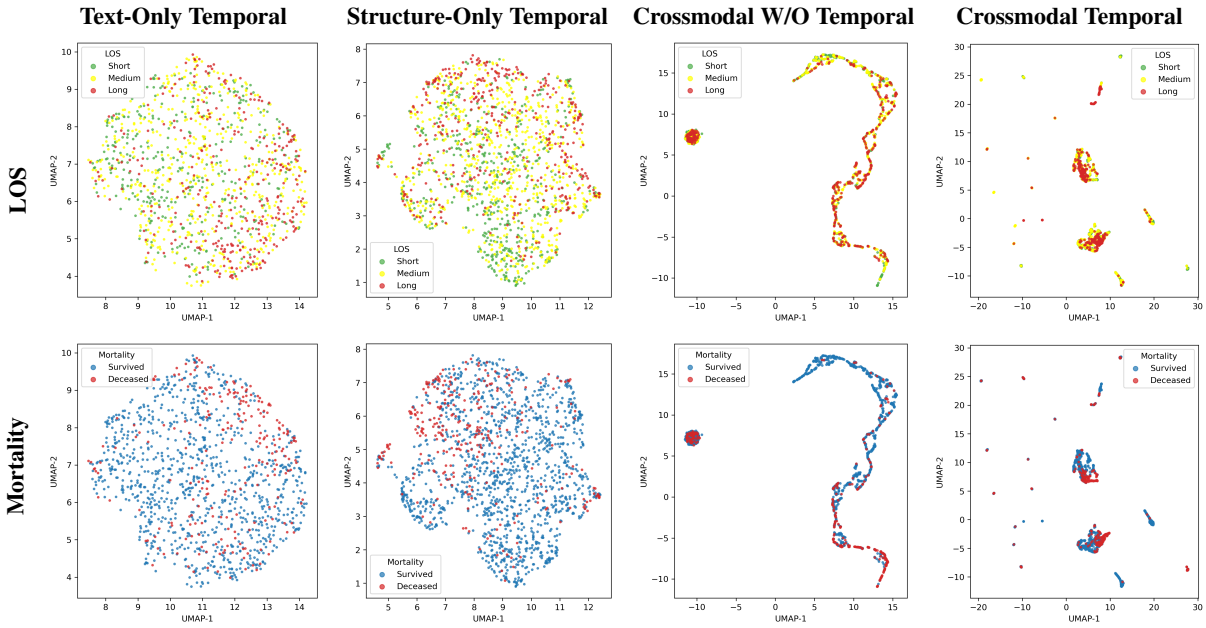

Figure 2: UMAP visualizations of admission-level trajectory embeddings under different modeling settings colorings by length of stay (Short: $\leq$6 , Medium: 7–14, and Long: $>$14 days) (top) and in-hospital mortality (bottom).

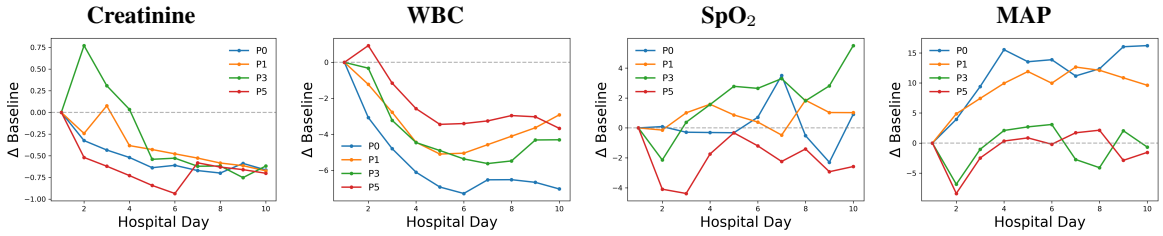

Figure 3: Normalized temporal trajectories of four key physiological variables across four clinical trajectory phenotype.

Clinical Trajectory Phenotypes Using the learned admission-level progression embeddings, we infer latent clinical trajectory phenotypes via fully unsupervised HDBSCAN clustering, yielding 16 distinct progression groups and automatically isolating 0.2% of admissions as noise corresponding to rare or atypical trajectories. Figure 5 (Appendix G) presents a UMAP visualization illustrating clear separation among progression patterns. A global characterization of each phenotype including cohort size, baseline demographics, outcome severity, and dominant clinical features, is summarized in Table 1.

To derive interpretable phenotype-level descriptors, we temporally aggregate QA-based note embeddings by averaging question-level polarity responses across hospitalization days, yielding trajectory-level QA profiles that capture the persistence and directionality of salient clinical signals. This analysis reveals substantial heterogeneity in disease severity and outcomes across phenotypes. Notably, phenotypes $P0$, $P5$, and $P6$ are associated with the highest mortality rates and prolonged hospital stays, with $P0$ representing the most adverse progression pattern.

We analyze phenotype-specific temporal dynamics using structured physiological variables. Figure 3 shows normalized trajectories of representative markers (*creatinine*, *WBC*, *SpO*$_2$, and *MAP*) across four major phenotypes. To mitigate bias from irregular sampling, trajectories are normalized per admission relative to each patient's first observation, emphasizing within-patient trends. Analysis is limited to the first 10 hospital days, where phenotype divergence is most pronounced. The results reveal heterogeneous recovery patterns, ranging from rapid normalization of inflammatory and renal markers to persistent dysregulation with di-

Table 1: Summary statistics of inferred clinical trajectory phenotypes (P).

P	N (%)	Age median [IQR]	LOS (days) median [IQR]	Mortality (%)	Dominated Clinical Characteristics
P0	195(11.7)	67.0[55.0, 78.0]	9.0[6.0, 13.0]	5.6	Sepsis with fungal and GI complications
P1	278(16.7)	73.5[61.0, 83.0]	11.0[7.0, 19.0]	36.0	Delayed recovery with cardiac abnormalities
P2	143(8.6)	65.0[49.5, 78.0]	6.0[5.0, 8.0]	1.4	Improving sepsis with renal dysfunction
P3	184(11.1)	66.0[53.0, 79.0]	9.0[7.0, 13.0]	3.3	Improving sepsis with multi-organ recovery
P4	130(7.8)	66.0[53.2, 78.0]	7.0[5.0, 9.0]	4.6	Sepsis recovery with acute renal failure
P5	281(16.9)	71.0[59.0, 80.0]	14.0[7.0, 13.0]	39.9	Delayed recovery after severe sepsis
P6	40(2.4)	65.5[59.8, 78.2]	3.0[3.0, 3.0]	40.0	Worsening septic shock, multi-organ failure
P7	38(2.3)	65.0[50.5, 80.0]	4.0[4.0, 4.0]	13.2	Infection resolution with hepatic recovery
P8	53(3.2)	71.0[57.0, 80.0]	5.0[5.0, 5.0]	15.1	Metabolic recovery during sepsis
P9	40(2.4)	70.5[57.8, 84.0]	6.0[6.0, 6.0]	15.0	Severe sepsis with multi-organ dysfunction
P10	37(2.2)	65.0[59.0, 79.0]	7.0[7.0, 7.0]	5.4	Resolving sepsis with renal–metabolic issues
P11	32(1.9)	72.0[65.8, 81.2]	8.0[8.0, 8.0]	6.2	Systemic recovery after multi-organ failure
P12	22(1.3)	68.5[62.2, 83.5]	9.0[9.0, 10.0]	4.5	Respiratory dysfunction during recovery
P13	20(1.2)	65.5[51.5, 83.5]	10.0[10.0, 10.0]	15.0	Renal recovery with respiratory dysfunction
P14	25(1.5)	72.0[65.0, 81.0]	11.0[11.0, 11.0]	4.0	Multi-organ recovery with neurologic improvement
P15	95(5.7)	66.0[53.0, 76.5]	15.0[13.0, 21.0]	13.7	Stable sepsis with liver dysfunction
Noise	47(2.8)	67.0[55.0, 78.0]	18.0[9.5, 29.0]	10.6	Atypical trajectories

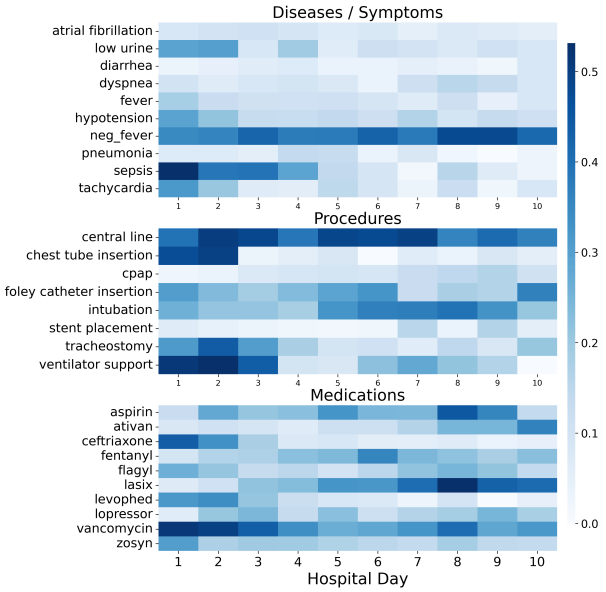

Figure 4: Heatmap visualization of phenotype P0.

vergent hemodynamic and oxygenation trends. To further assess interpretability, we visualize prototype clinical trajectories using QA-based semantic signals; Figure 4 summarizes the temporal prevalence of dominant conditions, procedures, and medications for phenotype $P0$.

Early Phenotype Prediction

Early trajectory prediction is used as a probing task to evaluate whether early clinical signals encode information consistent with later trajectory structure, rather than as a standalone clinical objective. Prefix embeddings constructed from the first 48 hours of hospitalization are used to predict inferred trajectory phenotypes via a lightweight MLP classifier. Table 2 reports performance across embedding variants using Accuracy and macro-F1 to account for class imbalance. Results show that prevalent phenotypes are often identifiable from early signals, whereas rarer phenotypes remain difficult to predict within the first 48 hours. This disparity indicates that early data captures coarse progression trends, while finer-grained trajectory distinctions emerge later or require additional temporal context.

Table 2: Early trajectory phenotype prediction performance.

Model (Inputs)	**Accuracy**	**Macro-F1**
GPT-5 (Notes)	0.45	0.29
ClinicalBERT (Notes)	0.49	0.25
ClinicalBERT + temporal	0.51	0.31
Our Model (Note)	0.51	0.48
Our Model (Structure)	0.55	0.50
Our Model (crossmodal)	**0.76**	**0.61**

6 Conclusion

In this paper, We present a multimodal representation learning framework for modeling disease progression from irregular longitudinal EHRs by integrating QA-based abstractions of clinical narratives with time-aware modeling of structured data. Experiments on a sepsis cohort from MIMIC-III show that jointly capturing interpretable semantic signals from text and explicit temporal dynamics yields more coherent and progression-aware trajectory representations. The resulting trajectory groupings are data-driven and serve to reveal progression structure in learned representations rather than define validated clinical phenotypes.

7 Limitations

Our proposed note encoding pipeline relies on LLM at multiple stages, including contrastive question generation, primary note abstraction, and the construction of labeled data for training the answer polarity model. While this enables scalable and interpretable representation learning, it also introduces dependence on LLM behavior, including potential hallucinations or inconsistencies. In the current work, we do not incorporate explicit human-in-the-loop validation or automated quality control mechanisms for assessing the reliability of LLM-generated questions and annotations, which we leave to future work.

Also, the inferred clinical trajectory phenotypes are discovered in a fully unsupervised manner and are evaluated primarily through quantitative outcomes and qualitative signal analysis. Determining whether these trajectory groupings correspond to clinically actionable or meaningful phenotypes ultimately requires expert clinical validation, which is beyond the scope of this study. Finally, our experimental evaluation is conducted on a single disease cohort (sepsis) derived from MIMIC-III. Although the proposed framework is disease-agnostic, additional validation across diverse clinical conditions and healthcare settings is necessary to assess its generalizability and robustness.

References

Ardavan Afshar, Ioakeim Perros, Haesun Park, Christopher Defilippi, Xiaowei Yan, Walter Stewart, Joyce Ho, and Jimeng Sun. 2020. Taste: temporal and static tensor factorization for phenotyping electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 193–203.

Yuri Ahuja, Yuesong Zou, Aman Verma, David Buckeridge, and Yue Li. 2022. Mixehr-guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record. *Journal of biomedical informatics*, 134:104190.

Paul S Aisen, Jeffrey Cummings, Clifford R Jack, John C Morris, Reisa Sperling, Lutz Frölich, Roy W Jones, Sherie A Dowsett, Brandy R Matthews, Joel Raskin, and 1 others. 2017. On the path to 2025: understanding the alzheimer's disease continuum. *Alzheimer's research & therapy*, 9:1–10.

Mohammad Al Olaimat, Serdar Bozdag, and Alzheimer's Disease Neuroimaging Initiative. 2024. Ta-rnn: an attention-based time-aware recurrent neural network architecture for electronic health records. *Bioinformatics*, 40(Supplement_1):i169–i179.

Khuder Alaboud, Imad Eddine Toubal, Butros M Dahu, Akram Abo Daken, Ammar Ali Salman, Nouha Alaji, Wael Hamadeh, and Ahmad Aburayya. 2023. The quality application of deep learning in clinical outcome predictions using electronic health record data: a systematic review. *South East. Eur. J. Public Heal*, pages 09–23.

Rawan AlSaad, Qutaibah Malluhi, Alaa Abd-Alrazaq, and Sabri Boughorbel. 2024. Temporal self-attention for risk prediction from electronic health records using non-stationary kernel approximation. *Artificial Intelligence in Medicine*, 149:102802.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ying An, Guanglei Cai, Xianlai Chen, and Lin Guo. 2023. Parse: A personalized clinical time-series representation learning framework via abnormal offsets analysis. *Computer Methods and Programs in Biomedicine*, 242:107838.

Michael Arias, Eric Rojas, Santiago Aguirre, Felipe Cornejo, Jorge Munoz-Gama, Marcos Sepúlveda, and Daniel Capurro. 2020. Mapping the patient's journey in healthcare through process mining. *International journal of environmental research and public health*, 17(18):6586.

Florian Becker, Age K Smilde, and Evrim Acar. 2023. Unsupervised ehr-based phenotyping via matrix and tensor decompositions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(4):e1494.

Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 432–440. SIAM.

Jessica K De Freitas, Kipp W Johnson, Eddye Golden, Girish N Nadkarni, Joel T Dudley, Erwin P Bottinger, Benjamin S Glicksberg, and Riccardo Miotto. 2021. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns*, 2(9).

Lipika Dey, Sudeshna Jana, Tirthankar Dasgupta, and Tanay Gupta. 2023. Deciphering clinical narratives-augmented intelligence for decision making in healthcare sector. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 11–24. IEEE.

Hossein Estiri, Zachary H Strasser, and Shawn N Murphy. 2021. High-throughput phenotyping with temporal sequences. *Journal of the American Medical Informatics Association*, 28(4):772–781.

Emily Getzen, Lyle Ungar, Danielle Mowery, Xiaoqian Jiang, and Qi Long. 2023. Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of Biomedical Informatics*, 139:104269.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sudeshna Jana, Tirthankar Dasgupta, and Lipika Dey. 2022a. Predicting medical events and icu requirements using a multimodal multiobjective transformer network. *Experimental Biology and Medicine*, 247(22):1988–2002.

Sudeshna Jana, Tirthankar Dasgupta, and Lipika Dey. 2022b. Using nursing notes to predict length of stay in icu for critically ill patients. In *Multimodal AI in healthcare: A paradigm shift in health intelligence*, pages 387–398. Springer.

Sudeshna Jana, Tirthankar Dasgupta, and Manjira Sinha. 2025. A data-driven approach to predicting and visualizing disease progression in hospitalized patients. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1–4.

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

Kevin B Johnson, Wei-Qi Wei, Dilhan Weeraratne, Mark E Frisse, Karl Misulis, Kyu Rhee, Juan Zhao, and Jane L Snowdon. 2021. Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93.

Hojjat Karami, David Atienza, and Anisoara Ionescu. 2024. Tee4ehr: Transformer event encoder for better representation learning in electronic health records. *Artificial Intelligence in Medicine*, 154:102903.

Justin Kauffman, Emma Holmes, Akhil Vaid, Alexander W Charney, Patricia Kovatch, Joshua Lampert, Ankit Sakhuja, Marinka Zitnik, Benjamin S Glicksberg, Ira Hofer, and 1 others. 2025. Infehr: Clinical phenotype resolution through deep geometric learning on electronic health records. *Nature Communications*, 16(1):8475.

Michelle Kang Kim, Carol Rouphael, John McMichael, Nicole Welch, and Srinivasan Dasarathy. 2023. Challenges in and opportunities for electronic health record-based data analysis and interpretation. *Gut and liver*, 18(2):201.

Bum Chul Kwon, Vibha Anand, Kristen A Severson, Soumya Ghosh, Zhaonan Sun, Brigitte I Frohnert, Markus Lundgren, and Kenney Ng. 2020. Dpvis: Visual analytics with hidden markov models for disease progression pathways. *IEEE transactions on visualization and computer graphics*, 27(9):3685–3700.

Yingbo Ma, Suraj Kolla, Dhruv Kaliraman, Victoria Nolan, Zhenhong Hu, Ziyuan Guan, Yuanfang Ren, Brooke Armfield, Tezcan Ozrazgat-Baslanti, Tyler J Loftus, and 1 others. 2024. Temporal cross-attention for dynamic embedding and tokenization of multimodal electronic health records. *arXiv preprint arXiv:2403.04012*.

Leland McInnes, John Healy, Steve Astels, and 1 others. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Tasha Nagamine, Brian Gillette, John Kahoun, Rolf Burghaus, Jörg Lippert, and Mayur Saxena. 2022. Data-driven identification of heart failure disease states and progression pathways using electronic health records. *Scientific Reports*, 12(1):17871.

Zlatana Nenova and Jennifer Shang. 2022. Chronic disease progression prediction: Leveraging case-based reasoning and big data analytics. *Production and Operations Management*, 31(1):259–280.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. 2019. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

José A Seoane, Ian NM Day, Tom R Gaunt, and Colin Campbell. 2014. A pathway-based data integration framework for prediction of disease progression. *Bioinformatics*, 30(6):838–845.

Ziyang Song, Qincheng Lu, He Zhu, David Buckeridge, and Yue Li. 2025. Trajgpt: Irregular time-series representation learning of health trajectory. *IEEE Journal of Biomedical and Health Informatics*.

Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony KH Tung, and Jun Yu. 2024. A general framework for producing interpretable semantic text embeddings. *arXiv preprint arXiv:2410.03435*.

Jasmin I Vesga, Edilberto Cepeda, Campo E Pardo, Sergio Paez, Ricardo Sanchez, Rafael M Sanabria, and 1 others. 2021. Chronic kidney disease progression and transition probabilities in a large preventive cohort in colombia. *International journal of nephrology*, 2021.

Fuying Wang, Feng Wu, Yihan Tang, and Lequan Yu. 2025. Ctpd: Cross-modal temporal pattern discovery for enhanced multimodal electronic health records analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6783–6799.

Table 3: Summary statistics of the sepsis cohort derived from MIMIC-III.

Statistic	Value
# patients	1,518
# hospital admissions	1,660
Age (mean ± std)	66.9 ± 16.4
Male / Female (%)	54.2 / 45.8
LOS days (median, [IQR])	9, [6, 14]
Note Category used	Nursing
Notes per admission (median, [IQR])	5, [4, 9]
Structured variables	27
Timestamp granularity	Each hospital day
Early prediction window	First 48 hours

Meina Wang, Roy S Herbst, and Chris Boshoff. 2021. Toward personalized treatment approaches for non-small-cell lung cancer. *Nature medicine*, 27(8):1345–1356.

Ruohan Wang, Zilong Wang, Ziyang Song, David Buckeridge, and Yue Li. 2024. Mixehr-nest: Identifying subphenotypes within electronic health records through hierarchical guided-topic modeling. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–8.

Jingge Xiao, Leonie Basso, Wolfgang Nejdl, Niloy Ganguly, and Sandipan Sikdar. 2024. Ivp-vae: modeling ehr time series with initial value problem solvers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16023–16031.

Xinlu Zhang, Shiyang Li, Zhiyu Chen, Xifeng Yan, and Linda Ruth Petzold. 2023. Improving medical predictions by irregular multimodal electronic health records modeling. In *International conference on machine learning*, pages 41300–41313. PMLR.

Yiye Zhang, Rema Padman, and Nirav Patel. 2015. Paving the cowpath: Learning and visualizing clinical pathways from electronic health record data. *Journal of biomedical informatics*, 58:186–197.

Yulong Zhou, Zhicheng Zhang, Jie Tian, and Shaoyun Xiong. 2020. Risk factors associated with disease progression in a cohort of patients infected with the 2019 novel coronavirus. *Ann Palliat Med*, pages 428–436.

A Dataset Statistics

Table 3 summarizes key statistics of the dataset used in our experiments.

B Entity-Based Note Embedding

As a weakly structured representation for inducing contrastive supervision in question generation, we construct an entity-based embedding of clinical notes that captures semantically salient clinical content while remaining interpretable. For each nursing progress note, we extract clinically relevant entities spanning three categories: (i) diseases and symptoms, (ii) implemented therapeutic procedures, and (iii) administered medications. Entity extraction is performed using the *LLaMA-3.1-8B-Instruct* model (Grattafiori et al., 2024) via a carefully designed prompt to ensure faithful, non-hallucinatory extraction.

To explicitly capture clinical negation, which is pervasive in clinical narratives (e.g., *"denies chest pain"*, *"no history of shortness of breath"*), the model is instructed to prefix negated mentions with `neg_`, treating affirmative and negated concepts as distinct entities. This preserves clinically aligned polarity information critical for downstream contrastive learning.

Prompt Template for Clinical Entity Extraction

Task Description: You are an expert clinical text analyzer. Extract *only* the entities explicitly mentioned in the given nursing note. Do not infer or assume information beyond the text. Entities to extract: 1. Diseases/Symptoms: - Include disease names, symptoms, and any abnormalities mentioned. - If a symptom or condition is explicitly stated as absent or negative for the patient, prefix it with "neg_". 2. Implemented Procedures: - Include only therapeutic or surgical procedures performed or planned (e.g., intubation, catheter insertion, surgery). - Do NOT include diagnostic tests like ECG, X-ray, MRI here. 3. Medications: - Include all medication names mentioned. Return the output in three separate lists in JSON format as shown below:

```
{
  "diseases_symptoms": [ ... ],
  "procedures": [ ... ],
  "medications": [ ... ]
}
```

Do not provide explanations, reasoning, or any additional text just the required output.

Moreover, we observed that entity extraction exhibits substantial synonymy and terminological variation (e.g., *"hemorrhage"* vs. *"bleeding"*). To normalize entity mentions, we map extracted entities to the UMLS Metathesaurus (Schuyler et al., 1993), which assigns each concept a unique Concept Unique Identifier (CUI). When an exact UMLS match is unavailable, we compute semantic

similarity between the extracted entity and candidate UMLS concepts using the *all-mpnet-base-v2* SBERT model (Reimers and Gurevych, 2019). Entity pairs with cosine similarity exceeding an empirically determined threshold of 0.9 are treated as equivalent. Entities that cannot be reliably mapped are assigned unique identifiers to ensure condition was overlooked.

Let $\mathcal{V}$ denote the resulting entity vocabulary. Each note n is represented as a ternary entity vector $E_n \in \{-1, 0, 1\}^{|\mathcal{V}|}$, defined as: $E_n = \langle f(e_i) \rangle_{i=1}^{|\mathcal{V}|}, \quad e_i \in \mathcal{V}$, where

$$f(e_i) = \begin{cases} 1, & \text{if } e_i \text{ is affirmatively mentioned in } n, \\ -1, & \text{if } e_i \text{ is mentioned in negated form,} \\ 0, & \text{if } e_i \text{ is not mentioned.} \end{cases}$$

This representation preserves clinically salient presence and polarity information while providing an interpretable, weakly structured embedding that serves as the basis for contrastive question generation.

C Contrastive Question Generation with LLMs

We generate contrastive clinical questions using the *LLaMA-3.1-8B-Instruct* model (Grattafiori et al., 2024). For each semantically coherent note cluster, we construct three anchor sets: *positive samples*, *hard negatives*, and *easy negatives*. Specifically, we sample five positive notes from each cluster (drawing from both high-density cores and boundary regions), five hard negatives from adjacent clusters, and five easy negatives from distant clusters. These anchor sets provide weak contrastive supervision by exposing clinically subtle and coarse distinctions between trajectories.

In our dataset, nursing notes are lengthy (averaging ∼2,000 tokens), making it infeasible to include all anchor notes directly within a single LLM prompt. To address this limitation, we first summarize each note using the same LLM, producing a concise yet faithful clinical abstraction. Summarization is performed with the following prompt, which enforces strict factuality and clinical focus.

Prompt Template for Nursing Note Summarization

Task Description: You are a clinical summarizer. Read the provided nursing note and write *one* clear, factual paragraph of approximately 250–300 words.
Requirements:

- Include *only* information explicitly stated in the note.
- Do not infer, speculate, or add new facts.
- Focus on clinical status, including respiratory function, vitals, hemodynamics, urine output, neurological status, intake/output, medications, interventions, and lines or tubes.
- Use concise, direct clinical language.
- If a clinical aspect is not mentioned, omit it.

Output format: Exactly one paragraph, no headings, bullet points, or explanations.

The resulting summaries are used as inputs for contrastive question generation. The LLM is instructed to generate binary (yes/no) clinical questions that yield consistent answers for all positive samples while producing opposite answers for both hard and easy negatives. To encourage clinically aligned discrimination, the model is guided to focus on disease states, symptoms, comorbidities, physiological conditions, interventions, and medications.

Prompt Template for Contrastive Question Generation

Task Description: You are a clinical question generator. You are given three groups of nursing note summaries:

- **Group A:** Clinically similar notes.
- **Group B:** Hard negatives—notes that appear similar but differ in key clinical aspects.
- **Group C:** Easy negatives—notes that differ more clearly.

Generate the 10 most discriminative yes/no clinical questions (each $\leq$12 words) that distinguish Group A from Groups B and C.
Instructions:

- Use only information present in the provided notes.
- Do not invent or infer clinical facts.
- Each question must be answerable from the note text.
- Questions should yield the *same* answer for all Group A notes and the *opposite* answer for all Group B and C notes.
- Prefer questions related to diseases, symptoms, comorbidities, clinical states, interventions, medications, or progression patterns.

Output format: A numbered list of 10 questions only, with no additional text.

D Answerability Gating Model

Contrastive questions are generated using a limited set of positive and negative anchor notes (Appendix C). As a result, many generated questions are not answerable from arbitrary notes outside the anchor sets, since the required clinical information may be absent. To avoid unreliable or spurious question-answer predictions, we introduce an *answerability gating model* that filters note-question pairs prior to polarity inference.

Let n denote a clinical note and q a contrastive question. We first encode both n and q using a shared BioClinicalBERT encoder, yielding dense contextual representations

$$\mathbf{h}_n = \mathrm{BERT}(n), \quad \mathbf{h}_q = \mathrm{BERT}(q)$$

We compute their semantic alignment via cosine similarity,

$$\mathrm{sim}(n, q) = \frac{\mathbf{h}_n^\top \mathbf{h}_q}{\|\mathbf{h}_n\| \|\mathbf{h}_q\|}.$$

In addition to semantic similarity, we incorporate explicit clinical grounding through entity overlap. Using the entity extraction procedure described in Appendix B, we extract sets of clinical entities from the note and the question, denoted by E_n and E_q, respectively. These entities include diseases, symptoms, procedures, and medications, with negated mentions treated as distinct entities. A note-question pair (n, q) is considered *answerable* if it satisfies either of the following conditions:

$$\mathrm{sim}(n, q) > \tau \quad \text{or} \quad \mathcal{E}_n \cap \mathcal{E}_q \neq \emptyset,$$

where τ is a similarity threshold empirically set to 0.7.

This dual criterion ensures that a question is retained if it is either semantically aligned with the note content or explicitly grounded in shared clinical entities. The gating model thus removes question-note pairs lacking sufficient semantic or clinical support, substantially reducing the number of candidates passed to the downstream polarity prediction model while preserving answerable cases. In practice, this filtering step eliminates approximately 20-30% of candidate pairs, yielding a more reliable and computationally efficient question answering pipeline.

E Answer Polarity Model

Following answerability filtering, we train an *answer polarity model* to predict the semantic polarity of answerable note-question pairs. The task is formulated as a three-class classification problem with labels $\{-1, 0, 1\}$ corresponding to *No*, *Cannot Answer*, and *Yes*, respectively.

Training Data Construction To construct supervision at scale, we generate pseudo-labeled training data using a pretrained LLM. For each clinical note in the dataset, we randomly sample 20 contrastive questions from the retained question set and prompt *GPT-5-mini* () to produce one of three responses: *Yes*, *No*, or *Cannot Answer*. This process yields a curated dataset of approximately 132K labeled note-question pairs, which we split into training, validation, and held-out test sets.

Model Architecture We initialize the polarity model from BioClinicalBERT and adopt a cross-encoder formulation. Each input pair is formatted as

$$\texttt{[CLS]}\ q\ \texttt{[SEP]}\ n\ \texttt{[SEP]},$$

where q is the contrastive question and n is the clinical note. The contextualized representation of the `[CLS]` token is passed to a linear classification head that outputs a probability distribution over the three polarity classes.

Training Objective The model is fine-tuned using categorical cross-entropy loss. Training is performed using the Adam optimizer with a learning

Table 4: Statistics of QA-based note embeddings.

Statistic	Value
Retained contrastive questions	2,424
Avg. answerable questions / note	959
Informative answers (% ± 1)	74.1%
Unanswerable answers (% 0)	25.9%

rate of 1×10^{-5}, batch size 4, and 10 epochs. Fine-tuning on 100K training samples requires approximately 9 hours on a single GPU.

Evaluation On the held-out test set, the polarity model achieves a macro-F1 score of 80% and an overall accuracy of 86%, indicating strong agreement with LLM-generated labels. These results demonstrate that the model reliably reproduces answer polarity while remaining computationally efficient for large-scale inference.

During inference, the trained polarity model is applied to all answerable note-question pairs to produce the final QA-based note embeddings used in downstream cross-modal and temporal modeling.

F Statistics of QA-Based Note Embeddings

Table 4 display the summary statistics of QA-based note embeddings.

G UMAP visualizations of Phenotypes

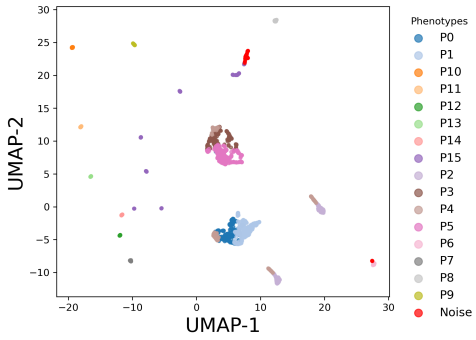

Figure 5: UMAP visualization of 16 clinical trajectory phenotypes.

H Experimental Setup

All experiments were conducted on a server with an NVIDIA Tesla V100 GPU (32 GiB), 9 vCPUs, and 60 GiB RAM. The model was implemented in PyTorch. The final hyperparameters are reported in Table 5.

I Structured Physiological Variables

Table 6 lists the 27 structured vital sign and laboratory variables used in our study. Variables are selected based on established sepsis severity scores (SOFA, qSOFA, SIRS) and supplemented with ICU-relevant inflammatory, metabolic, renal, hepatic, hematologic, and electrolyte markers commonly used in sepsis management. For days with multiple records of the same variable, we retain the *worst* value, defined as the measurement indicating greatest physiological derangement for sepsis.

Table 5: Hyperparameter settings used in the experimental evaluation.

Component	Hyperparameter (Value)
Cross-Modal Encoder	Note embedding dimension: 2424 Structured embedding dimension: 27×3 Number of structured variables: 27 Cross-attention heads: 4 Output dimension: 256 Batch size: 4 Epoch: 10 Optimizer: Adam Learning Rate: $1e-3$
Temporal Transformer	Transformer layers: 4 Attention heads: 4 Model dimension: 256 Time embedding dimension: 64 Contrastive temperature (τ): 0.07 Batch size: 16 Epoch:5 Optimizer: Adam Learning Rate: $1e-3$ λ_{aug}: 0.2
Clustering (HDBSCAN)	PCA components: 50 Minimum cluster size: 30 Minimum samples: 10 Distance metric: Euclidean Cluster selection method: EOM

J Example of Contrastive Question–Answer Generation

Figure 6 illustrates two nursing notes and the associated contrastive questions with their generated answers.

Table 6: List of structured vital sign and laboratory variables used in this study.

Parameter	Clinical Rationale	Worst-Value Criterion
Heart Rate	SIRS, hemodynamic instability	Higher (tachycardia)
Respiratory Rate	SIRS, respiratory distress	Higher (tachypnea)
Temperature	SIRS, infection severity	Distance from normal (hypo/hyperthermia)
Mean Arterial Pressure (MAP)	SOFA, shock severity	Lower (hypotension)
Glasgow Coma Scale (GCS)	SOFA, neurological dysfunction	Lower (impaired consciousness)
SpO_2	Oxygenation status	Lower (hypoxemia)
FiO_2	Respiratory support intensity	Higher (oxygen requirement)
White Blood Cell Count (WBC)	SIRS, inflammation	Distance from normal (leukocytosis/leukopenia)
Platelet Count	SOFA	Lower (thrombocytopenia)
Bilirubin	SOFA, hepatic dysfunction	Higher
Creatinine	SOFA, renal dysfunction	Higher
Blood Urea Nitrogen (BUN)	Renal dysfunction	Higher
Lactate	SOFA	Higher
AST	Hepatic injury	Higher
ALT	Hepatic injury	Higher
Albumin	inflammatory status	Lower
Hemoglobin	Oxygen-carrying capacity	Lower
INR	Coagulation abnormality	Higher
pH	Acid–base status	Distance from normal (acidosis/alkalosis)
HCO_3	Metabolic dysregulation	Lower
Glucose	Metabolic dysregulation	Distance from normal (hypo/hyperglycemia)
Sodium	Electrolyte imbalance	Distance from normal
Potassium	Electrolyte imbalance	Distance from normal
Chloride	Acid–base balance	Distance from normal
Calcium	Cardiac/electrolyte stability	Lower
Magnesium	Cardiac/electrolyte stability	Lower

Note A	Note B
pt. 90year old man, was admitted to the micu via the ew for elevated temp, elevated heart rate (afib) and hypotension.....pt has a good strong productive cough of thick yellow sputum. cxr done, and then down for a chest c-scan---lg right pleural effusion, also rt mass.....Cont triple abx vanco/flagyl/cefepine.Patient trached with 8.0 Portex. Pt. weaned to PSV 5, Peep 5, Fio2 40% yesterday.....	pt is a 63yom with admitted via EW with sepsis from infected Hickman (dialysis) catheter. Pt. had hx of MSSA, MRSA, VRE, and c-diff which was treated recently.....pt remains on vassopressin at .04 units /min, and levophed gtt has been slowly weaned to .13mcg/kg/min. maintaining mean pressure of greater than 65....sedated on propofol and fentanyl gtts. Lightened once at MN per Dr...Clear breath sounds bilaterally...Occ pac's noted, no ectopy, no afib...A/P chest showed parenceymal abnormality LL with persistent pulomnary edema heart normal size with no pleural effusion.

Q_Id	Questions	Answer A	Answer B
1	Is the patient receiving IV Vanco?	Yes	No
17	Is the patient's cardiovascular status concerning for rapid AFIB?	Yes	No
3	Does the patient have a pleural effusion?	Yes	No
23	Does the patient have a history of MSSA bacteremia?	No	Yes
36	Is the patient receiving IV Zosyn?	No	No
58	Is the patient's blood pressure stable on vasopressors?	No	Yes
64	Is the patient on fentanyl for pain?	No	Yes
83	Is the patient's blood pressure weaned off dopamine?	No	Can't answer
94	Is the patient's respiratory status concerning with thick yellow sputum?	Yes	No
144	Was the patient's respiratory status improved with bipap therapy?	No	No

Figure 6: Example of contrastive question–answer generation for two nursing notes.