Efficient EHR Foundational Models: A Mixture-of-Experts Approach for Patient Timeline Prediction

Sudhanva Manjunath Athreya Matthias Christenson Warren Woodrich Pettine

AI Summit 2025

Problem Statement

Hospitals face challenges in predicting critical patient outcomes like mortality and ICU stays. Traditional rule-based systems, unable to capture complex EHR patterns, are falling short. While Generative AI offers promise for predicting patient timelines from large EHR datasets, training these models is computationally expensive due to the high dimensionality and sparsity of EHR data.

Background & Motivation

- Clinical event prediction improves patient care and resource allocation.
- EHR systems cannot capture intricate semantic and temporal pattern in patient data.
- Foundational models show potential for predicting patient trajectories.
- High dimensionality and sparsity increase training costs.

Data Processing & Tokenization

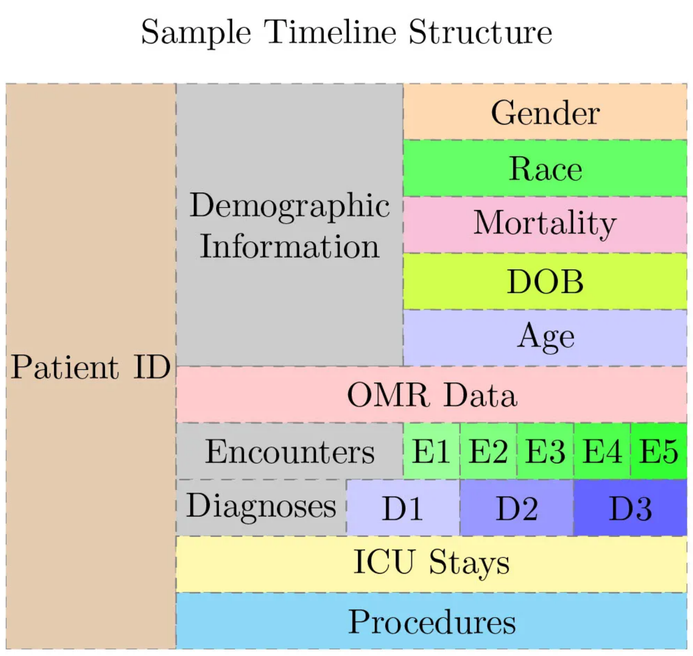


Figure 1. Patient EHR Timeline Representation

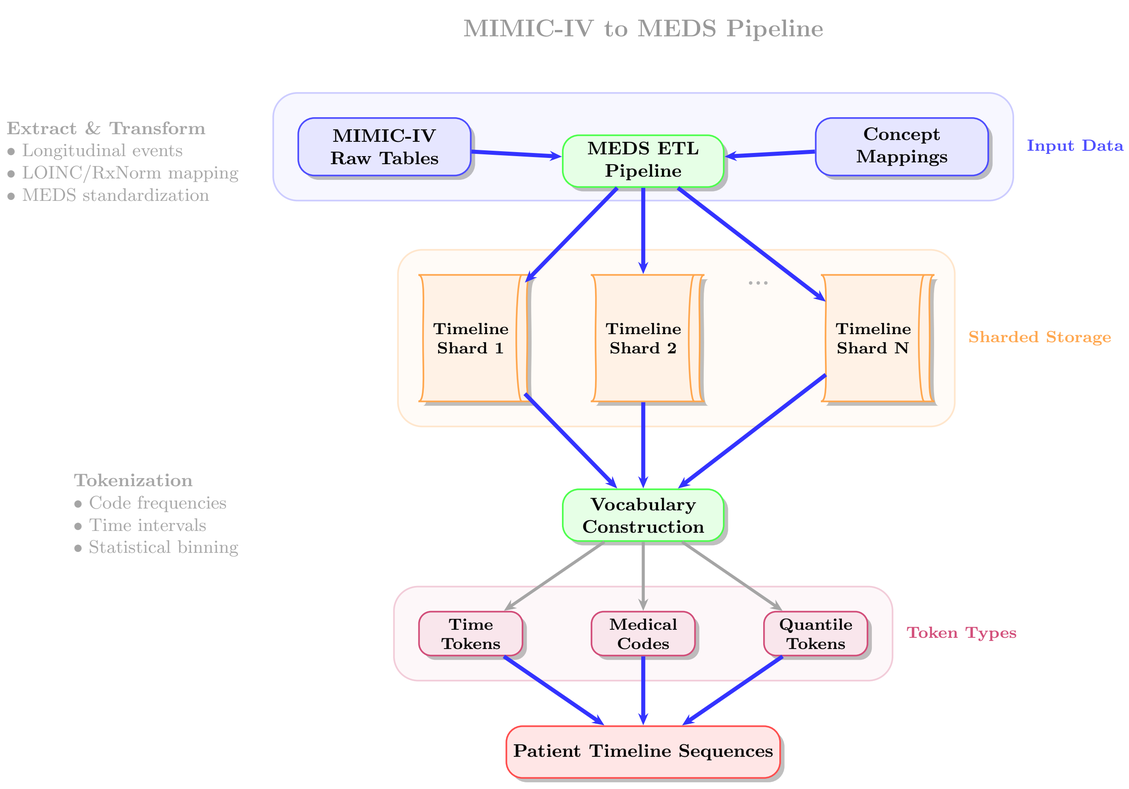


Figure 2. EHR Data Processing & Tokenization Pipeline

Our Approach: Mixture-of-Experts (MoE)

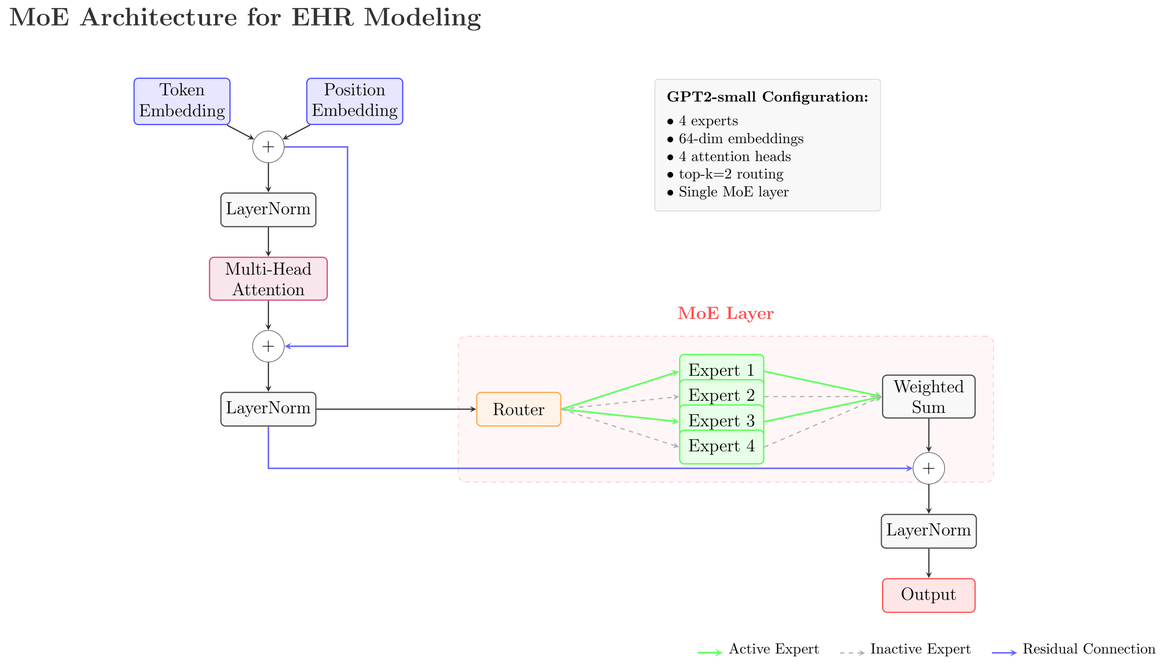
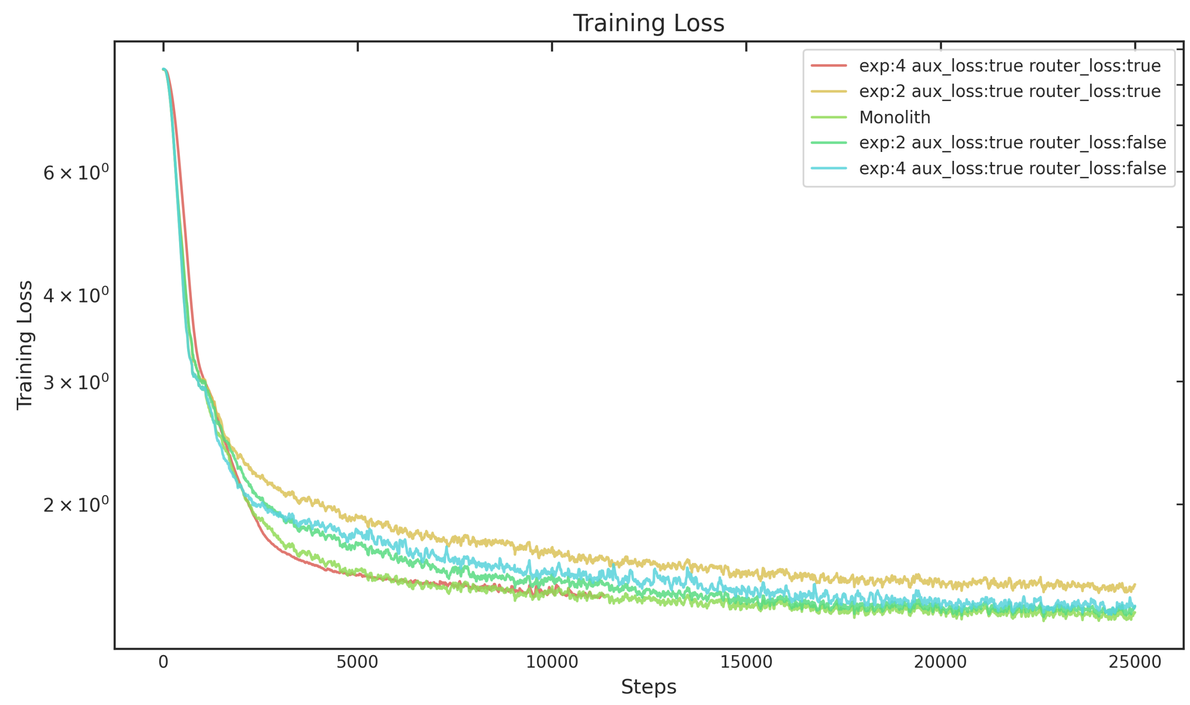


Figure 3. MoE Architecture for EHR Modeling

Experimental Results



Evaluation Metrics:

- **Clinical Performance:** ROC-AUC, accuracy, precision, and recall for clinical NTP tasks
- **Model Efficiency:** Parameter count comparison and computational complexity analysis (MACs and FLOPs operations)
- **Training Dynamics:** Convergence rate, training time, and resource utilization metrics
- **Inference Performance:** Batch processing latency, throughput, and memory consumption during validation

Configuration	Monolith	n_exp=4, aux=T, z_loss=F	n_exp=2, aux=T, z_loss=F	n_exp=4, aux=T, z_loss=T	n_exp=2, aux=T, z_loss=T
All Tokens (Top-3)					
Accuracy	**0.759**	0.752	0.746	0.735	0.721
Precision	**0.253**	0.251	0.249	0.245	0.240
Recall	**0.759**	0.752	0.746	0.735	0.721
MEDS_DEATH					
ROC-AUC	**0.997**	0.995	0.994	0.993	0.991
Accuracy (Top-3)	0.050	0.033	0.033	**0.139**	0.033
Precision (Top-3)	**0.273**	0.091	0.091	0.155	0.030
Recall (Top-3)	0.182	0.091	0.091	**0.241**	0.091
HOSPITAL_ADMISSION					
ROC-AUC	0.998	0.998	**0.999**	0.998	0.998
Accuracy (Top-3)	**0.859**	0.807	0.839	0.758	0.843
Precision (Top-3)	0.379	0.339	**0.388**	0.349	0.251
Recall (Top-3)	0.833	**0.834**	0.827	0.776	0.801
HOSPITAL_DISCHARGE					
ROC-AUC	**0.996**	0.995	**0.996**	0.995	0.995
Accuracy (Top-3)	0.590	0.559	**0.640**	0.578	0.546
Precision (Top-3)	**0.301**	0.282	0.258	0.250	0.241
Recall (Top-3)	0.622	0.591	**0.671**	0.585	0.552

HUNTSMAN MENTAL HEALTH INSTITUTE UNIVERSITY OF UTAH
HEALTH UNIVERSITY OF UTAH
KAHLERT SCHOOL OF COMPUTING
THE UNIVERSITY OF UTAH