Clinical Decision Support using Pseudo-notes from multiple streams of EHR Data

Simon A. Lee[1], Sujay Jain[1], Alex Chen[1], Kyoka Ono[1], Arabdha Biswas[1], Akos Rudas[1], Jennifer Fang[2-4], Jeffrey N. Chiang[1,5]

1. Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA
2. LA Health Services, Los Angeles, CA, USA
3. Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA, USA
4. Department of Emergency Medicine, University of California Los Angeles, Los Angeles, CA, USA
5. Department of Neurosurgery, University of California, Los Angeles, Los Angeles, CA, USA.

Corresponding Author: Jeffrey N. Chiang (njchiang@g.ucla.edu)

Funding

COI

Word Count: 4215
Abstract Word Count: 148
Tables: 1
Figures: 8

ABSTRACT
Electronic health records (EHR) contain data from disparate sources, spanning various biological and temporal scales. In this work, we introduce the Multiple Embedding Model for EHR (MEME), a deep learning framework for clinical decision support that operates over heterogeneous EHR. MEME first converts tabular EHR into "pseudo-notes", reducing the need for concept harmonization across EHR systems and allowing the use of any state-of-the-art, open source language foundation models. The model separately embeds EHR domains, then uses a self-attention mechanism to learn the contextual importance of these multiple embeddings. In a study of 400,019 emergency department visits, MEME successfully predicted emergency department disposition, discharge location, intensive care requirement, and mortality. It outperformed traditional machine learning models (Logistic Regression, Random Forest, XGBoost, MLP), EHR foundation models (EHR-shot, MC-BEC, MSEM), and GPT-4 prompting strategies. Due to text serialization, MEME also exhibited strong few-shot learning performance in an external, unstandardized EHR database.

Introduction

In recent years, increased access to Electronic Health Records (EHR) has enabled the development and application of clinically relevant artificial intelligence (AI) and machine learning (ML). For example, both traditional and cutting-edge machine learning techniques have been harnessed to augment medical image interpretation[1], drug discovery and delivery [2,3], diagnosis [4,5], and prognosis [6], to name a few [7,8]. Due to the large variety of bespoke clinical applications built upon clinical health data, recent efforts have turned to developing generalist AI for healthcare [9,10]. Foundation models (FMs), the basis of large, generalist AI, are pre-trained on massive amounts of diverse data which exhibit adaptability and effectiveness across numerous domains[11]. These models have been shown to be adaptable into the healthcare setting, exhibiting state-of-the-art performance in multiple settings [12].

The application of FMs to healthcare generally fits into one of two paradigms [13]. One approach augments widely-accessible large language models (LLMs) with clinical text (e.g., ClinicalBert[14], MedPaLM[15], GPT[11], etc), taking advantage of the general reasoning capabilities of these models. For example they have recently been able to generate discharge summaries from structured EHR without being trained on that particular task. However, continued adaptation of these models has been hampered by the fact that they are restricted to a text-based interface, making them incompatible with tabular EHR.

Another group of FMs are trained from scratch to operate upon sequences of discrete, structured items captured within the EHR (e.g., BEHRT[16] and its variants[17]). EHR FMs have been shown to exhibit better predictive performance than bespoke ML models. However, there are substantially less data available to develop EHR FMs, which casts doubt on their general utility across diverse healthcare populations[13]. In addition to the relative lack of publicly available EHR for developing EHR FMs, a data standard is yet to be adopted that harmonizes tabular EHR across institutions [18–20].

EHR are recorded in a variety of data types including numerical, categorical, and free-text, which traditional ML has struggled to jointly process. These issues are partially addressed by EHR FMs, which can be configured to process categorical codes and continuous measurements[21], but are limited by the need to harmonize these concepts. While EHR are commonly referred to and modeled by FMs as a single data type, these records span multiple biological scales and domains from laboratory measurements, to clinical interpretations and actions, to diagnostic codes. It is possible that this approach doesn't capture the underlying distributions given the high cardinality of EHR data and the relatively small amount of training data.

In this work, we present a modeling framework for EHR decision support that addresses the gaps above. First, we introduce clinical pseudo-notes, which convert tabular EHR into text. Our rationale is that text-serialization provides an alternative to concept harmonization and serves as an interface between EHR and LLMs. We leverage these pseudo-notes to develop the Multiple Embedding Model for EHR (MEME) for clinical decision support. MEME separately processes EHR concepts which are combined using self-attention. We demonstrate the

effectiveness of this approach through various prediction tasks in the emergency department setting, showing that our framework outperforms both traditional machine learning models and EHR FMs.

Results

Study Design and Cohort

This study was conducted retrospectively on datasets collected from the Beth Israel Deaconness Medical Center in Boston, USA[22] and the UCLA Health medical system in Los Angeles, USA (UCLA). From each database, deidentified electronic health records from emergency room visits were identified and extracted (MIMIC: n=400,019, UCLA: n=947,028) with additional details in Table 1. These were used to predict discharge and decompensation outcomes including Emergency Room disposition (EDdisp), discharge location (discharge), intensive care (ICU), and mortality as defined in Chen et al., 2024 [23]. The publicly available MIMIC database was used for model development and validation. We used a 70/15/15 split for the MIMIC Dataset treating each patient visit independently.

Table 1: Demographic and Clinical Characteristics of Patient Encounters at MIMIC and UCLA Hospitals: This table summarizes the demographic details such as median age, gender, and racial distribution, as well as clinical outcomes including emergency department disposition and details of ED decompensation among admitted patients. The data includes total patient encounters for MIMIC (n=400,019) and UCLA (n=947,028).

Patient Encounters	MIMIC (n=400,019)	UCLA (n=947,028)
Median Age [IQR]	56 [35, 71]	42 [18, 66]
Male (n, %)	195,189 (48.8%)	518,532 (54.8%)
Race (n,%)		
White	228,123 (57.0%)	583,925 (61.7%)
African American	76,798 (19.2%)	139,497 (14.7%)
Asian	18,528 (4.7%)	83,329 (8.8%)
Other	76,570 (19.1%)	140,277 (14.8%)
Outcomes (n,%)		
Ed Disposition	158,007 (39.5%)	239,598 (25.3%)
ED Decompensation (Only	**N = 158,007**	**N = 239,598**

Patient Encounters	MIMIC (n=400,019)	UCLA (n=947,028)
Median Age [IQR]	56 [35, 71]	42 [18, 66]
Male (n, %)	195,189 (48.8%)	518,532 (54.8%)
Race (n,%)		
White	228,123 (57.0%)	583,925 (61.7%)
African American	76,798 (19.2%)	139,497 (14.7%)
admitted patients)		
Discharge Location	70,945 (44.9%)	91,287 (38.1%)
ICU	31,127 (19.7%)	37,616 (15.7%)
Mortality	4,582 (2.9%)	7427 (3.1%)

Multiple Embedding Model for EHR (MEME)

The goal of our framework is to design a model that leverages off-the-shelf text models to represent EHR data effectively, addressing the challenges of variable-length inputs and the multistream nature of clinical records (e.g., triage information, medication info, vitals, etc). Streams of EHR data include, for example, diagnostic codes, prescription orders, and triage vitals, which represent separate biological and temporal scales. MEME processes each stream independently, embedding concepts separately to overcome token limit constraints (e.g., BERT's 512-token limit), thus preserving the integrity of patient data without truncation (See Methods).

Preprocessing: Conversion to Clinical Pseudo-notes

Electronic Health Records (EHR) are heterogeneous datasets encompassing various biological and temporal scales, represented across multiple tables in categorical, numerical, and textual formats. Integrating these data types presents additional challenges in terms of data harmonization and standards adoption. Instead, we perform text-serialization[24] in which tabular EHR are converted to text, which we refer to as clinical pseudo-notes.

While text generation with Large Language Models (LLMs) has been explored[25–27], persistent issues such as data hallucination pose significant challenges, as noted in[24]. Our approach uses a template approach (Figure 1), in which structured data is inserted into a pre-configured template. This resembles the manner in which the majority of clinical text is generated (e.g, SmartPhrases/DotPhrases from the Epic/EMR system[28]).

We process each stream independently, assigning a distinct embedding to each EHR data stream. This results in multiple paragraphs of clinical pseudo-notes, each containing a separate “domain” of EHR (diagnoses, encounter metainformation, medications, vitals, and information at triage). These embeddings are then concatenated and subjected to a self-attention layer, which synthesizes the entire patient context prior to decision-making (Figure 1; see Methods for additional details).

Multiple embedding for decision support

Our approach assigns distinct embeddings to each EHR stream, which are then concatenated and processed through a self-attention layer to synthesize the patient representation for decision-making. Prior work, such as ExBEHRT[17], along with our findings, demonstrates that this method improves performance by avoiding the truncation and ordering issues of single, heterogeneous embeddings. Figure 1 illustrates a schematic of the framework's workflow.

Embeddings for each pseudo-note paragraph are extracted using language foundational models, resulting in high-dimensional vectors that capture various aspects of a patient's medical history. These embeddings are then concatenated into a unified input vector for further processing. In the proceeding step, a self-attention layer analyzes the combined vector as a whole, capturing relationships between different medical concepts. The processed vector is then

passed through a classifier to predict outcomes such as ED Disposition or Decompensation, with the model optimized using a tailored loss function.

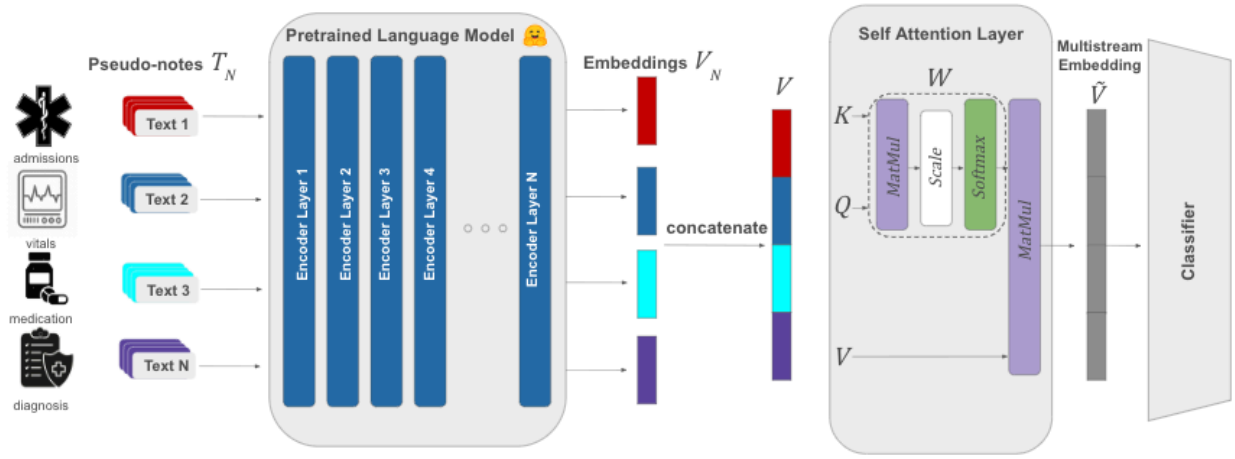
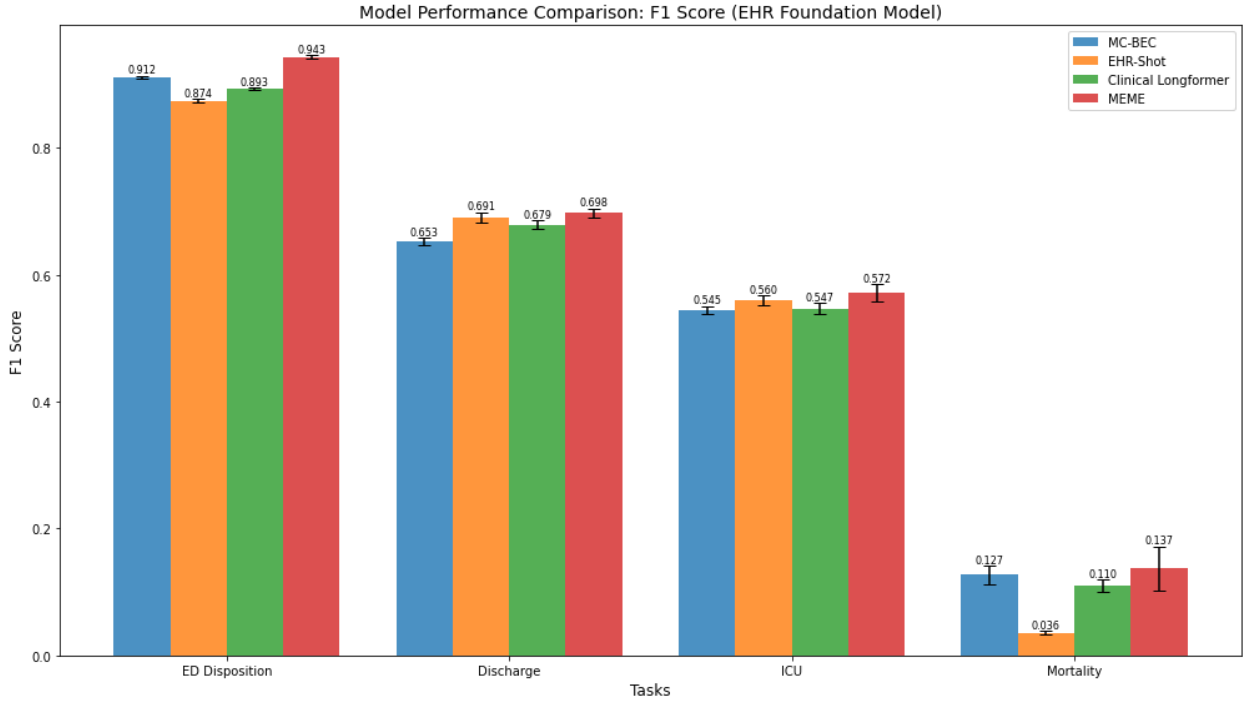
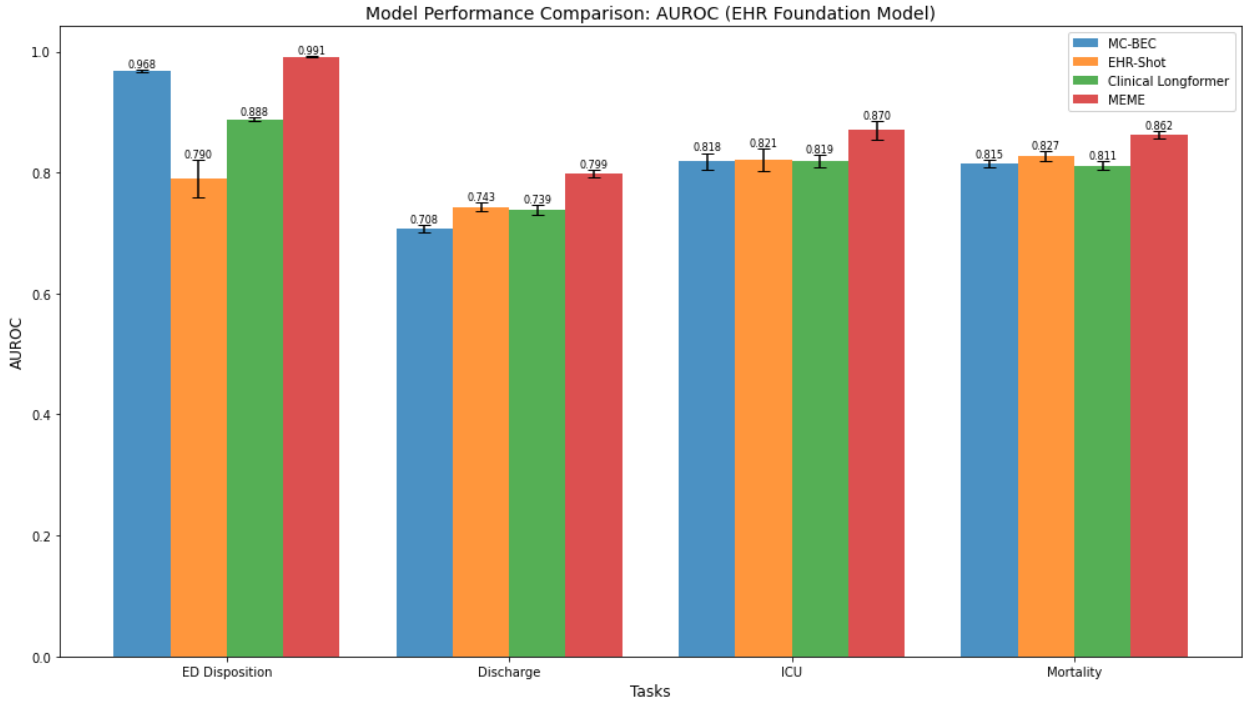


Figure 1: Overview of the Multiple Embedding Model for Electronic Health Records (EHR). This model integrates various input streams from distinct biological and temporal concepts of the EHR. Each concept is represented independently before being merged and processed through a self-attention layer. This multistream embedding is then passed through a Fully connected layer for downstream prediction.

This approach contrasts with existing efforts to develop foundational models for EHR representation [16,29,30]. We evaluated performance against representative foundational EHR models as well as baseline non-deep learning models trained from scratch[31]. Reference model comparisons were run within the MIMIC dataset due to data harmonization issues with the institutional database and quantified in terms of the Area Under the Receiver Operating Characteristic Curve (AUROC), the Area Under Precision-Recall Curve (AUPRC), and F1 scores. 95% confidence intervals were generated for each metric by resampling the test set 1,000 times.

MEME vs EHR Foundation Models





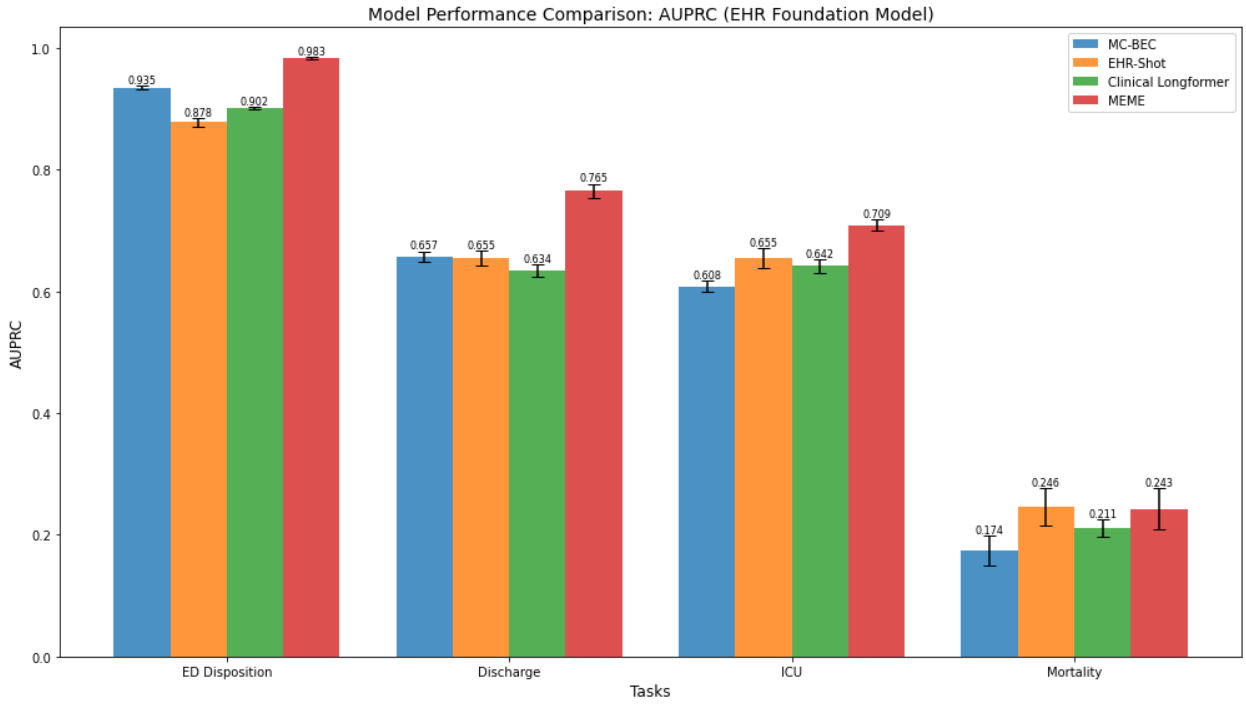
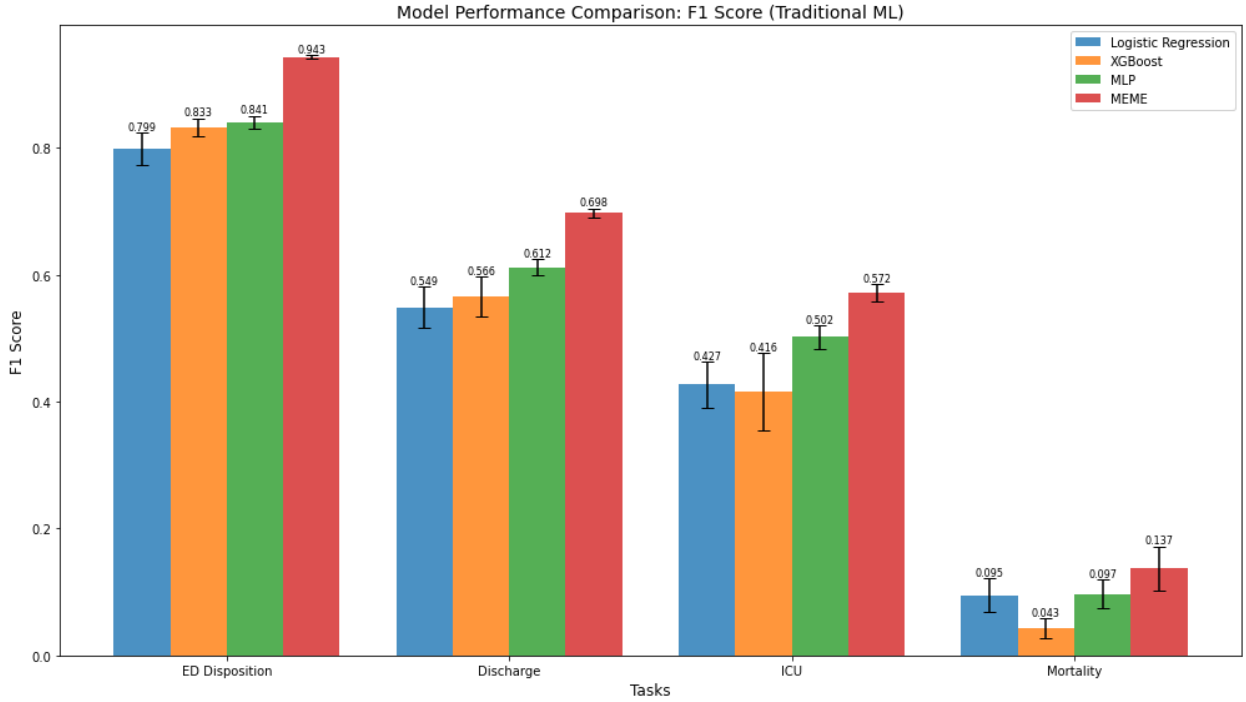


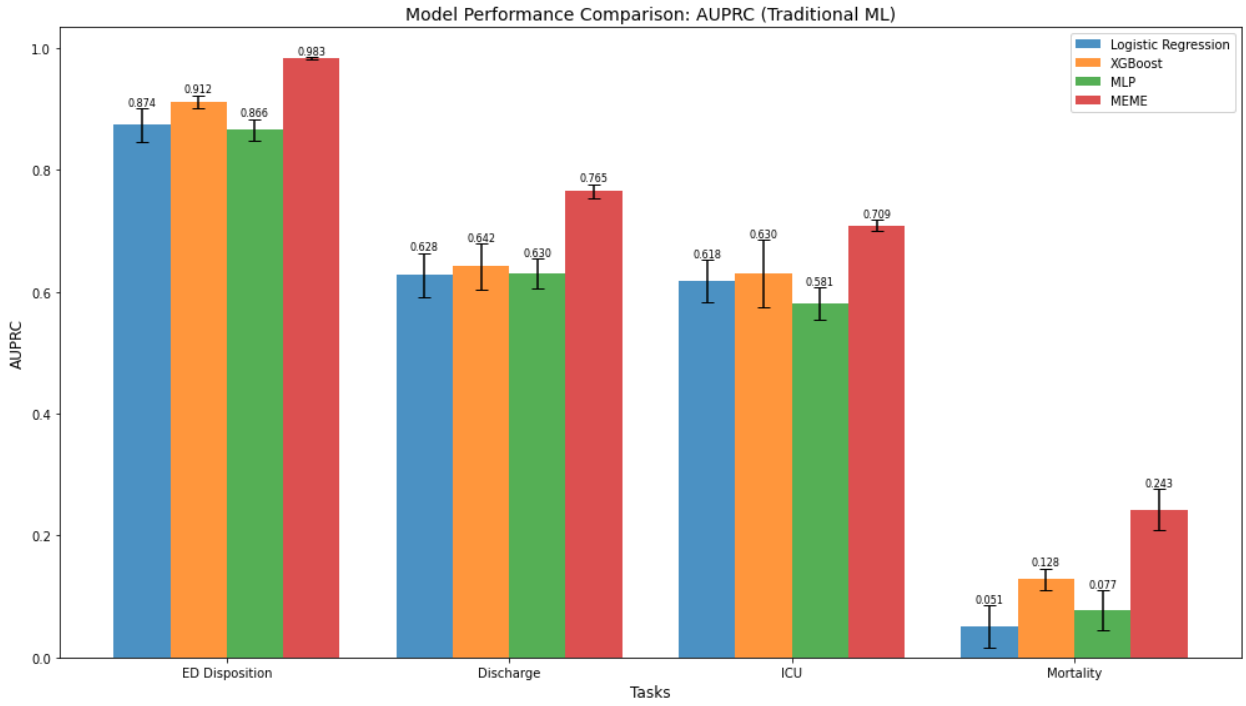
Figure 2a,b,c: Comparative Performance of EHR Foundation models: MC-BEC, EHR-Shot, and MEME Models Across Different Clinical Tasks. This bar chart displays the F1 Scores, Area under the Receiver Operating Characteristic (AUROC) and Area under the precision recall curve (AUPRC) for each model, assessed over various clinical tasks such as ED Disposition, Discharge, Tasks, ICU, and Mortality. Error bars represent the confidence intervals, highlighting the variability in model performance across tasks.

EHR-specific foundation models (EHR FMs) have been recently developed and have shown predictive capabilities across a variety of healthcare applications [13] We selected the following reference EHR FMs as representatives of the approach.

On the MIMIC validation set, MEME significantly outperformed EHR FMs in ED disposition as displayed in Figure 2. In the context of decompensation, MEME outperformed EHR FMs in all metrics when predicting ICU necessity and either outperformed or was statistically indistinguishable from EHR FMs when predicting mortality. We also show that by increasing the context window in the clinical longformer[32] (512 tokens vs 1024 tokens), it does not necessarily result in better performance supporting the added benefit of our framework design.

MEME vs traditional ML





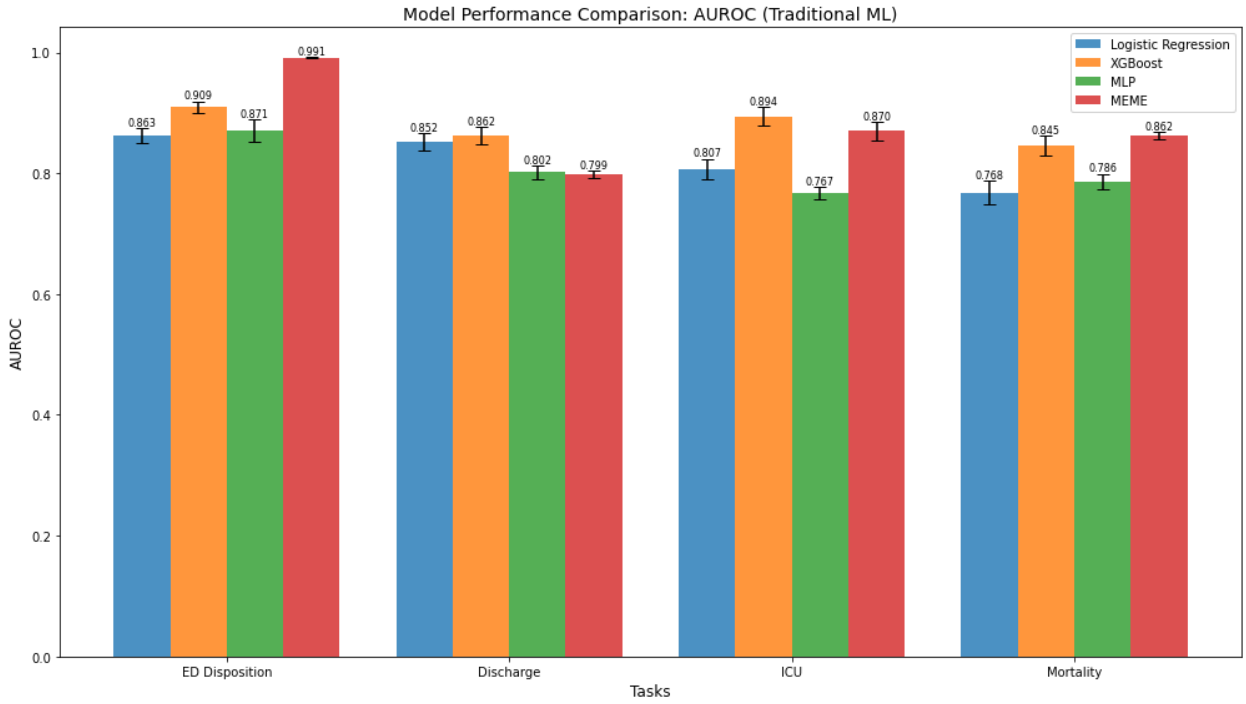


Figure 3a,3b,3c: Comparative Performance of MEME relative to Traditional ML techniques: Logistic Regression, XGBoost, MLP, MEME. This bar chart displays the AUPRC, AUROC, F1 for each model, assessed over various clinical tasks such as ED Disposition, Discharge, Tasks, ICU, and Mortality. Error bars represent the confidence intervals, highlighting the variability in model performance across tasks. All Statistics for the ED Traditional techniques including logistic regression and gradient boosting may be appropriate when there is a clear relation between input features and prediction targets, for example when decision protocols are governed by quantitative thresholds.

We evaluated MEME against a logistic regression, xgboost, and neural network model[31] operating over tabular EHR prior to pseudonote generation (Table/Figure 3a,3b,3c). MEME significantly outperformed these approaches in ED disposition. Evaluation on decompensation tasks were varied. The xgboost classifier outperformed MEME in terms of AUROC for discharge and ICU, and in terms of F1 for mortality. However, MEME significantly outperformed the same approaches in terms of AUPRC across all tasks. This could be due to differences in the incidence of these events, as discussed in[33].

Ablation Studies

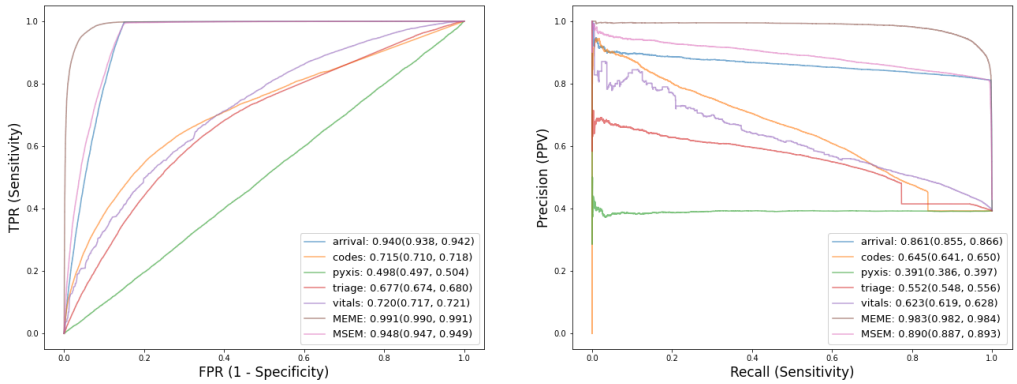


Figure 4: Ablation Study Comparing Different Model Variants on Area under the Receiver Operating Characteristic (AUROC) and Precision-Recall Curves (AUPRC). The left panel displays AUROC curves for independent concept models versus a Multiple Single Embedding Model (MSEM) and our MEME model. The right panel illustrates AUPRC curves, depicting the precision-recall relationship for the same models. Model performance metrics (AUC values) are annotated on both curves. It is evident that neither any single modality nor MSEM outperforms MEME.

MEME is composed of the combination of pseudo-notes as an interface between EHR and natural language LMs, and a multiple embedding approach in which EHR data domains are separately embedded. As shown above, the combination of these approaches achieves comparable or superior performance to alternative approaches for EHR modeling. We conducted the following ablation studies to characterize the contribution of the multiple embedding approach.

MEME was referenced against a single-modality embedding model (MSEM; see Methods), in which pseudo-notes were combined into one large text that produced a single embedding. This significantly compromised predictive performance in all scenarios tested, highlighting the importance of embeddings for separate modalities to MEME's performance[17].
An ablation study was also conducted to characterize the contribution of different EHR input modalities, such that only pseudo-notes from one data category at a time could be considered. Again, no single modality on its own approached MEME's performance. (Figure 8)

MEME vs LLM Prompting

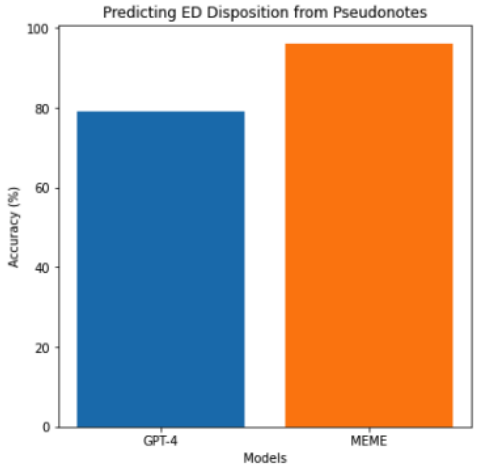


Figure 5: Comparing Model Accuracy for Predicting Emergency Department (ED) Disposition Using Pseudonotes against Zero-shot Prompting. The left panel showcases a code snippet illustrating how we prompted a GPT4 API. The right panel presents a bar graph comparing the accuracy of GPT-4 and MEME models in predicting ED dispositions. The graph highlights the significant performance disparity between the two models, with MEME outperforming GPT-4.

Given the emergent capabilities of generative AI models (e.g. GPT [11,34], LLaMA-2 [35], Claude, etc.), we investigated predictive performance of MEME relative to a zero-shot prompting approach[36]. We compared the MEME classifier and a zero-shot GPT4 API using 100 random samples to predict ED disposition based on pseudo-notes. We observed a notable difference in performance such that all EHR-specific models outperformed and GPT-4 in terms of Accuracy and F1 score, indicating the continued need for specialized models adapted to uncommon settings. We noticed from this experiment that training classifiers still outperform current generative models despite their general reasoning capabilities. We noticed a nearly 20% performance gap between these two models on both metrics with a 16% difference in accuracy and a 17% difference in f1 scores.

Multimodal embedding is compatible with evolving language models

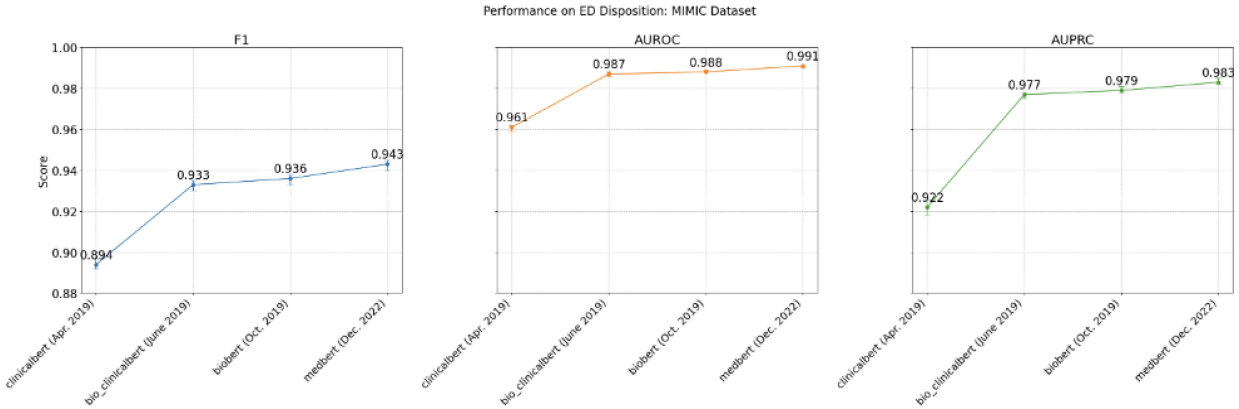


Figure 6: Longitudinal Performance Metrics of Foundation Models Predicting ED Disposition Using the MIMIC Dataset. This series of plots tracks the evolution of three key performance metrics—F1 score, Area Under the Receiver Operating Characteristic Curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC)—for foundational models across several years. The aim is to demonstrate that future foundation models are compatible with our framework, potentially leading to further improvements. Each plot shows a steady enhancement in model metrics over time, underscoring the effectiveness of utilizing open-source models that continue to get better.

The multimodal embedding approach is agnostic to the natural language model which embeds clinical pseudo-notes. We repeated the ED disposition experiment, testing several clinical language models in the MEME framework (Bio_ClinicalBERT[14] (June 2019), BioBERT[37] (October 2019), and MedBERT[38,39] (2022)). We found that advances in clinical language models translated to improvements in ED disposition.

Cross-institution generalization and adaptation

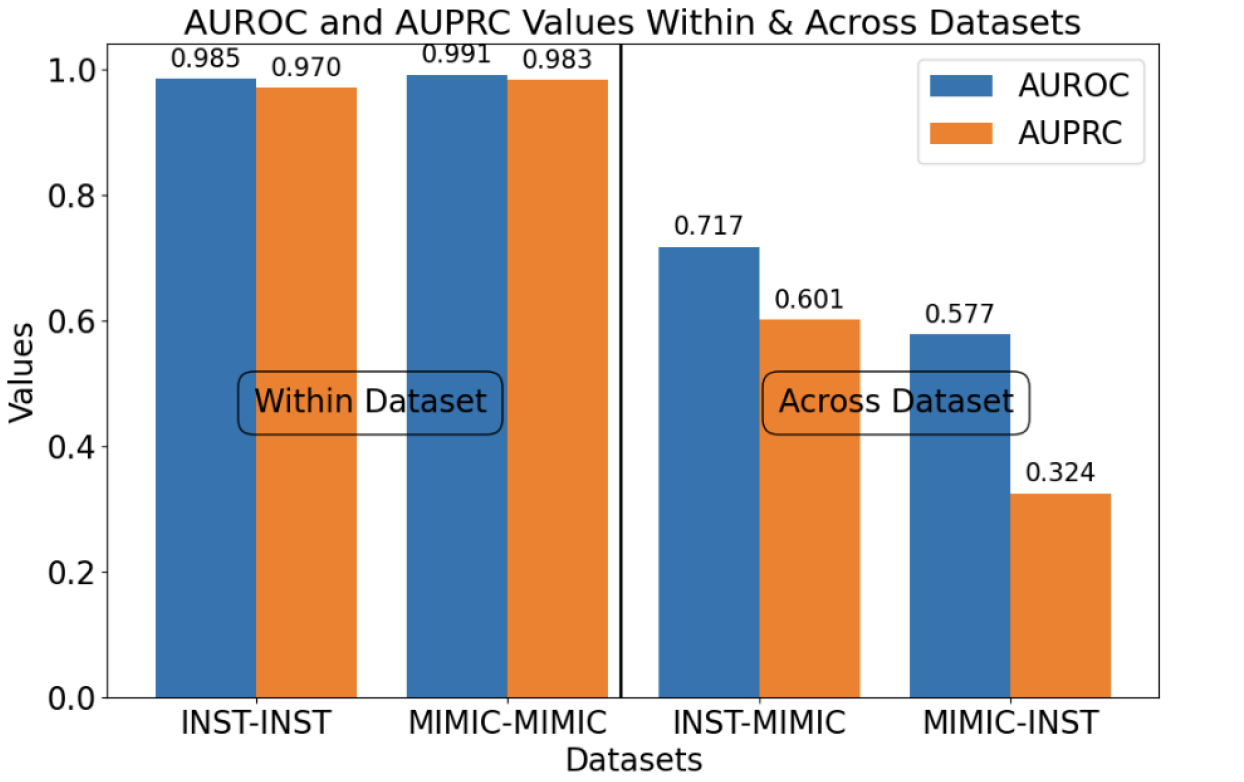


Figure 7. Performance of the ED Disposition task across and within datasets. We see noticeable performance dropoff in both AUROC and AUPRC when we test across sites.

MEME exhibited strong performance within individual institutions but showed poor generalizability when directly applied across different sites. This decline in performance is a common challenge in healthcare, where models trained on data from one institution often fail to generalize to others due to differences in patient populations, clinical practices, and data

collection methods[40]. In our cross-site experiments, training on one hospital's data and testing on others led to significant drops in F1, AUROC, and AUPRC scores.

However, few-shot learning offers a promising solution by enabling models to rapidly adapt to new environments with minimal data, improving generalization and robustness in the face of distribution shifts or out-of-distribution (OOD) data[41–43]. Many existing EHR foundation models, while powerful, struggle with real-world applications due to their lack of interoperability with proprietary databases[11]. In contrast, the pseudo-notes approach used in MEME enhances interoperability, allowing for generalization across proprietary datasets when combined with few-shot learning, making it a more practical tool in diverse clinical settings.

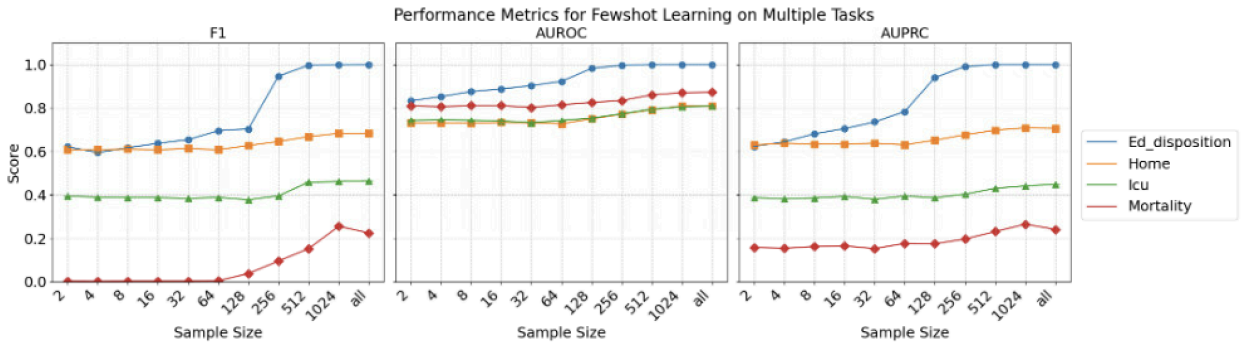


Figure 8: Performance Generalization via Few-Shot Learning. This set of graphs demonstrates the F1 score, AUROC (Area Under the Receiver Operating Characteristic Curve), and AUPRC (Area Under the Precision-Recall Curve) for few-shot learning models across multiple tasks as a function of increasing sample size. The plots illustrate steady improvements between 128-512 samples showcasing that this model can overcome OOD which is a current struggle of healthcare AI models.

To evaluate MEME's adaptability, we tested its performance on an external population from the UCLA Health system. Fine-tuning MEME for the same ED disposition and decompensation tasks, we varied the number of local training samples from 2 to 1024. MEME achieved near-maximal performance (AUROC, AUPRC, and F1) between 128 and 512 samples, consistent with previous findings for EHR foundation models like EHRshot[12]. This demonstrates MEME's potential for real-world applications where rapid adaptation to new data is crucial.

Discussion

In this work we introduce the Multimodal Embedding Model for EHR (MEME), a representation and decision-support framework for EHR. This approach uses pseudo-notes as an intuitive interface between structured electronic health data and foundational language models, and adopts a multi-stream approach to encoding EHR data domains. The combination of these approaches results in comparable or superior performance compared with canonical and modern machine learning approaches across decision support tasks around Emergency Department disposition and decompensation.

Pseudo Notes as an interface between EHR and Foundational Language Models

Our study revealed that using multiple sources of EHR information independently appears to have significant results. We generally see that MEME outperforms all models with considerable improvement over EHR-shot, and the three standard methods on the ED disposition task. We also noticed that XGBoost performs better on two of the decompensation tasks in terms of the AUROC metric, but this could be nuanced due to class imbalance [33]**.** We notice more subtle improvements in all other metrics across all models.

We designed our model to be compatible with the pseudo-notes design, which encodes separate biological and temporal scales. To test the algorithmic design of MEME, we compared it with several baselines from[33], ranging from traditional ML techniques to EHR foundation models. Our ablation study revealed that the multi-stream approach, which integrates multiple concepts, significantly outperforms individual models and the Multi-stream Single Embedding Model (MSEM). This method enables MEME to represent each EHR concept with high fidelity and dynamically combine them for inference using self-attention. Our comparative studies further demonstrate that MEME surpasses both single-stream models and EHR foundation model alternatives, highlighting its superior capability in handling the multifaceted nature of healthcare data and supporting our design choices. Questions regarding context length were also studied where we compared our MEME approach against the clinical-longformer [32] to motivate our framework design. We notice a considerable gap between these two methods, supporting our claims that different temporal and biological scales should be encoded separately instead of in a model with longer context length.

In addition to the performance advantages demonstrated by our experiments, the Multiple Embedding Model for EHR (MEME) offers several qualitative benefits in terms of portability and extendibility. Unlike EHR-specific models such as BEHRT[16], CHIRoN [44], and EHR-shot[29], which depend on evolving data standards and harmonization procedures for interoperability [18,45], MEME utilizes a natural language approach that is extendible to any data that can be text-serialized (e.g. [24,46,47]), providing a straightforward interface for serializing both public and proprietary EHR systems. This approach is more easily adopted by institutions and can gracefully handle changes in coding standards, leveraging general reasoning capabilities and increasing the medical domain knowledge captured by existing and emerging foundation language models (Figure 6). This framework not only promotes interoperability across diverse healthcare systems with varying protocols but also outperforms both EHR FMs and machine learning models which rely on harmonized structured formats, which are yet to be universally adopted.

MEME is adaptable across institutions

Healthcare AI models have often been criticized for failing to generalize across institutions[40]. We observed that MEME also displayed similar behavior. However, it has been shown that foundational models are more efficient to adapt to new scenarios [15,48]. We found that MEME, using language-based embedding models, approached ceiling performance using between 128 and 512 training examples for ED Decision support tasks, which is comparable to the EHR FM EHRshot in the few-shot learning setting[12]. Coupled with the interoperability benefits of not

requiring a data standard, this approach could be applied in settings where limited data annotations are available and the EHR are not recorded using a common data model.

Limitations

A limitation of this work is our inability to release our private institutional data, due to privacy restrictions and university policy. This highlights the significance of independent benchmarks, and underscores the necessity of external validation, for example benchmark datasets and tasks such as MC-BEC[23]. Additionally, our analysis was limited to a small set of hospital datasets and tasks from two sites, potentially not reflecting the full diversity of EHR systems. We did not investigate methods to harmonize different data schemas, which could affect the model's adaptability across diverse healthcare settings.

Conclusion

We describe a decision-support modeling framework which interfaces between structured electronic health data and foundational language models. This approach is adaptable across a variety of settings and is compatible with evolving foundational models, and may streamline the incorporation of modern AI into clinical decision support.

Methods

Data:

Our study sources de-identifed data from the publicly availableMedical Information Mart for Intensive Care (MIMIC)-IV v2.2 database[22], and UCLA Health [49]. This analysis was deemed non human-subjects research by the local institutional review board (IRB) due to its retrospective and deidentified nature. We detail the components of these databases to comprehensively explore the data inputs for our model.

MIMIC-IV ED [22]: This database is used for various downstream tasks, employing EHR concepts such as arrival information, which captures patient demographics and means of arrival; triage, documenting patient vitals and complaints at arrival; medication reconciliation (medrecon), detailing prior and current medications; diagnostic codes (ICD-9/10) for diagnoses; and measurements throughout the ED stay, including patient vitals and medications from pyxis. Data across these modalities are linked via unique visit or hospital admission IDs (Hadm_id) and associated with all prediction labels.

UCLA Database[49]**:** This database mirrors the MIMIC-IV in data modalities except for the absence of medication reconciliation (medrecon), with some variations due to different EHR system features. Our approach aims to make pseudo-notes across both databases closely resemble each other. Like MIMIC-IV, all concepts/streams in our UCLA data can be linked using a hospital admission ID and are also associated with all prediction labels.

In the MIMIC-IV database, we analyzed 400,019 unique visits, each associated with six modalities, contributing to a dataset size of approximately 2.4 million text paragraphs. For predicting ED disposition, we used the available data for training, validation, and testing with a set seed for reproducibility. For the decompensation prediction tasks, we utilized the subset of visits admitted to the hospital from the ED, resulting in a sample size of 158,010 patients. In the UCLA database, we analyzed a larger sample of 947,028 patients with five available modalities (excluding medrecon), resulting in approximately 4.75 million text paragraphs. All available data were used for the ED disposition task, and the 240,161 admitted patients were used for decompensation prediction. Further breakdowns can be found in our strobe diagrams in Table 1.

Benchmark

This paper focuses on binary prediction tasks related to Emergency Department (ED) disposition and decompensation, as defined in[23]. We evaluate our multistream method's effectiveness in both single and multilabel classification tasks, benchmarking it against other tabular-based and text-operating machine learning models. This assessment aims to highlight the performance advantages of using a text-based, multiple embedding strategy.

ED Disposition (Binary Classification): The first objective is to predict ED disposition, specifically where patients are sent after their Emergency Room visit, based on EHR measurements recorded during their stay. This is framed as a binary classification problem, distinguishing between patients discharged home and those admitted to the hospital.

ED Decompensation (Multilabel Binary Classification): The second objective involves analyzing the subset of patients admitted to the hospital and predicting various outcomes related to their ED visit. This is approached as a multilabel binary classification task, where the model predicts three ED outcomes simultaneously. The first task is to predict the patient's next discharge location, distinguishing between home and other facilities. The second task is to predict the need for Intensive Care Unit (ICU) admission. The final task is to predict in-patient mortality, specifically whether the patient dies during their hospital stay.

MEME

In the Results section, we introduced the Multiple Embedding Model for EHR (MEME). Here, in the Methods section, we provide a detailed explanation of how this framework progresses from processing pseudo-notes to generating predictions, outlining each step of the process comprehensively. This includes the transformation of raw data into structured embeddings, the application of self-attention mechanisms, and the integration of these embeddings into a predictive model, ensuring clarity at every stage of the pipeline.

Generating Embeddings

In the initial step of our model, we aim to generate embeddings for each EHR concept by feeding tokenized data into our foundational models' encoders, which produce rich, high-dimensional vector representations encapsulating various aspects of a patient's medical history. We choose to freeze the encoder layers, focusing on the training parameters of the subsequent layers dedicated to the prediction task. After generating embeddings for all concepts, we concatenate them into a unified input vector for further processing. This procedure can be mathematically represented as follows: In the model's first phase, modality-specific pseudo-notes are processed and structured into a tokenized format, denoted $D_{tokenized}$, which outlines a series of unique medical concepts or characteristics c_i derived from a patient's records. Each concept undergoes transformation via the foundation models encoder into a high-dimensional vector $\vec{v}_i$, offering nuanced, context-rich portrayals of each EHR concept and capturing complex clinical information. These vectors are then unified into a comprehensive vector $\vec{v}_{concat}$ through concatenation, laying the groundwork for our multimodal patient embeddings.

$$v_i = Foundation\ \vec{Model}(c_i)\ \forall c_i \in tokenized$$

$$\vec{V}_{concat} = Concatenate(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_n)$$

Self-attention Classifier

In the second step of our network, we introduce a new use case of a self-attention layer [50] designed to analyze the singular concatenated representation vector, $\vec{V}_{concat}$, as a unified entity. This approach arises from our intention to interpret aligned modalities collectively, rather than as

separate entities, allowing the network to operate comprehensively on the entire vector. It evaluates the relationships between elements within the vector, capturing patterns across different EHR concept vectors. The output from this layer is then directed through a fully connected layer, followed by a ReLU activation function, before being fed into the final classifying layer for prediction. This method, characterized by a unified analysis and attention-based processing, distinguishes our approach from traditional models and is pivotal to the enhanced predictive capabilities of our framework. Mathematically, this process involves transforming the input vector , $\vec{V}_{concat}$ into an attention vector $\vec{V}_{attention}$using the self-attention mechanism, further processing it through a fully connected (FC) layer and a Rectified Linear Unit (ReLU) activation to obtain a refined feature vector$\vec{V}_{fc}$, as outlined below:

$$V_{attention} = \vec{SelfAttention}(\vec{V}_{concat})$$

$$V_{fc} = ReLU(\vec{FC}(\vec{V}_{attention}))$$

$$\vec{z} = Classifier\ (\vec{V}_{fc})$$

The model leverages these refined features, $\vec{V}_{fc}$,, in a classifier to produce logits $\vec{z}$, subsequently processed to predict probabilities for ED Disposition or ED Decompensation tasks. The classifier's output is optimized by minimizing Cross Entropy Loss L, ensuring alignment of predicted probabilities $\hat{y}_i$ with true labels y_i. For multi-label tasks like ED

$$L = \sum_{i=1}^{n} \sum_{l=1}^{m} BCE(\sigma(z_{i,l}) y_{i,l})$$

Decompensation, each logit $\vec{z}_{i,l}$ undergoes individual sigmoid activation σ, and the model's training involves minimizing a tailored Cross Entropy Loss that aggregates binary cross-entropy losses across all labels for each observation, capturing the multi-label aspects of the data effectively.

Selecting Optimal Thresholds for F1, Precision and Recall

To select the optimal threshold for F1 and AUPRC (Area Under the Precision-Recall Curve) scores, we implemented a dynamic algorithm that samples thresholds from 0.00 to 1.00 in 1,000 discrete steps. This approach allows us to identify the threshold that maximizes the F1 score and AUPRC by evaluating model performance at each point. The algorithm dynamically adjusts and evaluates precision, recall, and F1 at each threshold, selecting the one that strikes the best balance between precision and recall for the F1 score, while optimizing the trade-off between sensitivity and precision for AUPRC. By using such fine granularity in threshold selection, the model ensures that the chosen threshold is optimal for both metrics, leading to better prediction performance.

References

1. Tajbakhsh, N. *et al.* Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med. Image Anal.* **63**, 101693 (2020).
2. Mullowney, M. W. *et al.* Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**, 895–916 (2023).
3. Farnoud, A., Ohnmacht, A. J., Meinel, M. & Menden, M. P. Can artificial intelligence accelerate preclinical drug discovery and precision medicine? *Expert Opin. Drug Discov.* **17**, 661–665 (2022).
4. Ávila-Jiménez, J. L., Cantón-Habas, V., Carrera-González, M. del P., Rich-Ruiz, M. & Ventura, S. A deep learning model for Alzheimer's disease diagnosis based on patient clinical records. *Comput. Biol. Med.* **169**, 107814 (2024).
5. Kumar, R. P. *et al.* Can Artificial Intelligence Mitigate Missed Diagnoses by Generating Differential Diagnoses for Neurosurgeons? *World Neurosurg.* **187**, e1083–e1088 (2024).
6. Khalighi, S. *et al.* Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *Npj Precis. Oncol.* **8**, 1–12 (2024).
7. Al Kuwaiti, A. *et al.* A Review of the Role of Artificial Intelligence in Healthcare. *J. Pers. Med.* **13**, 951 (2023).
8. Idowu, E. A. A., Teo, J., Salih, S., Valverde, J. & Yeung, J. A. Streams, rivers and data lakes: an introduction to understanding modern electronic healthcare records. *Clin. Med.* **23**,

409–413 (2023).

9. Zhao, W. X. *et al.* A Survey of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2303.18223 (2023).
10. Li, L. *et al.* A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs). Preprint at https://doi.org/10.48550/arXiv.2405.03066 (2024).
11. Radford, A. & Narasimhan, K. Improving Language Understanding by Generative Pre-Training. in (2018).
12. Guo, L. L. *et al.* A multi-center study on the adaptability of a shared foundation model for electronic health records. *Npj Digit. Med.* **7**, 1–9 (2024).
13. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. *Npj Digit. Med.* **6**, 1–10 (2023).
14. Alsentzer, E. *et al.* Publicly Available Clinical BERT Embeddings. Preprint at https://doi.org/10.48550/arXiv.1904.03323 (2019).
15. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
16. Li, Y. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci. Rep.* **10**, 7155 (2020).
17. Rupp, M., Peter, O. & Pattipaka, T. ExBEHRT: Extended Transformer for Electronic Health Records. in *Trustworthy Machine Learning for Healthcare* (eds. Chen, H. & Luo, L.) 73–84 (Springer Nature Switzerland, Cham, 2023). doi:10.1007/978-3-031-39539-0_7.
18. Makadia, R. & Ryan, P. B. Transforming the Premier Perspective® Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model. *eGEMs* **2**, 1110 (2014).
19. Medical Event Data Standard (MEDS): Facilitating Machine Learning for Health. in (2024).
20. McDermott, M., Nestor, B., Argaw, P. & Kohane, I. S. Event Stream GPT: A Data Pre-processing and Modeling Library for Generative, Pre-trained Transformers over

Continuous-time Sequences of Complex Events. *Adv. Neural Inf. Process. Syst.* **36**, 24322–24334 (2023).

21. Steinberg, E. *et al.* Language models are an effective representation learning technique for electronic health record data. *J. Biomed. Inform.* **113**, 103637 (2021).
22. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
23. Chen, E. *et al.* Multimodal Clinical Benchmark for Emergency Care (MC-BEC): A Comprehensive Benchmark for Evaluating Foundation Models in Emergency Medicine. *Adv. Neural Inf. Process. Syst.* **36**, 45794–45811 (2023).
24. Hegselmann, S. *et al.* TabLLM: Few-shot Classification of Tabular Data with Large Language Models. in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* 5549–5581 (PMLR, 2023).
25. Hegselmann, S. *et al.* A Data-Centric Approach To Generate Faithful and High Quality Patient Summaries with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2402.15422 (2024).
26. Ellershaw, S. *et al.* Automated Generation of Hospital Discharge Summaries Using Clinical Guidelines and Large Language Models. in (2024).
27. Yuan, D. *et al.* A Continued Pretrained LLM Approach for Automatic Medical Note Generation. Preprint at https://doi.org/10.48550/arXiv.2403.09057 (2024).
28. Textualization of Oral Epics. https://www.degruyter.com/document/doi/10.1515/9783110825848/pdf?licenseType=restricted#page=11.
29. Wornow, M., Thapa, R., Steinberg, E., Fries, J. & Shah, N. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. *Adv. Neural Inf. Process. Syst.* **36**, 67125–67137 (2023).
30. Hur, K. *et al.* GenHPF: General Healthcare Predictive Framework for Multi-Task

Multi-Source Learning. *IEEE J. Biomed. Health Inform.* **28**, 502–513 (2024).

31. Xie, F. *et al.* Benchmarking emergency department prediction models with machine learning and public electronic health records. *Sci. Data* **9**, 658 (2022).
32. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. Preprint at https://doi.org/10.48550/arXiv.2201.11838 (2022).
33. Saito, T. & Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE* **10**, e0118432 (2015).
34. OpenAI *et al.* GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
35. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).
36. Kojima, T., Gu, S. (Shane), Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *Adv. Neural Inf. Process. Syst.* **35**, 22199–22213 (2022).
37. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
38. Vasantharajan, C. *et al.* MedBERT: A Pre-trained Language Model for Biomedical Named Entity Recognition. in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* 1482–1488 (2022). doi:10.23919/APSIPAASC55919.2022.9980157.
39. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *Npj Digit. Med.* **4**, 1–13 (2021).
40. Goetz, L., Seedat, N., Vandersluis, R. & van der Schaar, M. Generalization—a key challenge for responsible AI in patient-facing clinical applications. *Npj Digit. Med.* **7**, 1–4

(2024).

41. Wang, Y., Yao, Q., Kwok, J. T. & Ni, L. M. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Comput Surv* **53**, 63:1-63:34 (2020).
42. Parnami, A. & Lee, M. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. Preprint at https://doi.org/10.48550/arXiv.2203.04291 (2022).
43. Liu, J. *et al.* Towards Out-Of-Distribution Generalization: A Survey. Preprint at https://doi.org/10.48550/arXiv.2108.13624 (2023).
44. Hill, B. L. *et al.* CHIRon: A Generative Foundation Model for Structured Sequential Medical Data. in (2023).
45. Spackman, K. A., Campbell, K. E. & Côté, R. A. SNOMED RT: a reference terminology for health care. *Proc. AMIA Annu. Fall Symp.* 640–644 (1997).
46. Dinh, T. *et al.* LIFT: Language-Interfaced Fine-Tuning for Non-language Machine Learning Tasks. *Adv. Neural Inf. Process. Syst.* **35**, 11763–11784 (2022).
47. Ono, K. & Lee, S. A. Text Serialization and Their Relationship with the Conventional Paradigms of Tabular Machine Learning. Preprint at https://doi.org/10.48550/arXiv.2406.13846 (2024).
48. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022).
49. Johnson, R. *et al.* The UCLA ATLAS Community Health Initiative: Promoting precision health research in a diverse biobank. *Cell Genomics* **3**, 100243 (2023).
50. Vaswani, A. *et al.* Attention Is All You Need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2023).