REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models

Yinghao Zhu[1 *], **Changyu Ren**[2 *], **Shiyun Xie**[1], **Shukai Liu**[2], **Hangyuan Ji**[2], **Zixiang Wang**[3], **Tao Sun**[2], **Long He**[1], **Zhoujun Li**[2], **Xi Zhu**[4], **Chengwei Pan**[1 †]

[1]Institute of Artificial Intelligence, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]School of Software, Beihang University, Beijing, China
[4]China Mobile Research Institute, Beijing, China

zhuyinghao@buaa.edu.cn, pancw@buaa.edu.cn

Abstract

The integration of multimodal Electronic Health Records (EHR) data has significantly improved clinical predictive capabilities. Leveraging clinical notes and multivariate time-series EHR, existing models often lack the medical context relevent to clinical tasks, prompting the incorporation of external knowledge, particularly from the knowledge graph (KG). Previous approaches with KG knowledge have primarily focused on structured knowledge extraction, neglecting unstructured data modalities and semantic high dimensional medical knowledge. In response, we propose REALM, a Retrieval-Augmented Generation (RAG) driven framework to enhance multimodal EHR representations that address these limitations. Firstly, we apply Large Language Model (LLM) to encode long context clinical notes and GRU model to encode time-series EHR data. Secondly, we prompt LLM to extract task-relevant medical entities and match entities in professionally labeled external knowledge graph (PrimeKG) with corresponding medical knowledge. By matching and aligning with clinical standards, our framework eliminates hallucinations and ensures consistency. Lastly, we propose an adaptive multimodal fusion network to integrate extracted knowledge with multimodal EHR data. Our extensive experiments on MIMIC-III mortality and readmission tasks showcase the superior performance of our REALM framework over baselines, emphasizing the effectiveness of each module. REALM framework contributes to refining the use of multimodal EHR data in healthcare and bridging the gap with nuanced medical context essential for informed clinical predictions.

1 Introduction

The advent of Electronic Health Records (EHR) marks a pivotal advancement in the way patient data is gathered and analyzed, contributing to a more effective and informed healthcare delivery system for clinical prediction Ma *et al.* [2023]; Gao *et al.* [2024]; Zhu *et al.* [2024c]; Zhang *et al.* [2024]; Liao *et al.* [2024]. This advancement is largely attributed to the utilization of multimodal EHR data, which primarily includes clinical notes and multivariate time-series data from patient records Zhang *et al.* [2022]; Wang *et al.* [2024]; Zhang *et al.* [2023a]. Such data types are integral to healthcare prediction tasks, mirroring the holistic approach practitioners adopt by leveraging various patient data points to inform their clinical decisions and treatment strategies, rather than depending on a single data source Xu *et al.* [2023]. Deep learning-based methods have become the mainstream approach, processing multimodal data to learn a mapping from heterogeneous inputs to output labels Choi *et al.* [2017]; Ma *et al.* [2018]; Zhang *et al.* [2022]. However, in contrast to healthcare professionals, who have a deep understanding of medical contexts through extensive experience and knowledge, neural networks trained from scratch lack these insights into medical concepts Miotto *et al.* [2018]. Without deliberate integration of external knowledge, these networks often lack the ability or sensitivity to recognize crucial disease entities or laboratory test results within the EHR, essential for accurate prediction tasks Zhu *et al.* [2024b]. In response, some recent studies have begun incorporating knowledge graphs to infuse additional medical insights into their analyses Ye *et al.* [2021]; Gao *et al.* [2022]. These graphs offer a supplementary layer of clinically relevant concepts, thereby enhancing the model's ability to provide contextually meaningful representations and interpretable evidence Yang *et al.* [2023]. Despite these advancements, formidable limitations remain, underscoring the imperative need for continuous research in integrating insights from knowledge graphs to refine and enhance the use of multimodal EHR data in healthcare.

Previous methods integrating external medical knowledge into EHR data analysis have focused on mining hierarchical and structured knowledge from clinical-context knowledge graphs. However, these approaches primarily extract

* Equal contribution, † Corresponding author.

medical concepts—entity names and their relationships into a graph—with limited direct contribution to predictive tasks (**Limitation 1**). Furthermore, they tend to extract entities only from structured data modalities, such as ICD disease codes, patient conditions, procedures, and drugs, neglecting the unstructured modalities. Although extracting knowledge from unstructured data is more challenging, both clinical notes and time-series modalities are more common and practical Rajkomar *et al.* [2018] (**Limitation 2**). With Large Language Models (LLMs) like GPT-4 Achiam *et al.* [2023] demonstrating strong capabilities in diverse clinical tasks Zhu *et al.* [2024b]; Wornow *et al.* [2023] and serving as large medical knowledge graphs (KGs) Sun *et al.* [2023], it is feasible to use LLM to bridge complex medical knowledge from KGs with multimodal EHR data. By prompting the LLM, GraphCare Jiang *et al.* [2023] constructs a GPT-KG on structured condition, procedure, and drug record data, with triples (entity 1, relationship, entity 2) and further employs graph neural networks for downstream tasks. This paradigm, however, faces the hallucination issue of LLMs, where incorrect or fabricated information may arise Zhang *et al.* [2023b] (**Limitation 3**).

To overcome these limitations, we propose utilizing LLM in a Retrieval-augmented Generation (RAG) approach Lewis *et al.* [2020]. The RAG process links the unstructured modalities and external KG with LLM's semantic reasoning capabilities Wang *et al.* [2023]. Despite its apparent simplicity, applying this intuition to clinical tasks presents technical challenges:

Challenge 1: How to extract entities from multimodal EHR data and match these entities with external KG consistently? Extracting entities from the diverse and complex formats of EHR data (including clinical notes and multivariate time-series data) presents a significant challenge. Moreover, unlike structured codes where it can directly compare the code-related entities' embedding with KG's entity, the entities extracted by LLM using prompts have hallucination issues. Accurately matching extracted entities with those in an external knowledge graph while eliminating the potential for hallucinations posed by LLMs is crucial for maintaining the integrity and reliability of the clinical prediction tasks Imrie *et al.* [2023].

Challenge 2: How to encode and incorporate retrieved knowledge with original data modalities? The extracted textual knowledge should be encoded using a sentence-level embedding model, thus posing a challenge in the selection of long-context supported models Xiao *et al.* [2023]. Additionally, effectively incorporating retrieved knowledge with the original data modalities to enhance prediction accuracy without compromising interpretability Ye *et al.* [2021] or introducing additional complexity into the model is vital as well.

To these ends, We propose REALM framework tackling the above limitations and challenges with the following approaches, which are our three-fold contributions:

- We design the RAG-driven multimodal enhancement framework for clinical notes and time-series EHR data (**Response to Limitation 1**). REALM leverages the capabilities of LLMs and professionally labeled external large medical knowledge graphs. We retrieve the medical entities by prompting LLM, match them in KG with detailed checking and alignment to ensure no hallucination (**Response to Limitation 3**). Apart from simple triples of entities, we also include much more knowledge and their relationships by extending the entities' definition and description and encode the long context medical knowledge into LLM embedding, which allows for capturing more complex semantic medical background knowledge that contains task-relevant insights (**Response to Limitation 2**).
- Methodologically, our RAG-driven entity extraction and matching process stands with a clinical standard that all knowledge comes from the professional medical knowledge graph (PrimeKG) with hallucination elimination and consistency guarantees. By carefully comparing LLM-generated entities with original data and employing threshold-based retrieval and review processes, we align the knowledge with external KGs. By extending the definition and description of entities beyond simple triples, REALM captures more complex semantic medical background knowledge. The overall process is designed to prevent hallucinations and preserve the high-level medical context from knowledge bases, ensuring the reliability of the extracted information and allowing for the inclusion of a broader range of knowledge and relationships (**Response to Challenge 1**). To infuse the extracted knowledge and with consideration of heterogeneity, we design an adaptive multimodal fusion network with self-attention and cross-attention mechanism that attentively learns each modality and fuses the final representation for downstream tasks (**Response to Challenge 2**).
- Our extensive experiments demonstrate REALM's superior performance on MIMIC-III mortality and readmission tasks and its effectiveness of our designed each module. Additionally, to meet the practical requirements of clinical use, we conduct an evaluation on model robustness to less training samples showing REALM's remarkable resilience against data sparsity. Moreover, the evaluation on quality of retrieved entities reflects the soundness of retrieved medical entities. Importantly, all operations in our REALM framework are conducted offline, ensuring privacy and data security.

2 Related Work

2.1 Multimodal EHR Learning

The evolution of medical technology has enabled the analysis of various medical modalities—ranging from clinical notes and time-series laboratory test data to demographics, conditions, procedures, drugs, and medical imaging. Noteworthy efforts in multimodal learning for healthcare include M3Care Zhang *et al.* [2022] which compensates for the missing modalities by imputing task-related information in the latent space through auxiliary information from similar patients. M3Care leverages a task-guided modality-adaptive similarity metric to effectively handle missing modalities without relying on unstable generative models. The work of Zhang *et al.* [2023a] further explored the irregularity of time

intervals in time-series EHR data and clinical notes via a time attention mechanism. Notably, Xu *et al.* [2023] introduced an innovative approach for joint learning from visit sequences and clinical notes, employing Gromov-Wasserstein Distance for contrastive learning and dual-channel retrieval to enhance patient similarity analysis. Lee *et al.* [2023] proposed a unified framework for learning across all EHR modalities, eschewing separate imputation modules in favor of modality-aware attention mechanisms.

Although the methods mentioned above perform well across multiple joint modalities, a common drawback is their limited consideration of incorporating clinical background information, wherein external medical knowledge could provide significant insights into the EHR data. Furthermore, the absence of semantic medical knowledge renders the training-from-scratch pipeline more challenging to converge, especially when data is scarce in practical clinical settings.

2.2 Incorporating External Knowledge for EHR

Addressing the need to blend clinical background knowledge with EHR data, numerous studies have leveraged medical knowledge graphs (KGs) to enhance the EHR data representation learning process, thereby augmenting predictive performance. Techniques such as utilizing the ancestor information of nodes within KGs have been employed to refine medical representation learning, as seen in GRAM Choi *et al.* [2017], which integrates hierarchical medical ontologies via a graph attention network. KAME Ma *et al.* [2018] builds on this by embedding ontology information throughout the prediction process, enriching the contextual understanding of models. Collaborative graph learning models, such as CGL Lu *et al.* [2021], explore patient-disease interactions and domain knowledge, while KerPrint Yang *et al.* [2023] focus on addressing knowledge decay on multiple time visits. The advent of Large Language Models (LLMs) as comprehensive knowledge bases Sun *et al.* [2023] offers new possibilities, exemplified by GraphCare Jiang *et al.* [2023], which creates a KG from structured EHR data for GNN learning, though it faces challenges related to content hallucination.

These efforts predominantly concentrate on extracting knowledge from structured medical data, overlooking the rich semantic information embedded in unstructured EHR data. This oversight limits the potential for fully leveraging the depth of knowledge contained within EHRs, highlighting the need for methodologies that encompass both structured and unstructured data modalities.

3 Problem Formulation

The electronic health records (EHR) dataset encompasses both structured and unstructured data, represented respectively by multivariate time-series data and clinical notes. For the purpose of analysis, these two modalities are initially treated independently to derive embeddings from the raw data matrix $\boldsymbol{X}$. Specifically, multivariate time-series data, denoted as $\boldsymbol{X_{TS}} \in \mathbb{R}^{T \times F}$, is characterized by T visits and F numeric features. Concurrently, clinical notes, represented as $\boldsymbol{X}_{Text}$, recorded at each patient visit. Accompanying these modalities, the temporal information is encapsulated in the timestamp vector $\boldsymbol{X}_{Time} \in \mathbb{R}^T$, with T signifying the respective visit times.

Furthermore, external knowledge graphs (KGs) are introduced to enhance the personalized representation of each individual patient. Information in the KG serve as a supplemental knowledge base reference.

The prediction objective is conceptualized as a binary classification task, encompassing the prediction of in-hospital mortality and readmission. Leveraging the comprehensive patient information derived from EHR data and KG, the model endeavors to predict specific clinical tasks. Formally, the prediction task is articulated as $\hat{y} = \text{Framework}(\boldsymbol{X}_{TS}, \boldsymbol{X}_{Text}, \boldsymbol{X}_{Time}, KG)$, where $\hat{y}$ represents the specific prediction label. Such formulation establishes a comprehensive framework for predicting clinical outcomes by amalgamating diverse data modalities and external knowledge sources.

The notations and their descriptions in the paper are shown in Table 1.

Table 1: *Notations symbols and their descriptions*

Notations	Descriptions
N	Number of patients
KG	External knowledge graphs
$\boldsymbol{X}_{TS}$	Time series data of one patient
$\boldsymbol{X}_{Text}$	Clinical records of one patient
$\boldsymbol{X}_{Time}$	Record time for certain modality of one patient
$\boldsymbol{X}_{RAG}$	Retrieved texts relative to time series or text data of one patient
$\boldsymbol{X}_{Time}$	Visiting timestamps of one patient
T	Number of visits for a certain patient
D	Embedding dimension of a single modal
F	Number of features in time series
$\boldsymbol{h}_i \in \mathbb{R}^{T \times d}, \boldsymbol{h}$	Representation of a single modality, fused to representation $\boldsymbol{h}$, d is each modal's embedding dimension
E_{TS}	Extracted entity set for one time series EHR data
E_{Text}	Extracted entity set for one clinical notes data
θ	Cosine similarity of two Embedding vectors
ϵ	Threshold for selecting anomalies in time-series data
η	Threshold for matching extracted entities with nodes in knowledge graph
s_i	z-score of i-th feature of one patient
$\boldsymbol{z}$	Final representation of a patient

4 Methodology

4.1 Overview

Figure 1 shows the overall framework architecture of our proposed REALM model. It consists of three main modules:

- **Multimodal EHR Embedding Extraction** applies GRU as embedding model for time series $\boldsymbol{X}_{TS}$ and LLM for text records $\boldsymbol{X}_{Text}$, which supports for long context inference

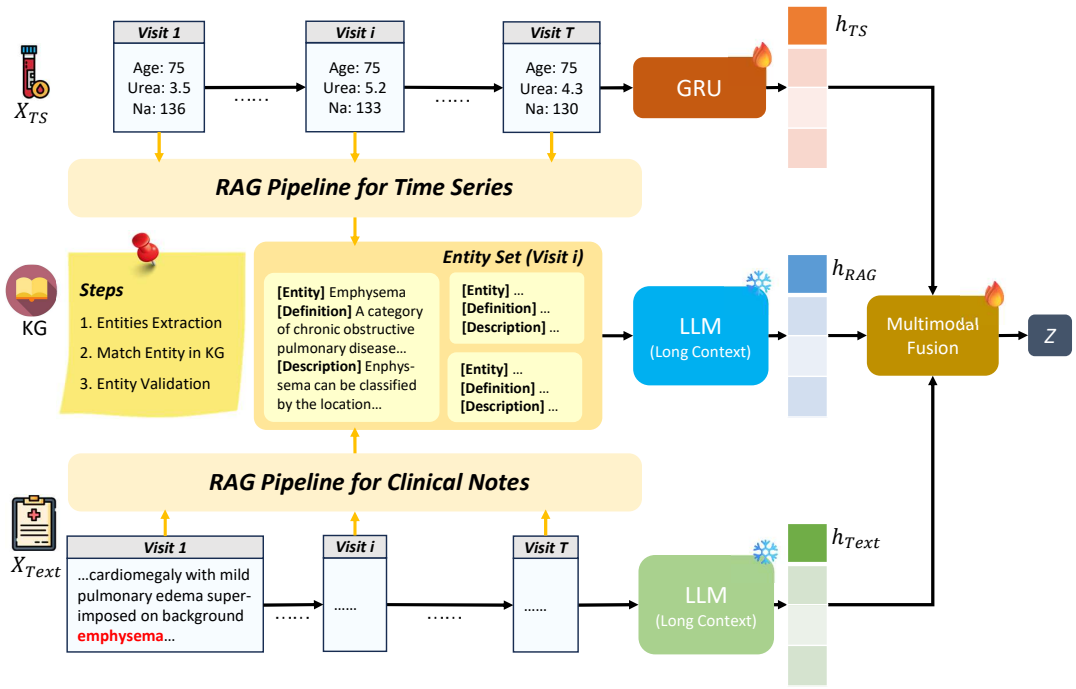

Figure 1: *Overall architecture of our proposed REALM framework.*

at once. Readable data are transferred into embeddings $\boldsymbol{h}_{TS}$ and $\boldsymbol{h}_{Text}$.

- **RAG-Driven Enhancement Pipeline** retrieves relevant knowledge raw input. We design a rule-based algorithm to find outlier features from time series $\boldsymbol{X}_{TS}$, and optimize LLM prompts to extract disease entities from clinical notes $\boldsymbol{X}_{Text}$. Then semantic based retrieval match extracted entities to relative nodes from KGs over threshold ϵ or η. After that we get external information $\boldsymbol{X}_{RAG}$, and encode them into $\boldsymbol{h}_{RAG}$ respectively.
- **Multimodal Fusion Network** gets embedding $\boldsymbol{h}_i$ from input modality $\boldsymbol{X}_i$ and fuses them in an adaptive way to get an enhances representation $\boldsymbol{z}$.

4.2 Multimodal EHR Embedding Extraction

We delve into the techniques used for embedding extraction from multimodal EHR, emphasizing the transformation from raw, human-readable inputs $\boldsymbol{X}$ to deep semantic embeddings $\boldsymbol{h}$ for a thorough analysis guided by the enhanced RAG.

When dealing with time-series data X_{TS}, we employ the Gated Recurrent Unit (GRU) network. GRU is a highly efficient variant of recurrent neural networks, capable of capturing the time dependencies in sequence data and encoding this time-related information into h_{TS}, the output from the time series encoder. We choose GRU due to its exceptional ability to model time in long sequence data and its potential to tackle long-term dependencies.

$$\boldsymbol{h}_{TS} = \mathrm{GRU}(X_{TS}) \tag{1}$$

For text records $\boldsymbol{X}_{Text}$, we incorporated a LLM encoder to obtain text embeddings $\boldsymbol{h}_{Text}$. The primary reason for choosing LLM as the heart of the text encoder is its outstanding capability to handle long text contexts. Although the BERT model excels in numerous natural language processing tasks, its maximum input length of 512 tokens can be a limitation, potentially leading to information loss when processing long contexts. LLM encoder can handle with longer input sequences, which making it a better fit for our detailed analysis of the rich textual information in EHR.

$$\boldsymbol{h}_{Text} = \mathrm{LLM}(\boldsymbol{X}_{Text}) \tag{2}$$

In the realm of EHR, the temporal dimension of patient visits plays a pivotal role, with each visit characterized by a unique timestamp, denoted as X_{Time}. To adeptly navigate the challenges posed by the irregular and asynchronous nature of time-series data within EHR, it is essential to have an embedding strategy that can seamlessly translate these discrete temporal markers into a meaningful, continuous vector space. To this end, we draw inspiration from the advanced techniques in multi-modal EHR analysis, where time-series data is often given precedence due to its critical significance.

Building upon the conventional Multilayer Perceptron (MLP) approach to embed time stamps $\boldsymbol{h}_{Time}$, we propose an enhanced method that leverages the sin/cos transformation, akin to the Transformer positional embedding mechanism. This approach not only captures the sequential order of visits but also preserves the cyclical continuity inherent in time-series data. By employing a sinusoidal function to encode time stamps, our model is endowed with the ability to discern the intricate inter-modality temporal relationships that are often neglected when time information is discarded. This sin/cos embedding harmonizes with the sophisticated attention mechanisms, enriching the model's capacity to prior-

itize relevant modalities and adapt to the dynamism of time-sensitive clinical tasks.

$$\boldsymbol{h}_{Time} = \text{MLP}(\boldsymbol{X}_{Time}) \tag{3}$$

By converting these three different types of data into compatible embeddings, our model lays a solid groundwork for the multimodal analysis of EHR. This strategy of embedding extraction sets the stage for further analysis tasks under the RAG framework, allowing us to accurately and comprehensively understand and analyze the complex information in EHR.

Naturally, RAG incorporates two RAG feature extraction submodules, each dealing with a different modality. These will be detailed in the following subsection.

4.3 RAG-Driven Enhancement Pipeline

Extract Entities from Multimodal EHR Data

To fully leverage the expert information in the knowledge graph to enhance prediction accuracy, we need to extract disease entities from time-series data and clinical notes and match them with the information in the graph. The disease entities set in the time-series data are denoted as $\boldsymbol{E}_{TS}$, and those in the clinical notes data can be directly denoted as $\boldsymbol{E}_{Text}$. We design two pipelines for each modality.

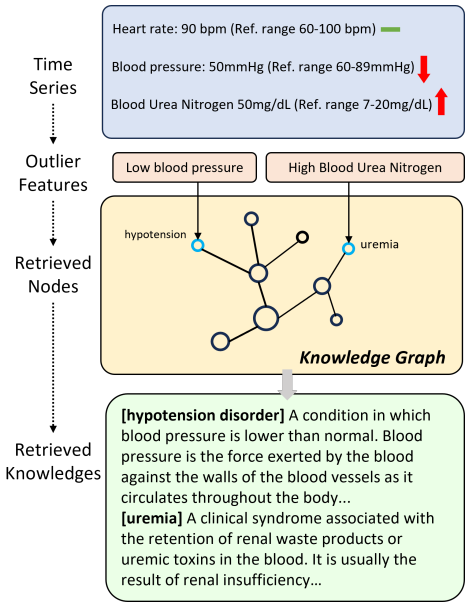

Figure 2: *RAG pipeline for time series EHR modality.*

RAG module for time series. Time series are a structured data including feature names and their values after clinical examination. Each feature name reflects parts of physical condition, which reminds us distinguishing patients through features out of reference range. As show in Figure 2, this record in total series shows a low Blood Pressure and high Blood Urea Nitrogen far beyond normal range. This reminds us the patient may suffer from hypotension and uremia. In fact, we can found these feature names in diseases defenitions and descriptions, and both lead to severe health risks.

Considering continuous numeric data have obvious distribution characteristics, we can find outlier values by calculating z-score of each feature, each seemed as an entity. There are mostly more than one entity (or outlier feature) in each patient, and some are missing values, so we only focus on those not empty. For each feature $\boldsymbol{X}_{TS_i}$, we can obtain mean value and standard deviation by their reference range, and calculate z-score of each feature as below, where s_i stands for z-score of i-th feature of one patient.

$$s_i = \frac{\boldsymbol{X}_{TS_i} - \text{mean}(\boldsymbol{X}_{TS_i})}{\text{std}(\boldsymbol{X}_{TS_i})} \tag{4}$$

Features over specified threshold (like 3-σ) are regarded as outlier ones, which means unhealthy physical conditions. We set ϵ as a threshold to screen out abnormal values, features with s_i greater than ϵ are regarded as abnormal entities and worth extraction. In order to set a reasonable threshold for clinical predictions, we divided a subset manually, and observe extracted entities under different ϵ, and determine one above which most entities are instructive.

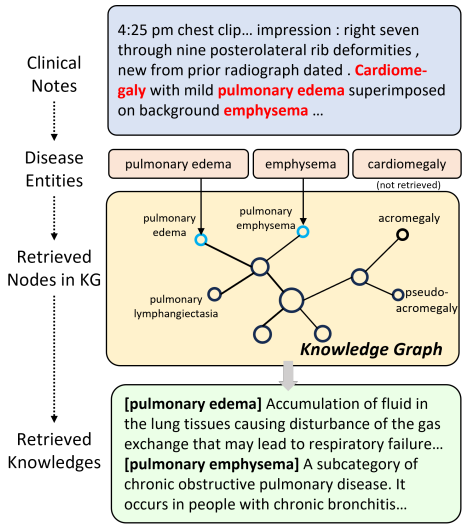

Figure 3: *RAG pipeline for clinical notes modality.*

RAG module for clinical text records. Due to context limitations of models like BERT, it may cause information loss when encoding clinical notes with BERT. LLM supports for longer context, but often introduce hallucination. So we utilize LLM as entities extractor with post processing:

1. **Entities Extraction:** To reduce LLM hallucination, we use one-shot as demostration and clear instruction in prompt, guiding LLM to focus only on disease entities appearing in raw notes. When calling LLM model once, sometimes LLM may cause failure without any entities returned, so we operate in multi rounds to enlarge current extracted entity set. In i-th round, we concat prompt $\boldsymbol{P}_{extract}$ and clinical text notes $\boldsymbol{X}_{Text}$ together as LLM input, and we can get a set of entities in output $\boldsymbol{E}^i_{Text}$, and update

total entities set with union of current one and last one.

$$\boldsymbol{E}^{i}_{Text} = \text{LLM}(concat(\boldsymbol{P}_{Extract}, \boldsymbol{X}_{Text})) \quad (5)$$

$$\boldsymbol{E}_{Text} := \boldsymbol{S}_{Text} \bigcup \boldsymbol{S}^{i}_{Text} \quad (6)$$

where $\boldsymbol{P}_{Extract}$ and $\boldsymbol{X}_{Text}$ represent the prompt to extract disease entities and clinical notes data respectively.

2. **Entities Refinement:** To mitigate hallucination issues of LLM, we design a post-processing procedure after extraction. This module consists of three steps: firstly, discard entities not appear in the original text; secondly, filter entities not in disease type using LLM; at last, delete duplicated entities in semantics. After that we get a illegal entities set, and delete them from last one. This procedure may lead to new empty set, so we should loop extraction above.

$$\boldsymbol{E}_{Text} := \boldsymbol{E}_{Text} - \boldsymbol{E}_{illegal} \quad (7)$$

We repeat step 1 and step 2 until convergence, to ensure the quality and quantity of extracted entities.

Match extracted entities with external KG

To accurately match the extracted entities with those in the knowledge graph, we employ dense vector retrieval based on semantics. First, we obtain embeddings of all KG nodes $Nodes$ with LLM. And we encode each entity in set $\boldsymbol{E}_{TS}$ or $\boldsymbol{E}_{Text}$ with the same LLM, to ensure embeddings align in the same vector space.

$$\boldsymbol{h}_n = \text{LLM}(n), n \in Nodes \quad (8)$$

$$\boldsymbol{h}_e = \text{LLM}(e), e \in E \quad (9)$$

Then we use current entity e as query, and compute cosine similarities between E_e and all embeddings of nodes in KG $\boldsymbol{h}_n$.

$$\theta^{n}_{e} = \frac{\boldsymbol{h}_n \cdot \boldsymbol{h}_e}{\|\boldsymbol{h}_n\|\|\boldsymbol{h}_e\|} \quad (10)$$

In our method, we set a threshold η , to judge whether two embeddings are similar enough. If the calculated cosine similarity is greater than η, it indicates that the disease entity is closely related to this node in KGs, and we regard related attributes from this node can help us with understanding diseases meanings.

To gain an appropriate threshold, we partition a subset and examined the matching status of entities under different thresholds, followed by manual expert verification.

Encode KG Knowledge

To fully utilize the information of matched entities in the graph, we encode them using LLM. Firstly, we concatenate each node details together in format like (entity name, entity definition, entity description). And we join multiple node details into one sequence with specified delimiter. Then we get references knowledge as a supplement information, and get its representation with LLM, also considering of long context.

$$\boldsymbol{h}_{RAG} = \text{LLM}(\boldsymbol{X}_{RAG}) \quad (11)$$

Additionally, when no entities found in the text, or no matched nodes in KG, we write an instruction text like "You are an experience doctor, please combine your background knowledge and patient's records to judge..." in replace of empty string instead of padding with zeros. In this way LLM can also encode an embedding containing instructive information to motivating more comprehensive understanding within the notes itself.

4.4 Multimodal Fusion Network

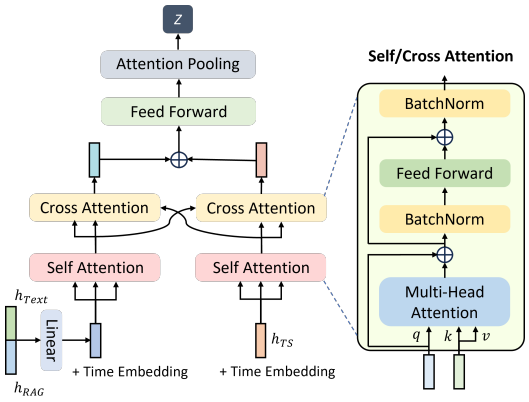

Figure 4: Fusion module. It combines multimodal embeddings with attention mechanism into a fused representation.

Currently, there are three learned hidden representations, denoted respectively as $\boldsymbol{h}_{TS}$, $\boldsymbol{h}_{Text}$, and $\boldsymbol{h}_{RAG}$. We first concatenate the hidden representations extracted from entities with those from the text, and then utilize MLP network to map them to a unified dimension.

$$\boldsymbol{h}'_{Text} = \text{MLP}(\text{Concat}(\boldsymbol{h}_{Text}, \boldsymbol{h}_{RAG})) \quad (12)$$

To better integrate information from different modalities, we proposed an attention-based fusion network mainly consisting of self-attention layers and cross-attention layers. Specifically, we first apply self-attention to each modality. Then we use the output of one modality as the query, and the output of the other modality as the key and value to compute cross-attention.

$$\begin{aligned} \tilde{\boldsymbol{h}}_{Text} &= \text{MHSA}(\boldsymbol{h}'_{Text} + \boldsymbol{h}_{Time}), \\ \tilde{\boldsymbol{h}}_{TS} &= \text{MHSA}(\boldsymbol{h}_{TS} + \boldsymbol{h}_{Time}), \\ \boldsymbol{h}_{Text} &= \text{MHCA}(\tilde{\boldsymbol{h}}_{Text}, \tilde{\boldsymbol{h}}_{TS}), \\ \boldsymbol{h}_{TS} &= \text{MHCA}(\tilde{\boldsymbol{h}}_{TS}, \tilde{\boldsymbol{h}}_{Text}) \end{aligned} \quad (13)$$

where MHSA represents multi-head self-attention, MHCA represents multi-head cross-attention, and $\boldsymbol{h}_{Time}$ represents time embedding. In addition, we apply residual connections and BatchNorm to every multi-head attention layer and Feed-Forward Network.

As a result, the outputs of the two cross-attention modules have carried information from both modalities. We further sum them up and use attention poooling layer to obtain the fused information.

$$\begin{aligned} \boldsymbol{z} &= \alpha * \boldsymbol{z}_{TS} + (1 - \alpha) * \boldsymbol{z}_{Text} \\ \boldsymbol{z}^{*} &= \text{AttnPool}(\text{MLP}(\boldsymbol{z})) \end{aligned} \quad (14)$$

where α is a learnable parameter and AttnPool refers to attention pooling.

Finally, the fused representation $\boldsymbol{z}^*$ is expected to predict downstream tasks. We pass $\boldsymbol{z}^*$ through a single-layer MLP network to obtain the final prediction results $\hat{y}$:

$$\hat{y} = \text{MLP}(\sigma(\boldsymbol{z}^*)) \tag{15}$$

The BCE Loss is selected as the loss function for the binary mortality outcome and readmission prediction task:

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \tag{16}$$

where N is the number of patients within one batch, $\hat{y} \in [0, 1]$ is the predicted probability and y is the ground truth.

5 Experimental Setups

5.1 Dataset, KG and Task Description

Sourced from the EHRs of the Beth Israel Deaconess Medical Center, MIMIC-III dataset is extensive and widely used in healthcare research. We adhere to the benchmark pipeline Gao *et al.* [2024]; Zhu *et al.* [2024a] for preprocessing time-series data. 17 lab test features (include categorical features) and 2 demographic features (age and gender) are extracted. To minimize missing data, we consolidate every consecutive 12-hour segment into a single record for each patient, focusing on the first 48 records. And we follow Khadanga *et al.* [2019] to extract clinical notes. We excluded all clinical notes lacking associated chart time and removed all patients without any notes. We randomly split the dataset into training (10776 samples), validation (1539 samples) and test (3080 samples) set with 7:1:2 percentage.

The external knowledge base we utilized is PrimeKG Chandak *et al.* [2023], which integrates 20 high-quality resources to describe 17,080 diseases with 4,050,249 relationships representing ten major biological scales, including disease-associated entities. Futhermore, PrimeKG extracts textual features of disease nodes containing information about disease prevalence, symptoms, etiology, risk factors, epidemiology, clinical descriptions, management and treatment, complications, prevention, and when to seek medical attention, which are highly relevant to the clinical prediction tasks.

We conduct in-hospital mortality prediction and 30-day readmission prediction task in our experiments. Both are binary classification tasks: predicting patient mortality outcomes (0: alive, 1: dead) and readmission likelihood (0: no readmission, 1: possible readmission).

5.2 Evaluation Metrics

We adopt the following evaluation metrics, which are widely used in binary classification tasks:

- **AUROC**: This metric is our primary consideration in binary classification tasks due to its widespread use in clinical settings and its effectiveness in handling imbalanced datasets McDermott *et al.* [2024].
- **AUPRC**: The AUPRC is particularly useful for evaluating performance in datasets with a significant imbalance between classes Kim and Hwang [2022].
- **min(+P, Se)**: This composite metric represents the minimum value between precision (+P) and sensitivity (Se), providing a balanced measure of model performance Ma *et al.* [2022].
- **F1**: The F1 score is particularly useful in scenarios where an equitable trade-off between precision and recall is desired Chinchor [1992].

All these four metrics are the higher the better.

5.3 Hyperparameters

The batch size is consistently set at 256. For all experiments, we report performance in the form of mean±std., where we adopt bootstrap strategy for 10 times.

We conduct a grid search for the baseline models. The hyperparameters for our REALM model are: a hidden dimension of 312 and a learning rate of 6e-4.

5.4 Baseline Models

EHR Prediction Models We include multimodal EHR baseline models (M3Care Zhang *et al.* [2022], MPIM Zhang *et al.* [2023a], UMM Lee *et al.* [2023], VecoCareXu *et al.* [2023]) and approaches that incorporating external knowledge from KG (GRAM Choi *et al.* [2017], KAME Ma *et al.* [2018], CGL Lu *et al.* [2021], KerPrint Yang *et al.* [2023]) as our baselines. Detailed description of each model is in Appendix.

Text Embedding Approaches we compare different text embedding approaches including BERT's [CLS] token Devlin *et al.* [2018], BGE-M3 Chen and Xiao [2024] and Qwen-7B's encoder Bai *et al.* [2023]. Detailed set ups are described in Appendix.

Multimodal Fusion Baselines To examine the effectiveness of our fusion network, we consider fusion methods: Add Wu and Han [2018], Concat Khadanga *et al.* [2019]; Deznabi *et al.* [2021], Tensor Fusion (TF) Zadeh *et al.* [2017], and MAG Rahman *et al.* [2020]; Yang and Wu [2021]. Detailed description is in Appendix.

6 Experimental Results

The performance of our REALM framework on in-hospital mortality and 30-day readmission prediction tasks on the MIMIC-III dataset is summarized in Table 2. Our approach consistently outperforms the baseline models. Specifically, REALM achieves a significant relative improvement in AUROC (1.09%, 2.06%), AUPRC (2.75%, 4.75%), min(+P, Se) (0.79%, 3.12%) and F1 scores (21.90%, 30.39%) with the best baseline model for both tasks, indicating its superior practical applicability in real-world clinical settings.

6.1 Ablation Studies

Comparing Each Modality with RAG-Enhancement

To understand the contribution of RAG-enhancement to each modality, we conducted an ablation study. The results, as illustrated in Table 3, reveal that the RAG-enhanced versions of both time-series and text modalities significantly improve the

Table 2: *In-hospital mortality and readmission prediction results on MIMIC-III.* **Bold** indicates the best performance. All metrics are multiplied by 100 for readability purposes.

Methods	Mortality Outcome Prediction				30-Day Readmission Prediction			
	AUROC	AUPRC	min(+P, Se)	F1	AUROC	AUPRC	min(+P, Se)	F1
MPIM	85.24±1.12	50.52±2.56	50.59±2.33	30.53±2.33	78.62±1.58	49.30±3.01	49.65±2.54	26.61±2.20
UMM	84.01±1.10	49.76±2.21	49.41±2.45	36.21±1.90	77.46±1.36	47.81±2.55	47.27±1.91	34.14±2.21
VecoCare	83.43±1.49	47.28±2.68	47.92±2.22	42.52±2.08	76.93±1.82	46.18±2.76	47.22±2.63	38.79±2.27
M3Care	83.33±1.24	47.86±2.33	49.96±1.99	24.81±2.62	76.80±1.55	46.29±2.62	45.38±2.32	21.51±2.23
GRAM	84.70±1.34	49.21±4.45	49.64±2.85	38.02±3.19	77.84±1.49	47.97±3.68	46.95±2.12	35.24±2.89
KAME	84.59±1.11	49.48±3.37	49.51±2.33	36.14±2.24	78.04±1.34	48.23±3.21	47.41±2.50	31.70±2.19
CGL	84.20±1.16	47.64±3.47	47.67±2.61	38.36±2.04	77.47±1.33	46.68±3.33	47.73±2.25	35.34±2.35
KerPrint	85.29±1.21	51.23±3.48	50.88±2.24	37.00±3.54	78.41±1.50	49.70±3.23	49.39±2.53	34.31±2.35
Ours (REALM)	**86.22±0.81**	**52.64±2.47**	**50.92±2.01**	**51.83±2.10**	**80.24±1.53**	**52.06±2.64**	**51.20±2.50**	**50.58±2.51**

model's performance. This confirms the hypothesis that enriching EHR data with external medical knowledge can effectively capture more complex semantic medical background knowledge, leading to more accurate clinical predictions.

Comparing Different Fusion Network

Our analysis extends to comparing the effectiveness of different fusion strategies for integrating time-series and text modalities. As shown in Table 3, our designed self- and cross-attention based adaptive multimodal fusion network outperforms all baseline methods in both tasks. This demonstrates the advantage of our fusion strategy in attentively learning and integrating modality-specific features for improved prediction performance. Moreover, with RAG-enhanced knowledge combining both modalities, our REALM method achieve the SOTA performance against all reduced versions.

Comparing Text Embedding Models

The impact of using different text embedding models on the performance of our REALM framework is also explored. Table 4 highlights that the Qwen-7B model, with its extensive training data and long-context support, significantly outperforms BERT and BGE-M3 in all metrics. This suggests that leveraging advanced large language models for embedding clinical notes can enhance the model's understanding of complex medical narratives.

6.2 Further Analysis

Robustness to Data Sparsity

To evaluate the robustness of our REALM framework against data sparsity, we conducted experiments by artificially reducing the dataset's completeness by 20%, 40%, 60% and 80%. As depicted in Figure 5, REALM demonstrates remarkable resilience, outperforming other recent SOTA models even under extreme data scarcity. This robustness is particularly crucial in clinical environments where large amount of data collection is often challenging, making REALM a valuable tool for real-world applications.

Evaluation of Quality of Retrieved Entities

We take the entities extracted by RAG pipeline as input to XGBoost model to calculate the importance of the entities, thereby indirectly measuring the contribution of the entities to the prediction task. Figure 6 shows the medical record of a patient's visit and the extracted disease type entities, among which "atrial fibrillation" and "sepsis" have the highest importance scores, followed by "osteomyelitis", and finally "renal failure". By examining the nodes and attribute information corresponding to each disease entity in the knowledge graph, we find that they are all relatively dangerous diseases in clinical practice, and patients have a high probability of experiencing mortality. This reflects the effectiveness of our proposed RAG-driven process.

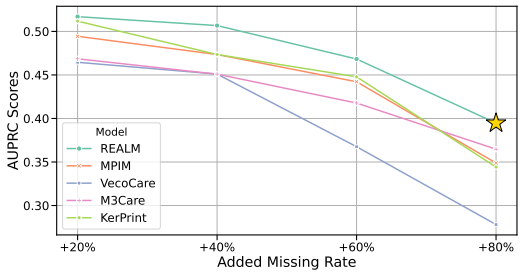

Figure 5: *AUPRC Performance across 4 Sparsity Levels on MIMIC-III mortality outcome task.* REALM exhibits better performance on multiple missing rate levels than recent SOTA baselines.

pt a 67 year old man with ckd v s/p renal transplantation , dm2 , cad s/p pcis and cabg , **atrial fibrillation** , and pvd s/p left toe amputation who originally presented to the hospital on with an infected left heel ulcer . his course has been complicated by progressive **renal failure** prompting discontinuation ... due to the patient s increasing nursing needs and concern for evolving **sepsis** from his heel **osteomyelitis** and his tenuous fluid balance , pt transferred to the micu for closer monitoring .

Figure 6: *Case study of retrieved entities in original clinical notes with importance score computed.* The deeper yellow background color denotes higher importance score.

7 Conclusions

In this work, we propose REALM, a RAG-driven multimodal EHR data representation learning framework that incorporates time-series EHR, clinical notes data and external knowledge graph for healthcare prediction. REALM framework comprehensively leverages LLM's semantic reasoning abil-

Table 3: *Ablation Studies results of 1) comparing each modality with RAG-enhancement, 2) comparing different multimodal fusion network.* **Bold** indicates the best performance. All metrics are multiplied by 100 for readability purposes.

Methods	Mortality Outcome Prediction				30-Day Readmission Prediction			
	AUROC	AUPRC	min(+P, Se)	F1	AUROC	AUPRC	min(+P, Se)	F1
TS only	83.43±1.08	48.70±3.04	46.72±2.10	37.38±2.94	77.63±1.38	48.11±3.23	47.41±2.08	33.40±2.91
TS+RAG_{TS}	84.22±0.98	49.80±3.15	48.35±1.91	41.10±2.95	78.02±1.37	48.36±2.98	47.31±2.38	34.39±2.73
Text only	80.11±1.69	40.54±3.51	41.05±3.27	33.96±2.35	74.57±1.86	40.99±3.52	42.49±3.10	30.87±2.50
Text+RAG_{Text}	81.01±1.52	42.92±3.43	42.51±3.02	45.13±2.44	74.48±1.91	43.38±3.70	43.46±3.18	40.01±2.91
TS+Text: `Add`	84.72±1.03	48.60±3.45	50.05±2.59	46.86±2.43	78.23±1.74	48.77±3.61	48.76±2.87	47.29±2.46
TS+Text: `Concat`	85.22±0.93	49.94±3.14	49.75±1.82	46.51±2.18	78.96±1.48	50.08±3.27	50.60±2.18	40.61±2.02
TS+Text: `TF`	84.13±1.24	49.06±3.38	50.21±2.88	37.54±3.05	77.16±1.96	47.64±3.60	48.17±2.29	31.86±2.65
TS+Text: `MAG`	84.75±0.97	50.31±2.71	48.58±2.42	45.81±2.20	78.04±1.58	49.26±2.86	48.88±2.37	45.30±2.43
TS+Text: `Ours`	85.18±0.95	50.68±2.64	47.90±2.27	49.81±2.37	78.79±1.47	49.69±2.92	48.91±2.57	49.94±2.36
Ours (REALM)	**86.22±0.81**	**52.64±2.47**	**50.92±2.01**	**51.83±2.10**	**80.24±1.53**	**52.06±2.64**	**51.20±2.50**	**50.58±2.51**

Table 4: *Ablation study results of suing different text embedding models.* **Bold** indicates the best performance. All metrics are multiplied by 100 for readability purposes.

Tasks	Metrics	BERT	BGE-M3	Qwen-7B
Out.	AUROC	83.66±1.34	84.72±0.97	**86.22±0.81**
	AUPRC	48.22±3.13	50.42±2.88	**52.64±2.47**
	min(+P, Se)	48.39±3.26	49.29±2.55	**50.92±2.01**
	F1	43.46±2.61	49.40±2.41	**51.83±2.10**
Read.	AUROC	76.55±1.89	78.03±1.63	**80.24±1.53**
	AUPRC	46.10±3.17	49.10±3.28	**52.06±2.64**
	min(+P, Se)	46.10±3.10	47.67±2.41	**51.20±2.50**
	F1	39.68±2.73	48.81±2.22	**50.58±2.51**

ity, long context encoding capacity, and knowledge graph's medical context. REALM framework achieve SOTA performance on MIMIC-III datasets' in-hospital mortality and 30-day readmission tasks, showcasing its effectiveness of incorporating knowledge from external knowledge bases. Our work marks a step towards more effective utilization of EHR data in healthcare, offering a potent solution to enhance clinical representations with external knowledge and LLMs.

Ethical Statement

This study, involving the analysis of Electronic Health Records (EHR) using the MIMIC dataset, is committed to upholding high ethical standards. The MIMIC dataset is a de-identified dataset, ensuring patient confidentiality and privacy. It is available through a data use agreement, underscoring our commitment to responsible data handling and usage. Our approach has been designed to minimize any potential harm and to ensure that our findings are as unbiased and fair as possible, taking into account the diverse and complex nature of medical data. We have also taken rigorous steps to ensure our research aligns with these values.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2022ZD0116401).