CARER - ClinicAl Reasoning-Enhanced Representation for Temporal Health Risk Prediction

Tuan Dung Nguyen[1], Thanh Trung Huynh[2], Minh Hieu Phan[3], Quoc Viet Hung Nguyen[4], Phi Le Nguyen[1§]

[1]School of Information and Communication Technology, Hanoi University of Science and Technology
[2]Swiss Federal Institute of Technology Lausanne (EPFL)
[3]Australian Institute for Machine Learning, The University of Adelaide
[4]School of Information and Communication Technology, Griffith University
dung.nt232198m@sis.hust.edu.vn, thanh.huynh@epfl.ch, vuminhhieu.phan@adelaide.edu.au, quocviethung.nguyen@griffith.edu.au, lenp@soict.hust.edu.vn

Abstract

The increasing availability of multimodal data from electronic health records (EHR) has paved the way for deep learning methods to improve diagnosis accuracy. However, deep learning models are data-driven, requiring large-scale datasets to achieve high generalizability. Inspired by how human experts leverage reasoning for medical diagnosis, we propose **CARER**, a novel health risk prediction framework that enhances deep learning models with clinical rationales derived from medically proficient Large Language Models (LLMs). In addition, we provide a cross-view alignment loss which aligns the "local" view from the patient's health status with the "global" view from the external LLM's clinical reasoning to boost the mutual feature learning. Through extensive experiments on two predictive tasks using two popular EHR datasets, our *CARER*'s significantly exceeds the performance of state-of-the-art models by up to 11.2 %, especially in improving data efficiency and generalizability [1].

1 Introduction

Electronic Health Records (EHRs) are valuable sources, widely adopted by clinicians globally to predict future health events, such as diagnosis (Choi et al., 2016b; Luo et al., 2020), mortality (Abedi et al., 2021), and readmission (Wang et al., 2023; Yang and Wu, 2021). EHR information can be divided into two categories: (1) structured data, encompassing ICD codes, laboratory measurements, and demographic details, and (2) unstructured, free-form clinical notes. Current methods in health risk prediction aggregates multimodal data by designing encoders to extract robust modality-specific features (Choi et al., 2016b), modeling temporal relationships across patient visits (Luo et al., 2020; Gao et al., 2020), or developing effective cross-modal fusion mechanisms (Luo et al., 2020; Xu et al., 2023, 2021).

Although these deep learning models set impressive benchmarks for health risk detection tasks, their data-driven nature results in several limitations. Firstly, these models exhibit low data efficiency, necessitating a diverse and extensive dataset to effectively capture medical patterns and relationships. In scenarios with limited data, their generalizability significantly diminishes. Second, they lack interpretability, which is pivotal for trustworthy clinical applications. This raises a crucial question: *is relying solely on EHR information adequate for accurate health risk prediction?*

Our motivation and challenges. Let's consider a common procedure physicians follow for disease diagnosis. Initially, they examine the EHRs, translate ICD codes to corresponding diseases, categorize lab values as normal or abnormal. Then they incorporate external knowledge from clinical experience and medical literature to synthesize the data, interpret the patient's condition, and diagnose health risks. The second step is referred to as *clinical reasoning*, a crucial process in formulating final health risk assessments. Contrary to the data-driven deep models, the ability to reason and integrate knowledge from various clinical sources enables generalizablity and alleviates the data scarcity problem in low-data domains. As such, how to integrate such a human-like clinical reasoning into a machine learning system remains an open question.

[1]Code is available at https://github.com/tuandung2812/CARER-EMNLP-2024.

[§]Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10392–10407
November 12-16, 2024

In recent years, state-of-the-art Large Language Models (LLMs) (Achiam, 2023; Jiang et al., 2023) have demonstrated substantial reasoning capabilities and efficacy across various tasks (Cabral et al., 2024; Lee et al., 2023; Liévin et al., 2024). Therefore, leveraging LLMs to provide reasoning capabilities, thereby advancing health risk prediction, presents a compelling approach. However, the implementation of this approach poses the following challenges:

- **C1.** LLMs are predominantly trained on general datasets and may not perform optimally in specialized domains such as medicine.
- **C2.** Despite their proficiency in textual inference, LLMs are less effective with numeric data. This limitation is significant considering the prevalence of numeric data such as ICD codes and lab values in EHRs.
- **C3.** Naively fusing features of EHRs and external LLM's knowledge is challenging. EHR data is a *local* view of the individual patient's conditions. In contrast, clinical reasoning provides a *global* view by integrating external knowledge and LLM's rationales. Semantic gap between the two creates challenges for feature fusions.

Our solution. To overcome the issues mentioned above, we propose *CARER* - **C**linic**A**l **R**easoning-**E**nhanced **R**epresentation for Temporal Health Risk Prediction (Fig. 1) - which leverages Chain-of-Thought (CoT) prompting to query clinical reasoning steps from multimodal EHR sources. To address **C1**, we employ Retrieval Augmented Generation (RAG) via external medical knowledge base to enrich the context of the reasoning process. Additionally, we utilize the CoT prompting technique to guide the reasoning process of the LLM. Regarding **C2**, we propose a verbalization mechanism to convert numerical data into text format, enriched with comprehensive semantics. To address **C3**, we introduce a cross-view alignment objective to enhance consistency between the *local* multimodal features and the *global* clinical rationale features.

Our contributions are three-fold as follows.

- We propose an LLM-assisted clinical reasoning mechanism that leverages RAG and CoT techniques to temporally reason patients' conditions from aggregated multimodal data. This clinical reasoning serves as an unified auxiliary modality which, enriches the available information and enhances the generalizability of health risk prediction. To the best of our knowledge, this is the first attempt to employ LLM-assisted reasoning in addressing the health risk prediction problem.
- We design a cross-view alignment loss, which aligns representations of EHR data and clinical reasoning to enhance cross-view consistency. As such, our technique mutually boosts the feature learning of the local view of raw patient's data and the global view of the external LLM's reasoning knowledge.
- Through comprehensive experiments, our CARER model surpasses state-of-the-art models by up to 11.2% on two diagnostic tasks across the MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) public benchmarks.

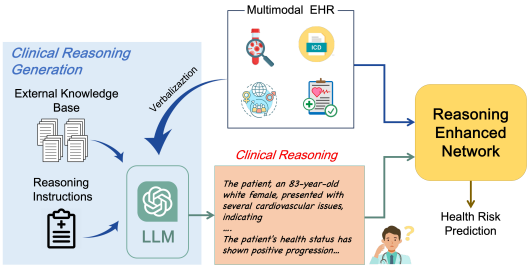

Figure 1: **CARER - ClinicAl Reasoning-Enhanced Representation for temporal health prediction**. We use CoT-based reasoning instructions and retrieval-augmented generation (RAG) via external medical knowledge base to prompt the LLM and extract reasoning about patients' disease progression.

2 Related Work

Many studies have applied deep learning models to predict diagnoses and health outcomes from temporal EHRs. We categorize the following methods into three primary approaches based on the information sources used for prediction: structured data, unstructured data or multimodal inputs, and the use of external knowledge.

Structured Data. Initial efforts focused on mining structured data like ICD codes for prediction. RETAIN (Choi et al., 2016b) uses a double RNN to process ICD sequences bi-directionally, followed by a two-level attention mechanism. Other methods, such as Dipole (Ma et al., 2017), Timeline (Bai et al., 2018), and T-LSTM (Baytas et al.,

2017), combine RNNs with attention. Time-aware models like T-LSTM (Baytas et al., 2017) and HiTANet (Luo et al., 2020) capture temporal patterns using time-decay and self-attention mechanisms. Later approaches use graph-learning mechanisms from visit-level or patient-level graphs to understand relationships between medical codes (Lu et al., 2022, 2021a; Yang et al., 2023; Choi et al., 2016a).

Table 1: Comparison of CARER and existing approaches.

Method	EHR data				External knowledge	Clinical reasoning
	ICD	Lab value	Demographic	Clinical note		
(Luo et al., 2020)	✓					
(Choi et al., 2016b)	✓					
(Lu et al., 2022)	✓				✓	
(Grundmann et al., 2022)	✓			✓		
(Xu et al., 2023)	✓			✓		
(Wang et al., 2023)	✓	✓	✓	✓		
(Xu et al., 2021)	✓	✓	✓	✓		
(Ye et al., 2021)	✓				✓	
(Yang and Wu, 2021)		✓	✓	✓		
(Zhang et al., 2023)		✓		✓		
(Gao et al., 2020)		✓				
(Zhu et al., 2024)		✓		✓	✓	
CARER (Ours)	✓	✓	✓	✓	✓	✓

Unstructured Data and Multimodality. Recent works incorporate unstructured data, mainly clinical notes, for prediction. CGL (Lu et al., 2021a) uses attention mechanisms to emphasize important words in medical text. VecoCare (Xu et al., 2023) integrates medical codes and text representations using two Transformer architectures and pretraining tasks for semantic alignment. Other recent studies combine multiple structured features (ICD codes, lab values, demographics, drug codes) with unstructured clinical notes. MUFASA (Xu et al., 2021) employs Neural Architecture Search for optimal multimodal network architecture, while MedHMP (Wang et al., 2023) uses a multimodal pretraining paradigm with tasks like reconstruction and contrastive learning for better representation.

External Knowledge. Multiple graph-based methods have exploited medical ontologies/knowledge graphs, like G-BERT (Shang et al., 2019), CGL (Lu et al., 2021a), Sherbert (Lu et al., 2021b) MetaCare (Tan et al., 2022) GRAM (Choi et al., 2016a) utilizing ICD-9 diseases hierarchy, or KerPrint utilizing medical knowledge graph containing multiple entities like diseases, lab tests, medications from SNOMED CT (Donnelly, 2006). Others utilize unstructured medical documents as auxiliary information, such as MedRetriever (Ye et al., 2021) and (Zhu et al., 2024). Tab. 1 compares our method with other state-of-the-art baselines. To the extent of our knowledge, we are the first to incorporate clinical reasoning for EHR's predictive modeling.

3 CARER

3.1 EHR Structure

Each patient's historical health record includes a sequence of hospital admissions $A = [A^1, \ldots, A^n]$, with each admission A^t $(t = 1, ..., n)$ contains multiple EHR modalities: categorical diagnosis codes I^t, lab values V^t, demographic features D^t, and clinical notes N^t. Diagnosis codes $I^t = [I_1^t, \ldots, I_d^t]$ include ICD-9 codes diagnosed by healthcare professionals. Lab values $V^t = [V_1^t, \ldots, V_m^t]$ represent continuous numerical data, e.g., blood glucose levels and blood pressure. Clinical notes $N^t = [N_1^t, \ldots, N_k^t]$ are written by medical professionals, and demographic features D^t include race, gender, and age. For simplicity, we omit patient-level and visit-level indices in the remaining sections.

3.2 Overview of CARER

Inspired by how human experts diagnose, we propose to extract clinical rationales by prompting LLMs using multimodal EHR data, relevant external medical documents, and CoT reasoning instructions. This clinical reasoning serves as an auxiliary modality that, when integrated with original EHR data, enhances contextual understanding and predictive accuracy.

Fig. 2 shows our CARER's framework, comprising three components. First, the multimodal encoding component utilizes standard encoders for extracting features from temporal EHRs between visits. Second, our proposed clinical reasoning module uses verbalized EHR data and medical texts from a knowledge database, employing CoT-based prompting techniques to generate clinical reasoning with LLMs (GPT-3.5). Third, the multiview alignment and fusion component leverages the proposed cross-view alignment loss to maximize consistency between two views (i.e., local multimodal features and global clinical reasoning) and fuses them to create predictive representation.

3.3 LLM-assisted Clinical Reasoning

This section describes our pipeline for constructing clinical reasoning with three steps: (1) verbalizing the EHR data, (2) retrieving pertinent medical documents, and (3) generating clinical reasoning.

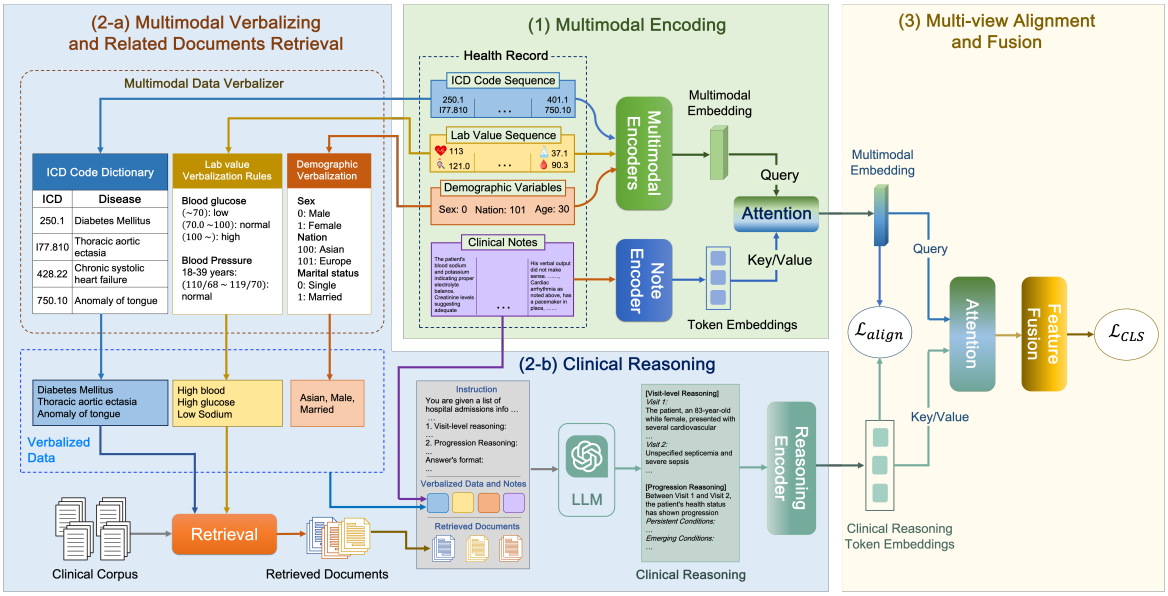

Figure 2: **Overview of CARER**. The clinical reasoning module constructs RAG and Chain of Thought (CoT) prompts using verbalized queries of multi-modal data (Subsection **2-a**) to generate clinical reasoning using LLM (Subsection **2-b**). Multi-view Alignment and Fusion component (Subsection **3**) uses cross-view alignment loss to align the global view from the clinical reasoning and the local view of raw multimodal EHR data modalities and subsequently fuses them to generate final predictions.

EHR Data Verbalization. The details of this process is illustrated in Block **2-a** in Fig. 2. For demographic variables D, we simply concatenate the variable descriptor (e.g., age, race, gender) and its value to obtain the verbalized input Q^D (i.e., "race: caucasian", "age: 80", "gender: male"). For the categorical ICD code I, we use the name from the ICD lookup table as the verbalized query Q^I. For the lab value V, we generate the verbalized input Q^V using a rule set f_{rule}, and determine three markers: *"low"*, *"normal"*, and *"high"*, based on statistical analysis of their value ranges. The verbalized query for a lab value v is then created by concatenating its indicator name (e.g., "blood glucose"), the numerical value (e.g., "115"), the measurement unit (e.g., "d/mL"), and the label (e.g., "high"). For instance, a lab value "blood glucose: 115 d/mL" is verbalized as "high blood glucose, 115 d/mL".

Medical Corpus. We derive the medical corpus from the medical knowledge base PrimeKG (Chandak et al., 2022). This knowledge base contains different types of medical entities like diseases, drugs, proteins, etc However, we only utilize 17,080 disease nodes in the knowledge graph. Each of these disease nodes contains descriptive textual features, including such as definition, symptoms, causes. In total, there are 14,252 textual node features, all of which we use to construct our final medical corpus.

Medical Document Semantic Retrieval. We develop a retrieval mechanism to obtain relevant documents for supplying contextual medical knowledge to LLMs, reducing hallucinations. This mechanism retrieves documents $\hat{\mathcal{P}}$ semantically similar to the patient's verbalized conditions Q. Using a Transformer pretrained with medical knowledge (Jin et al., 2023), we embed the documents and verbalized queries into a semantic space and compute cosine similarities between them. Pertinent documents $\hat{\mathcal{P}}$ that surpass similarity threshold $\beta = 0.95$ are retrieved.

Clinical Reasoning Generation. We create a Chain-of-Thought prompt based on temporal multimodal patients' data A and the retrieved medical knowledge $\hat{\mathcal{P}}$ to generate clinical reasoning $\mathcal{R}$ using LLM. Figure 3 provides a brief overview of our instructions, which guides the LLM to reason about individual visits, and cross-visit disease progressions. The details are as below.

Visit-level Reasoning: The LLM analyzes patients' health status visit-by-visit, using verbalized multimodal data. It provides a summary of current conditions, augmented by relevant retrieved documents to enhance accuracy and reduce hallucination. The instruction for this reasoning is denoted as $\mathcal{T}^v$.

Progression Reasoning: The progression reasoning instruction $\mathcal{T}^p$ directs the LLM to summarize health progression across visits, identifying persistent and improving conditions, emerging issues,

and changes in lab values or vital signs.

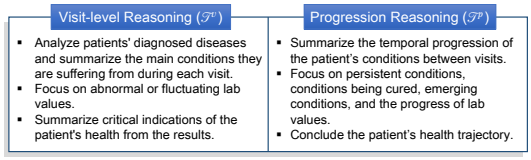

Figure 3: Visit-level and cross-visit progression reasoning.

We combine the instructions $\mathcal{T}^v$ and $\mathcal{T}^p$ with a description of the desired answer format $\mathcal{T}^f$ to form the full instruction template $\mathcal{T}$, i.e., $\mathcal{T} = [\mathcal{T}^v, \mathcal{T}^p, \mathcal{T}^f]$.

Patient Information Construction. We combine all the verbalized EHR data (i.e., Q^I, Q^V, and Q^D) with the clinical notes N to provide LLMs with patient's information, denoted as $\mathcal{S}$. This verbalized EHR data $\mathcal{S}$ is combined with the instruction template $\mathcal{T}$ and the retrieved documents $\hat{\mathcal{P}}$ to prompt LLMs (GPT-3.5). As such, we obtain clinical reasoning $\mathcal{R}$ based as follows:

$$\mathcal{R} = GPT\left(\mathcal{T}, \mathcal{S}, \hat{\mathcal{P}}\right).$$

The clinical reasoning $\mathcal{R}$ is fed into Clinical Longformer (Li et al., 2022), denoted as f^C, to derive the reasoning representation z_R:

$$z_R = [\mathbf{z}_{cls}, \mathbf{z}_{r_1}, \mathbf{z}_{r_2}, \ldots, \mathbf{z}_{r_n}] = f^C(\mathcal{R}).$$

3.4 Multimodal Encoding and Fusion

The multimodal encoder consists of m encoders, each embeds temporal information of a modality $M \in \{I, V, D, N\}$ into a representation space z_M. We use interval-aware Transformer (Li et al., 2022) for ICD code modality I, an LSTM (Baytas et al., 2017) for lab values V, an MLP for demographic data, and Clinical Longformer (Li et al., 2022) for clinical notes N. Finally, multi-modal features are obtained by summing features from m modalities: $z_E = \sum_{M \in \{I,V,D,N\}} z_M$.

Cross-view Alignment Loss. While the multimodal features z_E captures the local patterns and relationships presented in the patient's local data, the clinical reasoning z_R provides a more "global" view obtained from external documents and LLM's rationales. Thus, aligning these two views mutually boosts feature learning of the two encoders, and minimizes the semantic gap between them. To this end, we propose a cross-view alignment loss $\mathcal{L}_{\text{align}}$ to facilitate the fusion of the local multimodal features z_E and the global reasoning features z_R.

Given a mini-batch of size b, we have the batch-wise multimodal features $Z_E \in \mathbb{R}^{b \times d_{z_E}}$ and the corresponding batch-wise clinical reasoning features $Z_R \in \mathbb{R}^{b \times d_{z_R}}$. We compute matrices Q_E and Q_R ($\in \mathbb{R}^{b \times b}$), which represent the in-batch samples' similarities of multimodal features and clinical reasoning features, respectively:

$$Q_E = \frac{Z_E \cdot Z_E^\top}{\|Z_E\|}; \qquad Q_R = \frac{Z_R \cdot Z_R^\top}{\|Z_R\|}.$$

We define the alignment loss as the Frobenius norms between two similarity matrices Q_R and Q_E:

$$\mathcal{L}_{\text{align}} = \frac{1}{b^2}\|Q_E - Q_R\|_2. \tag{1}$$

Multi-view Feature Fusion. After aligning the "global" reasoning features and "local" multimodal features, we fuse them to generate the final diagnosis representation. To emphasize important reasoning tokens, an attention-pooling mechanism is applied to aggregate feature of the clinical reasoning token. Specifically, the multimodal features z_E are used as a query to query relevant and important reasoning tokens z_R:

$$\widetilde{z_R} = \text{softmax}\left(\frac{z_E W_Q (z_R W_K)^\top}{\sqrt{d_k}}\right)(z_R W_V),$$

where W_Q, W_K, and W_V are projection layers for the Query, Key, and Value, respectively. Multimodal features z_E and attention-pooled reasoning features $\widetilde{z_R}$ are concatenated and fed into a Multilayer Perceptron (MLP) to obtain a final representation of the patient's health history:

$$z = \text{MLP}([z_E, \widetilde{z_R}]).$$

A fully connected layer is applied on z to produce the prediction y_i' for each patient i. The cross-entropy loss is computed as below:

$$\mathcal{L}_{cls} = -\frac{1}{b}\sum_{i=1}^{b}\left(y_i^\top \log(y_i') + (1 - y_i)^\top \log(1 - y_i')\right),$$

where y_i' and y_i depict the predicted probability, and the ground truth, respectively. The final loss function is a combination of the classification loss $\mathcal{L}_{cls}$ and the cross-view alignment loss $\mathcal{L}_{\text{align}}$ weighted by γ:

$$\mathcal{L} = \mathcal{L}_{cls} + \gamma \mathcal{L}_{align}.$$

4 Performance Evaluation

In this section, we compare the performance of CARES with other diagnosis baselines on two EHR benchmarks to validate their predictive performance. Afterward, we conduct an in-depth analysis on our algorithm's generalizability and ablation studies on different components that made up our work.

4.1 Experimental Settings

Datasets. We conduct our experiments on two widely used Electronic Health Records datasets: MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) to validate the capabilities of our proposal. For each dataset, we only consider patients with more than one visit in their healthcare history, resulting in 7,493 patients for MIMIC-III and 85,155 patients for MIMIC-IV. As multimodal EHRs datasets, they contains four different types of input modalities: ICD Diagnosis codes (**I**), Continuous Lab values (**C**), Demographic information (**D**), and Clinical notes (**N**) (Xu et al., 2021). For MIMIC-III, we utilize all four data modalities. For MIMIC-IV, we exclude the clinical notes (**N**) because they are all "Discharge summaries" which typically contain direct indications of the diagnoses (Lu et al., 2021a).

Evaluation Scenarios. Following (Lu et al., 2022), we evaluate the predictive performance of the models over two following tasks: Heart Failure Prediction and Full Diagnosis Prediction. The former task is a binary classification, while the latter task is a multi-label classification. For each task, the final visit of each patient is used to construct the ground truth for predictive tasks, while all previous visits are utilized as the input to the model.

The datasets are split into Train/Validation/Test partitions, with a ratio of 70/10/20. We adopt standard evaluation metrics, including precision (P), Recall (R), weighted F1-score (F1), and AUC scores to evaluate the heart failure predictions. For full diagnosis predictions, employ top k recall (R@k), and weighted F1 score.

Baselines. We compare the performance of our proposed technique *CARER* against 8 recent methods: *RETAIN* (Choi et al., 2016b), *T-LSTM* (Baytas et al., 2017), *HiTANET* (Luo et al., 2020), *Chet* (Lu et al., 2022), *CGL* (Lu et al., 2021a), and *VecoCare* (Xu et al., 2023), *MUFASA* (Xu et al., 2021), *MedHMP* (Wang et al., 2023). *MUFASA* and *MedHMP* are the two recent methods that use all 4 modalities, like our model *CARER*. *VecoCare* and *CGL* can only integrate the ICD sequences (**I**) and the clinical notes (**N**), while the rest can only handle the ICD sequences.

Implementation Details. The output latent dimension of the ICD Sequence Encoder, Lab Values Encoder, and Demographic Information Encoder is set to be $R^d = 64$. For the Clinical Notes Encoder and Clinical Reasoning Encoder, we utilizes ClinicalLongformer (Li et al., 2022), a Transformer Encoder pretrained on a massive medical corpus and has the capability of processing long input text, up to 4096 tokens. The core LLM used in our framework is GPT-3.5 (*gpt-3.5-turbo*).

4.2 End-to-end Comparison

We report the performance of our proposed method, *CARER*, compared to other baselines on MIMIC-III and MIMIC-IV in the two tasks: Full Diagnosis Prediction and Heart Failure Prediction. The result is shown in Table 2. Overall, our proposed method outperforms other baselines in every metric for both tasks, demonstrating its effectiveness in incorporating and aligning clinical reasoning.

More specifically, *CARER* achieves a performance gain of around 4-11.2% against the best baseline *MedMHP* for the report metrics for Full Diagnosis Task, while those margins for Heart Failure Prediction Tasks tend to be less significant (0.96-5%). This might be because the Full Diagnosis Prediction Task requires more diverse and complex knowledge from multiple types of diseases and conditions; where our method thrives thanks to the retrieval and reasoning capabilities of the LLM core. On the other hand, our technique can achieve constantly the F1 score of nearly 0.8 in Heart Failure prediction for the both datasets, which are 3% better than the top baselines like *MedMHP* and *MUFASA*.

Among the baselines, *MedMHP* and *MUFASA* perform the best overall, as they can integrate all possible data modalities, similar to our proposed technique, *CARER*. VevoCare performs very well on the MIMIC-III dataset and achieves results nearly on par with *CARER*, but it is unable to generate results for MIMIC-IV, since VevoCare's pretraining task critically depends on the presence of medical notes (**N**). The four techniques *RETAIN*, *HiTANET*, *Chet*, and *T-LSTM*, which use only the ICD sequences, achieve results around 20% worse than *MedMHP* and even *CGL*, which justifies the importance of leveraging multi-modal EHR data,

Table 2: End-to-end performance comparison of the techniques.

Model	MIMIC-III							MIMIC-IV						
	Diagnosis			Heart Failure				Diagnosis			Heart Failure			
	R@10	R@20	F1	AUC	F1	P	R	R@10	R@20	F1	AUC	F1	P	R
RETAIN	26.60	34.35	20.13	82.71	72.20	70.96	73.44	28.40	34.46	24.89	83.17	72.71	72.03	72.94
T-LSTM	25.49	33.24	19.58	82.14	72.36	71.80	73.07	24.32	35.43	24.02	83.41	73.20	70.17	74.40
HiTANET	27.33	35.68	23.62	84.95	74.13	75.23	73.62	29.35	37.47	26.32	85.54	75.29	74.40	76.02
Chet	28.13	37.04	22.37	83.49	75.24	72.88	77.01	29.89	38.19	24.35	86.72	76.18	78.30	75.09
CGL	29.54	38.77	22.98	86.19	75.35	75.94	74.66	28.93	37.97	23.52	85.48	75.64	76.98	73.62
VecoCare	32.33	39.13	23.47	87.63	76.58	75.20	77.07	-	-	-	-	-	-	-
MUFASA	30.65	38.71	22.28	83.29	74.74	73.40	76.18	29.13	37.24	25.13	85.44	74.95	75.81	73.40
MedHMP	31.58	39.14	23.74	84.69	75.87	74.67	77.19	29.68	38.64	26.35	86.14	75.53	76.37	75.12
CARER	**32.90**	**42.34**	**26.40**	**87.92**	**78.61**	**77.38**	**79.60**	**31.46**	**40.50**	**28.41**	**88.60**	**77.80**	**77.32**	**78.65**

especially the informative but noisy clinical notes.

4.3 Data Efficiency

In the medical domain, labeled EHR data is often scarce due to low availability, strict ethical rules, privacy concerns, and the need for domain experts to label the data. We here investigate the data efficiency of the techniques against low-data regimes, reported in Figure 4. Specifically, we used only a portion of the MIMIC-IV training set ranging from 100%, 50%, 25%, to 10% for training the Heart Failure Prediction task. As shown, *CARER* outperforms all other methods for every level of training data. Additionally, we observe that the accuracy of other methods declines more rapidly than *CARER* as the amount of training data decreases. Our technique can still achieve the F1-score of 70% with only 10% of training data. This highlights the generalization capability of our technique, which incorporates clinical knowledge and reasoning instead of relying solely on data-driven deep learning methods.

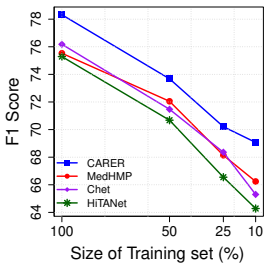

Figure 4: Performance comparison of CARER and some baselines with different portions of training data on MIMIC-III's Heart Failure Prediction.

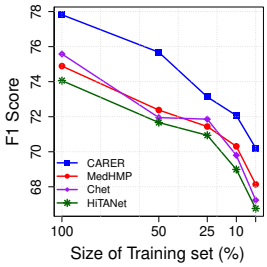

Figure 5: Cross-dataset generalization performance of CARER and baselines from MIMIC-IV to MIMIC-III for Heart Failure Prediction.

4.4 Generalizability Study

We here compare how well *CARER* and other methods generalize using a cross-dataset scenario. Specifically, we first pretrain predictive models on the MIMIC-III dataset (patients admitted from 2001 to 2012), with the Heart Failure Prediction task. Then, we take the pretrained model to continually finetune on MIMIC-IV dataset (patients admitted from 2013 to 2019) with different finetuning data proportions. We also test the generalization capabilities under zero-shot settings, where the predictive models are not given any training samples from the MIMIC-IV dataset. As shown in Fig. 5, *CARER* has better generalization capabilities compared with other methods with different amounts of MIMIC-IV finetuning data. Significantly, under the zero-shot settings, our method achieved an F1-score of 70.21, significantly outperforming the second-best baseline, MedHMP (68.04).

4.5 Ablation Study

We evaluated the design choices in our model by comparing it with other variants as follows:

- *w/o Alignment*: This variant removes the alignment loss $\mathcal{L}_{align}$ between multimodal data and clinical reasoning described in §3.4.
- *w/o Clinical Reasoning*: This variant removes all components related to clinical reasoning, essentially only using a Multimodal Encoder.
- *w/o Multimodal Encoder*: We remove the Multimodal Encoder described in §3.4 and use a MLP instead to achieve the classification.

Figure 6 shows the results of predictive modeling tasks on MIMIC-IV after removing specific components. First, we observe that removing the auxiliary

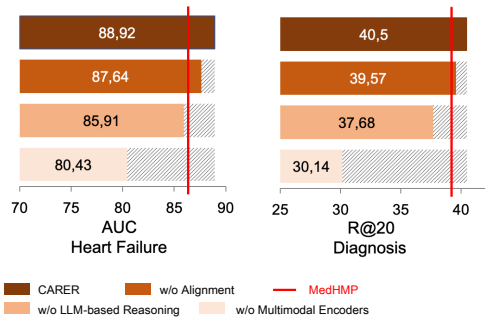

Figure 6: Diagnosis prediction and heart failure prediction for proposal variants on the MIMIC-IV dataset.

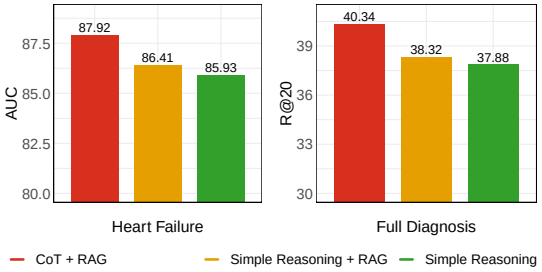

Figure 7: Full Diagnosis Prediction and Heart Failure Prediction results for different clinical reasoning instruction input on MIMIC-IV.

Table 3: Case Study: A diagnose of CARER with important sentences in the clinical reasoning (sentences with highest tokens' average attention weights). We highlight the words with the highest average token attention weights within the sentence in red.

Predictions	Atrial fibrillation (427.31), Malignant essential hypertension (401.0), and Chronic kidney disease, stage I (585.1)
Key sentences in the clinical reasoning (high attention weights *W*)	1. Atrial fibrillation remains a consistent diagnosis, indicating the patient continues to suffer from this chronic heart rhythm disorder. *(W: 0.000524)* 2. Acute kidney failure is another new and concerning diagnosis, suggesting a decline in kidney function, which may be associated with the infection or other acute stressors on the body. *(W: 0.000460)* 3. Acute on chronic diastolic heart failure and congestive heart failure suggest a worsening of cardiac function, potentially due to new heart attack or progression of underlying heart disease. *(W: 0.000442)* 4. The rise in blood creatinine from 1.0 mg/dL to 1.4 mg/dL is significant, as it suggests a decline in kidney function. *(W: 0.000428)*

alignment term reduces the performance of our proposed method, highlighting the necessity of properly aligning the local representation learned by the Multimodal Encoder with the clinical knowledge obtained from Retrieval and LLMs. Additionally, removing Clinical Reasoning altogether further worsens the predictive power, demonstrating that clinical reasoning is an essential source of information for our model.

However, the textual Clinical Reasoning alone is not sufficient for prediction, as shown by the significantly worse performance of the model *w/o Multimodal Encoder* on both tasks (significantly worse than the *MedHMP* baseline). This demonstrates that Clinical Reasoning should be used as an auxiliary source of information for conventional ML-based predictive models, not as a complete replacement.

We further examine the importance of the proposed Clinical Reasoning with Chain-of-Thought combined with RAG *(CoT + RAG)* proposed in *CARER*. We compare our proposal with two simpler variants:

- *Simple Reasoning:* uses a short instruction for LLM without detailed CoT instructions and additional information.
- *Simple Reasoning + RAG:* uses a short instruction for LLM without detailed instructions CoT instructions, but additional information from RAG is included.

We compare these variants on the MIMIC-IV datasets for both predictive tasks. The results are displayed in Figure 7. It can be seen that the performance dropped significantly when the clinical reasoning instruction is not provided, proving the necessity of a well-constructed and informative clinical reasoning thought process.

Necessity of fusing LLms and Deep Learning We investigate the effectiveness of fusing LLM's clinical reasoning representation with traditional Deep learning representation in this experiment, rather than relying solely on LLMs to make clinical predictions. In detail, after obtaining clinical reasoning text in Section 3.3, instead of fusing the clinical reasoning representation with Deep learning model, we ask GPT-3.5 to generate the predictions directly with the input being the clinical reasoning text.

The results for this experiment on MIMIC-IV

Table 4: Performance comparison between CARER and GPT-3.5 only on Diagnosis and Heart Failure Prediction on MIMIC-IV.

Method	Diagnosis			Heart Failure			
	R@10	R@20	F1	AUC	F1	P	R
GPT-3.5 only	5.21	5.74	4.09	-	0.533	0.676	0.590
CARER	29.88	39.64	25.70	87.71	0.794	0.776	0.812

is reported in Table 4. We observe that CARER achieves superior performance for both tasks compared with using GPT-3.5 for prediction. This demonstrates that while LLMs possess good clinical reasoning capabilities and knowledge, they still lack the ability to forecast health outcomes of fine-tuned Deep Learning models. Thus, this proves the necessity to fuse the representation of LLMs and Deep Learning models for more accurate predictions.

4.6 Robustness with other LLMs

Our proposed model, *CARER* mainly utilizes openAI's GPT-3.5 to generate the clinical reasoning on a patient's history. However, in real deployment scenarios, utilizing proprietary models might not be feasible in clinical settings, due to various reasons such as budget constraints, privacy requirements. We conduct experiments with open-source LLM backbones: Qwen2-7B (Yang et al., 2024), Mixtral-8x7B (Jiang et al., 2024) and Llama 3-8B (et al., 2024) as the clinical reasoning generator to demonstrate the robustness of our method to different LLMs. The results are reported in Table 5.
We observe that with other LLM clinical reasoning backbones, *CARER* still consistently achieve state-of-the-art perfomance compared with MedHMP, a strong baseline in previous experiments. This result demonstrated CARER's robustness to different LLMs, and does not rely on openAI's GPT models for good performance.

Table 5: MedHMP and CARER's performance (in **bold**) with different LLMs as clinical reasoning backbones on MIMIC-IV

Method	R@10	R@20	F1
MedHMP	31.58	39.14	23.74
GPT-3.5	32.90	42.34	26.40
Qwen2-7B	32.39	41.56	25.68
Mixtral-8x7B	33.42	41.27	26.89
Llama 3-8B	31.65	40.10	24.98

4.7 Interpretability Case Study.

We examine cross-attention weights of each token in the clinical reasoning text, and compute the average sentence-level attention, demonstrated in Table 3. The model predicts Chronic kidney disease diagnosis by reasoning about the elevated blood creatine. The transition from Acute kidney failure in past visits to *Chronic* kidney disease is explained by the elevated blood creatine, which is commonly observed in medical literature (Taner et al., 2010).

5 Conclusion

This paper introduces *CARER*, a health risk prediction framework that enhances multimodal deep learning models with clinical reasoning from Large Language Models (LLMs). Additionally, we introduce a multi-view alignment objective that enhances the consistency between the local view of patient-specific raw data and the global view of external LLM's reasoning. Extensive experiments on two predictive tasks using popular EHR datasets demonstrate that *CARER* significantly outperforms recent models, by up to 11.2%, particularly in enhancing data efficiency and generalizability.

6 Limitations

Despite the state-of-the-art results achieved by *CARER*, we acknowledge several limitations. First, we conducted our experiments on two health risk prediction tasks, however, our method can be applied to many other clinical prediction tasks: mortality prediction, hospital readmission prediction, medication recommendation. Second, we acknowledge the extensive computational cost and processing time of our method, due to utilizing large pre-trained text encoders and Large Language Models. This could pose challenges for healthcare facilities with limited computing resources in implementing our system, or in use cases where fast processing time is critical.

References

Vida Abedi, Venkatesh Avula, Seyed-Mostafa Razavi, Shreya Bavishi, Durgesh Chaudhary, Shima Shahjouei, Ming Wang, Christoph J. Griessenauer, Jiang Li, and Ramin Zand. 2021. Predicting short and long-term mortality after acute ischemic stroke using ehr. *Journal of the Neurological Sciences*, 427:117560.

Steven Adler Sandhini Agarwal Lama Ahmad Ilge Akkaya Florencia Leoni Aleman Diogo Almeida et al. Achiam, Josh. 2023. Gpt-4 technical report.

Tian Bai, Shanshan Zhang, Brian L. Egleston, and Slobodan Vucetic. 2018. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 43–51, New York, NY, USA. Association for Computing Machinery.

Inci M. Baytas, Cao Xiao, Xi Sheryl Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware lstm networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. 2024. Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians. *JAMA Internal Medicine*, 184(5):581–583.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2022. Building a knowledge graph to enable precision medicine. *bioRxiv*.

E. Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2016a. Gram: Graph-based attention model for healthcare representation learning. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016b. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Kevin P. Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279–90.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. 2020. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference 2020*, pages 530–540.

Paul Grundmann, Tom Oberhauser, Felix Gers, and Alexander Löser. 2022. Attention networks for augmenting clinical text with support sets for diagnosis prediction. In *Proceedings of the 29th international conference on computational linguistics*, pages 4765–4775.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *ArXiv*, abs/2201.11838.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Chang Lu, Tian Han, and Yue Ning. 2022. Context-aware health event prediction via transition functions on dynamic disease graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4567–4574.

Chang Lu, Chandan K. Reddy, Prithwish Chakraborty, Samantha Kleinberg, and Yue Ning. 2021a. Collaborative graph learning with auxiliary text for temporal event prediction in healthcare. *ArXiv*, abs/2105.07542.

Chang Lu, Chandan K Reddy, and Yue Ning. 2021b. Self-supervised graph learning with hyperbolic embedding for temporal health event prediction. *IEEE Transactions on Cybernetics*, 53(4):2124–2136.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019. Pre-training of graph augmented transformers for medication recommendation. In *International Joint Conference on Artificial Intelligence*.

Yanchao Tan, Carl Yang, Xiangyu Wei, Chaochao Chen, Weiming Liu, Longfei Li, Jun Zhou, and Xiaolin Zheng. 2022. Metacare++: Meta-learning with hierarchical subtyping for cold-start diagnosis prediction in healthcare data. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 449–459, New York, NY, USA. Association for Computing Machinery.

Basturk Taner, Ozagari Aysim, and Unsal Abdulkadir. 2010. The effects of the recommended dose of creatine monohydrate on kidney function. *NDT Plus*, 4:23 – 24.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2852, Singapore. Association for Computational Linguistics.

Yongxin Xu, Kai Yang, Chaohe Zhang, Peinie Zou, Zhiyuan Wang, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Vecocare: Visit sequences-clinical notes joint learning for diagnosis prediction in healthcare data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 4921–4929. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zhen Xu, David R So, and Andrew M Dai. 2021. Mufasa: Multimodal fusion architecture search for electronic health records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10532–10540.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Bo Yang and Lijun Wu. 2021. How to leverage the multimodal ehr data for better medical prediction? In *Conference on Empirical Methods in Natural Language Processing*.

Kai Yang, Yongxin Xu, Peinie Zou, Hongxin Ding, Junfeng Zhao, Yasha Wang, and Bing Xie. 2023. Kerprint: Local-global knowledge graph enhanced diagnosis prediction for retrospective and prospective interpretations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4):5357–5365.

Muchao Ye, Suhan Cui, Yaqing Wang, Junyu Luo, Cao Xiao, and Fenglong Ma. 2021. Medretriever: Target-driven interpretable health risk prediction via retrieving unstructured medical text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2414–2423.

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

Yinghao Zhu, Changyu Ren, Shiyun Xie, Shukai Liu, Hangyuan Ji, Zixiang Wang, Tao Sun, Long He, Zhoujun Li, Xi Zhu, et al. 2024. Realm: Rag-driven enhancement of multimodal electronic health records analysis via large language models. *arXiv preprint arXiv:2402.07016*.

Appendix

A Multimodal Encoder

We elaborate in greater detail on our chosen Multimodal Encoder.

ICD Sequence Encoder. For the ICD Sequence $I = [I^1, \ldots, I^n]$ with n being the number of visits. We encode the ICD Sequence using the time-aware Transformer architecture dubbed HiTANET (Luo et al., 2020). Besides taking the ICD sequence as input, HiTANET also employs a time-aware embedding mechanism, which takes in a sequence of time interval vectors $\delta = [\delta^1, \delta^2, \ldots, \delta^n]$, with δ_t represent the interval (in days) between the last visit n and the t-th visit. HiTANET learns and fuses the representation of the ICD sequence with the time interval information through embedding summation and an attention mechanism. To obtain the ICD sequence representation z_I with hidden size, we pass the ICD sequence and the time interval sequence through HiTANET (denoted as f^H):

$$z_I = f^H\left(\left[I^1, \ldots I^n\right], \left[\delta^1, \ldots, \delta^n\right]\right).$$

Continous values Encoder. Similar to ICD Encoder, we employ a time-aware architecture to learn the representation of the lab values sequence $V = [V^1, \ldots V^n]$, with n being the number of time steps. Let $\phi = [\phi^1, \ldots \phi^n]$ be the elapsed time (days) between each lab value vector. The architecture we use is T-LSTM (Baytas et al., 2017), which makes modifications to a standard LSTM cell so that the network can decide to memorize information based on the elapsed time with the previous time step. Denote the T-LSTM model as f^H, we can compute the feature representation of the lab value sequence as follows:

$$z_V = f^T\left(\left[V^1, \ldots V^n\right], \left[\phi^1, \ldots, \phi^n\right]\right).$$

Demographic variables. For static demographic features D, we simply embed them with a Multilayer Perceptron network:

$$z_D = \text{MLP}(D).$$

Structured Modalities Fusion. We first fuse the representation of three structured modalities: ICD sequence I, continuous values sequence V and demographic variables D. We then take the sum of the embedded representations of these three modalities:

$$z_S = z_I + z_V + z_D.$$

Clinical Notes. Given a clinical note $N = [N^1, N^2 \ldots N^k]$, with k being the number of tokens in the note, we embed this clinical note with the clinically pretrained Clinical-Longformer model (Li et al., 2022), denoted as function f^C:

$$z_N = [z_N^{cls}, z_N^1, \ldots z_N^k] = f^C([N^1, N^2 \ldots N^k]).$$

Multimodal Fusion. To fuse the representation of structured modalities z_S and the representation of clinical notes z_N, we first adopt an attention-pooling layer to place more focus on important segments within a long and noisy clinical note:

$$\widetilde{z_N} = \text{softmax}\left(\frac{z_S W_Q (z_N W_K)^\top}{\sqrt{d_k}}\right)(z_N W_V),$$

where W_Q, W_K, and W_V are projection layers for the Query, Key, and Value, respectively. Afterward, we project the attention-weighted representation of the clinical note $\widetilde{z_N}$ to the same dimension as the structured modalities' features z_S. We finally sum the structured modalities' feature vector with the projected clinical note feature vector to achieve multimodal fusion:

$$z_E = z_S + \text{MLP}(\widetilde{z_N}).$$

B Detailed Input Prompt

We present the details of our clinical prompt template in Figure 8, which consists of two components: the instruction, and the input data.

C Output Samples

We include some clinical reasoning samples generated by LLMs in Figure 10.

D Data Preprocessing

ICD Code.

For the ICD diagnosis codes, while MIMIC-III use the ICD-9 Code system for diagnosis annotation, MIMIC-IV uses a combination of ICD-9 and ICD-10, a newer diagnosis code system. These two code systems contain a large number of overlapping diseases. For example, the disease "Type 2 diabetes mellitus without complications" is represented as the code *250.0* in the ICD-9 system, and equivalently as *E11.9* in ICD-10. In order to have a consistent disease-code mapping across all experiments, we convert all ICD-10 codes in the

Clinical Reasoning Instruction

You are given a list of hospital admissions information of a patient, sorted by admissions time. The information includes the patient's demographic, diagnoses and lab values and some clinical note segments. You are required to summarize the health status and progression of that patient visit by visit in less than 2000 words.

1. **Visit-level reasoning** : First, analyze the information from each visit separately. Look through their diagnosed diseases and summarize the main conditions they are suffering from. Next, take a look at some of the lab values, pay attention to abnormal or fluctuating lab values, generate knowledge on the typical range of those lab values, and what the patients' results indicate about their health.

2. Progression Reasoning : Afterward, summarize and analyze the health progression of the patient's in-between visits. Pay attention to which types of conditions are persistent, which types of conditions are cured, which types are emerging, and the progression of lab values, especially abnormal ones, and generate reasoning on what those progressions mean to their health condition. Finally, draw the most important conclusions on the patients' health state.

Structure your answer in the following format
[Start Visit-level reasoning]
- Visit 1 (Reasoning on visit 1)
- Visit 2 (Reasoning on visit 2)
…
- Visit n (Reasoning on visit n)

[End Visit-level reasoning]

[Start Progression Reasoning]
Persistent Conditions (Reasoning on persistent conditions)
*** Emerging Conditions (Reasoning on emerging conditions)
*** Resolved Conditions (Reasoning on resolved conditions)
*** Lab values progression (Reasoning on lab values progression)
*** Conclusion (Final conclusion) ***
[End Progression Reasoning]

Figure 8: Clinical instruction template for LLM.

Patient's Input

Demographic: Demographic: Patient is a white female.
At the first visit, she was 83 years old

The following is this patient's diagnosis and treatment history:

Visit 1

ICD Diagnoses: Patient was diagnosed with the following:
- coronary atherosclerosis of native coronary artery
- diabetes mellitus type 2
- intermediate coronary syndrome

…

Lab values: The patients recorded the following lab values during their visit
- Blood Sodium is 139.0 mEq/L
- Blood Potassium is 4.1 mEq/L
- Blood Creatinine is 0.9 mg/dL

…

Some clinical notes available during this visit
The patient's blood sodium and potassium indicating proper electrolyte balance. Creatinine levels suggesting adequate kidney function … with potential complications affecting the heart's rhythm and structure.

…

Visit 2

…

The following are some **retrieved relevant documents**
- **Low blood sodium** : Hyperglycemia occurs when the concentration of sodium in your blood is abnormally low. Sodium is an electrolyte, and it helps regulate the amount of water that is in and around your cells …
- **Diabetes mellitus** : Type 2 diabetes is a condition that happens because of a problem in the way the body regulates and uses sugar as a fuel. That sugar also is called glucose. This long-term condition results in too much sugar circulating in the blood …

Figure 9: An example of patient's input multimodal data used for clinical reasoning generation. Demographic variables are highlighted with the color orange, ICD Diagnosis Codes are highlighted in red, verbalized continuous values are highlighted in blue, and retrieved documents in violet.

[Start visit -level reasoning]
Visit 1
The patient, an 83-year-old white female, presented with several cardiovascular diagnoses:
Coronary atherosclerosis of native coronary artery: This condition involves the buildup of plaque within the walls of the coronary arteries that supply the heart with blood. This plaque buildup can restrict blood flow or lead to
…
The lab results from this visit showed normal sodium, potassium, and creatinine levels
Visit 2
Two years later, at age 85, the patient returned with more serious conditions:
- Unspecified septicemia and severe sepsis: These diagnoses indicate a systemic infection that has triggered a dangerous inflammatory …
[End visit-level reasoning]

[Start Progression reasoning]
Persistent Conditions
- Atrial fibrillation remains a consistent diagnosis, indicating the patient continues to suffer from this chronic heart rhythm disorder.
…
Emerging Conditions
- Septicemia and severe sepsis are new diagnoses, indicating a recent serious systemic infection that can have a critical impact on overall health.
..
Cured or Unmentioned Conditions:
- There is no mention of coronary atherosclerosis or intermediate coronary syndrome in the second visit, which may indicate that these conditions were managed or stabilized, possibly through the aortocoronary bypass surgery and the placement of a cardiac pacemaker.

Lab values progression
- Sodium and potassium levels remained within normal ranges for both visits, suggesting that electrolyte balance has been maintained.
….
- The rise in blood creatinine from 1.0 mg/dL to 1.4 mg/dL is significant, as it suggests a decline in kidney function. This is consistent with the diagnosis of acute kidney failure during the second visit.
[End Progression reasoning]

Figure 10: An example of clinical reasoning output.

MIMIC-IV dataset into ICD-9, using the mapping available in this git repository. We discard all the ICD-10 codes that can't be converted to ICD-9.

Lab Values.
Following (Xu et al., 2021), we choose 48 significant continuous features from the electronic health records dataset, displayed in Table 6. The rule set to be used for verbalization is constructed by referencing medical documents and articles. A few examples of this rule set are demonstrated in Table 7. Before passing the continuous features to the Multimodal Encoder, we perform Min-Max normalization on these values.

Demographic Variables.
The three demographic variables used are gender, race (categorical values), and the age of the patient in their earliest visit (continuous values). Categorical variables are one-hot encoded, while the age is Min-Max normalized before going into the Multimodal Encoder.

Clinical Notes.
We concatenate all available clinical notes of the patient's history $[N_1, N_2, N_3, ..., N_i]$ into a single text N. As stated in Section 4, we discard the notes of type "Discharge summary" as they contain information indicative of the diagnosis. We follow the minimal preprocessing steps similar to those proposed in Clinical-Longformer (Li et al., 2022), the pretrained language model we utilize. Specifically, we removed all characters except for alphanumeric and punctuation marks, converting all letters to lowercase, and trimming any extra whitespace.

E Medical Corpus

We derive the medical corpus from the medical knowledge base PrimeKG (Chandak et al., 2022). This knowledge base contains different types of medical entities like diseases, drugs, proteins, etc However, we only utilize 17,080 disease nodes in the knowledge graph. Each of these disease nodes contains descriptive textual features, including such as definition, symptoms, causes. In total, there are 14,252 textual node features, all of which we use to construct our final medical corpus.

Table 6: The chosen lab values and their unit of measurement.

Observation Name	**Units**
Foley	ml
Hemoglobin [Mass/volume] in Blood	g/dl
Exhaled minute ventilation low	l/min
Heart Rate	bmp
Respiratory Rate	bmp
Present Weight (kg)	kg
Anion Gap	meq/l
Eosinophils	percent
PEEP SET	cm h2o
Apnea Interval	s
Urea Nitrogen	mg/dl
Potassium	meq/l
Temperature F	deg f
Arterial BP Mean	mmhg
SpO2	percent
Temperature C (calc)	deg f
Creatinine	mg/dl
Magnesium	mg/dl
Oxygen [Partial pressure] in Blood	mmhg
Phosphate	mg/dl
Arterial BP [Diastolic]	mmhg
Blood Flow	ml/min
NBP Mean	mmhg
Glucose	mg/dl
Hematocrit	percent
Phosphate	mg/dl
Bicarbonate	meq/l
Neutrophils urine	percent
Wbc count	k/ul
Calculated Total CO2	meq/l
O2 saturation pulseoxymetry	percent
Chloride	meq/l
Arterial BP [Systolic]	mmhg
Previous Weight (kg)	kg
Lymphocytes dif	percent
Monocytes	percent
PH	u
Weight Change (gms)	g
NBP [Diastolic]	mmhg
Arterial BP [Systolic]	mmhg
Cardiac output rate	l/min
Calcium [Moles/volume] in Serum or Plasma	mg/dl
Sodium	meq/l
NBP [Systolic]	mmhg
Tbili	mg/dl
FIO2	percent
Platelet Count	k/ul

Table 7: Example rules for some lab values.

Observation Name	Units	Categories
Anion Gap	meq/l	Low: < 3 Normal: 3 - 11 High: > 11
Apnea Interval	s	Normal: < 10 Abnormal: >= 10
Glucose	mg/dl	Low: < 70 Normal: 70 - 99 Prediabetic: 100 - 125 Diabetic: > 125
Heart Rate	bmp	Bradycardia: < 60 Normal: 60 - 100 Tachycardia: > 100
Sodium	meq/l	Low: < 135 Normal: 135 - 145 High: > 145
SpO2	percent	Low: < 95 Normal: 95 - 100
Creatinine	mg/dl	Low: < 0.7 Normal: 0.7 - 1.3 High: > 1.3
Potassium	meq/l	Low: < 3.5 Normal: 3.5 - 5.0 High: > 5.0
Magnesium	mg/dl	Low: < 1.7 Normal: 1.7 - 2.2 High: > 2.2
Temperature F	deg f	Low: < 96.8 Normal: 96.8 - 99.5 High: > 99.5
Wbc count	k/ul	Low: < 4000 Normal: 4000 - 11000 High: > 11000

F Experimental Environment and Settings

We conduct all of our experiments on a system with a single RTX 3090 GPU, 2 Intel Xeon Gold CPUs, and 120 GB Memory.

We use the PyTorch framework version 2.1.0 for our implementations, HuggingFace version 4.11.2, and the openAI version 1.25.1 library for LLMs.

The clinical reasoning texts generated by GPT-3.5 were obtained within a single run and then saved locally to be reused for later experiments. We only report the numerical result of a single run for every experiments, due to the extensive computational and time cost of the ChatGPT API and pretrained language models.

The input clinical notes and clinical reasoning texts are truncated and only the first 2048 tokens are chosen when passed through the Clinical-Longformer model. To reduce the computational cost, we freeze the 9 first encoder layers during training. Our model is trained for 100 epochs and the batch size is set to 8 for every experiment. the AdamW (Loshchilov and Hutter, 2019) optimizer is used for every experiment, with the learning rate being 0.0001 for the Heart Failure Prediction task, and 0.001 for Full Diagnosis Prediction.