RePrompT: Recurrent Prompt Tuning for Integrating Structured EHR Encoders with Large Language Models

Arya Hadizadeh Moghaddam, Drew Ross
Mohsen Nayebi Kerdabadi, Dongjie Wang, Zijun Yao*
University of Kansas, USA
{a.hadizadehm, drewross, mohsen.nayebi, wangdongjie, zyao}@ku.edu

Abstract

Large Language Models (LLMs) have shown strong promise for mining Electronic Health Records (EHRs) by reasoning over longitudinal clinical information to capture context-rich patient trajectories. However, leveraging LLMs for structured EHRs (e.g., standardized diagnosis and medication codes) presents two key challenges. First, translating time-stamped EHR sequences into plain text can obscure both temporal structure and code identities, weakening the ability to capture code co-occurrence and longitudinal regularities. Second, unlike cohort-trained predictive models that learn a shared, task-aligned representation space across patients, LLMs are often applied in a case-isolated inference setting where each patient is processed independently without leveraging population-level patterns. To address these challenges, we introduce **RePrompT**, a time-aware LLM framework that integrates structured EHR encoders through prompt tuning, without modifying underlying architectures. Specifically, RePrompT recurrently incorporates latent states from prior visits to preserve longitudinal information, and injects population-level information through trainable prompt tokens derived from a cohort-trained, task-aligned EHR encoder. Experiments on MIMIC-III and MIMIC-IV demonstrate that RePrompT consistently outperforms both EHR-based and LLM-based baselines across multiple clinical prediction tasks.

1 Introduction

Electronic Health Records (EHRs) capture comprehensive information on patients' diagnoses, procedures, and treatments across longitudinal clinical visits, and provide context-rich trajectories that enable data-driven clinical decision support systems (Choi et al., 2020; Jiang et al., 2023). While Large Language Models (LLMs) (Team et al., 2023) have shown promising results in EHR mining tasks, such as mortality and readmission prediction (Goyal et al., 2024; Gebreab et al., 2024), two significant challenges remain in effectively applying LLMs to structured EHR signals.

The first challenge arises from the difficulty LLMs face in capturing the temporal structure of EHRs when longitudinal data are linearized into plain text for input. A patient's history typically consists of multiple visits over time, and the evolving trajectory across these visits plays a critical role in determining downstream outcomes (Yang et al., 2021). For example, a history of chronic kidney disease (CKD) increases the likelihood of subsequent comorbidities such as cardiovascular disease (CVD) appearing in later visits (Bozkurt et al., 2016). However, converting sequential EHR into textual descriptions can obscure both temporal dependencies and discrete identities of clinical codes. Although inserting explicit separators in prompts to verbally denote visit boundaries can weakly encode temporality (Tan et al., 2024; Liu and Lapata, 2019), the model's ability to process structured EHR (Zaghir et al., 2024) still remains insufficient. A promising direction is to incorporate temporal awareness into prompt tuning by enabling the model to explicitly access latent states from prior visits. This design strengthens visit differentiation and supports modeling of longitudinal progression, while avoiding substantial modifications to the existing LLM architecture.

The second challenge (Meskó, 2023; Maharjan et al., 2024; Wang et al., 2025b) lies in the limited ability of LLMs to explicitly leverage population-level and task-specific representations for prediction. In traditional cohort-trained approaches (Choi et al., 2017; Zhang et al., 2020; Choi et al., 2016), models are optimized on a population of patients for a predefined clinical outcome. As a result, a shared, task-aligned representation space across patients enables the discovery of meaning-

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 26766–26778

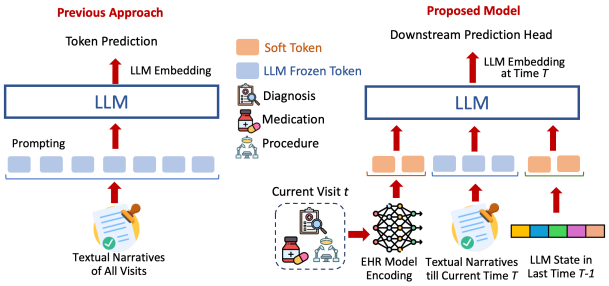

Figure 1: Illustration of the difference between existing approaches based on hard prompting and the proposed RePrompT framework. Unlike prior methods, RePrompT integrates both hard and soft prompting, with the soft prompts implemented through two strategies: struct-encoder and state-recurrent prompting.

ful patterns, such as disease co-occurrence, longitudinal progression, and ontology relationships that recur across peers for prediction support. In contrast, LLMs typically encode EHR information in a general-purpose manner and perform inference for each patient in a case-isolated setting. Without a shared, task-aligned patient representation space, LLMs lack an explicit mechanism to aggregate information from other patients to support prediction for a given individual.

A naive solution is to directly include similar patient profiles in the prompt, for example, via few-shot learning. However, the large scale and high dimensionality of modern EHR datasets make this approach impractical. A more promising direction is to integrate the complementary strengths of cohort-trained EHR models into LLMs. However, simple post-hoc fusion of embeddings from independently trained models is often suboptimal, as the representations are not jointly aligned. Recent advances in prompt tuning (Lester et al., 2021; Wu et al., 2023) provide an effective alternative. By introducing trainable prompt tokens grounded in representations learned from cohort-trained EHR encoders (Vu et al., 2021), LLMs can be adapted to incorporate patient-shared embeddings alongside context-rich clinical reasoning, which enables more principled structured EHR modeling.

To this end, as illustrated in Figure 1, we propose **RePrompT**, an adaptable LLM-based predictor that integrates structured EHR encoders through **Re**current **Prompt** **T**uning. Our proposed approach makes the following contributions:

- We develop a recurrent prompt tuning mechanism that allows an LLM to propagate visit-level EHR representation by reusing hidden states across time steps. This design mitigates the limitation of standard LLM inference, where the visit structure is only weakly encoded in plain text.
- We propose a framework that integrates general-purpose LLMs with cohort-trained, task-aligned EHR encoders by injecting structured EHR embeddings as trainable prompt tokens. By grounding the LLM in a shared patient representation space, our method enables population-aware modeling, in contrast to existing inference approaches that rely solely on text linearization.
- We conduct extensive experiments on two large-scale public benchmarks, MIMIC-III and MIMIC-IV, across readmission and mortality prediction tasks. Results show that the proposed approach consistently outperforms strong EHR-based and LLM-based baselines across different tasks and datasets.

2 Methodology

2.1 Problem Formulation

The EHR data for patient i is represented as a sequence of clinical visits $\{V_{i,t}\}_{t=1}^{T_i}$, where $V_{i,t}$ denotes the t-th visit in chronological order, and T_i is the total number of visits for patient i. Each visit $V_{i,t}$ consists of a set of medical codes, including diagnoses, medications, and procedures. Formally, the set of codes for visit t is defined as $V_{i,t} = \{x_{j,t}^i\}_{j=1}^{|V_{i,t}|}$, where $|V_{i,t}|$ denotes the number of medical codes recorded during that visit. In addition to medical codes, each visit also contains discharge notes represented as textual summaries. We denote the set of tokens in the discharge note for patient i at visit t as $\mathbf{C}_{i,t} = \{c_{j,t}^i\}_{j=1}^{|C_{i,t}|}$, where each $c_{j,t}^i$ corresponds to a discrete token. In this research, the terms "time-steps" and "visits" refer to the same concepts.

Task: Given a patient i with a sequence of visits, where each visit contains a set of medical codes $\{V_{i,t}\}_{t=1}^{T_i}$, and a corresponding discharge note $\{C_{i,t}\}_{t=1}^{T_i}$, the objective is to predict a specific healthcare outcome (e.g., mortality, readmission, or medication) in the next visit at T_i+1. This prediction is formulated as a binary or multi-label classification task for the target Y_{i,T_i+1}.

2.2 Model Summary

As shown in Figure 2, the proposed approach consists of three main modules: (1) Clinical Records Synthesis: Given a patient's medications, proce-

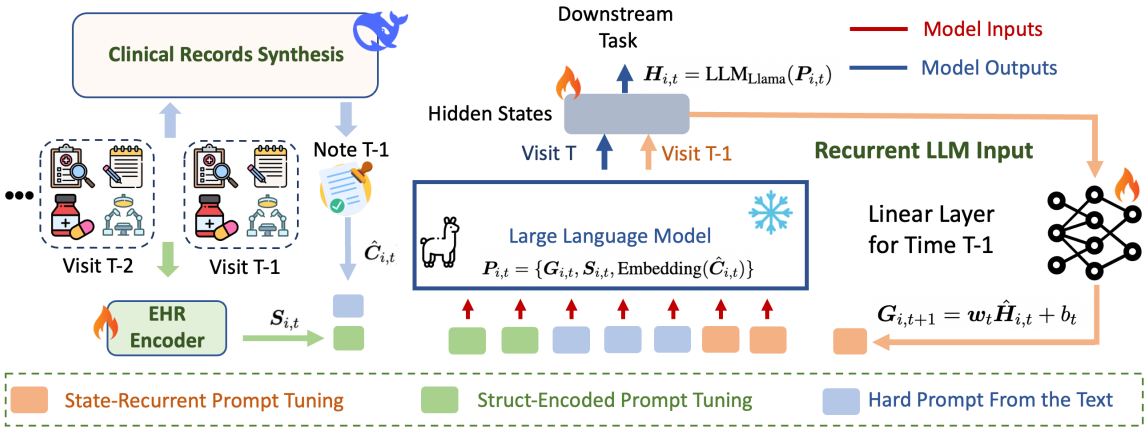

Figure 2: The framework of the proposed method. Medical codes and discharge notes are first used to generate patient summaries through Clinical Records Synthesis. Next, the structured medical codes in patient history are encoded into embeddings through a classic EHR encoder. Meanwhile, the LLM's hidden state in the previous time step is recurrently taken as a soft prompt for the current time step to capture longitudinal dependencies. Outputs from all prompting methods are combined as the input of the predictive LLM, and the LLM's state at the final visit is used for downstream binary classification tasks.

dures, diagnosis codes, and discharge notes, we first prompt a powerful and general-purpose LLM (e.g., DeepSeek) to synthesize a comprehensive patient summary. (2) State-Recurrent Prompt Tuning: We then use a local tunable LLM (e.g., Llama) to generate the hidden state from the previous time step and propagate it as a soft prompt to guide the generation at the current time step to capture longitudinal dependencies across visits. (3) Struct-Encoded Prompt Tuning: In parallel, we adopt a structured EHR encoder (Choi et al., 2016) to encode the sequential history of medical codes into dense representations, which serve as another set of soft prompts, allowing the tunable LLMs to incorporate shared structured patterns across different patient cases. Conditioned on both the synthesized patient summary and the two complementary soft prompting strategies, the proposed model generates a final representation for downstream classification tasks.

2.3 Clinical Record Synthesis

Structured EHRs consist of standardized medical codes, including medications, diagnoses, and procedures across patient visits. In addition, each visit (e.g., hospital stays) is accompanied by a discharge note that summarizes the patient's clinical course (Johnson et al., 2016, 2023). These notes are often verbose and noisy, containing redundant information and templated sections, which limit their usefulness for downstream modeling.

To address this, we employ a general-purpose LLM to denoise and synthesize discharge notes

> You are a clinical summarization assistant. Given structured medical data across multiple visits and discharge documentation, generate a concise but detailed patient summary. The summary must not exceed 700 tokens. Write in a professional, clinical style suitable for healthcare documentation. Explicitly address comorbidities, longitudinal trends across visits, and mortality considerations. Note that Visit 1 is the first visits and Visit N is the last visit and you should mainly focus on the last visit for its importance.
> Patient data for summarization (spanning multiple visits):
> - **** ICD-9 or 10 Diagnosis Codes ****: {diagnosis_codes}
> - **** ICD-9 or 10 Procedure Codes ****: {procedure_codes}
> - **** Medications NDC codes ****: {medication_codes}
> - **** Discharge Notes ****: {discharge_notes}
>
> LLM Agent
>
> Please generate a structured output with the following sections:
> 1. Longitudinal Patient Summary (≤500 tokens): Cohesive narrative integrating diagnoses, procedures, medications, and discharge details across multiple visits. Highlight disease progression or recurring conditions.
> 2. Risk & Mortality Considerations (≤100 tokens): Emphasize comorbidities, treatment history, complications, and longitudinal risk factors that may influence mortality.
> 3. Overall Clinical Impression (≤100 tokens): Concise synthesis of the patient's status, prognosis, and key follow-up considerations across visits.

Figure 3: The Prompt for clinical record synthesis.

together with structured medical codes into a concise patient summary. Specifically, we use DeepSeek-V3 (Liu et al., 2024) to summarize information from historical clinical notes and corresponding visit-level codes, as illustrated in Figure 3. The same synthesis prompt is applied at each visit t to produce a unified patient representation:

$$\hat{C}_{i,t} = \text{LLM}_{\text{DeepSeek}}(\{V_{i,r}, C_{i,r} | r <= t\}), \quad (1)$$

where $\hat{C}_{i,t}$ denotes the textual narratives of patient i at time step t.

2.4 Prompt Tuning for LLMs

In a standard language model, discrete input tokens are first mapped into continuous vectors through an embedding layer, which are then processed by the subsequent self-attention layers to

produce the model outputs. Achieving strong performance on a specific downstream task often requires adapting these models through fine-tuning on task datasets. Beyond conventional approaches such as full fine-tuning or parameter-efficient methods like LoRA (Hu et al., 2022), prompt tuning (Lester et al., 2021) offers an effective alternative. Rather than updating the LLM's internal weights, prompt tuning focuses on generating a small set of trainable continuous vectors, often called soft prompts, as the input embedding sequence to steer the model toward the target task. Because only the prompt parameters are optimized, prompt tuning reduces the cost and complexity of adaptation and is particularly appealing when the base LLM is large or not fully accessible, while still retaining strong flexibility across tasks.

Another practical advantage of soft prompts lies in their flexibility: since they are represented as trainable continuous variables, this method helps LLMs to be seamlessly integrated with existing neural architectures, allowing modular extensions without altering the original LLM parameters. Formally, the LLM input at time step t is:

$$\boldsymbol{P}_{i,t} = \{\boldsymbol{G}_{i,t}, \boldsymbol{S}_{i,t}, \text{Embedding}(\hat{\boldsymbol{C}}_{i,t})\}, \quad (2)$$

where $\boldsymbol{G}_{i,t} \in \mathcal{R}^{P \times D}$ corresponds to the *State-Recurrent Prompt Tuning* module, D is the hidden dimension of the LLM, and P represents the number of soft prompts for each module. $\boldsymbol{S}_{i,t} \in \mathcal{R}^{P \times D}$ is derived from the *Struct-Encoded Prompt Tuning* module, and $\text{Embedding}(\hat{\boldsymbol{C}}_{i,t}) \in \mathcal{R}^{N_{i,t} \times D}$ represents the token embedding associated with the *Clinical Record Synthesis*, where $N_{i,t}$ is the number of tokens for the patient summary for patient i at visit t. As shown in Equation 3, the representation $\boldsymbol{P}_{i,t}$ is subsequently fed into a predictor LLM, which produces the hidden state $H_{i,t} \in \mathbb{R}^{(2P+N_{i,t}) \times D}$. This hidden state is then used for the downstream classification.

$$\boldsymbol{H}_{i,t} = \text{LLM}_{\text{Llama}}(\boldsymbol{P}_{i,t}) \quad (3)$$

Since our primary focus is on embedding generation, we employ the Llama 3.1 1B model using the LLM2Vec framework, which adapts the language model specifically for high-quality embedding extraction (BehnamGhader et al., 2024).

2.5 State-Recurrent Prompt Tuning

LLMs have proven effective at converting textual information or prompts into high-level embedding representations, which can then be utilized to generate the next token. In existing literature, the input to an LLM is typically treated as a single document that is processed through multiple layers of self-attention, enabling the model to produce responses token by token. While powerful, this design introduces limitations in the context of EHR.

In EHR data, each patient has multiple visits, and within each visit, there is often a corresponding discharge note written by the physician. A straightforward approach to formatting sequential data for LLMs is to concatenate the visits into raw text with markers such as "Visit 1: ... Visit 2: ...". Although this method distinguishes visits through text description, the LLM still inherently treats the input as a single document and lacks an explicit mechanism to associate tokens with specific temporal dependencies.

To address this limitation, we propose State-Recurrent Prompt Tuning, an approach designed to better capture the structure of the visit level and longitudinal patient trajectories within EHR. Instead of aggregating all visit information as a single input to the LLM, we make the LLM process only one visit at a time, and output the token-level hidden state $H_{i,t}$ from the last layer, which is then aggregated through average pooling to form a visit-level hidden state $\hat{H}_{i,t}$. This state vector serves as a soft prompt that will be recurrently passed back to the same LLM to guide the generation of the hidden state for the subsequent visit, thereby enabling temporal continuity across visits.

As formulated in Equation 4, we first apply a linear transformation to the pooled hidden state, where $\boldsymbol{w}_t$ and b_t are trainable parameters, and $\boldsymbol{G}_{i,t} \in \mathbb{R}^{P \times D}$ shows the soft prompt embeddings, with P denoting the number of soft prompts.

$$\boldsymbol{G}_{i,t+1} = \boldsymbol{w}_t \hat{\boldsymbol{H}}_{i,t} + b_t \quad (4)$$

The output of this module, together with the following prompt tuning component will construct the soft prompt of the LLM.

2.6 Struct-Encoded Prompt Tuning

LLM-based prompting methods for EHR mining typically linearize extensive patient histories spanning multiple visits into a single, exhaustive per-patient input for LLMs to process (Meskó, 2023; Zaghir et al., 2024). While this formulation allows LLMs to access rich clinical narratives, it introduces two fundamental limitations. First, collapsing longitudinal records into static text hinders the model's ability to capture evolving patient

trajectories and disease progression over time, as temporal dependencies across visits are not explicitly represented (Zaghir et al., 2024). Second, such per-patient text representations prevent LLMs from effectively leveraging population-level and task-specific patterns that are critical for clinical prediction. Unlike cohort-trained EHR models, LLMs optimized with token- or sequence-level objectives do not enforce patient-centric alignment across the cohort, making it difficult to encode disease co-occurrence, shared ontologies, and longitudinal similarities among patients (Meskó, 2023). Consequently, clinically similar patients are not guaranteed to occupy nearby regions in the representation space, which leads to less contextually rich embeddings. Naively incorporating additional patient histories into the prompt is infeasible due to window constraints.

Therefore, we leverage structured encoders that learn patient representations from sequences of medical codes. Such models compress a patient's longitudinal history into a dense embedding that captures clinically meaningful patterns shared across patients. This representation can be injected into the LLM as a soft prompt. Among existing approaches, we adopt RETAIN (Choi et al., 2016) due to its effective use of dual-level attention and recurrent modeling for summarizing patient histories. Architectural details are provided in Appendix A.1. Given visit embeddings $\{\boldsymbol{V}_{i,j}\}_{j=1}^{t}$ RETAIN produces a patient representation $\boldsymbol{S}_{i,t}$, which is used as input to the LLM.

2.7 Prediction and Optimization

The output layer of an LLM is designed for next-token prediction. However, it can not reliably provide the calibrated probabilities needed for the classification tasks, as probabilities are treated as tokens and often suffer from hallucinations (Wang et al., 2024). To address this limitation, we introduce a classification head on top of the LLM, which takes the hidden state corresponding to the last visit input and maps it directly to one or multiple binary classes. Formally, as expressed in Equation 5, the model produces the output $\mathbf{y}_{i,T_{i+1}}$, which is then passed through the linear layer to generate scores. A threshold is then applied to yield the prediction $Y_{i,T_{i+1}}$.

$$\mathbf{y}_{i,T_{i+1}} = \boldsymbol{w}_{\text{output}}\hat{\boldsymbol{H}}_{i,T_i} + b_{\text{output}} \tag{5}$$

For optimization, we employ the Adam optimizer (Loshchilov and Hutter, 2017), and for the loss function, we use Binary Cross-Entropy loss in binary and multi-label classification tasks.

Table 1: Statistics of the EHR datasets.

Metric	MIMIC-III	MIMIC-IV
# of patients	7537	15874
# of patients with 2 visits	3622	4991
# of drugs per patient	79.30	113.55
# of diagnosis per patient	28.98	65.99
# of procedures per patient	7.37	9.08
# of visits per patient	1.65	3.11
# of patients with 1 visit	3622	4991
Positive rate for readmission	53.7%	53.5%
Positive rate for mortality	6.6%	1.3%

The trainable components include the EHR encoder, implemented using the RETAIN model (Equations 6-11), the linear layer that transforms the hidden representation at time $T_i - 1$ into the input representation at time T_i (Equation 4), and the output layer that maps the LLM output representation to the classification head (Equation 5). The LLaMA model remains frozen during both training and inference.

3 Experimental Setup

3.1 Datasets

In this study, we employ two real-world datasets to evaluate both mortality prediction and readmission prediction tasks:

- **MIMIC-III** (Johnson et al., 2016) is an open-access database containing health records of over 40,000 patients admitted to critical care units between 2001 and 2012. In this study, we focus on patients with multiple visits, aiming to predict the binary outcome.
- **MIMIC-IV** (Johnson et al., 2020) is a publicly available EHR dataset covering hospital admissions at Beth Israel Deaconess Medical Center from 2008 to 2019. It extends MIMIC-III with a clearer modular structure, richer clinical detail, and improved data provenance. MIMIC-IV contains data on over 380,000 unique patients across.

Both datasets are publicly available and have been thoroughly de-identified to comply with U.S. HIPAA regulations, which mandate the removal or modification of 18 types of personal identifiers. Their use in this study was conducted under the PhysioNet credentialed data use agreement.

3.2 Implementation Details

In this study, we focus on predicting two key healthcare outcomes: hospital readmission and

Model	MIMIC-IV				MIMIC-III			
	Readmission		Mortality		Readmission		Mortality	
	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC
Deepr	0.614	0.647	0.667	0.028	0.673	0.705	0.638	0.140
RETAIN	0.670	0.690	0.601	0.031	0.660	0.676	0.608	0.134
GRAM	0.578	0.609	0.633	0.028	0.627	0.660	0.626	0.139
GRASP	0.537	0.572	0.666	0.029	0.617	0.639	0.632	0.137
AdaCare	0.608	0.641	0.643	0.029	0.640	0.681	0.609	0.141
StageNet	0.656	0.691	0.664	0.032	0.676	0.702	0.633	0.142
Adore	0.605	0.641	0.656	0.024	0.661	0.693	0.629	0.138
ARCI	0.663	0.692	0.611	0.034	0.652	0.671	0.618	0.129
RePrompT	**0.706**	**0.728**	**0.673**	**0.036**	**0.688**	**0.719**	**0.646**	**0.152**

Table 2: Performance comparison on readmission and mortality prediction tasks using the MIMIC-III and MIMIC-IV datasets. Evaluation is conducted based on AUROC and PRAUC metrics.

mortality. Specifically, given information up to visit T_i, the model predicts the binary outcome at visit T_{i+1} for patient i with T_i prior visits or recommends medications on visit T_i based on diagnosis and procedures available from time 1 to T_i. Since these task requires longitudinal data, we excluded patients with only a single recorded visit.

For RePrompT and baselines, we use a greedy search approach to find the best hyperparameters for a comprehensive evaluation. We randomly split the data into 70% training and 30% testing sets and report the mean over three runs. We found that the optimal number of soft prompts is $P = 10$ or both modules, based on experiments with varying numbers of soft prompts for each component, which showed that this setting provides a favorable balance between performance and complexity. We also set the hidden dimension for the EHR model to 256 for the RETAIN model.

The dataset statistics are presented in Table 1. All experiments are implemented in Python, with PyTorch (Paszke et al., 2019) serving as the primary deep learning framework. In addition, RePrompT is fully compatible with the PyHealth framework (Yang et al., 2023), from which we use EHR baselines implementations. For LLM tuning we use the Hugging Face (Wolf et al., 2019) framework. We utilize a high-performance server equipped with three NVIDIA A6000 GPUs, 256 GB of RAM, and a 48-core CPU. We release the source code.[1] The computation time for a batch of patients is detailed in Appendix A.2.

In this research, we used two well-known threshold-independent classification metrics to comprehensively evaluate RePrompT, namely the AUROC and the PRAUC scores for readmission and mortality prediction.

3.3 Baselines

In this research, we utilize both well-known healthcare deep learning methods and LLM-based approaches. For deep learning methods, we use the following models (1) **Deepr** (Nguyen et al., 2016) represents each patient record as a sequence of coded events with time gaps and hospital transfers, then applies a convolutional neural network (2) **RETAIN** (Choi et al., 2016) incorporates a dual-RNN network to capture the interpretable influence of the visits and medical features for the prediction tasks. (3) **GRAM** (Choi et al., 2017) is a graph-based attention model that enriches EHR data with the hierarchical medical ontology, representing each concept as a weighted combination of its ancestors. (4) **GRASP** (Zhang et al., 2021) is a healthcare framework that improves EMR-based prediction by finding clinically similar patients. (5) **AdaCare** (Ma et al., 2020) is a health status representation model that captures both short- and long-term biomarker variations, adaptively emphasizes patient-specific risk factors. (6) **StageNet** (Ma et al., 2020) is a stage-aware neural network that learns disease progression via LSTM and integrates them with stage-adaptive convolution. (7) **ADORE** (Cheong et al., 2023) uses attention to adapt medical ontology category embeddings to EHR data for improved clinical prediction. (8) **ARCI** (Hadizadeh Moghaddam et al., 2024) disentangles coexisting temporal medical intents across sequential visits.

We have also conducted experiments on three different LLM-based baselines: (1) **Zero Shot** (Zhu et al., 2024), where GPT-5 (Wang et al., 2025a) is prompted to output probabilities for mortality and readmission prediction. (2)

[1] https://github.com/KU-AI4H/RePrompT

Model	MIMIC-IV				MIMIC-III			
	Readmission		Mortality		Readmission		Mortality	
	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC
Zero-Shot	0.512	0.545	0.591	0.011	0.505	0.537	0.501	0.069
Prompt-Tuning	0.575	0.608	0.607	0.019	0.634	0.663	0.611	0.132
COCONUT	0.581	0.611	0.605	0.021	0.639	0.670	0.612	0.136
RePrompT	**0.706**	**0.728**	**0.673**	**0.036**	**0.688**	**0.719**	**0.646**	**0.152**

Table 3: Performance comparison of LLM-based baselines on the MIMIC-III and MIMIC-IV datasets for both readmission and mortality prediction tasks on PRAUC and AUROC.

Prompt-Tuning, which introduces trainable soft prompts without relying on an EHR encoder and gets the probability of the "Yes" token from next-token prediction as the model output using the Llama 3.1 1B model. (Lester et al., 2021) (3) **COCONUT** (Hao et al., 2024), where a soft token is generated prior to predicting "Yes" or "No" to incorporate temporal information with the Llama 3.1 1B model, similar to our proposed approach.

4 Results and Discussion

This section presents the experimental analysis of comparisons between the proposed method and EHR and LLM-based approaches, ablation studies on the model and summarization, and evaluations using different EHR encoders.

4.1 Performance Comparison with EHR Baselines

Table 2 presents a comprehensive comparison between the proposed RePrompT framework and several well-established EHR baselines on both the MIMIC-III and MIMIC-IV datasets across two binary classification tasks. The results consistently highlight the superior performance of RePrompT across all evaluation metrics. In particular, when compared to RETAIN on the mortality prediction task, RePrompT achieves a substantial performance gain. This improvement stems from the use of a time-aware prompt tuning strategy that effectively links patient-specific EHR embeddings to the LLM to have more accurate modeling of longitudinal patient trajectories. Furthermore, against StageNet, the strongest baseline for MIMIC-IV mortality prediction, RePrompT demonstrates clear advantages. By integrating attention-aware RNNs with LLMs, our method surpasses the hybrid RNN-CNN architecture of StageNet, underscoring the benefit of incorporating language models into temporal EHR representations. Finally, the comparison with GRASP reveals that the time-aware LLM approach captures richer and more clinically meaningful information than methods relying solely on patient similarity during the embedding generation phase.

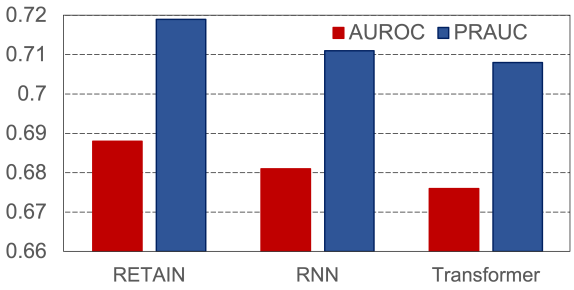

Figure 4: Performance comparison of different EHR encoders integrated into RePrompT. Results are reported using PRAUC and AUROC for the task of readmission prediction on the MIMIC-III dataset.

4.2 Performance Comparison for Different Integrated EHR Models

As illustrated in Figure 4, we conducted a series of complementary experiments focusing on the EHR backbone integrated within RePrompT for readmission and mortality prediction in the MIMIC-III dataset. In this analysis, patient embeddings generated by the EHR model were extracted and subsequently utilized as soft prompts within the LLM component, providing a more fine-grained examination of how different EHR architectures influence overall performance. Among the evaluated models, RETAIN achieved the highest performance, highlighting the effectiveness of its attention mechanism in emphasizing clinically relevant information from recent visits. For comparison, we also assessed two alternative configurations: one employing a standard LSTM architecture and another using a Transformer Encoder (Vaswani, 2017) as the EHR module. The Transformer-based approach performed the worst, likely because Transformer encoders have difficulty in modeling temporal dependencies across

Model	MIMIC-IV				MIMIC-III			
	Readmission		Mortality		Readmission		Mortality	
	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC
RePrompT w/o Both Modules	0.673	0.698	0.635	0.028	0.646	0.692	0.616	0.127
RePrompT w/o State-Recurrent	0.693	0.711	0.642	0.033	0.673	0.705	0.624	0.132
RePrompT w/o Struct-Encoded	0.698	0.722	0.665	0.031	0.676	0.712	0.637	0.148
RePrompT	**0.706**	**0.728**	**0.673**	**0.036**	**0.688**	**0.719**	**0.646**	**0.152**

Table 4: Ablation Studies of the proposed method RePrompT on the MIMIC-III and MIMIC-IV datasets for both readmission and mortality prediction tasks on PRAUC and AUROC.

successive patient visits, even when positional encodings are applied (Zhou et al., 2021).

4.3 Performance Comparison with LLM Baselines

Table 3 presents a comparison between the proposed approach and several well-established LLM-based baselines across both datasets. When the proposed approach is contrasted with the Zero-Shot Prompt Engineering method, our time-aware prompt tuning strategy achieves substantially higher performance on EHR prediction tasks, confirming that fine-tuned smaller models outperform larger models in specific tasks (Gao et al., 2023; Bucher and Martini, 2024). Furthermore, comparisons with prompt-tuning baselines show the critical importance of integrating soft prompts with the EHR model. Standard LLMs lack explicit knowledge of medical code co-occurrence patterns and thus fail to fully capture clinically meaningful embeddings. While chain-of-thought reasoning methods, such as the COCONUT approach, which also employs soft prompts, are generally advantageous, their performance here lags behind because conventional chain-of-thought reasoning does not adequately model temporal dependencies. In contrast, our time-aware chain-of-thought variant more effectively captures the evolving nature of patient trajectories, leading to superior performance.

4.4 Ablation Studies on RePrompT

Table 4 reports the ablation results used to assess the impact of each component in our framework. Removing either module leads to consistent performance degradation across all datasets and tasks, indicating that both components contribute meaningfully to the final model. The State-Recurrent Network produces the larger performance drop when excluded, suggesting that explicit modeling of temporal dependencies at the textual level plays a central role in improving predictive accuracy. Meanwhile, the gains associated with the Struct-Encoded module highlight the benefit of incorporating structured time-series information from multi-level EHR data. Taken together, these results show that the two modules capture complementary signals and jointly strengthen the model's ability to use multimodal clinical information.

4.5 Ablation Study on Input Summarization

To verify whether the improvement is mainly due to the DeepSeek-generated hard prompt, we removed DeepSeek and used only clinical notes as input to the Llama model. As shown in Table 5, the proposed method without the DeepSeek summarization still outperforms the RETAIN backbone, showing that the improvement comes from the framework itself, not only from DeepSeek summerization. The results also suggest that denoising the notes can improve performance.

5 Related Work

EHR-Based Predictive Models. Previous works have explored transforming EHRs into predictive representations for clinical decision support. Early deep learning models such as Deepr (Nguyen et al., 2016) bypass manual feature engineering by encoding records as sequences of discrete events, with convolutional networks detecting predictive clinical motifs for readmission risk. RETAIN (Choi et al., 2016) enhances interpretability with a reverse-time attention mechanism that highlights influential visits and variables, mimicking how clinicians review patient histories. To address data sparsity and domain alignment, subsequent methods incorporate external knowledge. GRAM (Choi et al., 2017) leverages hierarchical ontologies to produce knowledge-aligned embeddings, improving prediction for rare conditions. Hierarchical Attention Propagation (HAP) (Zhang et al., 2020) extends this idea by propagating attention bidirectionally across the ontology, capturing relationships among ancestors and descen-

Model	MIMIC-IV				MIMIC-III			
	Readmission		Mortality		Readmission		Mortality	
	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC	AUROC	PRAUC
RETAIN	0.670	0.690	0.601	0.031	0.660	0.676	0.608	0.134
RePrompT (w/o DeepSeek)	0.685	0.705	0.640	0.027	0.670	0.702	0.631	0.135
RePrompT	**0.706**	**0.728**	**0.673**	**0.036**	**0.688**	**0.719**	**0.646**	**0.152**

Table 5: Ablation Studies of input summerization on the MIMIC-III and MIMIC-IV datasets for both readmission and mortality prediction tasks on PRAUC and AUROC.

dants. While effective, such approaches do not naturally capture temporal dependencies across visits. GRASP (Zhang et al., 2021) embeds medical concepts into a unified semantic space using large language models, aligning semantically similar codes across datasets and mitigating coding inconsistencies. Patient heterogeneity and disease progression have been addressed by AdaCare (Ma et al., 2020), which models short- and long-term biomarker variations with multi-scale convolutions, and StageNet (Gao et al., 2020), which incorporates disease stage information with a stage-aware LSTM and adaptive convolutional module. Both improve prediction and interpretability but largely operate on structured features without integrating broader knowledge or reasoning capabilities. LINKO (Kerdabadi et al., 2025) uses LLM-initialized embeddings and dual-axis knowledge propagation, vertical within ontologies and horizontal across them, to capture both intra- and cross-ontology relationships. However, it does not fully exploit LLMs' potential for the related healthcare tasks.

LLM-Based Approaches. The growing capabilities of LLMs have motivated new approaches to clinical prediction. GraphCare (Jiang et al., 2023) constructs patient-specific knowledge graphs by combining structured knowledge bases with LLM outputs, using a bi-attention augmented GNN to enhance predictions across various predictive tasks. RAM-EHR (Xu et al., 2024) applies LLM-powered dense retrieval over multiple knowledge sources to augment patient representations, paired with consistency regularization to improve robustness. Previous works regarding zero-shot prompting (Zhu et al., 2024) show that prompts incorporating EHR-specific features enable LLMs to make effective predictions in few-shot scenarios. Instruction-based fine-tuning approaches, such as LlamaCare (Li et al., 2024), align general-purpose LLMs with clinical vocabulary and tasks, improving quality as judged by human evaluators. COCONUT (Chain of Continuous Thought) (Hao et al., 2024) enables reasoning directly in the LLM's latent space, exploring multiple inference paths rather than committing to a single chain-of-thought. This reduces premature commitment to a single trajectory and is promising for complex, high-stakes decision-making, such as differential diagnosis or treatment planning. However, current LLM methods still underutilize structured EHR data and fail to jointly model temporal dependencies and hierarchical relationships.

6 Conclusion

In this work, we addressed two fundamental limitations that arise when applying Large Language Models to Electronic Health Records: the lack of temporal awareness and the inability to capture patient-to-patient similarity patterns from raw text alone. To overcome these challenges, we introduced RePrompT, a time-aware and adaptable framework that integrates structured EHR representations with pretrained LLMs through soft prompt tuning. Experimental results on two large-scale clinical datasets, MIMIC-III and MIMIC-IV, demonstrate that RePrompT consistently outperforms both traditional EHR-based and standard LLM-based baselines.

7 Limitations

Despite promising experimental results, several limitations remain. First, the framework relies on the quality of the EHR data; domain shifts in other healthcare systems may affect generalizability. Second, future work is needed to extend the approach to more clinical prediction tasks.

8 Ethical Considerations

A potential risk and ethical consideration of this approach is that using non-anonymized or insufficiently de-identified EHR data may compromise patient privacy, underscoring the need for compliance with relevant data protection regulations.

References

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Biykem Bozkurt, David Aguilar, Anita Deswal, Sandra B Dunbar, Gary S Francis, Tamara Horwich, Mariell Jessup, Mikhail Kosiborod, Allison M Pritchett, Kumudha Ramasubbu, and 1 others. 2016. Contributory risk and management of comorbidities of hypertension, obesity, diabetes mellitus, hyperlipidemia, and metabolic syndrome in chronic heart failure: a scientific statement from the american heart association. *Circulation*, 134(23):e535–e578.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned'small'llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv preprint arXiv:2406.08660*.

Chin Wang Cheong, Kejing Yin, William K Cheung, Benjamin CM Fung, and Jonathan Poon. 2023. Adaptive integration of categorical and multi-relational ontologies with ehr data for medical concept embedding. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–20.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 606–613.

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.

Ze-Feng Gao, Kun Zhou, Peiyu Liu, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Small pre-trained language models can be fine-tuned as large models via over-parameterization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3819–3834.

Senay A Gebreab, Khaled Salah, Raja Jayaraman, Muhammad Habib ur Rehman, and Samer Ellaham. 2024. Llm-based framework for administrative task automation in healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–7. IEEE.

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. 2024. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1167–1168.

Arya Hadizadeh Moghaddam, Mohsen Nayebi Kerdabadi, Mei Liu, and Zijun Yao. 2024. Contrastive learning on medical intents for sequential prescription recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 748–757.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet.org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

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

Mohsen Nayebi Kerdabadi, Arya Hadizadeh Moghaddam, Dongjie Wang, and Zijun Yao. 2025. Multi-ontology integration with dual-axis propagation for medical concept representation. *Preprint*, arXiv:2508.21320.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Rumeng Li, Xun Wang, and Hong Yu. 2024. Llamacare: An instruction fine-tuned large language model for clinical nlp. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10632–10641, Torino, Italia. ELRA and ICCL.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. Adacare: Explainable clinical health status representation learning via scale-adaptive feature extraction and recalibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 825–832.

Jenish Maharjan, Anurag Garikipati, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Rahul Thapa, Qingqing Mao, and Ritankar Das. 2024. Openmedlm: prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. *Scientific Reports*, 14(1):14156.

Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: tutorial. *Journal of medical Internet research*, 25:e50638.

Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. 2016. Deepr: a convolutional net for medical records. *IEEE journal of biomedical and health informatics*, 21(1):22–30.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. 2024. Are language models actually useful for time series forecasting? *Advances in Neural Information Processing Systems*, 37:60162–60191.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*.

Cheng Wang, Gyuri Szarvas, Georges Balazs, Pavel Danchenko, and Patrick Ernst. 2024. Calibrating verbalized probabilities for large language models. *arXiv preprint arXiv:2410.06707*.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025a. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*.

Zixiang Wang, Yinghao Zhu, Huiya Zhao, Xiaochen Zheng, Dehao Sui, Tianlong Wang, Wen Tang, Yasha Wang, Ewen Harrison, Chengwei Pan, and 1 others. 2025b. Colacare: Enhancing electronic health record modeling through large language model-driven multi-agent collaboration. In *Proceedings of the ACM on Web Conference 2025*, pages 2250–2261.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. 2023. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in neural information processing systems*, 36:61060–61084.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. 2024. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765, Bangkok, Thailand. Association for Computational Linguistics.

Chaoqi Yang, Zhenbang Wu, Patrick Jiang, Zhen Lin, Junyi Gao, Benjamin Danek, and Jimeng Sun. 2023. PyHealth: A deep learning toolkit for healthcare predictive modeling. In *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2023*.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. *arXiv preprint arXiv:2105.02711*.

Jamil Zaghir, Marco Naguib, Mina Bjelogrlic, Aurélie Névéol, Xavier Tannier, and Christian Lovis. 2024. Prompt engineering paradigms for medical applications: scoping review. *Journal of Medical Internet Research*, 26:e60501.

Chaohe Zhang, Xin Gao, Liantao Ma, Yasha Wang, Jiangtao Wang, and Wen Tang. 2021. Grasp: generic framework for health status representation learning based on incorporating knowledge from similar patients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 715–723.

Muhan Zhang, Christopher R King, Michael Avidan, and Yixin Chen. 2020. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 249–256.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115.

Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M Harrison, Liantao Ma, and Chengwei Pan. 2024. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint arXiv:2402.01713*.

A Appendix

Model	Time (s)
Deepr	0.08
RETAIN	0.05
GRASP	0.14
AdaCare	0.03
GRAM	0.17
StageNet	0.07
Adore	0.19
ARCI	0.94
RePrompT	1.44

Table 6: Computational time analysis of the proposed method when processing a batch of eight patients.

A.1 RETAIN Encoder Details

For the Struct-Encoded Prompt Tuning we adopt RETAIN (Choi et al., 2016), a structured EHR encoder that summarizes sequential medical codes into dense patient embeddings using a dual-level attention mechanism.

Formally, RETAIN employs two sets of attention weights: visit-level attention $\{\boldsymbol{\alpha}_{i,j}\}_{j=1}^{t}$ and variable-level attention $\{\boldsymbol{\beta}_{i,j}\}_{j=1}^{t}$. Visit-level attention determines the relative importance of each visit embedding $\{\boldsymbol{V}_{i,j}\}_{j=1}^{t}$:

$$\{\boldsymbol{g}_{i,j}\}_{j=1}^{T_i} = \mathrm{GRU}_{\alpha}(\{\boldsymbol{V}_{i,j}\}_{j=1}^{T_i}) \tag{6}$$

$$\{\boldsymbol{\alpha}_{i,j}\}_{j=1}^{t} = \mathrm{Softmax}(\boldsymbol{w}_{\alpha}\{\boldsymbol{g}_{i,j}\}_{j=1}^{t} + b_{\alpha}) \tag{7}$$

Variable-level attention highlights the contribution of individual medical codes within each visit:

$$\{\boldsymbol{h}_{i,j}\}_{j=1}^{t} = \mathrm{GRU}_{\beta}(\{\boldsymbol{V}_{i,j}\}_{j=1}^{t}) \tag{8}$$

$$\{\boldsymbol{\beta}_{i,j}\}_{j=1}^{t} = \tanh(\boldsymbol{w}_{\beta}\{\boldsymbol{h}_{i,j}\}_{j=1}^{t} + b_{\beta}) \tag{9}$$

Final patient representation is computed based on both attention values:

$$\boldsymbol{k}_{i,t} = \sum_{j=1}^{t} \boldsymbol{\alpha}_{i,j}\boldsymbol{\beta}_{i,j} \odot \boldsymbol{V}_{i,j} \tag{10}$$

$$\boldsymbol{S}_{i,t} = \boldsymbol{w}_{\mathrm{enc}}\boldsymbol{k}_{i,t} + b_{\mathrm{enc}} \tag{11}$$

In the equations, $\boldsymbol{w}_{\alpha}$, b_{α}, $\boldsymbol{w}_{\beta}$, b_{β}, $\boldsymbol{w}_{\mathrm{enc}}$, and b_{enc} are trainable parameters. The resulting embedding $\boldsymbol{S}_{i,t}$ is used as a structured soft prompt for the LLM.

A.2 Computational Time Analysis

Table 6 reports the inference time required to generate predictions for a batch of eight patients. Although our model incurs higher latency than conventional deep learning baselines, the overall inference time remains practical for real-world deployment and is still fast enough to support timely prediction in realistic clinical settings.

A.3 Performance Comparison on Medication Recommendation Task

Table 7 presents the medication recommendation results on the MIMIC-III and MIMIC-IV datasets. Note that for medication recommendation, we do not group medications according to the ATC ontology, as our goal is to directly recommend specific medications rather than medication classes. RePrompT shows strong and competitive performance against the baseline methods across the reported metrics. These results show that integrating patient-specific EHR embeddings with the LLM through time-aware prompting provides useful clinical context for multi-label medication recommendation.

Model	**MIMIC-IV**		**MIMIC-III**	
	F1	Jaccard	F1	Jaccard
Deepr	0.222	0.184	0.249	0.156
RETAIN	<u>0.344</u>	<u>0.251</u>	0.251	0.158
GRAM	0.236	0.168	0.197	0.122
GRASP	0.264	0.189	0.150	0.093
AdaCare	0.244	0.175	0.205	0.129
StageNet	0.270	0.194	0.236	0.137
Adore	0.201	0.159	0.147	0.123
ARCI	0.308	0.245	<u>0.285</u>	<u>0.187</u>
RePrompT	**0.374**	**0.276**	**0.314**	**0.204**

Table 7: Performance comparison on the medication recommendation task on the MIMIC-IV and MIMIC-III datasets, evaluated by F1-score and Jaccard similarity.